Third International Symposium of Morphology (ISMo 2021)

Program and Abstracts

Nabil Hathout, Fabio Montermini and Juliette Thuilier

22–24 September 2021

Contents

Part I

Program

Schedule

Wednesday, 22 September 2021

13:45 *Opening*

14:00 Milena Belosevic and Sabine Arndt-Lappe. *Merci-Jens and Lösch-Leyen. The semantics of personal name compounds in German*

14:30 Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini and Giovanni Moretti. *Including the word formation Latin resource in the LiLa knowledge base*

15:00 M Silvia Micheli. *An extensive analysis of blends in contemporary Italian*

15:30 *Break*

16:00 Marine Wauquier and Olivier Bonami. *Social gender and derivational morphology: a distributional study of the gendered import of learned morphology in French*

16:30 Marcel Schlechtweg and Greville Corbett. *When the s remains / does not remain an s: Further explorations in the acoustics of the English plural suffix*

17:00 Nabil Hathout and Fiammetta Namer. *Derivational paradigmatic models put to test on some non-canonical phenomena*

Thursday, 9 September 2021

09:00 Jenny Audring. *Gestalts, impostors and semi-affixes: Boundary issues between phonology and morphology*

10:00 Borja Herce. *Stress and stem allomorphy in the Romance perfectum: emergence, typology, and motivations of a symbiotic relation*

10:30 *Break*

11:00 Sebastian Fedden, Matías Guzmán Naranjo and Greville Corbett. *Typological richness of the German gender system revealed by data mining*

11:30 Alice Missud and Florence Villoing. *What French eventive nominalizations without verbal bases tell usabout the salience of paradigmatic networks*

12:00 Natalia Bobkova and Fabio Montermini. *Quantitative approaches to suffixal rivalry in denominal adjective formation in Russian*

12:30 Laurence Labrune. *Featural linking elements*

14:00 Tim Zingler. *A diachronic approach to the formal idiosyncrasies of indexes*

14:30 Noah Diewald. *Wao Terero lexical suffixes: Bridging the lexicon and discourse*

15:00 Peter Arkadiev. *Between noun incorporation and "lexical affixation" in Abaza*

15:30 *Break*

16:00 Benjamin Macaulay. *Prosody-morphology interactions in Mantauran Rukai*

16:30 Irina Burukina. *Two ways to nominalize in Kaqchikel*

17:00 Berthold Crysmann. *Positional competition in Murrinh-Patha by rule composition*

Friday, 10 September 2021

09:00 Sabrina Piccinin and Serena Dal Maso. *The contribution of morphological skills to L2 reading comprehension*

09:30 Martina Penke. *Regular and irregular noun plurals in German individuals with Down syndrome*

10:00 Despina Stefanou, Madeleine Voga and Hélène Giraudo. *Exploring morphological connexions within the mental lexicon: evidence from speakers from diverse educational backgrounds*

10:30 *Break*

11:00 Maria Copot and Olivier Bonami. *Spare us the surprise: the interplay of paradigmatic predictability and frequency*

11:30 Alizée Lombard, Marine Wauquier, Cécile Fabre, Mai Ho-Dac, Richard Huyghe and Nabil Hathout. *Evaluating morphosemantic demotivation through experimental and distributional methods*

12:00 Gauvain Schalchli. *Measuring morphological productivity from frequency lists: the median threshold hypothesis*

12:30 Erich Round, Sacha Beniamine and Louise Esher. *The role of attraction-repulsion dynamics in simulating the emergence of inflectional class systems*

14:00 Xavier Bach. *Explaining patterns of suppletion in ordinal numerals*

14:30 Gilles Boyé. *ALLER et MOURIR oddities in French conjugation: il a été au spectacle à pied, il a mouru d'ennui du début à la fin*

15:00 Livio Gaeta. *At the core of morphological autonomy: inflectional classes as a residue, ballast, or resource?*

15:30 *Break*

16:00 Franck Floricic. *'Less than zero': the case of (Italo-)Romance vocatives and imperatives*

16:30 Clemens Poppe. *Lexical strata in Japanese and Korean and the notion of lexeme.*

17:00 Yui Suzuki. *A constructionist approach to the distinction between reduplication and repetition: A case study of Turkish*

17:30 *Closing*

Committees

Program committee

We would like to express our deepest gratitude to the members of the programme committee, for their expertise on the papers submitted to this conference. Without you, we would not have been able to put this programme together.

Nabil Hathout, Fabio Montermini and Juliette Thuilier (ISMo 2021 co-chairs)

Dany Amiot
Mark Aronoff
Harald Baayen
Sacha Beniamine
Olivier Bonami
Geert Booij
Gilles Boyé
Dunstan Brown
Basilio Calderone
Berthold Crysmann
Georgette Dal
Serena Dal Maso
Sebastian Fedden
Bernard Fradin
Giorgio Francesco Arcodia
Helene Giraudo
Nabil Hathout
Claudio Iacobini
Marianne Kilani-Schoch
Jean-Pierre Koenig
Stéphanie Lignon
Chiara Melloni
Petar Milin
Fabio Montermini
Fiammetta Namer
Vito Pirrelli
Jan Radimsky
Franz Rainer
Andrea Sims
Andrew Spencer
Pavel Štichauer
Gregory Stump
Anna Maria Thornton
Juliette Thuilier
Delphine Tribout
Kristel Van Goethem
Florence Villoing
Madeleine Voga

ISMo standing committee

Dany Amiot
Olivier Bonami
Gilles Boyé
Georgette Dal
Bernard Fradin
Hélène Giraudo
Nabil Hathout
Stéphanie Lignon
Fabio Montermini
Fiammetta Namer
Delphine Tribout
Florence Villoing

Acknowledgements

ISMo 2021 is organised by the CLLE laboratory (Université Toulouse Jean Jaurès & CNRS). We would like to thank our partner labs:

- Analyse et traitement informatique de la langue française (ATILF; Université de Lorraine & CNRS)

  `http://www.atilf.fr/`

- Laboratoire de linguistique formelle (LLF; Université de Paris & CNRS)

  `http://llf.cnrs.fr/`

- Modélisation, Dynamiques, Corpus (MoDyCo; Université Paris Nanterre & CNRS)

  `https://www.modyco.fr/fr/`

- Savoirs, Textes, Langages (STL; Université de Lille & CNRS)

  `https://stl.univ-lille.fr/`

Part II

Papers

Keynote

Gestalts, impostors and semi-affixes: Boundary issues between phonology and morphology

Jenny Audring
Leiden University

Intuitively, most linguists will agree that *-er* in *nicer* or *painter* is different from *-er* in *feather, slender* or *bother,* and that *boy-* in *boyhood* represents a different kind of unit than *boy-* in *boycott.* Yet, the boundary between morphological structure and 'mere' phonology can be difficult to draw. This issue, which touches on the question of what falls within the realm of morphology and what doesn't, will be discussed from various angles in this talk.

From a theoretical perspective, we see difficulties in classification and categorization, e.g. in the analysis of "semi-" or "pseudo-affixes" like German *-e* in *Treppe* 'stairs' (Eisenberg & Fuhrhop 2013: 209) or singleton affixes like English *-ric* in *bishopric* and *-ison* in *comparison,* as well as more generally with phonaesthemes (Kwon & Round 2015). From a cognitive perspective, the lines between phonology and morphology are drawn in the language user's mind, with the result that word-internal structure appears to be "intrinsically graded" (Hay & Baayen 2005) and speakers may show individual differences in the recognition and productive use of patterns (Dąbrowska 2012, De Smet 2016). Similar-looking structures can give rise to *gestalt* effects (Köpcke & Panther 2016), and indeed come to converge with affix patterns in various formal or semantic ways (Weidhaas & Schmid 2015). This suggests that, from a usage-based perspective, all recognizable structure can be considered beneficial for the storage and processing of words.

I will discuss a variety of data, mostly from Germanic, and interpret the observations in the light of a model of morphology based on lexical relations (Jackendoff & Audring 2020).

References

Dąbrowska, Ewa. 2012. Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism* 2(3). 219–253. https://doi.org/10.1075/lab.2.3.01dab.

De Smet, Hendrik. 2016. The root of ruthless. Individual variation as a window on mental representation. *International Journal of Corpus Linguistics* 21(2). 250–271.

Eisenberg, Peter & Nanna Fuhrhop. 2013. *Das Wort* (Grundriss der deutschen Grammatik, Vol. 1). Stuttgart: Metzler.

Hay, Jen & Harald R. Baayen. 2005. Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences* 9(7). 342–348. https://doi.org/10.1016/j.tics.2005.04.002.

Jackendoff, Ray S. & Jenny Audring. 2020. *The Texture of the Lexicon*. Oxford: Oxford University Press.

Köpcke, Klaus-Michael & Klaus Uwe Panther. 2016. Analytische und gestalthafte Nomina auf -er im Deutschen vor dem Hintergrund konstruktionsgrammatischer Überlegungen. In Andreas Bittner & Constanze Spieß (eds.), *Formen und Funktionen*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110478976-006.

Kwon, Nahyun & Erich R. Round. 2015. Phonaesthemes in morphological theory. *Morphology* 25(1). 1–27. https://doi.org/10.1007/s11525-014-9250-z.

Oral presentations

Between noun incorporation and “lexical affixation” in Abaza

Peter Arkadiev
Institute of Slavic Studies of the Russian Academy of Sciences &
Russian State University for the Humanities

1 Introduction

The polysynthetic Northwest Caucasian languages have been traditionally considered to lack noun incorporation (NI). Indeed, if NI is understood as productive and lexically unrestricted morphological compounding of the verbal root with the root of its (usually) patientive argument (cf. Baker 1988 or, more recently, Olthof 2020), then such phenomenon is clearly lacking in Northwest Caucasian. However, in many typological studies of NI (e.g. Mithun 1984, 2000, de Reuse 1994, Massam 2009) this phenomenon is defined more broadly to also include unproductive and restricted noun-verb compounding as well as cases when the relation of the nominal root to the verbal root is distinct from the patient.

What the Northwest Caucasian languages, especially Abaza and Abkhaz, are rich in is the so-called preverbs, i.e. verbal prefixes expressing spatial meanings. From a typological perspective, the Northwest Caucasian preverbs can be linked to the notion of “lexical affixes”, which have very concrete semantics resembling that of roots (Mithun 1997; Mattissen 2004: 190–194)., denoting e.g. body-parts, salient artifacts or natural objects, locations, and various adverbial notions (Mattissen 2006: 297–333).

Historically, the Northwest Caucasian preverbs mostly go back to incorporated nouns with relational or locational semantics (Lomtatidze 1983), exemplifying a cross-linguistically well-attested path of development (Mithun 1984: 885–887; Mithun 1997: 365-366; Kinkade 1998; Mattissen 2006). I shall argue, however, that some of these preverbs can be considered incorporated nouns synchronically as well and that in general these elements form a cline from more root-like to more affix-like behaviour.

The data for this paper comes from Abaza (ISO 639-3 abq), which is spoken by ca. 35 000 people in the Karachay-Cherkess Republic in the Russian North Caucasus (for an outline of Abaza grammar in English see Lomtatidze et al. 1989, O’Herin 2002). The system of preverbs in Abaza and their combinatorics with verbal roots has been amply described in Klychev (1995), which, together with a number of published texts, is the primary source of examples in this paper.

2 Noun incorporation in Abaza

Abaza boasts more than a hundred simplex and complex spatial preverbs. They occur in the middle of the prefixal template being preceded by the absolutive cross-reference prefixes, subordinators and applicatives and separated from the verbal root by the ergative and indirect object prefixes and markers of negation and causative (1).

(1) *a-wasa* *a-š’acara* *jə-**la**-wə-m-sa-n*
DEF-sheep DEF-lawn 3SG.N.ABS-**LOC.mass**-2SG.M.ERG-NEG-shave-IMP
‘Do not shear the sheep on the lawn.’ (Klychev 1995: 145)

While many of the Abaza preverbs do not have apparent lexical cognates or can be traced back to lexical (usually nominal) roots only etymologically, some have clear synchronic counterparts among nouns. Two subclasses of these can be singled out, which I shall discuss in turn.

2.1Body-part nouns

A considerable number of preverbs is constituted by body-part nouns, a feature Abaza shares with other Northwest Caucasian languages as well as cross-linguistically (Fleck 2006; Mattissen 2006: 310–315; Massam 2009: 1090). Some of these are clearly grammaticalized and have replaced their original meaning with that of spatial configuration, cf. *qa* 'head' (2).

(2)	*a-č'ḳʷən*	*a-ʒəχ'*	*d-a-**qa**-ĉ-ṭ*
	DEF-youth	DEF-spring	3SG.H.ABS-3SG.N.IO-**LOC.above**-sleep(AOR)-DCL

'The boy fell asleep over the spring of water.' (Klychev 1995: 258)

However, alongside the more grammaticalized preverbs like *qa-* in (2) there exist a number of preverbs that retain their body-part meaning, e.g. *lakta-* 'face' (3) or *qʷda* 'neck' (4). Note that such nouns are accompanied by a personal prefix cross-referencing their notional possessor that becomes an indirect object of the verb.

(3)	*a-saba*	*ʕa-rə-**lakta**-ṗl-əw-n*
	DEF-dust	CSL-3PL.IO-**LOC.face**-pour-IPF-PST

'Dust was pouring onto their faces.' (Klychev 1995: 154)

(4)	*arq̇an-g'əj*	*ʕ-a-**qʷdə**-j-χ-χə-n*
	rope-ADD	CSL-3SG.N.IO-**LOC.neck**-3SG.M.ERG-take-RE-PST

'He took the rope off its (the stallion's) neck.' (Abaza Tales 2015: 142)

Some of such body-part nouns undergo metaphorization, cf. the two uses of *naṗə-* 'hand' (this noun always occurs with the preverb *ça-* 'under') in (5) and (6).

(5)	*a-kʷṭaʁ'*	*s-**naṗə**-ça-p.č-ṭ*
	DEF-egg	1SG.IO-**LOC.hand**-LOC.under-break(AOR)-DCL

'The egg broke in my hands.' (Klychev 1995: 170)

(6)	*də-r-**naṗə**-ça-ŝa-ṭ*	*a-haǯ'rat-kʷa*
	3SG.H.ABS-3PL.IO-**LOC.hand**-LOC.under-fall(AOR)-DCL	DEF-robber-PL

'[A man] was attacked by (lit. fell under the hands of) robbers.' (AbLu 10:30)

2.2 Non-relational nouns

Another common lexical source of Abaza preverbs are non-relational nouns denoting salient landmarks such as *čʕʷa-* 'oven', *gara-* 'cradle' (7), *čḳara-* 'courtyard', *qʷa-* 'ashes' (8) etc.

(7)	*a-sabəj*	*d-**gara**-l-gʷa-n*	*d-ǯəkʷə-l-χ-ṭ*
	DEF-child	3SG.H.ABS-**LOC.cradle**-3SG.F.ERG-put-PST	3SG.H.ABS-LOC.out-go-RE(AOR)-DCL

'She put the child into the cradle and went away.' (Klychev 1995: 67)

(8)	*ajnəẑ*	*d-**qʷa**-la-j-gʷa-ṭ*
	DEF + giant	3SG.H.ABS-**LOC.ashes**-LOC.mass-3SG.M.ERG-put(AOR)-DCL

'He threw the giant into the ashes.' (Klychev 1995: 273)

When incorporated, such nouns may, like body-part nouns, refer to a concrete landmark, as in (7) and (8), but may also function as *sui generis* "verbal classifiers" (Aikhenvald 2000: 149–171; Mithun 1984: 863ff) corresponding to a referential landmark expressed by a full noun phrase whose root can be both identical to the preverb (9) or different (10).

(9) *rə-**čʕ^{w}a*** *d-**čʕ^{w}a**-pχa-ṭ*
3pl.io-**oven** 3sg.h.abs-**loc.oven**-get.warm(aor)-dcl
‘He warmed himself at their oven.’ (Klychev 1995: 213)

(10) *wadərʕ^{w}ana* *a-**ḳ’adəg^w*** *wə-**čḳara**-l-ṗ*
then def-courtyard 2sg.m.abs-**loc.yard**-go-npst.dcl
‘Then you’ll enter their courtyard.’ (Abaza Tales 2015: 85)

2.3 Affinity of incorporated nouns with preverbs

As already said, incorporated nouns in Abaza fall into one distributional class with other spatial preverbs, many of which, even if diachronically descending from nouns, are highly grammaticalized and desemanticized. Moreover, some of the incorporated noun roots discussed above have developed a vowel alternation distinguishing between essive/lative vs. elative forms (12), similarly to many preverbs (11) (Avidzba 2017), which testifies to their greater integration into the system of verbal spatial prefixation.

(11) a. *a-q̇anž̍’a* *a-ʕ^{w}ara* *j-**ta**-pssʕa-χ-ṭ*
def-crow 3sg.n.io-nest 3sg.n.abs-**loc.in**-fly-re(aor)-dcl
‘The crow flew back into its nest.’ (Klychev 1995: 197)
b. *a-warba* *a-ʕ^{w}ara* *j-**tə**-pssʕa-ṭ*
def-eagle 3sg.n.io-nest 3sg.n.abs-**loc.in.elat**-fly(aor)-dcl
‘The eagle flew out of its nest.’ (Klychev 1995: 205)

(12) a. *aʕ^w* ***čʕ^{w}a**-l-ga-ṭ*
def + trough **loc.oven**-3sg.f.erg-carry(aor)-dcl
‘She brought the trough to the oven.’ (Klychev 1995: 211)
b. *d-**čʕ^{w}ə**-r-ga-χ-ṭ*
3sg.h.abs-**loc.oven.elat**-3pl.erg-carry-re(aor)-dcl
‘They carried him out of the oven.’ (Klychev 1995: 218)

3 Discussion

If Abaza indeed has noun incorporation, then this is a typologically rather peculiar case of NI. First, while NI occurs with body-part nouns and a number of non-relational nouns, the class of nouns allowing incorporation is apparently closed. The productivity of NI in terms of verbs allowing it is hard to assess; for some preverbs, Klychev (1995) lists hundreds of verbs combining with them, while for others just a dozen combinations are recorded. Second and most importantly, in all cases incorporated nouns serve as spatial modifiers of the verbal root and never denote affected patients or instruments. This contrasts with incorporated nouns and lexical affixes described in the typological literature. Finally, incorporated nouns in Abaza fall into one class with unequivocal spatial prefixes themselves mostly originating from nouns, which suggests that there have been several successive waves of NI in Abaza.

Abbreviations

1 — 1st person; 2 — 2nd person; 3 — 3rd person; abs — absolutive; add — additive; aor — aorist; csl — cislocative; dcl — declarative; def — definite; elat — elative; erg — ergative; f — feminine; h — human; imp — imperative; io — indirect object; ipf — imperfect; loc — locative; m — masculine; n — non-human; neg — negation; npst — non-past; pl — plural; pst — past; re — refactive; sg — singular.

Acknowledgment

This work is financially supported by the Russian Science Foundation, grant # 17-18-01184.

References

AbLu 2013: *АжвабыжьбзихIвагIв ЛукIа йшгIайхIвауа* [Gospel of Luke translated into Abaza]. Cherkessk.

Abaza Tales 2015: *Абаза турыхква* [Abaza Tales]. Mineral'nye vody: Alashara.

Aikhenvald, Alexandra Yu. 2000. *Classifiers*. Oxford: Oxford University Press.

Avidzba, Asmat V. 2017. *Lokal′nye preverby v abxazskom i abazinskom jazykax* [Local preverbs in Abkhaz and Abaza]. Ph.D. Dissertation, Sukhum.

Baker, Mark C. 1988. *Incorporation. A theory of grammatical function changing*. Chicago, London: The University of Chicago Press.

de Reuse, Willem. 1994. Noun incorporation. In: Ron Asher (ed.), *The Encyclopedia of language and linguistics*, Vol. 5, 2842–2847. Oxford: Pergamon Press.

Fleck, David W. 2006. Body-part prefixes in Matses: Derivation or noun-incorporation. *International Journal of American Linguistics* 72(1). 59–96.

Kinkade, M. Dale. 1998. Origins of Salishan lexical suffixes. In: *Papers for the 33rd International Conference on Salish and Neighboring Languages*, 266–295. Seattle: University of Washington.

Klychev, Rauf N. 1995. *Slovar′ sočetaemosti lokal′nyx preverbov s suffiksoidami i glagol′nymi kornjami v abazinskom jazyke* [The collocational dictionary of locative preverbs with suffixoids and verbal roots in Abaza]. Cherkessk: Karačaevo-čerkesskoe knižnoe izdatel′stvo.

Lomtatidze, Ketevan V. 1983. Osnovnye tipy lokal′nyx preverbov v abxazskom i abazinskom jazykax [Main types of local preverbs in Abkhaz and Abaza]. In Nurja T. Tabulova & Raisa X. Temirova (eds.), *Sistema preverbov i poslelogov v iberijsko-kavkazskix jazykax* [System of preverbs and postpositions in Ibero-Caucasian languages], 10–13. Cherkessk.

Lomtatidze, Ketevan, Rauf N. Klychev & B. George Hewitt. 1989. Abaza. In: B. George Hewitt (ed.), *The indigenous languages of the Caucasus. Vol. 2. The North West Caucasian languages*, 91–154. Delmar, N.Y.: Caravan.

Massam, Diane. 2009. Noun incorporation: Essentials and extensions. *Language and Linguistics Compass* 3(4). 1076–1096.

Mattissen, Johanna. 2004. A structural typology of polysynthesis. *Word* 55(2). 189–216.

Mattissen, Johanna. 2006. The ontology and diachrony of polysynthesis. In: Dieter Wunderlich (ed.), *Advances in the theory of the lexicon*, 287–353. Berlin, New York: Mouton de Gruyter.

Mithun, Marianne. 1984. The evolution of noun incorporation. *Language* 60(4). 847–894.

Mithun, Marianne. 1997. Lexical affixes and morphological typology. In: John Haiman, Joan Bybee & Sandra Thompson (eds.), *Essays on language function and language type*, 357–372. Amsterdam, Philadelphia: John Benjamins.

Mithun, Marianne. 2000. Incorporation. In: Geert Booij, Christian Lehmann & Joachim Mudgan (eds.), *Morphology: A handbook on inflection and word formation*, Vol. 1, 916–928. Berlin: Walter de Gruyter.

O'Herin, Brian. 2002. *Case and agreement in Abaza*. Arlington: SIL International.

Olthof, Marieke. 2020. *Incorporation: Constraints on variation*. Amsterdam: LOT Publications.

Explaining patterns of suppletion in ordinal numerals

Xavier Bach

Trinity College, University of Oxford

1 Introduction

All languages of the world present at least some cardinal numerals. A large number of languages also present derived series of numerals, mainly ordinals expressing the position of a referent in a sequence, multiplicatives, distributives, and group numerals (Veselinova 2020). These series of numerals are most commonly based on and derived from cardinal numerals Veselinova 2020); the most studied among them are ordinals, for which suppletion of the lower numerals is commonplace (Veselinova 1997).

Suppletion is the phenomenon of two word forms within a paradigm presenting two different roots. In the case of ordinals, it is suppletion in a derivational paradigm, but it has similarities with inflection in that there is a series, and the derivation is relatively systematic. Suppletion is found in a large number of languages of the world (see Brown et al. 2003 for an overview, Veselinova 2006 for verbs, and Vafaeian 2010, 2013 for nouns and adjectives).

2 The distribution

According to the WALS sample of 321 languages (Stolz & Veselinova 2013), a majority of languages that have ordinals also have suppletion, at least for 'first', as shown in blue on the map below. 12 languages show a pattern 'first two three', 54 languages present an alternation between suppletive and non suppletive 'first', 110 languages present a pattern 'first, twoth, threeth', and finally 61 languages present suppletion for at least two terms (Stolz & Veselinova 2013). That amounts to 244 languages out of 321 presenting suppletion for ordinals, or 76%, a figure in keeping with previous estimates (Veselinova 1997 found 69% of languages with suppletion for 'first').

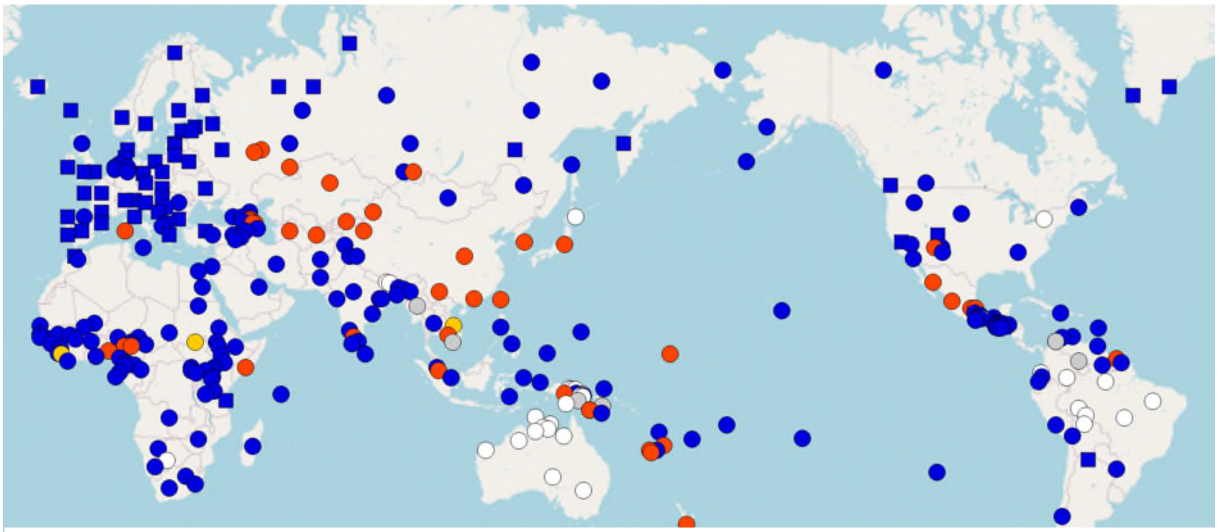


Figure 1. Suppletion in ordinals in the WALS sample (adapted from Stolz & Veselinova 2013). In blue are all the languages with suppletion for 'first' (blue squares indicate languages with further suppletion for 'second' or above).

By comparison, the most frequently suppletive verbs in the WALS sample for suppletion of 193 languages (Veselinova 2006) are verbs meaning 'come' and/or 'go' (grouped together in the data in Veselinova 2006), suppletive in 108 languages, i.e. 56% of the sample, and 'be/exist' suppletive in 69 languages, i.e. 36%. Suppletion is attested in a large number of languages (Brown et al. 2003, Corbett 2007), but it is quite exceptional for three quarters of languages in a large sample to present suppletion for the same item, mainly one and first. Such a distribution cannot be due to chance, and calls for explanation.

3 Existing proposals

A range of explanations for the prevalence of suppletion in ordinal series have been put forward in the literature. I examine three here, which are found to be insufficient.

3.1Veselinova (1997): grammaticalization

In a seminal typological study of suppletion in ordinals, Veselinova (1997) puts forth a framework where "suppletion is viewed as part of the broader process of grammaticalization during which certain lexical meanings turn out to be more suitable than others for expressing some more abstract meanings" (Veselinova 1997:430). She attributes the prevalence of suppletion for 'first' and 'second' to the restricted range of different diachronic source candidates for these words in various language groups: mainly 'front, before, precede', 'begin, early', 'leader' or a borrowing for 'first', and 'other', 'behind, back' or 'next, following' for 'second' (Veselinova 1997:443-444). Veselinova explains very well how these developments occur, but it is less clear why there is such a frequency of suppletion.

3.2Wier (ms): relative token frequency

Most suppletive words have high frequency in corpora (Hippisley et al. 2004). It is possible that frequency effects act on the preservation of suppletion across time, just as they play a role in the preservation of irregularity more generally.

In recent unpublished work, Wier (ms) examines suppletion in the languages of the Caucasus. In these languages, suppletion for derivational numeral series is restricted to ordinals. Thus in Kartvelian, 'first' is suppletive in Georgian, and 'first' and 'second' in Svan. Wier then asks why suppletion is readily found in languages of the world for the lower numerals, and his answer is that these items are more frequent than higher numerals. He shows for English that in a diachronic corpus the relative frequency of each ordinal is rather stable: 'first' is always more frequent than 'second', which in turn is more frequent than 'third', etc... in broad conformity to Zipf's law. The greater frequency of lower ordinal numerals justifies for him that they be more often suppletive.

3.3Barbiers (2007): mathematical properties of 'one'

Based on an analysis of Dutch, Barbiers (2007) sets out to explain the obligatory suppletion of 'first' based on the mathematical properties of the cardinal numeral 'one'. He shows that ordinal 'first' in Dutch has a different syntax from other ordinal numerals, in the same way as 'one' is different from higher cardinals. Barbiers treats 'one' and 'many' as indefinites, but ordinals as inherently definite, and claims that suppletion occurs to prevent a clash of feature in definiteness. Barbiers analysis fails to account for the fact that some languages have derived, non-suppletive ordinals for 'first' as well as the others (all the languages in red on the map of the WALS sample).

4 A novel proposal: dual series competition

Ordinals are words used to place referents or events in a sequence. There is, in all languages, one 'ordinal'that is never derived from a cardinal numeral, and this is the word for 'last'. I propose a unified account for this fact and the common suppletion for 'first', in that there are two series of words used to place referents in a sequence: on one hand a series derived from cardinal numerals, and on the other hand a pair of antonyms meaning 'first' as opposed to 'last'. Languages can thus lexicalize from both series, with 'first' at the intersection of the two series, where competition between the two patterns can occur.

This hypothesis makes a number of predictions. The first is that there could be languages with no derived ordinals which still lexicalize 'first' and 'last', and such languages are indeed found. For example, Australian languages are famous for not presenting ordinals, but in Warray there is a word for 'first' and another for 'last'. The WALS lists 12 such languages (Stolz & Veselinova 2013).

This is not to say that all languages lexicalize 'first' in this way, and one expects to find languages where the derivational ordinal rules out lexicalization of other items, as well as languages where the lexicalization of 'first' as an antonym to 'last' ousts the expected derivation, a situation that is found in the WALS sample: 41 languages only present the derived form (e.g. Somali), and 110 languages have suppletion only for first, the dominant pattern in the sample (e.g. Basque).

The really interesting category is the following. In the case of competition between two patterns for a specific item, one should expect to find cases which have both the derived, expected form, and the form lexicalized from 'last': this is indeed a robustly attested pattern, as 54 languages of the WALS sample present both forms. The hypothesis of competition between two series accounts sucessfully for such cases, as well as for cases where one of the two patterns 'wins over'. An example of such a case is Turkish, with cardinal *bir* 'one' presents both the derived ordinal *birinci* and the suppletive *ilk* 'first' (Stolz and Veselinova 2013).

5 Further evidence supporting dual series competition

The last prediction providing evidence in favour of a dual series competition concerns the etymology of suppletive forms. The competition hypothesis predicts that words for 'first' and 'last' should be taken from similar semantic fields, and originally be antonyms, in the languages that present suppletion, or which have 'first' and 'last' but no other ordinals. A preliminary survey finds that this is indeed what happens in a number of cases. For example, in Warray, a language that lexicalizes 'first' and 'last' without presenting other ordinals, the word for 'first' is derived from a word meaning 'front', and the word for 'last' is derived from a root meaning 'behind' (Harvey 1999:56).

In other examples, the morphological derivation of the words is shared. Thus, in Dutch, the suppletive form for 'first', *eerste*, is the superlative of *eer* 'early', just as *laatst* 'last' is the superlative of *laat* 'late' (Barbiers 2007:861), which shows a shared pattern of lexicalization for the two terms, beyond their original, opposite, lexical semantics. Similarly, in French, *premier* 'first' is derived from Latin *primarius* ultimately meaning 'in front', and *dernier* 'last' is derived from Latin *deretro* 'after, at the back'. Both words also present an identical suffix due to analogy between the two words (TLFi 'dernier'), compared to other ordinals derived from cardinal numerals with the suffix *–ième*, such as *troisième* 'third'.

6 Conclusion

Several explanations have been proposed for the frequency of suppletion in ‘first’ for ordinal series. I bring a novel approach by examining the neglected issue of the lexicalization of ‘last’, the only ordinal never derived from a numeral. The hypothesis of competition between two series, one derived from cardinal numerals, the other one built on the lexical pattern for ‘last’ correctly predicts the properties observed in a range of system, and in particular the frequent presence of both a derived and an underived ordinal for ‘first’ within the same language.

References

Barbiers, Sjef. 2007. Indefinite numerals ONE and MANY and the cause of ordinal suppletion. *Lingua* 117. 859-880.

Brown, Dunstan, Marina Chumakina, Greville G. Corbett & Andrew Hippisley. 2003. Surrey Suppletion Database. University of Surrey. http://dx.doi.org/10.15126/SMG.12/1

Corbett, Greville G. 2007. Canonical Typology, Suppletion, and Possible Words. *Language* 83(1). 8-42.

Harvey, Mark. 1999. *Warray grammar*. Ms.

Hippisley, Andrew, Marina Chumakina, Greville G. Corbett & Dunstan Brown. 2004. Suppletion: frequency, categories and distribution of stems. *Studies in Language* 28(2). 387-418.

Stolz, Thomas & Ljuba Veselinova. 2013. Ordinal numerals. In Matthew Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/53, Accessed on 2021-03-31.)

TLFi = *Tresor de la langue française informatisé*. ATILF – CNRS & Université de Lorraine. http://www.atilf.fr/tlfi

Vafaeian, Ghazaleh. 2010. *Breaking paradigms: A typological study of nominal and adjectival suppletion*. Stockholm: Stockholms universitet.

Vafaeian, Ghazaleh. 2013. Typology of nominal and adjectival suppletion. *STUF – Language Typology and Universals* 66(2). 112-140.

Veselinova, Ljuba. 1997. Suppletion in the derivation of ordinal numerals: A case study. In B. Bruening (ed.), *Proceedings of the Eighth Student Conference in Linguistics (SCIL-8)*, 71-92. Cambridge, MA: MITWPL.

Veselinova, Ljuba. 2006. *Suppletion in verb paradigms: bits and pieces of the puzzle*. Amsterdam, Philadelphia: Benjamins.

Veselinova, Ljuba. 2020. Numerals in Morphology. *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press.

Wier, Thomas. ms. Suppletion in the Languages of the Caucasus. Ms, University of Tbilisi.

Merci-Jens and *Lösch-Leyen*
The Semantics of Personal Name Compounds in German

Milena Belosevic & Sabine Arndt-Lappe
Trier University

1 Introduction

This paper examines German determinative compounds with a personal name as their second component and is based on 532 different word types in context from the microblogging platform Twitter. Consider the examples below.

1. Impfstoff-Bestellung: „Der Verdacht, dass Deutschland ein Unternehmen bevorzugt haben könnte“ Die Daten hierfür sind leider schon wieder von **Berater-Ursula**'s Handy gelöscht worden[1].
   Vaccine ordering: There is a suspicion that Germany might have preferred one company" Unfortunately, this information has already been deleted from **Advisor-Ursula**'s cell phone.

2. Können wir den **Laber-Lindner** nicht einfach mal nicht einladen? Den will doch keiner mehr hören[2].
   Could we just not invite **Babble-Lindner**? No one wants to hear him anymore.

Head constituents in all compounds refer to individuals (by first name *Ursula* in (1), by family name *Lindner* in (2). All compounds are proper names. The compound modifiers contribute some important properties of the name bearer or of events in which the name bearer is involved. Note that modifiers maybelong to different syntactic categories: *Berater* (1) is a noun, *laber*(*n*) (2) is a verb. As we will see below, the syntactic category of the first constituent does not affect the analysis.

In contrast to compounds with a proper name in a first position (cf. Koptjevskaja-Tamm 2009, Alexiadou2020), compounds headed by personal names (henceforth: PN-compounds) have received very little attention in the literature. This is, in part, due to their alleged marginality. For instance, compounds of this type in German account for only 0.2 % of the data in Ortner et al. 1991. Similarly, Kürschner (2020) finds only 0.9 % of proper name compounds among nicknames. According to Wildgen (1981), the meanings of PN-compounds are mainly characterised by their high degree of context-dependency (e.g. *Krisen-Strauß*). Furthermore, their interpretation is supported by other linguistic means from the context which also denote the name bearer.

In the present paper we will provide systematic analysis of the meanings of 532 different PN-compounds in context that were extracted from the microblogging platform Twitter. Contrary to what has been proposed by Wildgen (1981) and Kürschner (2020), we will argue that PN-compounding is not as marginal as is often assumed. Not only is the process very productive in naturalistic usage data representing informal language use (like social media data). Also the claim that their meanings are context-dependent and, hence, unpredictable should be relativized. Using a frame-semantic approach, we will show that meanings can be generalized according to different types of both extra-linguistic and semantic knowledge, which determine the meaning relations between the proper name and the common noun. Such relations are predictable. Decoding the meaning of an PN-compound thus involves accessinga semantic frame (often an event frame), on the basis of contextual and encyclopedic knowledge about the name bearer, and determining the

[1] https://twitter.com/Gerd581/status/1353140494801514503?s=20
[2] https://twitter.com/joergbartz/status/1350720500515942401?s=20

relationship between the constituents on the basis of the semanticframe structure (slot filling). We argue that both steps are not substantially different from determinative compounds headed by common nouns. What sets the latter apart from PN-compounds, however, is twothings: One is that, unlike general semantic frames, the slot-filling operation with a name leads to an interpretation of the frame mostly in terms of a specific event. Another is that aspects of this specific event then become available for the interpretation of the pragmatic function of the compound as a nickname.

2 Methods and data

532 PN-compound types with the names of politicians as head were extracted from the microblogging platform Twitter and annotated for their semantic properties. Table 1 provides an overview of the data.

word class of the modifier	percentage	example	gloss
common noun	83.8 %	*Geldkoffer-Schäuble*	'money case Schäuble'
proper name	7.6 %	*Schweden-Greta*	'Sweden Greta'
verb	4.5 %	*Laber-Lindner*	'babble Lindner'
adjective	4.1 %	*Dummgabriel*	'stupid Gabriel'

Table 1: PN-compound types

Annotation concerned (a) relevant semantic frames and their slots, and (b) the pragmatic function of thecompounds in context.

3 Frames for PN-compounds

In order to identify the extra-linguistic patterns, we analysed different types of knowledge evoked by the proper name head of the compound within the theoretical framework of frame-based word-formation theory (cf. Löbner 2013; Kotowski et al. 2021; cf. Olsen 2019: 117ff. for an overview of psycholinguistic approaches in which extra-linguistic knowledge plays an important role in compound meaning construction, cf. esp. Benczes 2006). We argue that proper name components of compounds evoke different types of knowledge about name bearers. These are encyclopedic and discursive knowledge, e.g. about the individual's history or their actions, and cotext based knowledge. These then serve as anchors for the activation of the relevant semantic frame (cf. Bonami et al. 2021 for a similar, scenario-based proposal). PN compounds are thus similar to what Löbner (2013) terms 'frame compounds'. The parallel nature of our PN compounds and Löbner's frame compounds, however, provides a challenge to Löbner's idea that this type of meaning construction is restricted to compoundsheaded by words denoting artifacts and to relations linking such artifacts to their affordances.

In the first step, we paraphrased each attestation considering the linguistic, cotextual, and contextual aspects as well as encyclopedic knowledge about the name bearer on the basis of the context in which the compound occurs. Based on the paraphrase, we annotated the frame and its frame elements accordingto the classification in German FrameNet[3]. The analysis of our 532 PN compounds yields eight frames (cf. Table 2).

frame	example	gloss	percentage
ACTIVITY	Kopftuch-Claudia	'headscarf Claudia'	26.9 %
MENTAL_PROPERTY	Dummlindner	'stupid Lindner'	25.7 %

[3] https://gsw.phil.hhu.de/framenet/

ENFORCING	Dosen-Jürgen	‘tin Jürgen’	13.4 %
SERVING_IN_CAPACITY	Finanzscholz	‘finance Scholz’	9.9 %
PEOPLE_BY_ORIGIN	Bayern-Toni	‘Bavaria Toni’	9.1 %
PREDICAMENT	Berater-Ursula	‘advisor Ursula’	9.1%
EXPRESSING_ PUBLICLY	Eiskugel-Jürgen	‘scoop Jürgen’	5.2 %
MEDICAL_CONDITIONS	Ischias-Schulz	‘sciatica Schulz’	0.7 %

Table 2: Distribution of frames in the corpus

Let us illustrate how we identified frames, using *Villen-Spahn* as an example. The name component *Spahn* evokes knowledge about the German Minister of Health Jens Spahn (e.g. appearance, function, origin, actions, events in which he was involved, statements, political decisions). In the attestation *Villen- Spahn*, the lexeme *Villen* is one frame element of the frame BUY, which becomes accessible through knowledge about a discursive event (*Spahn has bought an expensive villa in Berlin*). This discursive knowledge is central for the interpretation of the compound, as *Spahn has bought a villa* and not as *Spahn lives in a villa* or *Spahn has sold his villa*. Therefore, we annotated the attestation with the frame COMMERCE_BUY from FrameNet. The frame BUY comprises two frame elements GOODS and BUYER, which are then filled by the components of the compound, *Spahn* and *Villen*. The example shows that knowledge about a specific discursive event is crucial in identifying the frame. In a second step, we grouped similar frames (e.g. COMMERCE_BUY, COLLABORATION, BORROWING) to a more general framePREDICAMENT, in order to find patterns of semantic relations on a more abstract level. The frame PREDICAMENT contains knowledge about the name bearer's involvement in big political affairs, which turns out to be relevant for 9.1% of our compounds. Another interesting example is the frame EXPRESSING_PUBLICLY, in which the modifier is part of a statement made by the name bearer. *Eiskugel*in *Eiskugel-Jürgen*, for example, is a part of the statement of the green politician Jürgen Trittin, in whichhe compared the future cost of renewable energy to the price of a ‘scoop of ice cream’ (*Eiskugel*).

4 Pragmatic function: nicknames

The frame analysis does not account for the pragmatic function of PN-compounds. We find that they aremostly evaluative (cf. Štekauer 2015, Barbaresi & Dressler 2020), referring to and evaluating, for example, events that have damaged their reputation (*Berater-Ursula*), that have functioned as an emblemof their stance on a political question (*Kopftuch-Claudia*), or that have characterised their political actions (*Dummlindner*). Unlike nicknames, PN-compounds are not a permanent part of the personal name since they are created in order to fulfill certain communicative functions in a text, which is often a negative evaluation, but can also be mocking and exaggeration (cf. *Dumm-dumm-Katha*).

5 Summary and conclusion

Summarizing, PN compounds function as proper names. As the analysis of the twitter data shows, theirformation is based on several extra-linguistic, but predictable knowledge-based patterns which provideaccess to relevant semantic frames. The fact that they formally correspond to determinative compoundswhile still having the referential properties of names makes the link between their formal characteristicsand their semantic interpretation a highly interesting object of study. The paper has shown that the studyof PN compounds in informal language usage may shed new light on the role of discourse-based knowledge in the generation of compound meaning.

References

Alexiadou, Artemis. 2020. On the morphosyntax of synthetic compounds with proper names: A casestudy on the diachrony of Greek. *Word Structure* 13/2. 189–210.

Barbaresi, Lavinia Merlini & Wolfgang U. Dressler. 2020. Pragmatic explanations in morphology. In Vito Pirelli, Ingo Pirelli, & Wolfgang U. Dressler (eds.), *Word Knowledge and Word Usage*, 406–451.Berlin & Boston: De Gruyter Mouton.

Benczes, Réka. 2006. *Creative compounding in English: the semantics of metaphorical and metonymical noun-noun combinations*. Amsterdam: Benjamins.

Bonami, Olivier, Louise McNally & Denis Paperno. 2021. *The meaning of derivation: Relations andscenarios*. Paper presented at the 43. Annual Conference of the Linguistic Society of Germany (DGfS) (Workshop: The Semantics of Derivational Morphology), Freiburg. https://www.linguistik.uni- freiburg.de/43rd-annual-conference-of-the-german-linguistic-society-dgfs/program/arbeitsgruppen.pdf.

Koptjevskaja-Tamm, Maria. 2009. Proper-name nominal compounds in Swedish between syntax andlexicon. *Rivista di Linguistica / Italian Journal of Linguistics* 21. 119–148.

Kotowski, Sven, Sabine Arndt-Lappe, Natalia Filatkina, Milena Belosevic & Audrey Martin. 2021. The semantics of personal name blends in German and English. Manuscript submitted for publication.

Kürschner, Sebastian. 2020. Nickname formation in West Germanic: German Jessi and Thomson meet Dutch Jess and Tommie and English J-Bo and Tommo. In Gunther de Vogelaer, Dietha Koster & Torsten Leuschner (eds.): *German and Dutch in Contrast*, 15–47. Berlin: de Gruyter.

Löbner, Sebastian. 2013. *Understanding Semantics*. New York: Routledge.

Olsen, Susan. 2019. Semantics of compounds. In Claudia Maienborn, Klaus Heusinger & Paul Portner (eds.), *Semantics - Interfaces*, 103–142. Berlin, Boston: de Gruyter.

Ortner, Lorelies, Elgin Müller-Bollhagen, Hans Wellmann, Maria Pümpel-Mader & Hildegard Gärtner. 1991. *Substantivkomposita*. Berlin: de Gruyter.

Štekauer, Pavol. 2015. Word-formation processes in evaluative morphology. In Nicola Grandi & Lívia Körtvélyessy (eds.), *Edinburgh handbook of evaluative morphology*, 43–60. Edinburgh: Edinburgh University Press.

Wildgen, Wolfgang. 1981. *Grundstrukturen und Variationsmöglichkeiten bei Eigennamenkomposita: Komposita mit den Eigennamen Schmidt und Strauß als Konstituenten in Wahlkampfberichten des Spiegels*. Trier.

Quantitative approaches to suffixal rivalry in denominal adjective formation in Russian

Natalia Bobkova, Fabio Montermini

CLLE-ERSS, CNRS & Université de Toulouse Jean Jaurès

1 Introduction

The derivation of adjectives from nouns is a complex issue in Russian morphology, as these lexemes display a great deal of variation in the range of suffixes employed. Consequently, they constitute a good testing ground for the study of the competition between rival derivational strategies for the same syntactic and semantic function (Lindsay and Aronoff 2013; Aronoff 2016; Bonami and Thuilier 2018, among others).

The strategies used to derive adjectives from nouns in Russian are varied. Švedova (1980), for instance, enumerates more than 25 suffixes, which have various degrees of productivity. The main three adjectival suffixes are *-n-*, *-sk-* and *-Ov-*[1], all other suffixes may be considered as their extended variants, for instance, *-esk-*, *-ičesk-* are variants of *-sk-*, whereas *-ičn-* is a variant on *-n-* (Bobkova and Montermini 2019).

Recent developments in derivational morphology (cf. Plénat 2011, Roché 2011 among others) consider that various types of constraints (phonological, morphological, semantic, pragmatic, etc.) display a complex interaction, resulting in the choice of one of the rival suffixes, or in the emergence of doublets:

(1)	*slesar'*	'locksmith'	↔	*slesar**n**(yj) / slesar**sk**(ij) / slesar**ev**(yj)*
	dinamika	'dynamics'	↔	*dinamič**esk**(ij) / dinamič**n**(yj)*
	simmetrija	'symmetry'	↔	*simmetr**ičesk**(ij) / simmetr**ičn**(yj)*
	bojec	'fighter'	↔	*bojc**ovsk**(ij) / bojc**ov**(yj)*[2]

In this paper we are particularly interested in the rivalry between the following productive suffixes, as they frequently appear with the same nominal bases:

(2) *-n- / -sk- / -Ov-*
-esk- / -n-
-ičesk- / -ičn-
-Ovsk- / -Ov-

We aim to establish properties of nominal bases which allow a distinction between these suffixes, regardless of doublets. The choice of one or the other of the suffixes is accounted for by scholars (Švedova 1980, Hénault 2016) by either purely phonological factors, semantic or lexico-morphological ones:

[1] The notation *-Ov-* indicates the variation of the vowel of this suffix, capital *O* may correspond to orthographically different surface forms, <o> or <e>.

[2] Note that the examples of *dinamika* and *bojec* display 2 cases of stem allomorphy: the first one concerns a mutation of the last phoneme of the stem ('dinamik'-'dinamič'); the second – vowel-zero alternation ('bojec'-'bojc'). Both phenomena are typical of Russian language, however, they can not be described by productive morphophonological processes in synchrony.

(3) *-n-* tends to form more qualitative adjectives, whereas *-sk-* is used to form more relational ones;

-Ov- appears with inanimate base nouns, *-Ovsk-* choses to combine with animate ones;

-esk- privileges nouns with stems ending with velars;

-ičesk- appears in particular in lexemes of foreign origin, and consequently also with lexemes containing specific suffixes / combining forms (e.g. *-ija*, *-izm*, *-ik*, etc.).

Our goal is to use quantitative approaches to reveal the main predictors (constraints) which result in the choice of a particular suffix. We show the results of 2 models based on a multifactorial analysis: logistic regression and decision trees. Since both allow an easy visualisation of the list of predictors (the most important features for the choice of the suffix, and – in case of decision trees – the visualization of the procedure of classification), we expect to highlight the properties of the base nouns which can motivate the choice of a particular affix.

2 Data and methodology

To perform our analysis, we extracted the adjectives from the National corpus of Russian language (https://ruscorpora.ru/), proceeded to manual cleaning and automatically reconstructed the bases for each adjective. Our final data set is composed of 4351 entries. Since the competition between the affixes listed above is driven by a complex combination of factors, the base nouns were annotated according to some of their properties:

(4) phonological: last phoneme of the stem, length of the base noun in syllables, stress position;

morphological: inflexional class;

semantic: animacy (Thuilier 2012), which combines in different ways such properties as [±common], [±human], [±concrete];

etymological: native or loanword.

The properties listed in (5) form the list of predictors for both models.

The data were further divided into two subcorpora: the highest frequent lexemes (>100; 2275 entries) and hapaxes (frequency 1; 2076 entries) lexemes. Dal & Namer (2012) show for instance that very low-frequency lexemes, if observed on a large scale, are likely to be good indicators of the creative use speakers do of morphological constructions, since they are less likely to have undergone phenomena of lexicalization and thus to be formally and/or semantically opaque.

The main goal of dividing the data into two subcorpora is to build a statistical model which learns the adjectival formation from the high frequency subcorpus (training set) and to evaluate how well the same model can apply its knowledge to predict the suffix in the low frequency subcorpus (test set). The training set was randomly divided into a proper training set and a dev set – for evaluation of the model on high frequency data as well.

Since a large number of predictors can introduce noise in models, we perform feature selection for every suffix as well.

3 Results

First, we observe descriptive statistics and data distribution between high and low frequency subcorpora for emerging tendencies.

The distribution of *-n-/-sk-/-Ov-* is even between both subcorpora, there are more lexemes in high frequency data set, comparing to low frequency one. The same tendencies are observed for *-esk-/-n-* distribution.

As for *-ičesk-/-ičn*, they are equally distributed between the subcorpora, although the lexemes are more numerous in the low frequency data set, especially with *-ičesk-*; we can hypothesize that this suffix is productive in synchrony and is used by speakers more often than other ones to form new adjectives.

-Ovsk-/-Ov- represents another interesting case for study: the proportions of both suffixes are inversed in 2 subcorpora. If in high frequency lexemes *-Ov-* is attested more often than -*Ovsk-*, in low frequency lexemes it is the opposite: the adjectives formed with *-Ovsk-* are more numerous than with *-Ov-*. We would expect to find proper nouns, of Russian and foreign origin, among noun bases – since their inventory can be potentially unlimited, this can explain the high productivity of *-Ovsk-* among low frequency lexemes.

3.1Rivalry of -*n-/-sk-/-Ov-*

Since the choice here is made between 3 suffixes, we used a multinomial logistic regression to evaluate the results. Both the logistic regression and decision trees can apply quite well the tendencies learned from the training set on dev and test sets, with an accuracy of 72 and 61, respectively. The main constraints which interact here and determine the choice of the suffix are animacy (*-n-* choses common abstract nouns, *-sk-* combines with common human and proper non-human, and *-Ov-* privileges common concrete nouns); to a lesser extent - the length of stem in syllables (whereas *-n-* and *-sk-* choses polysyllabic bases, *-Ov-* has a clear preference for monosyllabic ones) and the last phoneme of the stem (*-n-* privileges dental consonants, *-sk-* and *-Ov-* both combine more often with alveolars, *-Ov-*, in its turn, has also a preference for velars).

3.2Rivalry of -*esk-/-n-*

Both binomial logistic regression and decision trees provide excellent results in classification for both dev and test set with accuracy of 97 and 92. This proves that the same tendencies are preserved between the lexicalized adjectives formed with these suffixes and new emerging adjectives. Phonological constraints are the strongest: stress position is crucial to determine the choice of the suffix: *-esk-* is chosen more often for nouns where the antepenultimate syllable is stresses; *-n-* combines with nouns where the ultimate and penultimate syllables are stressed. Another phonological factor involved to determine the choice of the suffix is the last phoneme of the stem: *-n-* combines with dentals whereas *-esk-* with velars. Etymological factor is also important, however to a lesser extent: *-n-* privileges native stems more often than foreign, whereas the tendency is the opposite for *-esk-*.

3.3Rivalry of -*ičesk-/-ičn-*

Comparing to the previous competing suffixes, both statistical models perform more poorly to solve the rivalry between *-ičesk-* and *-ičn-*: the accuracy on the dev set is 95, and the generalization to the test set is 82. According to the models, semantic constraints prevail (animacy), the phonological factor is present again (last phoneme of the stem). However, if we take a closer look on misclassified data, we can see that *-ičesk-* was classified correctly in almost all the cases in dev and test sets, the large majority of misclassified data concerns -*ičn*-. Unfortunately, the tendencies learned by the models cannot shed light on the rivalry between these two suffixes.

3.4 Rivalry of -*Ovsk*-/-*Ov*-

As mentioned above, there is more data for testing the models than for training them. Unsurprisingly, the models can learn tendencies for high frequency lexemes and apply the knowledge quite well for the data coming from the same distribution: the accuracy on dev set is 94; as for classification on test set, the performance of both models drop, the accuracy is only 75. However, the conclusions about the main predictors can be made: semantic factors constitute the main constraint (-*Ov*- combines mostly with common abstract and common concrete nouns, -*Ovsk*- choses common human and proper human nouns); the phonological factor can also play a role (-*Ov*- privileges stems ending with velars whereas the is no clear preference for -*Ovsk*-).

4 Discussion

Our study based on statistical models allowed us to identify the constraints determining the choice of different rival suffixes forming denominal adjectives in Russian. The strongest constraints concern phonology, semantics and etymology of the base noun. Morphological factors, such as inflectional class, play a less significant role. The results of our study show that the factors often cited in the literature are good predictors for the choice of the suffix. Our results may contribute to improve the list of the best predictors for the choice of the affix, and to order these predictors according to their force in imposing the choice of an affix. The use of statistical models needs some precautions: the issues of inequalities in distributions as well as the lack of data should be addressed. The models used for the study capture only formal properties of base nouns and do not allow to take into account details concerning the semantics of derived adjectives. For this study we did not include in the list of parameters the type of corpus adjectives appear in (literature, newspapers, oral texts, poetry, etc.). We keep for further investigations the inclusion of the type of corpus among the predictors, the study of semantics of the adjectives using distributional methods and a more detailed study of existing doublets as well.

References

Aronoff, M. 2016. Competition and the lexicon. In A. Elia, C. Iacobini & M. Voghera (Eds), *Livelli di Analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della Società di linguistica Italiana*. Roma: Bulzoni, 39-52.

Bobkova, N. & F. Montermini (2019). Suffix rivalry in Russian: what low frequency words tell us, *Mediterranean Morphology Meetings* 12:1-17.

Dal, G. & F. Namer. Faut-il brûler les dictionnaires? ou comment les ressources numériques ont révolutionné les recherches en morphologie. In *SHS Web of Conferences*, volume 1, pages 1261–1276. EDP Sciences, 2012.

Lindsay, M. & M. Aronoff. 2013. Natural selection in self-organizing morphological systems. In N. Hathout, F. Montermini & J. Tseng (Eds.), *Morphology in Toulouse. Selected prodeedings of Décembrettes 7*. Munich: Lincom Europa, 133-153.

Bonami, O. & J. Thuilier. 2018. A statistical approach to rivalry in lexeme formation: French -iser and -ifier. *Word Structure* 11(2): 4-41.

Hénault, C. & S. Sakhno. 2016. Čem supermarket-n-yj lučše supermarket-sk-ogo? Slovoobrazovatel'naja sinonimija v russkix ad"ektivnyx neologizmax po dannym Interneta. In Branko Tošović & Arno Wonisch (eds.), *Wortbildung und Internet, xxx–xxx*. Graz: Institut für Slawistik.

Plénat M. 2011. Enquête sur divers effets des contraintes dissimilatives en français. In M. Roché, G. Boyé, N. Hathout, S. Lignon et M. Plénat, *Des unités morphologiques au lexique*, Hermes-Lavoisier, Paris, pp.145-190.

Roché M. 2011. Quel traitement unifié pour les dérivations en -isme et en -iste ?. In M. Roché, G. Boyé, N. Hathout, S. Lignon et M. Plénat, *Des unités morphologiques au lexique*, Hermes-Lavoisier, Paris, 69-143.

Švedova, N. 1980. *Russkaja grammatika*. Moskva: Nauka.

Thuilier, J. 2012 *Contraintes préférentielles et ordre des mots en français*. PhD thesis, Université Paris-Diderot-Paris VII.

ALLER et MOURIR oddities in French conjugation: *il a été au spectacle à pied, il a mouru d'ennui du début à la fin*

Gilles Boyé
Université Bordeaux-Montaigne, CNRS CLLE

French verbal inflection presents classical cases of deviation from canonical inflection (Corbett, 2005): suppletion (Corbett, 2007), overabundance (Thornton, 2019) and defectiveness (Sims, 2015). In this paper, we look at some oddities in the paradigms of ALLER 'to go' and MOURIR 'to die' and the interactions between defectiveness, suppletion and overabundance and their consequences for morphological theory.

1 The data

At first glance, ALLER 'to go' is not defective and it is a straightforward example of suppletion with both idiosyncratic stems (e.g. *i* for the future and conditional forms: *i-rai, i-ras, ...*) and forms (e.g. *vont* for the present 3rd person plural). It does not display the types of overabundance of POUVOIR 'can/may' (*je peux/puis*) limited to one form of a lexeme, of ASSEOIR 'to sit' (*je m'asseois/assieds, nous nous asseyons/assoyons* limited to some stems of a lexeme or that of the verbs in *-ayer* with fluctuating yods such as BALAYER 'to sweep' (*je balaie/balaye*).

ALLER	1sg	2sg	3sg	1pl	2pl	3pl
ind prs	va	va	va	alɔ̃	ale	**vɔ̃**
ipfv	alɛ	alɛ	alɛ	aljɔ̃	alje	alɛ
pst	ale	ala	ala	alam	alat	alɛʁ
fut	**i**ʁe	**i**ʁa	**i**ʁa	**i**ʁɔ̃	**i**ʁe	**i**ʁɔ̃
sbjv prs	aj	aj	aj	aljɔ̃	alje	aj
ipfv	alas	alas	ala	alasjɔ̃	alasje	alas
cond prs	**i**ʁɛ	**i**ʁɛ	**i**ʁɛ	**i**ʁjɔ̃	**i**ʁje	**i**ʁɛ
imp prs	—	va	—	alɔ̃	ale	—
	inf	prs part	pst part			
non-fini	ale	alɑ̃	ale	ale	ale	ale

However, there are two observations that complicate the situation a little. First, there is an idiomatic use of S'EN ALLER 'to leave' whose paradigm does not follow from the combination of the French special clitics and the conjugation paradigm of ALLER. The examples in (1) are both attested in French with different sociolinguistic value.

(1) He left.
  a. %Il s'est en allé. (informal)
  b. %Il s'en est allé. (formal)

(1a) could be interpreted as *en* being reanalyzed as a prefix on ALLER in the informal variety but the imperative forms used by all speakers display the special clitic in its usual position which is not compatible with a verb S'ENALLER.

(2) Leave!
  a. va-t'en !
  b. *en vas-toi !

This seems a case of incomplete reanalysis rather than overabundance. But, in the formal register, there is evidence of suppletive overabundance: as shown below in (3), there are alternative forms for the simple past borrowed from the ÊTRE paradigm.

(3) He went looking for the police.
  a. Il s'en alla chercher la police. (neutral)
  b. %Il s'en fut chercher la police. (formal)

The forms in (3b) are rare but well established in French literature. This particular overabundance in the simple past does not spread to ALLER itself but ALLER also borrows forms of ÊTRE The use of *avoir été* (literraly 'to have been') in examples such as (5b) is stigmatized in formal French invoking the fact that ÊTRE is not a verb of movement.

(4) He goes to the station. (movement)
  a. Il va à la gare.
  b. *Il est à la gare.

(5) He went to the station. (movement)
  a. Il est allé à la gare.
  b. Il a été à la gare.

But the expressions in (4) have completely different meanings[1], while those in (5) are synonymous.[2] Moreover, *avoir été* is compatible with a movement specific complement unlike *être*:

(6) He goes from London to York.
  a. Il va de Londres à York.
  b. *Il est de Londres à York.

(7) He went from London to York.
  a. Il est allé de Londres à York.
  b. Il a été de Londres à York.

We see this as a suppletion in the paradigm of ALLER, creating overabundance in the compound tenses of the movement verb. Of course, ALLER also functions as an auxiliary for near future in the present and the imperfective indicative making it a defective auxiliary compared to ÊTRE and AVOIR, which combine with all simple tenses.

While the preceding phenomenon focuses on two extremely frequent verbs (ÊTRE 'to be' 32000 occ/Mw, ALLER 'to go' 10000 occ/Mw), the case of MOURIR 'to die' is different, centering on marginal data concerning a less frequent verb (920 occ/Mw). The inflection paradigm of MOURIR as represented in references such as the Bescherelle (Arrivé, 1997) is neither suppletive, nor overabundant, nor defective. Despite this display of apparent simplicity, two complications arise.

[1] *Il est à la gare* is correct with the interpretation *he is at the station* using the verb ÊTRE.

[2] *il a été à la gare* could also be understood as ÊTRE and mean *he has been at the station (at some point)* in some specific contexts.

MOURIR	1sg	2sg	3sg	1pl	2pl	3pl
ind prs	mœʁ	mœʁ	mœʁ	muʁɔ̃	muʁe	mœʁ
ipfv	muʁɛ	muʁɛ	muʁɛ	muʁjɔ̃	muʁje	muʁɛ
pst	muʁy	muʁy	muʁy	muʁym	muʁyt	muʁyʁ
fut	muʁʁe	muʁʁa	muʁʁa	muʁʁɔ̃	muʁʁe	muʁʁɔ̃
sbjv prs	mœʁ	mœʁ	mœʁ	muʁjɔ̃	muʁje	mœʁ
ipfv	muʁys	muʁys	muʁy	muʁysjɔ̃	muʁysje	muʁys
cond prs	muʁʁɛ	muʁʁɛ	muʁʁɛ	mu**ʁʁj**ɔ̃	mu**ʁʁj**e	muʁʁɛ
imp prs	—	mœʁ	—	muʁɔ̃	muʁe	—
	inf	prs part	pst part			
non-fini	muʁiʁ	muʁɑ̃	mɔʁ	mɔʁ	mɔʁt	mɔʁt

The simpler one is that the prescribed forms of the 1st and 2nd person plural of conditional present *nous mourrions* and *vous mourriez* have a phonological make-up that poses a problem for French phonology. In the future and the conditional present, the double *r* must be pronounced as a geminate contrary to what usually occurs in French, where a double *r* can be pronounced either as a simple *r* or a geminate. In practice, the forms of indicative imperfective and conditional present offer a contrast:[3]

(8) *mourrait* conditional present vs *mourait* indicative imperfective

a. il mourait : muʁɛ/*muʁʁɛ

b. il mourrait : *muʁɛ/muʁʁɛ

Because this constraint on the double *r* is so limited, a context where a double *r* is followed by a yod appears only in the 1st and 2nd person plural of conditional present of these few verbs. With this distribution of data, most speakers feel uneasy speaking out the forms leading to competition between forced articulation, alternative repairs and avoidance.

(9) We would die.

a. nous [muʁʁjɔ̃] (forced pronunciation)

b. nous [muʁəʁjɔ̃] (schwa insertion, analogy on BOURRER ‘to stuff’)

c. nous [muʁiʁjɔ̃] (infinitive stem borrowing, analogy on PARTIR ‘to leave’)

d. avoidance defectiveness

Another complication appears in a completely different context with the past participle of MOURIR. Like ALLER, MOURIR has a pronominal form SE MOURIR with a different aspectual value. Where MOURIR ‘to die’ is an achievement, SE MOURIR ‘to be dying’ is an activity:

(10) MOURIR (achievement)

a. ‘he dies’

i. Il meurt.

b. ‘he died’

i. Il est mort. (standard French)

ii. ??Il est mouru. (playful joke)

iii. ??Il a mouru. (child’s mistake)

[3] This is also the case for verbs of the inflectional classes of COURIR ‘to run’ and ACQUÉRIR ‘to acquire’

(11) SE MOURIR (activity)

a. ‘he is dying of shame’
   i. Il se meurt de honte.

b. ‘he was dying of shame’
   i. *Il s’est mort de honte.
   ii. %Il s’est mouru de honte.
   iii. avoidance defectiveness

In the case of the activity, the *mort* past participle seems to be strongly associated with a state and cannot be used for the activity. The remaining choice is between *mouru* as an “overabundant” participle or avoidance. For the speakers using *mouru*, it seems that the activity past participle extends to MOURIR in contexts with the same activity reading:

(12) MOURIR: ‘he was dying of shame during the whole evening’ (activity)
   a. Il mourait de honte pendant toute la soirée. (indicative imperfective)
   b. %Il a mouru de honte pendant toute la soirée. (compound past)

2 Analysis outline

The data presented shows cases of suppletion, overabundance and defectiveness but in various ways depending on the frequency of the paradigm cells targeted.

The MOURIR phonological difficulty with *mourrions, mourriez* stands somehow outside the domain of morphology along with the defectiveness of the *-eur/-rice* derivations on stems ending in *s* (e.g. prédécesseur/*prédécessrice ‘predecessor’) but in a slightly different way. The morphological output is just as clear but the phonological trap it falls in is different. While there is a clear constraint for the avoidance of *sr* in derived words, the rarity of the double *r* plus yod summons an insecurity for the resolution and the possible avoidance of conflictual decisions, which leads to alternate forms or defectiveness. For the S’EN ALLER with the quasi-prefix, there would a simple solution extending the paradigmatic analysis of Bonami & Boyé (2007) to the pronominal verb S’EN ALLER and to include *va-t’en, allons-nous-en, allez-vous-en* in the portemanteaux rules of their PFM analysis (Stump, 2001) given the high frequency of the imperative forms.

The most interesting cases lie with the compound tenses and the alternate past participles. In the case of ALLER compound tenses (*être allé* vs *avoir été*), one could suggest a defective alternate verb ÊTRE, synonym with ALLER and restricted to compound tenses, avoiding overabundance in the spirit of Acquaviva (2008), but we would rather treat *été* as an overabundant suppletive past participle for ALLER along the lines defended by Thornton (2018). The same would apply to the activity MOURIR/SE MOURIR licensing a special form of the past participle, either no past participle or *mouru* in place of *mort*. The frequencies of *été* as the past participle of ALLER and the potential for the use of (SE) MOURIR as an activity are vastly different, which could explain in turn the disparity between the suppletion in one case and the insecurity in the other.

Both these cases would be similar to the *hay* form of HABER ‘to have’ in Spanish, which replaces the standard *ha* (present 3SG) when it is used as an existential. All these cases could be analyzed with pairs of lexemes sharing almost the same flexeme in the sense of Fradin & Kerleroux (2003). To account for these special variations, the flexeme would have to accomodate localised variation for the realization of a stem and some of its selection properties like in the cases of both ALLER and MOURIR, the alternate past participle has to use a different auxiliary to form its compound tenses.

References

Acquaviva, Paolo. 2008. *Lexical plurals: A morphosemantic approach.* Oxford University Press.

Arrivé, Michel. 1997. *La conjugaison pour tous* Bescherelle. Hatier.

Bonami, Olivier & Gilles Boyé. 2007. French pronominal clitics and the design of paradigm function morphology. In Geert Booj, Luca Ducceschi, Bernard Fradin, Emiliano Guevara, Angela Ralli & Sergio Scalise (eds.), *On-line proceedings fo the fith mediterranean morphology meeting*, 291–322. Bologna: Università degli Studi di Bologna.

Corbett, Greville G. 2005. The canonical approach in typology. In Zygmunt Frajzyngier, Adam Hodges & Rood David S. (eds.), *Linguistic diversity and language theories*, vol. 25, 25–49. Amsterdam/Philadelphia: John Benjamins Publishing.

Corbett, Greville G. 2007. Canonical typology, suppletion and possible words. *Language* 83(1). 8–42.

Fradin, Bernard & Françoise Kerleroux. 2003. Troubles with lexemes. In Geert Booij, J de Cesaris, Sergio Scalise & Angela Ralli (eds.), *Topics in morphology. selected papers from the third mediterranean morphology meeting*, 177–196. Barcelona: ULA-Universitat Pompeu Fabra.

Sims, Andrea D. 2015. *Inflectional defectiveness*, vol. 148. Cambridge University Press.

Stump, Gregory T. 2001. *Inflectional morphology. A theory of paradigm structure.* Cambridge: Cambridge University Press.

Thornton, Anna M. 2018. Troubles with flexemes. In Olivier Bonami, Gilles Boyé, Georgette Dal, Hélène Giraudo & Fiammetta Namer (eds.), *The lexeme in descriptive and theoretical morphology*, 303–321. Berlin: Language Science Press.

Thornton, Anna M. 2019. Overabundance: a canonical typology. In *Competition in inflection and word-formation*, 223–258. Springer.

Two ways to nominalize in Kaqchikel

Irina Burukina
Research Centre for Linguistics ELKH &
Eötvös Loránd University

1 Overview

The paper examines deverbal nominalization patterns in the Patzún variety of Kaqchikel (PK; K = other dialects as documented in grammars; Mayan, Guatemala; ergative, VOS). Adopting Moulton's (2014) analysis for nominalization in English in terms of existential closure that builds upon the Grimshaw (1990), we argue that, where in English the same morpheme can be used to create result and event nominals, Kaqchikel disambiguates between the two uses: the nominalizer -V*n/m* existentially closes the event argument in result nominals and the nominalizer -V*y/j* existentially closes the internal argument in event nominals. We further show that the two nominalizers can exceptionally be combined to create a nominal predicate that is used in the periphrastic perfective construction. The paper contributes to the discussion of nominalization in Mayan languages, continuing the line of research in Coon (2013), Imanishi (2020), Coon and Royer (2020), and Burukina (2021), i.a.

2 Theoretical background

Adhering to the Distributed Morphology framework and following Grimshaw's (1990) classification for deverbal nominals, Moulton (2014) proposes that a nominalizer head (i) selects either a larger Aspect/Event phrase or just a root projection as its complement, and (ii) existentially closes either the internal argument or the event argument within it, which results in an event/result nominal, respectively. In English, the same nominalizer often performs both functions, cf. *assign**ment** of problems in an hour* (event) vs. *the assign**ment** is on the table* (result).

3 Result vs. Event nominalization in Kaqchikel

3.1 The patterns of derivation

In Kaqchikel, one productive pattern of deverbal derivation is **-V*n/m*** (PK/K) nominalization. -V*n/m* nominalizer is used exclusively to create result nominals out of transitive verbs (1).[1]

(1) *Result nominals used as arguments*

x-Ø-qa-tz'ët	ri	(oxi')	ru-loq'-**on**	ri	Maria.
CMP-ABS3S-ERG1P-see	DET	three	ERG3S-buy-NMZ	DET	Maria

'We saw the (three) thing(s) that Maria had bought.'

We argue that nouns such as *-loq'on* in (1) are derived when a root (*-loq'*) that has an internal argument (IntA) and an event argument is combined syntactically with a nominalizer -V*n*/*m*. The main function of the nominalizer -V*n*/*m* is to existentially close the event argument giving rise to the interpretation 'x such as there was an event of (buy)ing it.' This is schematized in (2); for the sake of simplicity we mark the root projection as VP.

(2) *Existential closure of an **event** argument* (adapting Moulton 2014)

a. $[\![[n\ \exists]]\!] = \lambda P_{<e<s,t>>}.\lambda x.\exists e[P(x)(e)]$

b. $[_{nP}\ \text{n -V}\underbrace{\textit{n/m}\ [_{VP}\ \text{Root}}_{\exists}\ \text{IntA.variable}]]$

[1] The data discussed in the paper come from the variety of Kaqchikel spoken in Patzún, Guatemala, and were collected during the author's fieldwork in 2018 – 2020. We would like to thank Irma Yolanda Raquec Teleguario, Filiberto Patal Majzul, Gloria Imelda Ixim, Gloria Esmeralda Canú Chirix, and Celsa Teleguario for sharing with us their knowledge of the language. The research is supported by the Hungarian National Research, Development and Innovation Office under the grant NKFI 129921, and by Jacobs Research Funds grants (2018 – 2020).

The argumental -V*n/m* nominals normally appear with a determiner; we assume that the addition of a determiner renders the nominal type *e*. The nominalizer **-V*y/j***, in contrast, creates simple event nouns: -*loq'* 'buy' → *loq'oj* 'act of buying' (3).

(3) *Event nominals used as arguments*

n-Ø-qa-rayi-j	ri	loq'-**oj**	pa	ka'i'	ramaj.
ICMP-ABS3S-ERG1P-desire-DTV	DET	buy-NMZ	in	two	time

'We want to buy something at two o'clock.'

Following Moulton (2014), we assume that the role of -V*y/j* is to existentially close the IntA variable (4); cf. a similar analysis recently proposed for agentive nominals in Chuj by Coon and Royer (2020).

(4) *Existential closure of an* ***internal*** *argument* (adapting Moulton 2014)

a. $[\![[n\ \exists]]\!] = \lambda P_{<e<s,t>>}.\lambda e.\exists x[P(x)(e)]$

b. $[_{nP}$ n -V*y/j* $[_{VP}$ Root IntA.variable]]
∃

In the full version of the paper we will show the results for event vs. result diagnostics applied to -V*y/j* vs. -V*n/m* nominals.

3.2 External arguments

Assuming that external arguments (ExtA) are projected by a separate head, v, we further propose that in Kaqchikel, similarly to English, the same nominalizer can combine with structures of different sizes: VP or vP. There is, however, no case available within the verbal part: absolutive is uniformly assigned by Infl (Coon et al. 2014), while ergative is assigned by transitive Voice to an ExtA in Spec,vP or by Poss to a possessor in Spec,nP. Hence, the ExtA within a nominal cannot be a referential DP; thus, adding -V*n/m* or -V*y/j* on top of a vP leaves us with two variables. In this case, the ExtA variable (PRO) can be controlled by a possessor merged above the n head, as in ***nu****-loq'oj* 'my act of buying' (5); see Imanishi (2020) and Burukina (2021) discussing control approaches to nominalization in Kaqchikel.

(5) *Controlling the external argument inside nominals*

$[_{PossP}$ DP$_i$ $[_{Poss'}$ Poss $[_{nP}$ **-V*y/j*** $[_{vP}$ PRO$_i$ $[_{v'}$ v $[_{VP}$ Root IntA.variable]]]]]]
∃

4 Deverbal nominals used predicatively

-V*n/m* nominals can also be used predicatively due to the presence of an internal argument variable; compare (6a) and (6b) to (6c), which involves a non-verbal adjectival predicate. In this case, they must be determiner-less and the derived nP is directly predicated of the DP merged in Spec,PredP; the derivation is schematized in (7).

(6) *Result nominals used as predicates*

a.	(röj)	e-qa-loq'-**on**	rije'.
	we	ABS3P-ERG1P-buy-NMZ	they

'We have bought them.'
Literally: 'They are the result of (our) buying.'

b.	(rije')	e-loq'-**on**.
	they	ABS3P-buy-NMZ

'They have been bought.'

c.	(rije')	e-tz'uyül
	they	ABS3P-seated

'They are seated.'

(7) *Non-verbal predication*

$[_{PredP}$ *rije'* $[_{Pred'}$ Pred^0 $[_{nP}$ n **-V*n/m*** $[_{VP}$ *loq'*]]]]

BACKGROUND ASSUMPTIONS behind (7): (i) Subjects of non-verbal predicates are projected as external arguments, in Spec,PredP (Levin et al. 2021); (ii) non-verbal predication is stative and incompatible with (in)completive aspect morphology (Coon and Preminger 2009); (iii) there is no overt copula in Kaqchikel (Patal Majtzul et al. 2000).

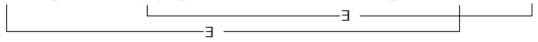
A -V*y/j* nP can also be used predicatively in such periphrastic perfective constructions (8a); however, in these cases -V*y/j* must be combined with the -V*n/m* nominalizer. First, -V*y/j* selects a vP, closing the IntA varible. Second, -V*n/m* nominalizer is added to existentially close the event argument

(8b). The result gets the reading ‘y such as there was an event of (buy)ing something by them’; this nominal is then predicated of a DP argument, in parallel to (7) .

(8) *Combining the two nominalizers*

a. (röj) oj-loq’-**oy-on**.
we ABS1P-buy-NMZ-NMZ
‘We have bought something.’ (Literally: ‘We are buyers of something.’)

b. [$_{\text{nP}}$ n -V*n/m* [$_{\text{nP}}$ n -V*y/j* [$_{\text{vP}}$ ExtA.variable [$_{\text{VP}}$ Root IntA.variable]]]]



Constructions such as (6/8) are literally translated as ‘they are [(our) result of (buy)ing]’ and ’we are [buyers of something]’, and are normally interpreted as perfective ’we have bought them/something’: they can be combined with *already*-type modifiers and cannot be continued with ‘… but we didn’t finish’ or ‘… and we are still doing that’. Analysing perfective forms as nominal further allows us to draw a parallel between perfective and progressive in Mayan languages; see Laka (2006), Coon (2013), and Imanishi (2020) on progressive in Mayan involving eventive nominal predicates.

References

Burukina, Irina. 2021. On the nature of arguments in event nominals. *Proceedings of the Linguistic Society of America* 6(1). 996–1008.

Coon, Jessica. 2013. *Aspects of split ergativity*. Oxford: Oxford University Press.

Coon, Jessica, Pedro Mateo Pedro & Omer Preminger. 2014. The role of case in A-bar extraction asymmetries: Evidence from Mayan. *Linguistic Variation* 14(2). 179–242.

Coon, Jessica & Justin Royer. 2020. Nominalization and selection in two Mayan languages. In Artemis Alexiadou and Hagit Borer (eds.), *Nominalization: 50 Years on from Chomsky’s Remarks*, 139–169. Oxford: Oxford University Press.

García Matzar, L. P. & J. O. Rodríguez Guaján. 1997. *Rukemik ri Kaqchikel Chi’: Gramática Kaqchikel*. Editorial Cholsamaj.

Imanishi, Yusuke. 2020. Parameterizing split ergativity in Mayan. *Natural Language & Linguistic Theory* 38. 151–200.

Levin, Theodore, Paulina Lyskawa & Rodrigo Ranero. 2021. Optional agreement in Santiago Tz’utujil Mayan is syntactic. *Zeitschrift für Sprachwissenschaft* 39(3). 329–355.

Moulton, Keir. 2014. Simple event nominalization. In Ileana Paul (ed.), *Cross-linguistic investigations of nominalization patterns*. 119–144. Amsterdam: John Benjamins.

Patal Majtzul, Filiberto. 2007. *Rusoltzil ri Kaqchikel: Diccionario bilingüe estándar Kaqchikel illustrado*. OKMA.

Patal Majzul, F., L. P. García Matzar & C. I. Espantzay Serech. 2000. *Rujunamaxik Ri Kaqchikel Chi’: Variación Dialectal En Kaqchikel*. Editorial Cholsamaj.

Spare us the surprise: the interplay of paradigmatic predictability and frequency

Maria Copot — Université de Paris
Olivier Bonami — Université de Paris

1 Background

There is a generally recognised relationship between frequency and the paradigmatic predictability of a word form: word forms that are paradigmatically unpredictable (such as suppletive or otherwise highly irregular forms) tend to be frequent, while infrequent word forms tend to be highly paradigmatically predictable. In other words, it is within very frequent lexemes or very frequent paradigm cells that we tend to find unpredictable word forms. When unpredictability of form is encountered outside these contexts, there is a diachronic push to regularise it (the praeterite of English HELP used to be *holp*, now regularised to *helped*), or for the whole context to fall out of use (see the ongoing decline of the Italian *passato remoto*).

This relationship is rooted in the communicative function of language, and the way this interacts with memory: the more high-frequency a syntactic word is, the more it can afford to have an unpredictable form, because its frequency ensures that its phonological form is highly active in memory and thus easily accessible. On the flip side, low frequency words are more likely to be easily predictable from other members of the paradigm: if a word is already syntagmatically uncertain (low-frequency words are tautologically an unexpected way to continue the average utterance), it's unlikely to tolerate additional uncertainty on the paradigmatic axis (Filipović Đurđević & Milin, 2019).

Syntagmatic predictability is well known to facilitate access during language use (Hale, 2001; Levy, 2008; Frank, 2013). Less is known about the role of word form predictability (though see Milin et al. (2009) for an overview from a paradigmatic perspective). We set out to test the hypothesis that, at parity of lexeme frequency, less paradigmatically predictable word forms will be used less frequently, as they are less easily accessible than their counterparts (a host of causal factors can be invoked here: for example, less predictable forms are a case of rare behaviour, so their neighbourhood will be less populated, making access more difficult). This general tendency is predicted to be absent for very high lexeme frequency: as very frequent lexemes are overall highly accessible, so are their individual phonological word forms.

2 Motivation

With the goal of better understanding the relationship between frequency and predictability, we perform a corpus study in which we attempt to predict the frequency of a word type based on the frequency of the lexeme it belongs to, and its paradigmatic predictability. We hope to provide an operationalisation of form predictability that is

- empirical: it is derived bottom-up from morphological data, and lines up with how predictability is characterised in other domains of language.
- paradigmatic: it makes use of the paradigmatic structure of morphological data, in a way that emulates emerging evidence about how speakers exploit said paradigmatic structure in language use.

- continuous (as a corollary property, falling out from the other two). We therefore make the falsifiable prediction that we don't expect the effect of predictability to be categorical.

The hypothesis we wish to test is that at parity of lexeme frequency, words that are less paradigmatically predictable will be used less frequently, since they are more difficult to access than their predictable counterparts, due to the low type frequency of the pattern they instantiate. Because we are investigating the effect of predictability and lexeme frequency at the level of individual words, we expect that predictability will be weighted differently at different levels of lexeme frequency: if the overall frequency of a lexeme is high, then its predictability will matter less – since frequent word forms have their own representation in the mental lexicon, the speaker does not need to predict them in order to use them, but rather they just need to retrieve them from memory. We expect all these to show up as gradient effects, partly due to the impact of all sorts of other factors on the accessibility of mental representations, but chiefly because we believe the effect to truly be gradient.

3 Operationalising Predictability

We adopt an information-theoretic view of paradigmatic predictability (Ackerman et al., 2009), whereby a word is predictable inasmuch as its shape is unsurprising given the rest of its paradigms and the distribution of inflectional patterns in the language. More precisely, we use the Qumín package (Beniamine, 2018) to identify, for all pairs of words (w_1, w_2) filling the same two paradigm cells (c_1, c_2), the alternation pattern relating these two cells. From this we can estimate the conditional probability of the word in c_2 having the shape w_2 given that the word in c_1 has the shape w_1, based on the statistical distribution of patterns relating c_1 and c_2. The PARADIGMATIC SURPRISAL of w_2 in c_2 given knowledge of c_1 is the negative logarithm of this conditional probability: the more frequent the pattern relating w_1 and w_2 is among viable alternatives, the lower the surprisal. We use the paradigmatic surprisal of a word form filling a paradigm cell given knowledge averaged over all possible predictors as our estimation of paradigmatic predictability.

In the present case study on French, all calculations rely on applying Qumín to the full paradigm of the 4951 nondefective verbs in the Flexique database (Bonami et al., 2014).

4 Surprisal and frequency

In order to investigate the relationship between form predictability (operationalised as paradigmatic surprisal) and frequency (at the level of the lexeme, the cell, and the word form), we perform a corpus study on the French verbal system. For frequency data, we extracted word form and lexeme frequency from FrCoW (Schäfer & Bildhauer, 2012). Whenever lexeme annotations were missing, we converted the token into the most appropriate lexeme given the POS tag using Levenshtein distance.

Because French conjugation exhibits widespread syncretism, for many paradigm cells, it is not possible to estimate frequency reliably. We hence decided to focus on those cells where a sizeable portion of the lexicon (at least 250 lexemes) uses a form with no homograph documented in the GLÀFF (Hathout et al., 2014)). We also excluded cells out of current usage such as the past subjunctive, for which attestations might be archaic or ironic. After this filtering, 14 cells are left for modeling. Separate bayesian poisson models were fitted to each cell, each predicting token frequency based on lexeme frequency, average surprisal, and their interaction. For the reasons discussed at the start of the section, we predict

- lexeme frequency to always have a positive coefficient - tautologically the more frequent a lexeme, the more frequent the words that belong to it
- surprisal to always have a negative coefficient - once lexeme frequency is taken into account, words that are harder to predict should be used less.
- the interaction coefficient should have a positive sign - we expect that for high values of lexeme frequency, surprisal should progressively matter less, since the language user's task is to remember the form rather than predicting it.

5 Results

These predictions are largely borne out. Three cells are exceptional: the INFINITIVE, the PRESENT PARTICIPLE and the IMPERFECT 3SG. For these cells, at least one of the coefficients involving surprisal is either very small and of unexpected monotonicity, or with an effect indistinguisheable from 0. Importantly, this exceptional behaviour is attested in what are the three most frequent cells under consideration. We hypothesise that because these cells are so frequent, the decreasing importance of surprisal doesn't just hold for the most frequent lexemes but rather it applies to most items in the entire cell: because the cell (and therefore words within it) is so frequent, its word forms are easily accessible directly, which diminishes the speaker's reliance on deducing it based on other forms of its paradigm.

Cell	**Lexeme freq.**	**Surprisal**	**Interaction**
FUT.1SG	0.9935	–0.3783	0.0675
FUT.2SG	1.0771	–0.2306	0.0447
FUT.3SG	1.1764	–0.0261	0.0073
FUT.1PL	0.9693	–0.1932	0.0415
FUT.2PL	1.1072	–0.3368	0.0647
FUT.3PL	1.1466	–0.0040	0.0088
COND.3SG	1.2509	–1.0392	0.1835
COND.1PL	1.2544	–1.7739	0.2876
COND.2PL	1.2583	–2.7622	0.4486
COND.3PL	1.2312	–1.3889	0.2404
IPFV.3SG	1.1707	–0.0441	–0.0010
IPFV.3PL	0.9352	–0.5588	0.0959
PRS.PTCP	0.5916	0.0545	0.0053
INF	0.9438	0.0620	–0.0089

🟥 Unexpected coefficient sign
🟦 95% Credible interval overlaps with zero

Coefficient values by cell

6 Conclusion

This work proposes an operationalisation of form predictability that is empirical, gradient and inherently paradigmatic. In the corpus study described, paradigmatic surprisal appears to capture well language users' reticence to employ forms that are hard to predict at parity of lexeme frequency. The study also provides insight into the relationship between form predictability and frequency: for very frequent lexemes and paradigm cells, form predictability matters progressively less to the language user since the frequent word form, no matter how unpredictable, already has a representation in memory and does not need to rely chiefly on being deduced based on paradigmatic information.

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.

Beniamine, Sacha. 2018. *Typologie quantitative des systèmes de classes flexionnelles*: Université Paris Diderot dissertation.

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.

Filipović Đurđević, Dušica & Petar Milin. 2019. Information and learning in processing adjective inflection. *Cortex* 116. 209–227. doi:https://doi.org/10.1016/j.cortex.2018.07.020. `https://www.sciencedirect.com/science/article/pii/S0010945218302375`. Structure in words: the present and future of morphological processing in a multidisciplinary perspective.

Frank, Stefan L. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science* 5(3). 475–494. doi:https://doi.org/10.1111/tops.12025. `https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12025`.

Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second meeting of the north American chapter of the association for computational linguistics*, `https://www.aclweb.org/anthology/N01-1021`.

Hathout, Nabil, Franck Sajous & Basilio Calderone. 2014. GLÀFF, a Large Versatile French Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177. doi:https://doi.org/10.1016/j.cognition.2007.05.006. `https://www.sciencedirect.com/science/article/pii/S0010027707001436`.

Milin, Petar, Dusica Filipovic Durdevic & Fermin Moscoso del Prado Martin. 2009. The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from serbian. *Journal of Memory and Language* 60. doi:10.1016/j.jml.2008.08.007.

Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the eighth international conference on language resources and evaluation*, 486–493.

Positional competition in Murrinh-Patha by rule composition

Berthold Crysmann
CNRS, Laboratoire de linguistique formelle

In this talk, I shall address positional competition between subject and object agreement markers in Murrinh-Patha, a polysynthetic Non-Pama-Nyungan language of Australia. The data discussed here are taken from Nordlinger (2010, 2015).

Verbs in Murrinh-Patha minimally consist of a lexical stem (open class) and a classifier stem (CS) from a set of 38 classifier stem paradigms. Together, these two stems express basic lexical meaning. While the lexical stem (in slot 5) is uninflected, the classifier stem (in slot 1) differentiates TAM as well as subject agreement.

1	2	3	4	5	6	7	8	9
CS.SUBJ.TAM	SUBJ/OBJ NUM	RR	IBP	**LEX-STEM**	TAM	ADV	SUBJ/OBJ NUM	ADV

Figure 1: Murrinh-Patha position classes (Nordlinger, 2015)

In addition to inflection by means of the classifier stem, Murrinh-Patha verbs are inflected with a number of discrete markers, organised into a positional template, as shown in Figure 1. Of particular interest for this paper are slots 2 and 8, where exponents of subject and object agreement can be found.

Agreement marking operates along up to four inflectional dimensions (illustrated by the paradigm of object agreement markers in Table 1): the language distinguishes four numbers (singular, dual, paucal, plural) and three persons, including a distinction between inclusive and exclusive for first person non-singular cells. Additionally, Murrinh-Patha marks a rather unique category of non-sibling in the dual and the paucal. Exponents of this category are differentiated for gender, which is otherwise not marked in the verb. Furthermore, the paucal is only distinguished for non-siblings. With siblings, paucal and plural are non-distinct. Another peculiarity of the non-sibling marker pertains to its morphotactics: while all other exponents of object agreement surface in slot two, the dual and paucal non-sibling markers are realised discontinuously in slot 8 (in the case of direct object agreement).

			1		2	3
			INCL	EXCL		
SG			N/A	-ngi	-nhi	∅
DU	NSIB	M	-nhi	-nganku + nintha	-nanku + nintha	-(pu)nku + nintha
		F	-nhi	-nganku + ngintha	-nanku + ngintha	-(pu)nku + ngintha
	SIB		-nhi	-nganku	-nanku	-(pu)nku
PC	NSIB	M	-nhi + neme	-nganku + neme	-nanku + neme	-(pu)nku + neme
		F	-nhi + ngime	-nganku + ngime	-nanku + ngime	-(pu)nku + ngime
	SIB		-nhi	-ngan	-nan	-(pu)n
PL			-nhi	-ngan	-nan	-(pu)n

Table 1: Object agreement markers

Subject agreement (cf. Table 2) is quite similar to object agreement, despite the difference in exponence: while object agreement is realised by discrete markers in slots 2 and 8, subject agreement is realised fusionally as part of the classifier stem (slot 1) plus discrete markers for non-sibling (slot 2/8) and for the non-future dual (slot 2). Another difference pertains to dual non-sibling marking: with direct object markers, the person/number exponent (slot 2) is syncretic with the person/number exponent of the sibling dual, whereas for subjects the classifier stem is syncretic with the singular.

			1		2	3
			INCL	EXCL		
SG			N/A	bam	dam	bam
DU	NSIB	M	thubam	bam + nintha	dam + nintha	bam + nintha
		F	thubam	bam + ngintha	dam + ngintha	bam + ngintha
	SIB		thubam	ngubam + ka	nubam + ka	pubam + ka
PC	NSIB	M	thubam + neme	ngubam + ka + neme	nubam + ka + neme	pubam + ka + neme
		F	thubam + ngime	ngubam + ka + ngime	nubam + ka + ngime	pubam + ka + ngime
	SIB		thubam	ngubam	nubam	pubam
PL			thubam	ngubam	nubam	pubam

Table 2: Subject agreement (non-future sub-paradigm for classifier stem *see(13)*)

1 Positional competition

As discussed above (cf. also Figure 1), the positions for the affixal markers of subject agreement overlap with those for object marking, so the central question is to how conflict is actually resolved. Murrinh-Patha witnesses two strategies: displacement of the subject marker, and omission.

The first case of positional competition relates to the subject non-sibling markers *nintha/ngintha*. When marking subject agreement, these markers surface in slot 2, if available, i.e. before the lexical stem.[1]

(1) bam- **-ngintha-** ngkardu
SUBJ.3.SG-CS.SEE(13).NFUT NON-SIB.F.DU see
'They (dual non-sibling) saw him/her.' (Nordlinger, 2010)

However, if object agreement is overtly realised (any cell other than 3rd singular), slot 2 receives the object person/number marker and the subject non-sibling dual marker must surface in slot 8 instead, i.e. after the lexical stem, cf. (2).

(2) bam- -ngi- ngkardu **-ngintha**
3.SUBJ.SG-CS.SEE(13).NFUT 1.SG.OBJ see SUBJ.DU.NON-SIB.F
'They (dual non-sibling) saw me.' (Nordlinger, 2010)

Given the fact that subject and object non-sibling markers are syncretic, and that object non-sibling markers are also realised in slot 8, non-sibling marking may end up ambiguous as to whether it refers to the subject or the object, cf. the examples from Nordlinger (2015) below.

(3) ma- -nanku- -rdarri- purl -nu- **-ngintha**
1.SUBJ.SG-CS.HANDS(8).FUT OBJ.2.DU/PC back wash FUT NON-SIB.F.DU
'I will wash your (female dual non-sibling) backs.' *or*
'We (two exclusive female non-sibling) will wash your (dual sibling) backs.'

[1] The paucal non-sibling marker *-neme/-ngime* are always realised in slot 8.

In (3), *ngintha* may either refer to the object, leaving subject agreement solely marked by the singular classifier stem, yielding singular. Alternatively, singular stem and dual non-sibling marker jointly express first person exclusive female non-sibling dual, leaving the object marker in slot 2 to express sibling dual.

What is important about realisation of the subject dual non-sibling markers is that realisation in slot 8 is only ever licit when slot 2 is blocked by another exponent. If slot 2 is free, subject *ngintha/nintha* must surface there.

The second case relates to the dual/paucal number marker *ka* which appears in slot 2 in the non-future, as shown in (4a,b) from Nordlinger (2010). Note that in the non-future, as opposed to other TAM categories, the dual and plural stems are syncretic.

(4) a. pubam- **-ka-** -ngkardu
3.DU/PL-CS.SEE(13).NFUT -DU/PC.NFUT see
'They (dual sibling) saw him/her.'

b. pubam- **-ka-** -ngkardu- -ngime
3.DU/PL-CS.SEE(13).NFUT DU/PC.NFUT see PC.NON-SIB.F
'They (paucal, female, non-sibling) saw him/her.'

c. pubam- -nhi- -ngkardu
3.DU/PL-CS.SEE(13).NFUT 2.SG.O see
'They (two/paucal/plural siblings) saw him/her.'

d. pubam- -ngkardu
3.DU/PL-CS.SEE(13).NFUT see
'They (plural) saw him/her.'

Again, in the case of overt object marking (4c), subject marking in slot 2 becomes unavailable. In contrast to the dual non-sibling markers, there is no alternate realisation for *ka*, even if a suitable position (like slot 8) happens to be unoccupied. Instead *ka* is simply dropped, possibly leading to ambiguity between dual and plural, as shown in (4c). Note that without a competitor in slot 2, only a non-dual interpretation is possible (4d).

2 Realisational morphology

As argued by Nordlinger (2010), the high degree of overlapping exponence, involving discontinuous surface positions provides evidence against a morpheme-based view, favouring instead a templatic realisational perspective. Ever since Stump (1993), position class systems have provided core evidence for an inferential-realisational approach to morphology. However, no formal analysis has yet been developed for the data at hand. Indeed, positional competition in Murrinh-Patha poses some non-trivial challenges: rule block systems, such as PFM (Stump, 1993), ensure maximal independence of rules of exponence in different rule blocks, which is good for multiple exponence, but does not lend itself easily to capture the exclusive disjunction between subject sibling marking in slots 2 and 8. While an ambifixal rule block (Stump, 1993) may serve to capture the positional alternation of subject agreement as prefixation vs. suffixation to the lexical stem, it cannot capture the dependence on overt realisation of object agreement, which must be introduced in a different rule block. In more recent work, Stump (2017) proposes rule conflation as a means to compose complex realisation rules from more elementary building blocks. However, conflation is inherently constrained to strict adjacency. What we need, however, for Murrinh-Patha is the exact opposite, namely composition of rules in order to model the discontinuous dependency between exponents of subject agreement in slot 8 on the presence of exponents in slot 2.

I shall therefore build on Information-based Morphology (Crysmann & Bonami, 2016), an alternative approach to inferential-realisational morphology that represents position class information as a first class property of exponents, such that realisation rules can simultaneously license multiple, even discontinuous exponents. Furthermore, realisation rules are organised in a cross-classifying inheritance hierarchy, such that complex rules can be built from partial descriptions by means of unification.

The formal analysis I propose captures positional dependencies by means of complex rules that simultaneously license exponents for subject and object agreement in slots 1, 2 and 8. These rule constraints from which these complex rules are built are organised into three dimensions (INI, MID, FIN), such that each composed rule must inherit from exactly one rule constraint in each dimension. The first dimension (INI) consists of stem selection rules that introduce suitable classifier stems in slot 1, according to subject agreement and TAM specifications. The second (MID) and third dimension (FIN) jointly describe the range of affixal realisations for both subject and object agreement. The rule constraints in the FIN dimension describe the shape and position of the non-sibling markers, which are always final in this complex: while the paucal markers *neme/ngime* are restricted to slot 8, the dual markers *nintha/ngintha* are underspecified for their exact surface position. Yet, they do require that slot 2 be non-empty. Alongside these exponence constraints, there is a purely morphotactic rule constraint that captures the situation where no second affixal marker is present, restricting exponents to slots 1 and 2. The MID dimension finally provides exponence rule constraints for the initial affixal markers, including dual *ka* and the object person/number markers (*nhi*), which are all constrained to slot 2. However, these rules are open to combine with exponents contributed by the FIN dimension. In addition, the MID dimension provides two purely morphotactic constraints: one constraint that leaves slot 2 empty to license bare classifier stems (cf. (4d)) and paucal non-sibling subject markers (slot 8), and finally a constraint to receive a marker in slot 2. These constraints are actually sufficient to derive the distribution of non-sibling marking: while *neme/ngime* are always in slot 8, whether or not slot 2 is filled, *nintha/ngintha* surface in slot 8, if in combination, or in slot 2 when there is no other marker that can fill slot 2. Finally, the distribution of dual/paucal *ka* is governed by both Paninian and positional competition: if no object markers are present, the availability of a specific dual/paucal form, cf. (4a) restricts the bare dual/plural classifier to denote plural, cf. (4d). However, there is no more specific form that could preempt the *combination* of subject and object agreement, yielding (4c).

To conclude, the study of positional competition in polysynthetic languages like Murrinh-Patha highlights a basic requirement for realisational morphology: the possibility to compose rules of exponence and to be able to do so in a discontinuous fashion.

References

Crysmann, Berthold & Olivier Bonami. 2016. Variable morphotactics in Information-based Morphology. *Journal of Linguistics* 52(2). 311–374. doi:10.1017/S0022226715000018.

Nordlinger, Rachel. 2010. Verbal morphology in Murrinh-Patha: Evidence for templates. *Morphology* 20. 321–341.

Nordlinger, Rachel. 2015. Inflection in Murrinh-Patha. In Matthew Baerman (ed.), *Handbook of inflection*, Oxford: Oxford University Press.

Stump, Gregory T. 1993. Position classes and morphological theory. In Geert E. Booij & Jaap van Marle (eds.), *Yearbook of morphology 1992*, 129–180. Dordrecht: Kluwer.

Stump, Gregory T. 2017. Rule conflation in an inferential-realizational theory of morphotactics. *Acta Linguistica Academica* 64(1). 79–124.

Wao Terero lexical suffixes: Bridging the lexicon and discourse

Noah Diewald
Ohio State University

In this talk I discuss a formal treatment of a subcomponent of the Wao Terero lexical suffix system, adjectival classifiers. Wao Terero (Glottocode waor1240) is a linguistic isolate spoken in the Ecuadorian Amazon. Data is from ongoing fieldwork. Lexical suffixes are bound elements that provide nominal meanings to their host constructions (Sapir 1911). In the context of my fieldwork, I utilize a fragment methodology, where grammatical models encode hypotheses that can be tested through elicitation. To support this effort, I have developed Lexical Proof Morphology (LP), a theoretical framework embedded in a constructive logic (Coquand & Huet 1988). The theory is particularly concerned with interface issues and is compatible with the tenets of Word and Paradigm Morphology (WP) (Robins 1959). It seamlessly interfaces with existent, broad coverage theories of syntax (Pollard & Worth 2015) and dynamic semantics (Martin & Pollard 2012). This integration is necessary for a treatment of Wao Terero lexical suffixes. Lexical suffixes have predictable but polysemous lexical semantic meanings. Some may be used as classifiers, where they play a role in anaphoric constructions. In these cases, the polysemy offered by the lexicon must be narrowed to include only those qualities compatible with a referent. Therefore, the dynamic semantic context must limit lexical variability. At the same time, discourse provides another domain where classifier constructions receive multiple interpretations. These are driven by the dichotomy created by information introduction and information reference. This means that morphological forms sit at the intersection of multiple interpretations in two semantic domains. On the one hand, this relationship evokes classic realizational assumptions. On the other, as will be made clear, the dynamic context recruits lexical meaning into diverse interpretative contexts, such that realization also behaves as a conduit between two semantic domains. LP makes the relationship between intrinsic, lexical semantic meanings and extrinsic, dynamic meanings explicit using a unique, proof-theoretic realizational architecture.

The Wao Terero lexical suffix system is complex. There are roughly 35 suffixes. These may occur with nearly every part of speech – including demonstratives, nouns, adjectives, verbs, question words, and others. Each construction type comes with particular quirks and semi-productive nuance. For the sake of simplicity, I focus on productive, transparent adjectival constructions but it is helpful to get a taste of the broader system. As examples (1) and (2) show, where lexical suffix glosses are in bold, lexical suffixes may be used in constructions that have compound and incorporation-like meanings. This is despite the fact that the suffixes do not correspond to the stems of free words. Meanings may also be classifier-like (Peeke 1968), as in (2) and (3), depending on the construction type and context. Though the classifier constructions in these examples may resemble grammatical agreement, classifiers are optional. They are acceptable in many contexts but may occur only occasionally in spontaneous speech.

(1)	*kewe-ñabo*	(2)	*Onom-po*	*kem-po-tabopa.*	(3)	*Ñene-po*	*wipo*	*impa.*
	cassava-**leaf**		body-**hand**	cut-**hand**-1.past		big-**canoe**	canoe	copula
	‘cassava leaf’		‘I cut my hand.’			‘The canoe is big.’		

No comprehensive formal treatments of lexical suffixes exist in the literature but there are some outline proposals. Wiltschko (2009) sketches a Distributed Morphology (DM) (Halle & Marantz 1993) treatment of a similar system in a Salishan language. Her proposal takes advantage of the hybrid nature of DM to treat lexical suffixes in an item and arrangement manner. Specifically, the suffix and its host are roots below a root node: $[\sqrt{\text{root}}[\sqrt{\text{host}}][\sqrt{\text{suffix}}]]$. As (2) and (3) show, lexical suffixes do not inhabit syntactic argument positions, otherwise they would block the occurrence of overt nominal arguments. Placing lexical suffixes in sub-root non-argument positions allows for this quality of lexical suffix behavior according to DM-like assumptions.

Wiltschko's representation of lexical suffixes as roots in hierarchical configurations is inadequate for Wao Terero and likely other languages. There are two reasons for this. One is lexical semantic and the other is due to dynamic semantics. The lexical issue is easily explained. Wao Terero lexical suffix constructions are highly polysemous. The form-meaning correspondences of DM roots assume meaning underspecification, which allows for some quasi-polysemous meaning variation. Wao Terero patterns exhibit true polysemy, multiple meanings (Copestake & Briscoe 1995), rather than underspecification.

(4)	*ñene-we*	(5)	*ñene-mo*	(6)	*ñene-mpo*
	big-**plant/tree/pole**		big-**eye/fruit/face**		big-**canoe/hand/finger**
	'big (plant/tree/pole)'		'big (eye/fruit/face)'		'big (canoe/hand/finger)'

Polysemy is particularly clear in adjectival classifier constructions, where the lexical suffix imposes a selectional restriction on an argument, which may be explicit in syntax or supplied by discourse. This is the case in (4), (5) and (6), where the selectional restriction varies. This means that (6) is an appropriate answer to the questions, "What is her hand like?" or "What is her canoe like?" but not a question about a plant.

The second problem posed by a root-configuration approach requires consideration of the discourse context. The hierarchical root schema is far too simplistic to predict the diversity of lexical suffix construction interpretations in discourse, in particular their role in introducing information and their role as anaphora. Adjectival classifier constructions may play (at least) three discourse roles. In role 1 the construction modifies a noun when a referent is introduced into discourse (see (3)), analogous to *A short boat exists*. In both role 2 and 3, the construction is anaphoric. In role 2 only the adjectival information is new information, analogous to *The boat is short.* In role 3 both the adjectival and classifier content are part of the descriptive content of the anaphor, *It is the short boat.*

The diagnostic for establishing that adjectival constructions can serve each of these roles involves negation. In formal pragmatics it has been observed that the descriptive content of an anaphoric expression is not offered for acceptance or rejection in discourse (Roberts 2010). The interlocutors presuppose its validity. This means that if one negates an expression containing an anaphor, the descriptive content of that anaphor will not be negated. For example, the 'boat' meaning in the anaphoric expression *the boat* falls outside the scope of negation in *It isn't the boat.* One can felicitously follow up with *The boat is black.* This would not be the case in the non-anaphoric case *It isn't a boat.* Presenting a full paradigm of these diagnostics for Wao Terero constructions will not fit within the confines of this abstract. Role 2 is perhaps the most interesting case because the diagnostic predicts a split interpretation of meaning components, where the adjectival meaning is negated but not the classifier's descriptive content. This can be seen in (7), which is an answer to the question "Is the canoe short and red?" In it *okampo*, 'short', is negated, but not the descriptive content of the classifier.

(7)	*Obatawe*	*wii*	*oka-mpo*	*inamai*	*impa.*
	red	not	short-**canoe**	not	is

'It is red but not short.'

The proposal of a particular phrase-like structure cannot speak to these discourse phenomena. The DM proposal fails to predict anything like it, as would similar schemata proposed in popular construction morphologies (Booij 2010). Truly accounting for form-meaning correspondences requires a linkage to some engine of interpretation, such as dynamic semantics, where meanings are composed and their entailments can be verified.

LP is a multi-paradigm theory, roughly in the mold of (Sadler & Spencer 2001). A morphological paradigm space exists, which is a non-symmetrical taxonomic space of triples (mc, mf, lx), called form entries, where mc is a purely morphological category, called an m-cat, mf is a morphological form, and lx is a lexeme identifier. LP provides a declarative system for defining this space that does not make direct reference to syntactic or semantic categories. An mc tends to be named after formal (phonological) characteristics of a form. So ***ñenempo***, 'big (canoe, hand, fore-paw etc.)', has a category of po. This category may look redundant here but in inflectional systems with a high degree of allomorphy, these categories span diverse forms. It is also important to note that the category is for the whole of the form, not just the suffix. A form entry for ***ñenempo*** is $(po, \text{ñenempo}, \textsc{ñene})$. The lexical entries, called signs, of the system constitute the syntax-facing paradigm. There is no intermediary notion of cells but there is a family of relations between form-entries and signs called form-sign mappings that provide generalizations over paradigmatic structure. These are declarative, natural deduction-style rules. A simplified example is below. It lacks important but distracting technical details.

$$\frac{(mc, mf, lx) \qquad class(lx) \leq adj \qquad meaning(s, lx, mc)}{(mf, \text{RefAdj}, (\lambda P_1 P_2 . P_1(x)/P_2(x))(\pi_1 s)(\pi_2 s))}$$

This says that given a form-entry with a lexeme of a particular class – in this case adjectival – when there is a meaning s, corresponding to the lexeme and m-cat, there is a sign with the corresponding form-entry's morphological form, the syntactic category RefAdj, for referential adjective, and a meaning derived from s. This form-sign mapping could be seen as the analog of a *cell* for adjectives that have a referential sub-part. The sign it is used to prove can be seen as the realization of that cell for a lexeme.

The meaning of the resulting sign is complex. For adjectives of this type, s is a pair of an adjectival predicate and classifier predicate. The πs are projection functions for accessing these elements. The slash notation, $P_1(x)/P_2(x)$, is essentially an annotation to avoid going into details of the dynamic semantic theory. It is intended to communicate that the two predicates exist in different scopal relationships to logical operators, such as negation. P_2, the classifier meaning, should be outside negation's scope. The value of the variable x is supplied by discourse.

An important part of the form-sign mapping is the predicate $meaning$, which is true when there is a meaning s for lx and mc. That is to say, it is true when a lexeme of a particular category has a meaning. The existence and value of s is determined by proof. In order to provide such a proof further axioms and theorems must be provided.

The lexical semantic system assumed here is minimal. This is because elaborating that system is a matter of ongoing research into Wao Terero lexical patterns. In the system assumed here, intrinsic meanings are associate with a lexeme as pairs $(\textsc{ñene}, \mathsf{big})$, $(\textsc{ñene}, \mathsf{fat})$, etc. These are axioms of the lexical semantic theory. The intrinsic meanings associated with categories are likewise given as axioms, (po, hand), (po, canoe). Then rules are provided to describe the licit combinations of these meanings.

$$\frac{(lx, P_1) \qquad (mc, P_2) \qquad class(lx) \leq adj \wedge mc \leq classifier}{meaning((P_1, P_2), lx, mc)}$$

The first elements of the premise in the example rule above refer to the previously mentioned pairs. The notion that lexemes belong to classes has already been introduced. It is also the case that morphological categories are hierarchically ordered, as indicated by the use of $mc \leq classifier$. In this case the m-cat hierarchy ensures that only the lexical suffixes that distribute like classifiers take part in the rule.

An example proof using the rule would be:

$$\frac{(\text{ÑENE}, \text{big}) \qquad (po, \text{canoe}) \qquad class(\text{ÑENE}) \leq adj \land po \leq classifier}{meaning((\text{big}, \text{canoe}), \text{ÑENE}, po)}$$

The result of this proof can be used to provide the meaning to the form-sign mapping above.

$$\frac{(po, \text{ñenempo}, \text{ÑENE}) \qquad class(ene) \leq adjectival \qquad meaning((\text{big}, \text{canoe}), \text{ÑENE}, po)}{(ene, \text{RefAdj}, \text{big}(x)/\text{canoe}(x))}$$

The point of all this is that when this sign is composed with other signs, the restriction on the referent is clear. Lexical ambiguity is still available based on the multiplicity of homophonous signs that may be proven. This captures the notion that when a speaker makes an assertion about some previously mentioned *canoe*, they are not also vaguely making a statement equally applicable to some *hand* or other compatible referent.

Realizational mechanisms have been provided in two domains, with the lexical semantics directly feeding the dynamic semantics. This is the bridge between the intrinsic and extrinsic. Polysemy is navigated by the $meaning$ predicate, while the multiplicity of discourse interpretations is handled by form-sign mappings. This realization apparatus utilizes meaning representations that are plausible, justified and interpretable based on practices in state of the art semantic theories.

References

Booij, Geert. 2010. *Construction morphology*. Oxford University Press.

Copestake, Ann & Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics* 12. 15–67.

Coquand, Thierry & Gérard Huet. 1988. The calculus of constructions. *Information and Computation* 76(2). 95–120. https://doi.org/10.1016/0890-5401(88)90005-3.

Halle, Morris & Alec Marantz. 1993. Distributed morphology and the pieces of inflection. In Kenneth L. Hale & Samuel J. Keyser (eds.), *The View from Building 20*, 111–176. Cambridge, MA: MIT Press.

Martin, Scott & Carl Pollard. 2012. A higher-order theory of presupposition. *Studia Logica*: An International Journal of Symbolic Logic 100. 727–751. https://doi.org/10.1007/s11225-012-9427-6.

Peeke, M. Catherine. 1968. *Preliminary grammar of auca*. Indiana University PhD.

Pollard, Carl & Chris Worth. 2015. Coordination in linear categorial grammar with phenogrammatical subtyping. In *Proceedings for the esslli 2015 workshop on empirical advances in categorial grammar (cg 2015)*, 162–182.

Roberts, Craige. 2010. Retrievability and definite noun phrases. In *Chicago workshop on semantics and philosophy*.

Robins, Robert H. 1959. In defence of WP. *Transactions of the Philological Society* 58(1). 116–144.

Sadler, Louisa & Andrew Spencer. 2001. Syntax as an exponent of morphological features. In *Yearbook of Morphology 2000* (YOMO (Yearbook of Morphology)), 71–96. Springer.

Sapir, Edward. 1911. The problem of noun incorporation in American languages. *American Anthropologist* 13(2). 250–282.

Wiltschko, Martina. 2009. √root incorporation: Evidence from lexical suffixes in Halkomelem Salish. *Lingua* 119(2). 199–223.

Typological richness of the German gender system revealed by data mining

Sebastian Fedden
Université Sorbonne Nouvelle,
University of Surrey

Matías Guzmán Naranjo
Universität Tübingen

Greville G. Corbett
University of Surrey

1 Introduction

In recent years linguistic typology has increasingly profited from computational methods; the hope is to discover patterns in large data sets more quickly and more accurately than would be possible for a human researcher. This is commonly known as ‘data mining’. A linguistic system which could benefit from such an approach is German gender.

2 A typological gem

The German gender system is a gem among the assignment systems found in the world, for the complexity of its interacting semantic, morphological and phonological assignment principles. As fast as it offers partial results it raises new questions. This is the more remarkable since there are just three gender values: masculine, feminine, and neuter.

(1)	a.	ein	neu-er	Wagen
		a[NOM.M/N.SG]	new-NOM.M.SG	car(M)[NOM.SG]
		‘a new car’		
	b.	ein-e	neu-e	Kutsche
		a-NOM.F.SG	new-NOM.F.SG	coach(F)[NOM.SG]
		‘a new coach’		
	c.	ein	neu-es	Fahrrad
		a[NOM.M/N.SG]	new-NOM.N.SG	bicycle(N)[NOM.SG]
		‘a new bicycle’		

Furthermore, the basic semantic assignment rules are relatively straightforward. Much more challenging are (i) the relation between gender and inflection class (see August 1975; Pavlov 1995; Bittner 1999; Kürschner & Nübling 2011) and (ii) the phonological assignment rules, investigated by Köpcke (1982) and Köpcke & Zubin (1983) among others.

3.1 Semantics

Sex-differentiable nouns, i.e. nouns which refer to male or female humans, or to male or female (higher) animals, are assigned gender on the basis of biological sex: e.g. *der Mann* ‘man’, *die Frau* ‘woman’, *der Eber* ‘wild boar’, *die Bache* ‘wild sow’. In addition there are various non-core semantic assignment rules, some of which are highly specific and yet surprisingly robust.

3.2 Word formation

Those German nouns which are morphologically complex are governed by the Last Member Principle (*Letzt-Glied-Prinzip*, see Köpcke & Zubin 1984: 28-29, and references there): the gender of the whole word is determined by the gender of the last element. In compounds the last element is a word with its own gender value. For example *der Mutterschutz* ‘maternity’ consists of the feminine first member *die Mutter* ‘mother’ and the masculine last member *der*

Schutz ‘protection’; by the Last Member Principle it is masculine. Derivational affixes are similarly associated with a gender value, which is assigned to the derived word irrespective of the gender of the base (if this is a noun). For example, the suffix *-schaft* derives feminine nouns, e.g. *die Freundschaft* ‘friendship’ from the masculine noun *der Freund* ‘friend’, or *die Landschaft* ‘landscape’ from the neuter noun *das Land* ‘land’.

3.3 Inflection

For many instances, gender can unambiguously (or nearly unambiguously) be predicted from inflection class, e.g. all nouns which inflect like *die Lampe* ‘lamp’, are feminine. Then, as a reduced prediction, there are several inflection classes whose nouns can be masculine or neuter but not feminine. For instance, we can predict that *Knopf* ‘button’ cannot be feminine based on its paradigm.

3.4 Phonology

Köpcke (1982) and Köpcke & Zubin (1983) establish a number of phonological rules to account for the gender of monosyllabic nouns. For example, almost all monosyllabic nouns starting with the cluster /kn/ are masculine (93%), e.g. *der Knopf*, ‘button’, *der Knick* ‘crease’, the only exception being the neuter noun *das Knie* ‘knee’. The majority of nouns which end in the clusters /ft/, /xt/ or /çt/ are feminine (64%), e.g. *die Zunft* ‘guild’, *die Frucht* ‘fruit’, *die Sicht* ‘visibility’. And in general, the more consonants a monosyllabic noun has in its onset or coda, the higher the probability that the noun is masculine.

This body of research has demonstrated clear regularities in the assignment of gender to German nouns. And yet, despite this progress in understanding parts of the system, and the great typological interest of German gender, no attempt has been made to analyse the system as a whole.

3 Pitfalls in the analysis of German gender

In analysing a system as complex as German there are at least three potential pitfalls:

1. cherry picking: observations of alleged regularity are sometimes based on few examples and the overall applicability of these regularities is left unexplored;
2. generalizations without a baseline: thus a prediction of a particular gender value for, say, 35% of the nouns is hardly remarkable if 35% of the nouns overall are of that gender; without a clear baseline we do not know how successful a rule is compared to pure chance;
3. not allowing for overlapping factors: given that phonological, morphological and semantic properties may make the same gender value more probable, making a claim for a particular generalization (e.g. phonological) requires us also to eliminate the possible effects of morphology and semantics.

4 Data mining

To avoid these pitfalls and make progress towards a holistic analysis of the German gender system, we mine a database of more than 30,000 German nouns from WebCELEX (Baayen et al. 1995), coded for gender, frequency, phonological shape, inflection class, and derived/compounded status. We have cleaned this database, and we have added semantic information (human, animal, object, abstract, mass) and frequency (based on the COW corpus, Schäfer 2015). We then built a series of analogical models using Extreme Gradient Boosting trees (similar to Guzmán Naranjo 2020), including different combinations of predictors (morphology, semantics, phonology, inflection class). The baselines in this dataset are approximately 35% masculine, 45% feminine and 20% neuter.

Our choice of using a Boosting Tree model (Chen and Guestri 2016) is purely pragmatic, this type of model has been shown to work very well for this type of task (Bonami and Pellegrini, forth.). In our case, the models learn to predict the gender of a given noun based on a set of predictors.

To include the morphological predictors, inflection class predictors and hand-annotated semantic predictors is a simple matter of adding factors to the model. For phonology and semantics we use a technique based on similarity neighbourhoods. For phonology we calculate a right-hand-side weighted Levenshtein distance between all nouns (we use the phonological transcription). For semantics we induce gender-neutralized semantic vectors using Word2Vec from the COW corpus and calculate a cosine distance matrix between all nouns. With these distance matrices we extract for each noun the nearest five neuter neighbours, nearest five feminine neighbours and nearest five masculine neighbours (once for semantics and once for phonology). We then use these distance values as our phonological and semantic predictors. The intuition behind this technique is that the gender of a noun depends (in part) on how similar phonologically or semantically the noun is to other neuter, masculine and feminine nouns. While using a simple *K*-nearest neighbours algorithm also works, we found that our implementation performed considerably better.

For all models we report the 10-fold cross-validation accuracy. In cross-validation we divide our dataset in 10 groups; and fit a model leaving out one of the 10 groups, we then try to predict the gender of the nouns in the omitted group. We repeat this process for all groups.

5 Results

The overall accuracy results (Figure 1) show clearly that the system is anything but arbitrary. The combined factors reach a predictive success of 96% (top line of Figure 1). Individual factors are also strong predictors, most notably phonological shape and inflection class. The German gender assignment system – while complex and unusual – represents a typologically well-known type: a combination of semantic and formal (morphological/phonological) assignment principles (Corbett 1991).

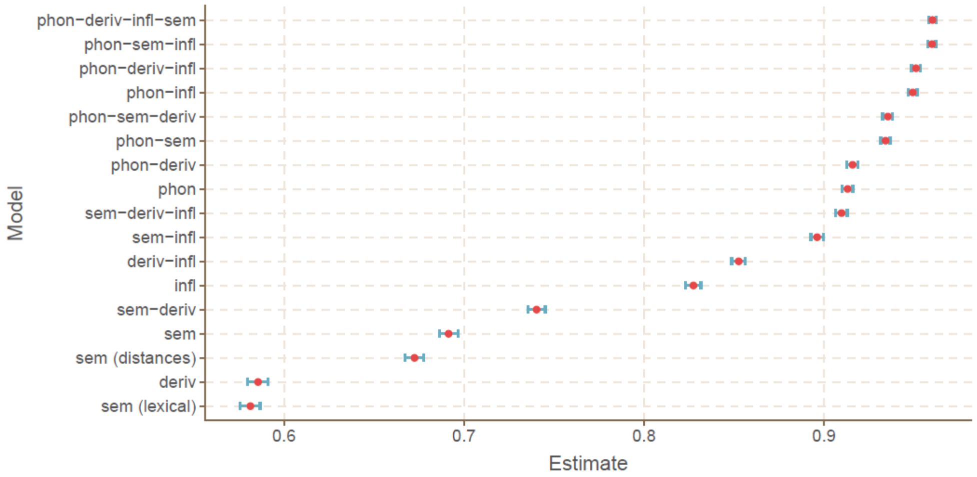


Figure 1. Accuracy and uncertainty intervals by model
Abbreviations: sem - semantics, phon - phonological shape of the stem, deriv - complex (derived) nouns, infl - inflection class

6 Conclusions

Our conclusions relate first to German gender, where we see increasingly clearly the interlocking regularities of the system. We hope to reduce the ill-informed comments still made about German gender, sometimes even by linguists. Second, we make a larger point by showing how typologists can benefit from data mining. From both sides of the collaboration, it is important to keep asking what the generalizations which are established actually mean. In earlier work on gender assignment there was an emphasis on distinguishing regularities from each other, to establish which was responsible for a particular assignment. The current work takes a broader view of inter-connected and mutually reinforcing regularities. And for this, German is indeed a typological gem.

References

Augst, Gerhard. 1975. *Untersuchungen zum Morpheminventar der deutschen Gegenwartssprache*, Forschungsberichte des Instituts für deutsche Sprache Mannheim 25, Tübingen: Gunter Narr.

Baayen, R. Harald, Richard Piepenbrock & Leon Gulikers. 1995. The CELEX Lexical Database (CD-ROM), Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Bittner, Dagmar. 1999. Gender classification and the inflectional system of German nouns. In Barbara Unterbeck (ed.), *Gender in Grammar and Gognition*, Part 1: *Approaches to Gender*, 1–23. Berlin: Mouton de Gruyter.

Bonami, Olivier & Matteo Pellegrini. *Derivation predicting inflection. The role of families, series and morphotactics*. Paper presented at the *19th International Morphology Meeting*, Vienna, February 2020.

Chen, Tianqi & Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.

Guzmán Naranjo, Matías. 2020. Analogy, complexity and predictability in the Russian nominal inflection system, *Morphology* 30. 219–262.

Köpcke, Klaus-Michael. 1982. *Untersuchungen zum Genussystem der deutschen Gegenwartssprache*, Tübingen: Niemeyer.

Köpcke, Klaus-Michael & David A. Zubin. 1983. Die kognitive Organisation der Genuszuweisung zu den einsilbigen Nomen der deutschen Gegenwartssprache. *Zeitschrift für germanistische Linguistik* 11. 166–182.

Köpcke, Klaus-Michael & David A. Zubin. 1984. Sechs Prinzipien für die Genuszuweisung im Deutschen. Ein Beitrag zur natürlichen Klassifikation. *Linguistische Berichte* 93. 26–50.

Kürschner, Sebastian & Damaris Nübling. 2011. The interaction of gender and declension in Germanic languages. *Folia Linguistica* 45. 355–388.

Pavlov, Vladimir. 1995. *Die Deklination der deutschen Substantive. Synchronie und Diachronie*, Frankfurt a. M.: Peter Lang.

Schäfer, Roland. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of Challenges in the Management of Large Corpora (CMLC-3)*, 28–34.

‘Less than Zero’: the Case of (Italo-)Romance Vocatives and Imperatives

Franck Floricic

Université de Paris 3 - Sorbonne Nouvelle & LPP (CNRS)

Imperatives and Vocatives have long been neglected as a category, probably due to their speech-rooted character, a feature which has contributed to leave them outside morphological and syntactical research (cf. however Maiden 2007, Swearingen 2011 and Maiden, Swearingen & O’Neill 2009 for a notable exception). Typological studies on Imperatives have however been proposed on several occasions in recent years (see Khrakovskij (2001), Aikhenvald (2017), etc. which offer an overall picture of the properties of this category). The aim of this contribution is to analyze the case of two related categories: Romance truncated Vocatives (cf. the Italo-Romance Vocatives in (1)) and a sub-class of Romance Imperatives, namely the allomorphic reduced Imperatives some of which are presented in (2) (cf. Huber-Sauter 1951: 65sqq.). The data in (1a-d), which have been recorded among speakers of the Regional Italian spoken in Sardinia, are typically found in the Central-Southern part of Italy. They show that these Vocative forms may retain nothing else than the stressed C(C)V syllable of the noun (cf. Rohlfs 1966: 448-449, §§317-318) or all the segmental material up to the stressed syllable. Needless to say, the segmental shape of these Vocatives may vary according to the Regional Italian taken into account (cf. the data of the Taviano Salentine Italian discussed by Kenstowicz 2019):

(1)

	Full form	Vocative		Full form	Vocative
a.	'pjɛ:ro	'pjɛ	f.	an'to:njo	an'to
b.	'sandro	'sa	g.	te'rɛ:za	te'rɛ
c.	'silvja	'si	h.	fran'ʧɛsko	fran'ʧɛ
d.	'fraŋko	'fra	i.	kor'ra:do	kor'ra
e.	salva'tɔ:re	'tɔ (< 'tɔ:re)	j.	ʤu'zɛppe	ʤu'zɛ

(2)

	French		Italian
a.	ekut > kut ‘listen!’	a.	as'pɛtta > as'pɛ ‘wait!’
b.	atɑ̃ > tɑ̃ ‘wait!’	b.	'gwarda > 'gwa ‘look!’
c.	ʁəgaʁd > gaʁd ‘look!’	c.	'tjɛni > 'tjɛ ‘hold!’
d.	aʁɛt > ʁɛt ‘stop!’	d.	'tɔʎʎi > 'tɔ ‘take!’
		e.	as'kɔlta > as'kɔ ‘listen!’
		f.	'mɔstra > 'mɔ ‘let me see!’

Spanish Sardinian

a. ˈtɔma > ˈtɔ ‘take!’
b. ˈmira > ˈmi ‘look!’

a. ˈtɛnɛ > ˈtɛ ‘hold’
b. ˈmira > ˈmi ‘look!’
c. ˈnara > ˈna ‘say’
d. ˈtɔkka > ˈtɔ ‘go on!’

The questions to be addressed here are of various nature: a) can the competing Imperative forms in (2) be analyzed in terms of ‘overabundance’? Observe *en passant* that examples such as Italian gua’ / guarda ‘look!’ or aspè / aspetta ‘wait!’ clearly contradict Thornton’s claim that in Italian “(…) only verbs which have an asyllabic stem have overabundance in the 2SG.IMP cell » (Thornton 2019 : 249); b) do the Imperative and Vocative forms listed in (1)-(2) obey any prosodic template? c) is the monosyllabic shape of some of these reduced Imperatives due to frequency effects, a thesis defended among others by Mańczak (2004)? d) to which extent these reduced forms can be said to obey the ‘Phonetic Laws’ of the respective languages? Needless to say, it has long been recognized that frequency can lead to drastic phonetic reductions (cf. Schuchardt 1885 on the question of frequency and Sound Laws). It was pointed out – at least since Pott (1852) – that the highly irregular aspect of verbs like Italian *andare*, French *aller*, Span. *andar*, Occitan / Catalan *anar*, Rheto-Rom. *la / na*, etc. was due to frequency effects (cf. as well Mańczak 1980, 1982, 2001, 2008, etc.). It will be shown that monosyllabic Imperatives and Vocatives owe their reduction to another parameter: they belong to the “appeal sphere”. On the one hand, this speech-rooted feature is clearly responsible for the reduction of these forms: they cannot be said to obey any Prosodic Template and they may violate the Minimality Constraint of the languages in which such a constraint is active (cf. Floricic & Molinu 2012). It will be shown that the speech-rooted feature of Vocatives and Imperatives also is responsible for their “marked” status. The markedness of Imperatives and Vocatives will be discussed, and it will be argued that it cannot be reduced to frequency effects (cf. Haspelmath (2006)). It has repeatedly been observed that Imperatives may show transcategorial shifts, and the development of ‘discourse markers’ or attention-getting devices from Imperatives is widely attested typologically (cf. Aikhenvald 2017: 24). It will be shown that such a transcategorial shift is not only responsible for the reduction of various Imperatives; it is also responsible for the semantic change they may be subject to (cf. the case of Italian ˈtɔ ‘take!’ from ˈtɔʎʎi ‘take away’, which parallels Latin *em* ‘there!’ from *eme* ‘take!’ (< *emere* ‘to acquire, buy, purchase) (cf. Umceta Gómez 2017).

References

Aikhenvald, Alexandra. 2010. *Imperatives and commands*. Oxford: Oxford University Press.

Floricic, Franck. 2002. La morphologie du Vocatif: l’exemple du sarde. *Vox Romanica* 61, 151–177.

Floricic, Franck & Lucia Molinu. 2003. Imperativi ‘monosillabici’ e ‘Minimal Word’ in italiano ‘standard’ e in sardo. In Actes du *XXXV Congresso internazionale di Studi della SLI*.

"Il verbo italiano - Approcci diacronici, sincronici, contrastivi e didattici" (Paris, 20 - 22 septembre 2001), 343–357. Roma: Bulzoni.

Floricic, Franck & Lucia Molinu. 2012. Romance monosyllabic imperatives and markedness. In Thomas Stolz, Nicole Nau & Cornelia Stroh (eds.), *Monosyllables. From phonology to typology*, 149–172. Berlin: Akademie Verlag.

Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of linguistics* 42 (1), 25–70.

Huber-Sauter, Margrit. 1951. *Zur Syntax des Imperativs im Italienischen.* Affoltern am Albis: J. Weiß.

Kenstowicz, Michael. 2019. The analysis of truncated Vocatives in Taviano (Salentine) Italian. *Catalan Journal of Linguistics* 18, 131–159.

Khrakovskij, Victor. S. 2001 (ed.). *Typology of Imperative Constructions*. München: Lincom Europa (*LINCOM Studies in Theoretical Linguistics* 09).

Maiden, Martin. 2007. On the morphology of Italo-Romance Imperatives. In Delia Bentley & Adam Ledgeway (eds.), *Sui dialetti italoromanzi. Saggi in onore di Nigel B. Vincent*, 148–164. King's Lynn, Norfolk: The Italianist.

Maiden, Martin, Andrew, Swearingen & Paul O'Neill 2009. Imperative morphology in diachrony: evidence from the Romance languages », in Monique Dufresne, Fernande Dupuis & Etleva Vocaj (eds.), *Historical Linguistics 2007. Selected papers from the 18th International Conference on Historical Linguistics (Montréal, 6-11 august 2007)*, 99–108. Amsterdam / Philadelphia: John Benjamins.

Mańczak, Witold. 1980. Irregular Sound Change due to Frequency in Latin. *Language Sciences* 2, 62–68.

Mańczak, Witold. 1982. *Fonetica e morfologia storica dell'italiano*. Kraków: Uniwersytet Jagiellonski.

Mańczak, Witold. 2004. Certaines formes de l'impératif en italien et en sarde. In Marcela Świątkowska, Roman Sosnowski & Iwona Piechnik (eds.), *Maestro e Amico*. Miscellanea in onore di Stanisław Widłak. Mistrz i Przyjaciel. Studia dedykowane Stanisławowi Widłakowi, 231–234. Kraków: Wydawnictwo UJ.

Mańczak, Witold. 2008. *Linguistique générale et linguistique indo-européenne*. Kraków : Polska Akademia Umiejętności.

Pott, Friedrich A. 1852. Plattlateinisch und romanisch. *Zeitschrift für Vergleichende Sprachforschung* 1, 309–350.

Rohlfs, Gerhardt. 1966. *Grammatica storica della lingua italiana e dei suoi dialetti.* Fonetica. Torino: Giulio Einaudi.

Schuchardt, Hugo. 1972. On sound laws: against the neogrammarians. In Theo Vennemann & Terence H. Wilbur (eds), *Schuchardt, the neogrammarians, and the transformational theory ofphonological change: four essays*, 41–72. Frankfurt am Main: Athenäum Verlag. Firstpublished (1885), *Über die Lautgesetze: gegen die Junggrammatiker*. Berlin: Oppenheim.

Swearingen, Andrew. 2011. *Romance Imperatives. Syncretism, irregularity, autonomy.* Oxford : University of Oxford.

Thornton, Anna. 2019. Overabundance: A Canonical Typology. In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler & Hans Christian Luschützky (eds), *Competition in Inflection and Word-Formation* (Studies in Morphology, volume 5), 223–258. Cham: Springer.

Umceta Gómez, Luis. 2019. Discursive and pragmatic functions of Latin *em*. In Camille Denizot & Olga Spevak (eds.), *Pragmatic Approaches to Latin and Ancient Greek*, 63–82. Amsterdam: John Benjamins Publishing Company.

At the core of morphological autonomy: inflectional classes as a residue, ballast, or resource?

Livio Gaeta
University of Turin

1 Inflectional classes as a residue

Inflectional classes (= ICs) can be held to constitute the essence of morphological autonomy as they cannot be reduced to any other language component in terms of phonologically conditioned alternations, syntactically determined distribution, or semantically driven class assignment. In this sense, they are an irreducible residue and warrant for the autonomous status of morphology (cf. Aronoff 1994: 46, 166 a. o.). On the other hand, they clearly profile language-specific patterning while their cross-linguistic relief remains limited as they are not found or only marginally present in several languages.

2 Ways for lexical groupings

To be sure, ICs do not represent the only way for grouping words within a certain lexical class. To make just one example, we can divide intransitive verbs in English on the basis of their syntactic behavior. Unaccusatives display conjoined past participles, while unergatives don't: *The girl arrived* / **slept yesterday is my sister*. On the other hand, inflectional properties can also interact in a crucial way with such a syntactically based grouping. For instance, German unaccusatives typically select the auxiliary BE for the present perfect (cf. *ist angekommen* 'has arrived'), while unergatives take HAVE (cf. *hat geschlafen* 'has slept'). In other words, the paradigm of an unaccusative and of an unergative verb in German crucially differ in this property.

2.1 The role of analytic constructions in morphological paradigms

However, the property of different auxiliary selection is not normally used to distinguish inflectional behavior, i.e. IC membership. Accordingly, one does not normally consider *verrosten* 'to roast' and *putzen* 'to clean' as belonging to two different ICs in spite of their different auxiliary selection in the present perfect, cf. *sind verrostet* vs. *haben geputzt*, where exactly the same bundle of morphological features {ind., pres. perf., 3rd ps., pl.} is spelled out. Notice that assuming different ICs for different word forms associated with the same set of morphological features is the normal strategy adopted in a typical IC distinction like in the preterite of two German verbs like *schlafen* 'to sleep' and *schlagen* 'to hit', cf. respectively *schliefen* and *schlugen*, corresponding to the same feature set {ind., pret., 3rd ps., pl.}.

In addition, excluding the analytic pieces of an inflectional paradigm from the calculus of IC assignment is also due to the fact that verbs can select different auxiliaries in dependence of the different argument structure which is concretely selected in a certain syntactic environment. Accordingly, a German verb like *fahren* 'to go' shows unaccusative behavior and selects BE (cf. *Hans ist gestern nach München gefahren* 'Yesterday, Hans went to Munich'), but can also appear in a transitive frame where it selects HAVE: *Hans hat gestern seine Frau nach München gefahren* 'Yesterday, Hans drove his wife to Munich'. Furthermore, one and the same verb can display different auxiliary selection as shown by the verb *schließen* 'to close' which combines with both auxiliaries: *Die Metzgerei ist* / *hat geschlossen* 'the butcher shop is / has closed'. Notice that this variation does not crucially resemble the phenomenon of so-called overabundance, namely when two forms occupy one and the same slot (cf. the past participle of the German verb *weben* 'to weave': *gewebt* / *gewoben*).

From a lexical / lexicological point of view, two options are possible for treating such cases. On the one hand – and this is the mostly adopted solution – one simply dismisses the issue considering the different analytic combinations as resulting from the syntactic implementation of the lexeme. Accordingly, the different usages are at best considered as single specifications of the same vocabulary entry, independently of any morphological relief. On the other hand, it is possible to conceive the different usages as relating to different lexemes standing in a derivational relation via a process of conversion (cf. García Velasco & Hengeveld 2002). This latter option accounts for the empirical operation observed in *schließen* which consists in a change of the argument structure whereby the patient / direct object is promoted to a subject while the original agent / subject is totally demoted and even inexpressible.

2.2 On the possible and impossible interactions of syntax and morphology

The provisional conclusion of this brief discussion amounts to saying that IC assignment has normally been taken to represent an exclusive morphological phenomenon in which the morphosyntactic context does not seem to play any role. Even more than this, however. For instance, Corbett (2012: 61, see also Corbett & Baerman 2006) emphasizes that syntactic information is only indirectly relevant for IC assignment to the extent that an IC assignment rule like the following has to be expressly rejected:

> "*Verbs which inflect according to inflectional class II take a preceding direct object; others take a following direct object".

This complies with a Morphology-Free Syntax Principle (= MFSP) which maintains that strictly morphological information like ICs is generally inaccessible to syntax and to syntactic processes (cf. Zwicky 1992). To come back to our German examples with unaccusative verbs, the MFSP prevents auxiliary selection from influencing IC assignment in the following hypothetical terms:

> *Verbs which inflect according to the strong IC take the auxiliary BE; others take HAVE.

In fact, in German we observe a complete independence of auxiliary selection and IC membership: as already mentioned above, both *fahren* and *verrosten* select BE but belong respectively to the strong and to the weak IC.

3 Exploiting syntax for preserving morphology

In contrast to Corbett's and Zwicky's view, however, the generality of the MFSP cannot be taken for granted a priori, and needs in fact empirical validation. In this connection, an interesting development is found in a variety of Highest Alemannic spoken in Gressoney, a Walser German island of north-western Italy (cf. Zürrer 2009). There, the strong/weak IC membership depends on the morphosyntactic environment in which a verb occurs. In particular, most verbs belonging to the etymological strong class follow the I_S-Class-Rule:

> Verbs which inflect according to the inflectional class I_S display the strong suffix in the past participle when the latter is used in constructions where the auxiliary BE is selected, while they take the weak suffix when the participle appears in constructions where HAVE is found.

Accordingly, a verb like *schribe* 'to write' shows two different past participles in clear dependence of the morphosyntactic environment in which it occurs: *éscht gschrében* 'is / was written' vs. *hät gschrébet* 'has written / wrote'. It must be added that past participles taking BE generally display subject agreement, while participles taking HAVE don't, independently of the IC: *ennéra halb stòn ésch z'bròt bach-en-z* / **bach-et-z gsid* 'within one half hour the bread(N.SG) has been baked-N.SG vs. *de ma wò hannensch noch hientoa schwoarz brot bach-et* / **bach-en-z* 'the men who still have baked the brown bread occasionally'. Notice that only few etymological strong verbs don't display this alternation and show the weak suffix throughout all environments, as for instance *erfénne* 'to invent' which has the forms *éscht* / *hät erfònnet*: *de freezer éscht noch nid erfònn-et-e gsid* 'the fridge(M.SG) has not yet been invented-M.SG' and *de lehrer hät d'mòsék erfònn-et* 'the teacher has invented the music'. This is similar to what is found with the other two weak verb classes, for instance with II-class verbs like *publiziere* 'to pulish': *éscht* / *hät publiziert*, or with III-class verbs like *entwécklò* 'to develop': *éscht* / *hät entwécklòt*. In addition, it must be emphasized that also etymological weak verbs like like *spreite* 'to spread' have adopted the I$_{S}$-Class-Rule and display the forms: *éscht gspreiten* 'is / was spread' / *hät gspreitet* 'has spread / spread'.

The distribution of the participles in Gressoney is particularly interesting because it results from a language change which does not have a reductive effect on the Germanic strong and weak classes in contrast to what is commonly observed in the rest of the family. As is well known, etymological strong verbs normally shift to the weak class, as shown by the Middle English preterite *healp* which is remodeled as *helped*, etc., while the opposite change is only sporadically found (cf. Fertig 2020: 207). As an extreme case of this general tendency, Afrikaans has completely dispensed with the morphological ballast provided by different ICs and verbs follow the same inflectional pattern.

3.1 Verbal periphrases and complexification

Far from being reductive, the change observed in Gressoney shows that in fact IC assignment has grown in complexity to the extent that most IC I verbs have developed two different ways of forming the past participle depending on a clear distribution. Periphrases containing the auxiliary BE trigger strong inflection of the past participle, while the selection of HAVE implies weak inflection of the participle. The latter is promoted to a general property of the system to the extent that it is uniformly adopted throughout all verbal classes. This brings along a clear advantage in terms of what Wurzel (1984) labels extra-morphological motivation of ICs. Accordingly, with the exception of a handful of verbs which only display the weak form, most IC I verbs are now characterized by a selective form of the participle in dependence of the morphosyntactic environment. Since in periphrastic constructions taking HAVE the past participles remain uninflected, the focal difference between IC I verbs and the others is overtly – i.e. via inflectional markers – expressed only where the past participles display agreement markers. It's this extra-morphological motivation – i.e. the occurrence of BE triggering agreement – which decides for IC membership and is expressed by the strong form of the past participle.

3.2 Syntax-driven complexification as a response to ballast

Such a syntax-driven complexification can be seen as a response to the general tendency towards the reduction of the strong/weak class distinction found in the Germanic languages. Such a change counteracted the loss of IC distinction which is completely dismissed in Afrikaans as a useless ballast, where this loss has left behind a considerable number of residues of the strong IC in terms of allomorphic variants of the participle when the latter is used as a predicative adjective: *Dit is* (**deur die polisie*) *verbode* 'That is forbidden (*by the police)', in contrast to its use in the passive periphrasis: *Dit is deur die polisie verbied* 'That has been/was forbidden by the police'. Instead of the chaotic and totally idiosyncratic picture observed in Afrikaans (cf. Donaldson 1993: 259), I class verbs in Titsch display a clear-cut distribution, where the distinction has acquired a new extra-morphological motivation provided by the morphosyntactic environment in connection with the occurrence of overt agreement.

4 ICs as ballast or resource?

Such a dialectic tension between dismissing ICs as a useless ballast or exploiting them as an important resource within the lexicon (cf. Enger 2014 for a discussion) lies behind the actual distribution of ICs in Titsch. The solution adopted there, which clearly stands in contrast to the massive reduction observed in Afrikaans, is interesting because it exploits information of syntactic nature which is generally considered to be outside of the reach of ICs and actually provides empirical evidence that the latter need not necessarily be the case: ICs can also be accessed by and wired to their morphosyntactic environment. This paves the way for a new consideration of periphrastic structures within inflectional paradigms (cf. in this regard Spencer 2001, Ackerman, Stump & Webelhut 2011).

Finally, Titsch is characterized – like Afrikaans – by massive language contact and is even exposed to significant processes of language decay. However, this did not bring about a simplification leading to the loss of a dysfunctional morphological ballast. The other aspect of the sociolinguistic milieu in which Titsch is immersed, namely its isolation in Romance-speaking surroundings, is likely to have favoured the processes of remotivation leading to the actual complex distribution of ICs. In this light, contact does not necessarily imply simplification, but can also lead to complexification if accompanied by isolation (cf. Baechler 2016).

References

Ackerman, Farrell, Gregory T. Stump & Gert Webelhuth. 2011. Lexicalism, periphrasis and implicative morphology. In Robert D. Borsley, and Kersti Börjars (eds.), *Non-transformational theories of grammar*, 325–358. Oxford: Oxford University Press.

Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes*. Cambridge, MA: MIT Press.

Baechler, Raffaela. 2016. Inflectional complexity of nouns, adjectives and articles in closely related (non-)isolated varieties. In Raffaela Baechler & Guido Seiler (eds.), *Complexity, isolation, and variation*, 15–39. Berlin & Boston: De Gruyter Mouton.

Corbett, Greville G. 2012. *Features*. Cambridge: Cambridge University Press.

Corbett, Greville G. & Matthew Baerman. 2006. Prolegomena to a typology of morphological features. *Morphology* 16. 231–246.

Donaldson, Bruce C. 1993. *A Grammar of Afrikaans*. Berlin & New York: Mouton de Gruyter.

Enger, Hans-Olav. 2014. Reinforcement in inflection classes: Two cues may be better than one. *Word Structure* 7.2 153–181.

Fertig, David. 2020. Verbal Inflectional Morphology in Germanic. In Michael T. Putnam & B. Richard Page (eds.), *The Cambridge Handbook of Germanic Linguistics*, 193–213. Cambridge: Cambridge University Press.

García Velasco, Daniel & Kees Hengeveld. 2002. Do we need predicate frames? In Ricardo Mairal Usón & María Jesús Pérez Quintero (eds.), *New Perspectives on Argument Structure in Functional Grammar*, 95–123. Berlin: Mouton de Gruyter.

Spencer, Andrew. 2001. The Paradigm-Based Model of Morphosyntax. *Transactions of the Philological Society* 99. 279–313.

Wurzel, W. U. (1984). *Flexionsmorphologie und Natürlichkeit*. Berlin: Akademie-Verlag.

Zürrer, Peter. 2009. *Sprachkontakt in Walser Dialekten. Gressoney und Issime im Aostatal (Italien)*. Stuttgart: Steiner.

Zwicky, Arnold M. 1992. Some choices in the theory of morphology. In Robert Levine (ed.), *Formal Grammar: Theory and Implementation*, 327–371. New York: Oxford University Press.

Derivational paradigmatic models put to test on some non-canonical phenomena

Nabil Hathout
CLLE, CNRS & Université Toulouse Jean Jaurès

Fiammetta Namer
ATILF, Université de Lorraine & CNRS

A growing body of work discusses the benefits of a paradigmatic description of derivational morphology (Bochner, 1993; Van Marle, 1985; Bauer, 1997; Štekauer, 2014; Hathout & Namer, 2018, 2019) and in particular in the analysis of several non-canonical constructions (Corbett, 2010) . The aim of this talk is to highlight the features that characterize and distinguish the paradigmatic models of derivational morphology by putting to test four of them on a variety of non-canonical phenomena including (*i*) form-meaning discrepancies, a minimum prerequisite they must meet; (*ii*) defectiveness, suppletion and *n*-uplets which are difficult to capture by WFRs; (*iii*) a comparison of the ability of the models to account for the paradigmatic dimension of these phenomena including the explicit representation of derivational families, the distinction between abstract and concrete paradigms, and paradigm generalization.

1. Data. The models will be compared on the French paradigm (1) proposed by Bonami & Strnadová (2019). The paradigm contains three families aligned semantically. The semantic contrasts between the lexemes included in the aligned cells of each family are identical: the first column contain location nouns, the second one nouns denoting people whose activity is related to the location, the third one, relational adjectives of the location nouns, and the fourth one, verbs that denote the action of (metaphorically) moving something in the location. On the formal level, the lexemes of the last three columns are formed by concurrent processes (*-ant*/*-ier*; *-al*/*-aire*; *-iser*/*en-*/*in-*). The paradigm also contains suppletive forms in the second and third families. Finally, the third family contains a doublet composed of *emprisonner* formed on the location noun *prison* and *incarcérer* formed on the same stem as the relational adjective *carcéral*. Of course, this study is only an illustration of the capabilities of the models tested and we would need to use a more complete dataset to compare the ability of the models to account for a more representative set of challenging phenomena in morphology.

(1)

commerce 'commerce'	*commerçant* 'shopkeeper'	*commercial* 'commercal'	*commercialiser* 'to commercialize'
école 'school'	*écolier* 'schoolboy'	*scolaire* 'educational'	*scolariser* 'to send to school'
prison 'prison'	*prisonnier* 'inmate'	*carcéral* 'of prison'	*emprisonner* 'to imprison' *incarcérer* 'to imprison'

2. Models. In this talk, we compare the analysis of (1) in four models: Construction Morphology (CxM) of Booij (2010), the Cumulative Patterns (CP) of Bochner (1993) (B93), the Paradigmatic Systems (PS) of Bonami & Strnadová (2019) (BS19) and ParaDis of Namer & Hathout (2020). **CxM** is based on three devices: a multiple inheritance hierarchy; construction schemas which describe the internal structure of lexemes; second-order schemas that describe indirect relations between lexemes of the same complexity (Booij & Masini, 2015). Not being explicitly designed to account for derivational paradigmatic phenomena, this notion plays a secondary role in this theory. In particular, morphological families could be described by extended second-order schemas to fully instantiated constructions. **B93**'s model introduces two structures, the cumulative sets (CSs) which are sets of lexical items that belong to the same morphological family and the CPs which are sets of schemas that generalize the relations that hold in a collection of CSs or CPs. The model is said to be cumulative because any subset of a CS (resp. a CP) is itself a CS (resp. a CP). **BS19** define PSs as a set of semantically aligned families of the same size. The family members are sets of lexical items. Alignment of families consists in superposing the elements which present the same contrasts of meaning with the other members of their families. In **ParaDis**, the paradigmatic representation is distributed over three levels of representation in order to enable a separate description of the formal, categorical and semantic regularities that exist in the paradigm. The three levels of representation contain paradigms that are in correspondence with the paradigms of a fourth level, the morphological level. At all levels, concrete paradigms are superpositions of families. Abstract paradigms define graphs. The graphs defined by the abstract formal, categorical and semantic in correspondence with a morphological paradigm may be different in shape and size.

3. Discrepancies. In the first family in (1), *commercialiser* presents a form-meaning discrepancy because its meaning directly depends on that of the noun *commerce* ('to put in commerce') and its form is coined by suffixing *-iser* to the form of the adjective *commercial*. **CxM** can account for this asymmetry by means of formal and semantic indexes (2) used in order to dissociate form and meaning within the paradigm. **B93** describes this partial family by means of the CS {*commerce*, *commercial*, *commercialiser*} and accounts for the mismatch in the same way as CxM. In **BS19**, all paradigms are considered to be complete graphs, semantically and formally. BS19's paradigms being morpho-semantic, the discrepancy is just ignored. In **ParaDis**, the morphological paradigm which contains *commercialiser* is in correspondence with a formal (abstract) paradigm and a semantic one. These paradigms define graphs of different shapes. The formal one is a complete graph but the semantic one is not because the meanings of *commercialiser* and *commercial* are not directly connected.

(2) $<[\text{kɔmers}]_{\text{N}i} \leftrightarrow [\text{SEM}]_i> \;\approx\; <[[\text{X}]_{\text{N}i}\ \text{-jal}]_{\text{A}j} \leftrightarrow [\text{of}\ [\text{SEM}]_i]_j> \;\approx$
$<[[\text{Y}]_{\text{A}j}\ \text{-iz}]_{\text{V}k} \leftrightarrow [\text{to put in}\ [\text{SEM}]_i]_k>$

4. Defectiveness, suppletion, *n*-uplets. The families of *école* and *prison* are DEFECTIVE because they lack relational adjectives derived from the noun, unlike *commerce*∼*commercial*. In **CxM**, **B93** and **ParaDis**, defective families are not distinguished from the other ones. On the other hand, defectiveness is explicitly represented by empty sets in **BS19**. Moreover, Bonami & Strnadová (2019) point out that the gaps are correlated with the presence in these families of relational adjectives formed on suppletive stems: *scolaire*, *carcéral*. In **CxM** and **B93**, SUPPLETIVE STEMS are described by means of independent variables (X and Y*-aire* standing for *école* and *scolaire*). In CxM, the second order schema $X \approx Y$*-aire* generalizes the more specific schema $X \approx X$*-aire* used for example to describe the *déficit*∼*déficitaire* derivation. In **BS19**, suppletives form do not have particular representation in the SPs. In **ParaDis**, the family of *école* (resp. *prison*) is a lexical family made up of two formally homogeneous morphological families: {*école*, *écolier*} and {*scolaire*, *scolariser*}. Both families are in correspondence with one and the same semantic family. In **CxM**, the DOUBLET *emprisonner/incarcérer* is represented by a second-order schema where the two constructions share the same semantic representation. Alternatively, one could use two partially redundant second-order schemas: *prison*≈*prisonnier*≈*carcéral*≈*emprisonner* and *prison*≈*prisonnier*≈*carcéral*≈*incarcérer*. **B93** may account for the doublet in the same way with either a CS of 5 lexemes or two partially redundant CSs with 4 lexemes each. In **BS19**, the doublet is described by a set of two lexemes which represents one member of the family. In **ParaDis**, doublet emerge from (*i*) the superposition of two formally homogeneous morphological paradigms: one contains [*prison*, *prisonnier*, *emprisonner*], and the other [*carcéral*, *incarcérer*], and (*ii*) the fact that *emprisonner* and *incarcérer* are in correspondence with a same cell in the same semantic paradigm.

5. Paradigms and generalizations. CxM second-order schemas are minimal abstract paradigms defined as associations of (two) construction schemas of the same complexity. However, strictly speaking, CxM does not contain paradigms. These can nevertheless be described with generalized second order schemas between more than two constructions that may have different complexity. For example, paradigm (1) can be described as in (3) with a generic second-order schema which generalizes the constructions and relations of the family of *commerce* to the formal variations present in the three families of the paradigm. (3) is an abstract paradigm. CxM does not have a device to superpose families into concrete ones. On the other hand, CxM is redundant: (3) is complemented with more specific schemas between subsets of words of the families of (1). It therefore provides accurate representations of all the "local" relations along with a global description of the paradigm in which these are embedded. However, the local and global descriptions are not formally connected in the inheritance hierarchy because the schemas are of different sizes. A description of (1) in the same vein as B93's (4,5,6) is also possible in CxM.

(3) $< [\mathrm{X}]_{\mathrm{N}i} \leftrightarrow [\mathrm{SEM}]_i > \; \approx \; < [[\mathrm{X}]_{\mathrm{N}i}\ \text{-suff1}]_{\mathrm{N}j} \leftrightarrow [\text{whose activity is related to } [\mathrm{SEM}]_i]_j > \; \approx$
$< [[\mathrm{Y}]\ \text{-suff2}]_{\mathrm{A}k} \leftrightarrow [\text{of } [\mathrm{SEM}]_i]_k > \; \approx \; < [\mathrm{Z}]_{\mathrm{V}l} \leftrightarrow [\text{to put in } [\mathrm{SEM}]_i]_l >$

In **B93**, CPs describe abstract paradigms but the model cannot represent the superposition of CSs or CPs. Simplicity being the main objective of the B93, CPs stay close to the data. To this aim, each of the three families of paradigm (1) is described by a specific CP (4,5,6) which "locally" generalizes only one families.

(4) {[X,N,Z], [X-ɑ̃,N,'whose activity is related to Z'], [X-jal,A,'of Z'], [X-jaliz,V,'to put in Z']}

(5) {[X,N,Z], [X-je,N,'whose activity is related to Z'], [Y-ɛʁ,A,'of Z'], [Y-aʁiz,V,'to put in Z']}

(6) {[X,N,Z], [X-je,N,'whose activity is related to Z'], [Y-al,A,'of Z'], [ɑ̃-X,V,'to put in Z'], [ɛ̃-Y,V,'to put in Z]}

In **BS19**, the situation is reversed. SPs are concrete paradigms made up of aligned families but the model does not explicitly include abstract paradigms. Families are aligned according to meaning contrasts only, regardless of their formal variations (stem suppletion or affix competition). It is also possible to align families of different sizes by adding empty sets (resp. putting several lexemes in a single set) in order for them to fit into larger (resp. smaller) paradigms as in (7). This makes BS19 a very flexible model.

(7)	{commerce}	{commerçant}	{commercial}	{commercialiser}
	{école}	{écolier}	{scolaire}	{scolariser}
	{prison}	{prisonnier}	{carcéral}	{emprisonner, incarcérer}

ParaDis is more complete than the three previous models because formal, categorical and semantic regularities are described separately and then mapped into the morphological level. Each level of representation contains families and paradigms. Paradigms are superpositions of families with identical contrasts and are therefore totally homogeneous at the three levels of representation (formal, categorical and semantic). The morphological level contains two sorts of paradigms: (*i*) homogeneous morphological paradigms in correspondence with a single paradigm in each of the three levels of representation; (*ii*) derivational paradigms which are superpositions of morphological paradigms. The latter account for particular generalizations like the identity of the semantic contrasts in (1). The analysis of (1) involves five morphological paradigms highlighted in (8) with different colors. The five morphological paradigms are in correspondence with five distinct formal paradigm. On the other hand, all five are in correspondence with one categorical paradigm and one semantic paradigm which accounts for the identity of the meaning contrasts in the three families of (1). In sum, the analysis of (1) in ParaDis unfolds all the specific regularities it contains and then reconstructs the full paradigm by superposing the unfolded morphological paradigms.

(8)

commerce	commerçant	commercial	commercialiser
école	écolier		
		scolaire	scolariser
prison	prisonnier		emprisonner
		carcéral	incarcérer

6. Conclusion. All the models considered in this study account in a more or less precise way for the non canonical phenomena illustrated by (1). However, we have seen (*i*) that CxM does not provide explicit representations of paradigms; (*ii*) that only abstract paradigms can be described in B93 and that the paradigmatic structure is in large part determined by the formal variations; (*iii*) that in contrast, BS19 contains only concrete paradigms structured according to meaning contrasts and that the formal variations are secondary; (*iv*) that ParaDis gives a precise account of the paradigmatic regularities by separating and articulating the description at four levels (formal, categorial, semantic and morphological). We have also seen that ParaDis is the only model that provides both concrete and abstract paradigms.

References

Bauer, Laurie. 1997. Derivational paradigms. In *Yearbook of morphology 1996*, 243–256. Springer.

Bochner, Harry. 1993. *Simplicity in generative morphology*. Berlin & New-York: Mouton de Gruyter.

Bonami, Olivier & Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29(2). 167–197.

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

Booij, Geert & Francesca Masini. 2015. The role of second order schemas in the construction of complex words. In Laurie Bauer, Lívia Körtvélyessy & Pavol Štekauer (eds.), *Semantics of complex words*, vol. 47, 47–66. Heidelberg: Springer.

Corbett, Greville G. 2010. Canonical derivational morphology. *Word Structure* 3(2). 141–155.

Hathout, Nabil & Fiammetta Namer. 2018. Defining paradigms in word formation: concepts, data and experiments. *Lingue e Linguaggio* 17(2). 151–154.

Hathout, Nabil & Fiammetta Namer. 2019. Paradigms in word formation: what are we up to? *Morphology* 29(2). 153–165.

Namer, Fiammetta & Nabil Hathout. 2020. ParaDis and Démonette – from theory to resources for derivational paradigms. *The Prague Bulletin of Mathematical Linguistics* 114. 5–33.

Van Marle, Jaap. 1985. *On the paradigmatic dimension of morphological creativity*. Dordrecht: Foris.

Štekauer, Pavol. 2014. Derivational paradigms. In Rochelle Lieber & Pavol Štekauer (eds.), *The oxford handbook of derivational morphology*, 354–369. Oxford: Oxford, Oxford University Press.

Stress and stem allomorphy in the Romance perfectum: emergence, typology, and motivations of a symbiotic relation

Borja Herce
University of Zurich

The (syntagmatic or paradigmatic) predictability of some morphological properties from others has become a very active domain of study in morphology in recent years (see Ackerman & Malouf 2013, Blevins et al. 2016, etc.). In general, more predictability is equated with greater simplicity and, as such, it could be expected to constitute a shaping force in diachrony. Changes where two traits become aligned in the paradigm, or in the lexicon would thus seem to demand an explanation along these lines (see e.g. Herce 2020). In Romance verb inflection, this is notably the case of perfective stem allomorphy and stress:

	Latin			Italian		
	IPF	PERF	PLUP.SBJV	IPF	PERF	PLUP.SBJV
1SG	koˈ**kweː**bam	ˈ**ko**ksiː	kokˈ**si**ssem	kwoˈ**ʧe**vo	ˈ**kɔ**ssi	kwoˈ**ʧe**ssi
2SG	koˈ**kweː**baːs	kokˈ**si**stiː	kokˈ**si**sseːs	kwoˈ**ʧe**vi	kwoˈ**ʧe**sti	kwoˈ**ʧe**ssi
3SG	koˈ**kweː**bat	ˈ**ko**ksit	kokˈ**si**sset	kwoˈ**ʧe**va	ˈ**kɔ**sse	kwoˈ**ʧe**sse

Table 1: Partial paradigm of 'cook' at two stages in Romance

Unlike in Latin (see Table 1), rhizotony (i.e. root-stress) and the former perfectum stem constitute in contemporary Romance purely morphological traits (see Maiden 2018), as they no longer correlate to any well-defined semantic or phonological environment. Whereas in Latin the two traits were completely independent of each other and orthogonal, we no longer find this perfect cross-classification anywhere in Romance (see Table 2), where the perfectum root and rhizotony can now predict each other to some extent in all varieties:

		Latin		Romanian		Italian		Friulian	
Perfectum stem		+	-	+	-	+	-	+	-
Rhizotony	+	✓	✓	✓	✗	✓	✗	✗	✗
	-	✓	✓	✓	✓	✗	✓	✗	✓

Table 2: The relation between perfective rhizotony and perfectum stem in Romance

Although all trait combinations were attested in Latin, a big asymmetry did exist nonetheless according to the number of verbs that displayed each of them:

	+ Stem allomorphy	- Stem allomorphy
+ rhizotony	57.6% e.g. *dīcō*	11.5% e.g. *vertō*
- rhizotony	0.7% e.g. *quaerō*	30.1% e.g. *petō*

Table 3: The relation between perfective rhizotony and allomorphy in Latin

In the light of the frequency of the traits (Table 3) in the 300 most frequent Latin verbs (LatInFlex1.1, Pellegrini & Passarotti 2018), one may feel tempted to interpret the Romance developments in Table 2 as driven by language-users' necessity to predict these unmotivated morphological traits in the absence of extramorphological cues. In a context where +rhizotony +allomorphy, and -rhizotony -allomorphy were the most frequent combinations, the emergence of a perfect predictability relation may not be unexpected.

Although this might have been an important factor, it can only be part of the story. The class -allomorphy +rhizotony (e.g. *vertō*) was not infrequent in Latin but has been completely eliminated from every single contemporary Romance variety. The opposite is found in the class of verbs +allomorphy -rhizotony, which was extremely infrequent in Latin but is still encountered occasionally in Romance. A strong bias is observed, thus, only against +rhizotony -allomorphy but not against other combinations. My proposal will be that the reason for this might be found in homophony avoidance pressures within the paradigm. Arhizotony and/or a dedicated stem alternant unmistakably identify a form as 'past', however, the absence of both properties would not, and would give rise to very "uncomfortable" past-present diagonal syncretisms:

	caber 'fit'		*decir* 'say'		pseudo-*caber* 'fit'		pseudo-*decir* 'say'	
	PRS	PRET	PRS	PRET	PRS	PRET	PRS	PRET
1SG	ˈkepo	ˈkupe	ˈdiɣo	ˈdixe	ˈkepo	ˈkabe	ˈdiɣo	ˈdiθe
2SG	ˈkabes	kuˈpiste	ˈdiθes	diˈxiste	ˈkabes	kaˈbiste	ˈdiθes	diˈθiste
3SG	ˈkabe	ˈkupo	ˈdiθe	ˈdixo	ˈkabe	ˈkabo	ˈdiθe	ˈdiθo

Table 4: Actual (left) and hypothetical (right) partial paradigms of two Spanish verbs

No-shared-value homophony like the one illustrated in Table 4 is costly in language processing (MacGregor et al. 2015) and is sometimes even actively avoided by defectiveness (see Baerman 2011). This Romance data seem to show that morphological properties that give rise to suboptimal configurations might also be diachronically dispreferred, which would be understandable (only?) under the discriminative role attributed to morphology in abstractive models (Blevins et al. 2016).

References

Ackerman, Farrell, & Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89, 3: 429–464.

Baerman, Matthew. 2011. Defectiveness and homophony avoidance. Journal of Linguistics 47, 1: 1-29.

Blevins, James P., Farrell Ackerman, Robert Malouf & Michael Ramscar. 2016. Morphology as an adaptive discriminative system. Morphological metatheory: 271-302.

Herce, Borja. 2020. Alignment of forms in Spanish verbal inflection: the gang poner, tener, venir, salir, valer as a window into the nature of paradigmatic analogy and predictability. Morphology.

MacGregor, Lucy J., Jennifer Bouwsema & Ekaterini Klepousniotou. 2015. Sustained meaning activation for polysemous but not homonymous words: Evidence from EEG. Neuropsychologia 68: 126-138.

Pellegrini, Matteo & Marco Passarotti. 2018. LatInfLexi: an Inflected Lexicon of Latin Verbs. In Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018).

Maiden, Martin. 2018a. *The Romance Verb: Morphomic Structure and Diachrony*. Oxford: Oxford University Press.

Featural Linking Elements

Laurence Labrune
Université Bordeaux Montaigne & CNRS UMR 5263 CLLE

1 Word compounding devices in languages

Languages make use of a variety of devices to signal word compounding, ranging from full phonological sequences (corresponding to full morphemes) to supra-segmental features. Five basic types of compounding processes, based on the formal structure of the compounding devices they use, can be identified: segmental, sub-segmental, supra-segmental, stem suppletive and void (= absence of overt compounding marker), as shown in Table 1.

Table 1. Types of compounding devices across languages (an overlook)

language	*example*	*translation*	*device*	*reference*
A. SEGMENTAL : one phoneme or more				
French	pomme-**de**-terre	potato	*de* (preposition)	
Japanese	otoko-**no**-ko	boy	*no* (enclitic part.)	
Movima	maropa-**n**-di	papaya seed	*-n-* (linking cons.)	Haude 06
Dutch	pann-**en**-koek	pancake	*-en-* (linking el.)	
Russian	hleb-**o**-zavod	bread factory	*-o-* (linking vowel)	Ralli 08
B. SUB-SEGMENTAL: one feature				
Japanese	kawa-**g**ishi	river side	[+voice]	
Korean	p’allɛ-**p’**inu	laundry soap	[+tense]	Labrune 99
Slave	tsá-**dh**éh	beaver skin	[+voice]	Rice 89
Nivkh	c^{h}o-**x**erqo	catch fish	[+cont]	Shiraishi 06
Nêlêmwa	pw**ã**-jam	candlenut tree nut	[+nas]	Bril 04
Basque	su-**p**azter	fire corner	[-voice]	Labrune 14
Malagasy	satro-**p**otsi	white hat	[-cont]	Keenan & Polinsky 98
C. SUPRA-SEGMENTAL: specific tone, stress or accent pattern				
Etsako (Ekpheli dial.)	uno-efa **HH**-LL	father’s mouth	associative H tone	Akinlabi 96, 11
Tibetan	see-yöö **H**-H	intellectual	elimination of tonal contour in 1st syll. and change from L to H in 2nd syll.	Meredith 90 in Kenstowicz 94
English	bl**á**ck-mailer	blackmailer	initial stress	
Japanese	kawa-**á**sobi	river game	accent on initial μ of 2nd element	
D. STEM SUPPLETIVE: allomorphic or substractive process				
French	**franc**o-anglais	franco-English	*français* ‘French’	
German	**Schlitt**-schuh	skid shoe (skate)	*schlittern* ‘slid’	P.c. by anon. reviewer
Basque	**be**t-azal	eye-lid	*begi* ‘eye’	Labrune 14
Japanese	**ama**-kaze	rainy wind	*ame* ‘rain’	Labrune & Irwin 2021
E. NO OVERT MARKING: but word order relevant				
French	papier-toilette	toilet paper	Head-Modifier	
Japanese	niwa-tori	rooster	Modifier-Head	
Mandarin Chinese	chōŋ-diànqì	electric charger	Modifier-Head	

Note 1: in Table 1, the hyphen denotes the boundary between the constituents of the compound, regardless of the orthographic conventions of the language under consideration.

Note 2: two (or more?) of these devices may be combined in one compound, as in *franco-anglais*, which resorts to types A and E (linking vowel *-o* + shortened allomorph *franc-*), or *kawa-gishi*, which resorts to type B and C (sub-segmental feature + new accent pattern). In addition, several different linking elements may co-exist in one language.

This paper will focus on the second type of compounding devices occurring in determinative compounds (mainly nominal). Such sub-segmental elements will be labelled as *Featural Linking Elements* and defined as follows:

(1) Featural Linking Elements: a definition
A Featural Linking Element (henceforth FLE) is a sub-segmental morphological element which occurs at the boundary between two constituents of a compound, which lacks referential value, and whose function is to signal composition. It is inherently defective, and prototypically involves a consonant or vowel alternation that can be characterized phonologically as one floating feature; in some less prototypical cases it involves a modification in segmental quantity (for instance consonant gemination), or more than one feature, or the realization of a full segment resulting from default filling of an empty position.

2 Aims of talk and research questions

The aim of this talk is to document FLEs across languages and to assert their relevance as morphological objects. I will first present and discuss in more detail examples from a number of languages which arguably possess FLEs: Slave, Movima, Kanamari, Malagasy, Nivkh, Nêlêmwa, Japanese, Korean, Basque and Malayalam. I will also provide a general characterization of the properties of FLEs, comparing them with the other types of compounding devices identified in Table 1. The main research questions which will be addressed are:
- what are the properties of FLEs?
- what is the difference between FLEs and some other linguistic processes which come close to them but are not quite like them, for instance free-standing linking elements, sub-segments, featural affixes (Akinlabi 1996, 2011, Trommer undated), consonant mutation (Wolf 2007), sandhis, etc. ?
- how do morphology and phonology interact in FLEs?
- what type of theoretical issues do FLEs raise?

3 Formal properties of FLEs

Eight formal properties which stand out as characteristic of FLEs have been identified. These eight properties are, presumably, characteristic of FLEs cross-linguistically, and can be viewed as signalling their existence in a given language, thus helping us identify them in a more principled way. However, some of these properties are also found in segmental and tonal linking elements.

a) LOCATION: an FLE is implemented at the boundary between two constituents of a compound (this is also a defining property of free-standing segmental linking elements).

b) SIZE AND PHONOLOGICAL NATURE: an FLE is inferior to a full phoneme in size in its underlying representation. It is inherently incomplete, consisting of one (or sometimes two interrelated) feature/s, or of a prosodic position. It behaves like an autosegment (in the sense of Zoll 1998).

c) LICENSOR: because of its incompleteness, FLEs need a phonological licensor to be realized. The phonological host or licensor can be a full segment or, in some cases, an empty structural position.

d) CONDITIONS OF REALIZATION: The surface realization of the FLE obeys a 'no host, no marker' condition: i.e., in the absence of a proper licensor, the marker fails to be realized. This occurs, for instance, in Japanese *rendaku* which can be represented as a [+voice] FLE (cf. *kawa-**g**ishi* in Table

1): when the second element begins with a consonant that cannot be voiced (either because it is already voiced, or because it has no voiced counterpart in the system), the [+voice] *rendaku* FLE cannot be expressed at the surface level. This also happens with supra-segmental linking elements: e. g. if the association of a high tone to the initial syllable of the second element of a compound is the exponence of a linking element, this linking element receives no exponent if the syllable in question is already high.

e) PREDICTABILITY OF SURFACE FORM: FLEs may receive different surface realizations, depending on their host/licensor, but the crucial point is that the final surface realization is always predictable from the host. In contrast, what is *not* predictable is whether the marker will be inserted or not (see property h below).

f) CONVERGENCE: The result of FLE insertion often resembles the result of the application of certain post-lexical rules or constraints found in the language. A consequence of this is a certain amount of surface opacity, because it is not always clear whether or not a consonant alternation occurring at the boundary between the two elements of a compound is an instance of an FLE or not. For example, in Japanese, it is sometimes impossible to decide whether one is dealing with *rendaku* or post-nasal voicing (Labrune 2012). A tentative explanation would be that some (or all?) FLEs developed out of the morphologization of a phonological process. This is a question that will be further investigated during my talk.

g) MULTI-DIMENSIONALITY: FLE occurrence is very strongly constrained by a variety of morphological, phonological (prosodic and segmental), lexical, etymological, semantic, syntactic and sociolinguistic factors, which interact with each other in a highly complex manner. FLEs are thus multidimensional elements. This is characteristic of linking elements in general.

h) INHERENT VARIABILITY: FLEs appear as fundamentally inconsistent, irregular and variable. This apparently inconsistent character seems to constitute a rather common property of linking elements (see for instance Kürschner & Szczepaniak 2013; Ralli 2008), but it is particularly conspicuous in the case of FLEs. It is explicable by their conditions of realization (see property d), i.e. FLEs are morphological elements whose realization is heavily dependent on phonology and largely determined by the phonological nature of the host. It is also an indirect consequence of the convergence phenomenon in f). On the one hand, the marker cannot be realized in a great number of phonological contexts due to the phonological conditions that constrain its implementation (see d), but on the other hand, an FLE often looks like it is present even when it is not, due to the convergence phenomenon. These two facts are arguably instrumental in allowing a large variability for FLE exponence.

4 Claims

As linguistic objects which exist in between morphology and phonology, FLEs seem to have escaped the attention of morphologists and phonologists. My claim is that FLEs are morphological objects that represent an intermediate stage between fully segmental linking elements like the German *fugenlaut* or the linking vowels of Greek or Russian, and supra-segmental ones. Like segmental linking elements, FLEs have segmental exponence but, like prosodic elements, they are underlyingly dependent on a host and lack autonomy. All three types of linking elements exhibit a number of similarities in their morphological behaviour, in their functions, in the type of processes that they trigger, and in their conditions of application. They essentially differ at the level of their phonological essence and nature. Another claim that will be put forth is that although FLEs seem to be absent from Indo-European languages, they are not rare or anecdotal in the languages of the world. As such, I argue that FLEs should be recognized in their own right, alongside other types of compound markers which have received more descriptive and theoretical attention in cross-linguistic and typological research.

Acknowledgments: I wish to thank three anonymous reviewers for fruitful comments on the first version of this abstract.

Bibliography

Akinlabi, Akinbiyi. 1996. Featural affixation. *Journal of Linguistics* 32. 239–289.

Akinlabi, Akinbiyi. 2011. Featural affixes, in *The Blackwell companion to phonology* vol. IV, M. van Oostendorp, C. J. Ewen, E. Hume, K. Rice (eds), Wiley-Blackwell, 1945-1971.

Anderson, Stephen R. 1985. Typological distinctions in word formation, in Timothy Shopen (ed) *Language typology and syntactic description, vol. III, Grammatical categories and the lexicon* (1st Edition), Cambridge University Press, 3-56.

Bril, Isabelle. 2004. Nêlêmwa, in Pierre Arnaud (ed), *Le nom composé : données sur seize langues*, Lyon : Presses universitaires de Lyon : 185-220.

Fabb, Nigel. 1998. Compounding, in A. Spencer, and A. M. Zwicky (eds), *the Handbook of Morphology*, Oxford: Blackwell, 66-83.

Haude, Katharina. 2006. *A grammar of Movima*, Phd thesis, Radboud Universiteit, Nijmegen.

Keenan, Edward L. & Polinsky Maria. 1998. Malagasy, in Spencer, Andrew & Zwicky Arnold M. (eds), *The handbook of morphology*, Oxford / Malden: Blackwell, 563-623.

Kenstowicz, Michael. 1994. *Phonology in generative grammar*. Cambridge, Mass. & Oxford UK: Blackwell.

Kürschner, Sebastian & Szczepaniak, Renata (eds). 2013. *Linking elements - origin, change, and functionalization,* special issue of *Morphology*, 23:1, February 2012.

Labrune, Laurence. 1999. Variation intra- et inter-langue: morpho-phonologie du *rendaku* en japonais et du *sai-sios* en coréen, *Cahiers de Grammaire* 24, 117-152.

Labrune, Laurence. 2012. *The phonology of Japanese* (The phonology of the world's languages), Oxford: Oxford University Press.

Labrune, Laurence. 2014. Featural linking elements in Basque compounds, *Morphology* 24:4, 377-405.

Labrune, Laurence. 2016. Rendaku in cross-linguistic perspective. In T. Vance & M. Irwin (eds). *Sequential voicing in Japanese*. Amsterdam / Philadelphia: John Benjamins, 195-233.

Labrune, Laurence & Irwin, Mark, 2021. Japanese apophonic compounds. *Journal of Japanese Linguistics* 37-1, 25-67.

Lieber, Rochelle & Štekauer. Pavol. 2009, *The Oxford handbook of compounding*, Oxford: Oxford University Press.

Meredith, Scott. 1990. *Issues in the phonology of prominence*. Cambridge, Mass. : MIT Phd Dissertation.

Ralli, Angela. 2008. Compound markers and parametric variation, *STUF Language typology and universals* 61, 19-38.

Rice, Keren. 1989. *A grammar of Slave,* Berlin: Mouton de Gruyter.

Shiraishi, Hidetoshi. 2006. *Topics in Nivkh phonology*, Phd thesis, Groningen University.

Štekauer, Pavol, Salvador, Valera, & Körtvélyessy, Livia. 2012. *Word-formation in the world's languages, a typological survey*, Cambridge: Cambridge University Press.

Trommer, Jochen, undated. Featural affixes: The morphology of phonological features. Ms: http://www.uni-leipzig.de/~featuralaffixes/antrag.pdf.

Wolf, Mathew. 2007. *For an autosegmental theory of mutation*. University of Massachusetts Amherst. Ms, http://roa.rutgers.edu/files/754-0705/754-WOLF-0-0.PDF.

Zoll, Cheryl. 1998. *Parsing below the segment in a constraint based framework*, Stanford: CSLI publications.

Evaluating morphosemantic demotivation through experimental and distributional methods

Alizée Lombard
Université de Fribourg

Marine Wauquier
Université de Paris
LLF, CNRS

Cécile Fabre
Université Toulouse Jean Jaurès
CLLE, CNRS

Nabil Hathout
Université Toulouse Jean Jaurès
CLLE, CNRS

Lydia-Mai Ho-Dac
Université Toulouse Jean Jaurès
CLLE, CNRS

Richard Huyghe
Université de Fribourg

1 Introduction

Demotivation as the obliteration of the morphosemantic relation between a base word and a derivative is part of the lexicalization process (Bauer, 1983; Lipka, 1992; Brinton & Traugott, 2005). It is caused by idiosyncrasic factors such as onomasiological needs, lexical competition, diachronic change, etc. Although morphosemantic demotivation is in itself a gradual process, as shown for example by Roché (2004) in the analysis of French nouns suffixed with *-ier*, it has rarely been evaluated or quantified as such.

Focusing on the morphosemantic demotivation of French nouns derived from verbs, our goal in this presentation is to investigate methods that can be used to measure morphosemantic demotivation. We present and compare two empirical methods, based respectively on experimental and distributional approaches, in order to determine to what extent judgements and corpus-based assessments of demotivation are correlated.

2 Selection of the materials

We define three conditions on base-derivative pairs, identified as C1 for no motivation (e.g. *partage* ‘sharing’/*partir* ‘leave’), C2 for ongoing demotivation (e.g. *créature* ’creature’/*créer* ‘create’), and C3 for synchronically motivated pairs (e.g. *rasoir* ‘razor’/*raser* ‘shave’). Each condition is populated with pairs extracted from various corpora and lexical resources such as FRCOW16A, Anagrimes[1] and *Trésor de la Langue Française*.

The selected pairs satisfy the formal and semantic criteria of (i) a diachronically attested link between the base and its derivative, and (ii) the absence (in C1 condition) or presence (partial in C2 and complete in C3 conditions) of a semantic relationship between the base and its derivative in synchrony. The selection and classification of the materials in C1/C2/C3 was carried out jointly by six linguists in order to obtain balanced samples, not only in terms of item number, but also of frequency range and suffix representation. Due to their rarity, C1 pairs have served as a reference for selecting materials in the other conditions. 26 pairs were finally selected in each condition. These include a variety of suffixes (*-ade, -ance, -oir, -eur, -ette*) and semantic types (action, object, property, location, etc.).

[1] Tool exploiting the French Wiktionary, accessible online at the url `https://anagrimes.toolforge.org/`.

3 Experimental approach

3.1 Design

The aim of the experiment is to measure the degree of demotivation based on French native speakers' judgements about the semantic proximity of the 78 verb-noun pairs[2]. It took the form of two surveys (39 stimuli each, about 15 minutes long) completed online by 411 Bachelor students in the humanities. In order to control for sociological factors, we chose to analyze only the data from native speakers who were no older than 25 years old, which resulted in 150 and 159 responses per survey (N = 309, M_{age} = 19.4, $Range_{age}$ = [17, 25]).

Each pair was presented separately to participants who where asked to evaluate the proximity between the meaning of the base and that of the derivative on a scale from 0 (unrelated meanings) to 6 (highly related meanings). Participants could also indicate if they did not know one or both words presented in each stimulus. The corresponding data were excluded from the analysis (652 trials, i.e. 0.058% of the data). According to our hypotheses, demotivated pairs (C1) should elicit lower scores of proximity than motivated ones (C3). Semi-demotivated pairs (C2) should yield intermediate scores between those of C1 and C3.

3.2 Results

As shown in Figure 1, the experimental results clearly support our hypotheses. Demotivated pairs (C1) obtain the lowest scores (44% of score 0), whereas motivated pairs (C3) have the highest ones (56% of score 6). Scores assigned to semi-demotivated pairs (C2) are rather evenly distributed, which reveals the heterogeneity of this group.

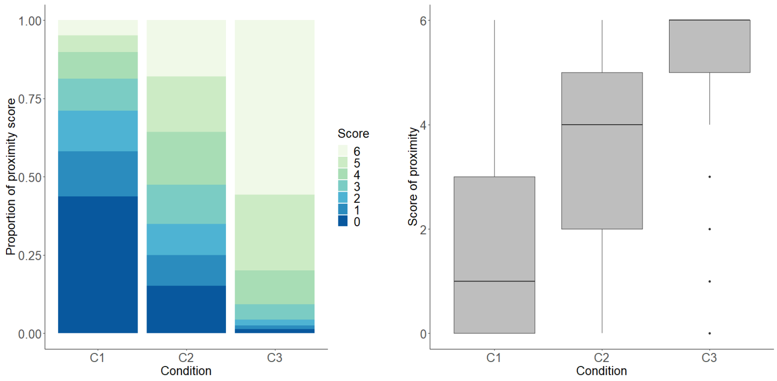


Figure 1: Distribution of experimental scores per condition

These results are analysed through a mixed ordinal logistic regression. The chosen model determines the proximity score as a function of the experimental condition (C1, C2 or C3). It also includes random intercepts per participant and per verb-noun pair. The significance of the predictor effect is determined by a log-likelihood ratio test between models with or without the predictor. The condition appears to have a highly significant effect on the proximity score (p < 2.2e-16). Speakers' intuitions are thus consistent with the experts' judgements.

[2] The experiment was preregistered (including hypotheses, procedure, materials, and analysis plan) on the OSF platform: `https://osf.io/fbtr6/?view_only=f3e66f3aa5dc4a029d6b059cd2f7039b`

4 Distributional approach

4.1 Design

We use distributional semantics to automatically quantify the semantic similarity between derived nouns and base verbs in the C1, C2 and C3 pairs. Distributional Semantics Models (DSMs) provide a vectorial representation of meaning based on the cooccurrences of a given word in a corpus. In the resulting vector space, the distance between vectors approximates the degree of similarity between the corresponding words.

We compute three distributional measures to estimate the demotivation between a base and its derivative: the proximity score P, the rank of the base in the neighborhood of the derivative (rankB), and the rank of the derivative in the neighborhood of the base (rankD). We expect C3 pairs to display a higher proximity score (i.e. closer to 1), and a higher rank (i.e. closer to 1st rank) than C1 pairs. C2 pairs should display in-between values both in terms of proximity score and ranking.

These measures are computed from a vector space concatenating 5 DSMs trained with Word2Vec (Mikolov et al., 2013) on the Wikipedia lemmatized corpus, consisting of 900 million words. The DSMs training parameters are: CBOW, Negative Sampling, frequency threshold of 5, window size of 5.

4.2 Results

Results are in line with our hypotheses. As shown in Figure 2, the proximity scores increase as the semantic motivation of the pairs increases. The median proximity score is higher for C3 pairs than for C1 pairs, and the median score for C2 falls in between. While C2 median score is close to that of C1, C2 pairs display a much higher dispersion, showing the heterogeneity of C2 as a class, compared to C1 and C3. Similar observations can be made with respect to rankB and rankD depending on the degree of motivation. Figure 2 shows that C3 derivatives appear higher in the base neighborhood than C1 and C2 derivatives. C2 pairs also display a higher dispersion than C1 and C3 pairs.

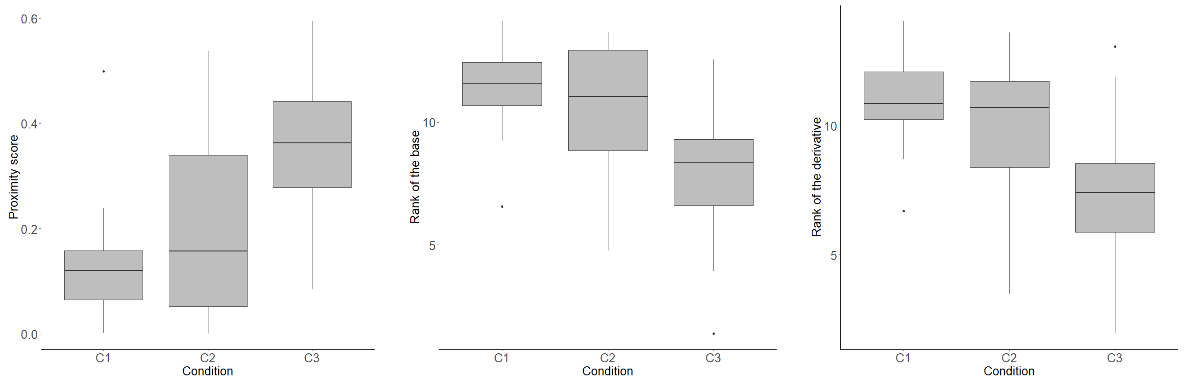


Figure 2: Proximity score (left), ranking (log) of the base in the derivative neighborhood (middle) and ranking (log) of the derivative in the base neighborhood (right) per condition

5 Discussion

The study highlights the psycholinguistic and distributional reality of morphosemantic demotivation, and confirms its gradual properties. Both approaches converge with respect to higher proximity of motivated pairs (C3) than of demotivated pairs (C1). They also converge in analysing C2 as an intermediate case between C1 and C3, showing in addition a wider variation of C2 items. These observations are consistent with the presumed scalarity of demotivation. We analysed the correlation between experimental and distributional results through a mixed ordinal logistic regression. The model determines the experimental proximity score as a function of the distributional proximity and includes random intercepts per participant and per pair. The effect of the predictor is significant (p = 6.76e-06), showing that the distributional approach is clearly in line with speakers' judgements.

While both measures converge, they also differ to some extent. First, the experimental method seems to provide more accurate results than the distributional one. As can be seen in Figure 3, the experimental score (on the y-axis) allows for a better differentiation of the three conditions than the distributional proximity (on the x-axis). Second, some discrepancies can be observed with respect to the assessment of the demotivation of some specific pairs. For instance, there is a strong disagreement between both scores for pairs such as *peignoir* 'bathrobe'/*peigner* 'comb' (C1) (with a high distributional score, but a low experimental one) and *mouvance* 'movement'/*mouvoir* 'move' (C1) (with a low distributional score, but a high experimental score). These results suggest that in some cases, distributional similarity might be too coarse to grasp the level of demotivation of specific items.

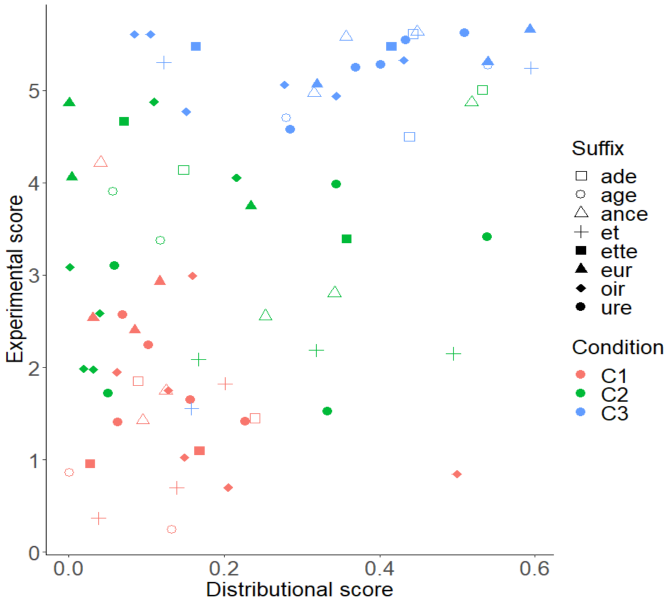


Figure 3: Experimental (0 to 6) and distributional (0 to 1) scores per pair

Although the experimental approach appears to be more accurate, it is more costly and more difficult to extend to a larger scale, especially for quantifying demotivation of previously non-evaluated pairs. By contrast, basing the analysis on the distributional approach would have the advantage of automaticity. It would however need some refinement to neutralize the discrepancies observed between the two methods.

References

Bauer, Laurie. 1983. *English word-formation*. Cambridge University Press.

Brinton, Laurel J & Elizabeth Closs Traugott. 2005. *Lexicalization and language change*. Cambridge University Press.

Lipka, Leonhard. 1992. Lexicalization and institutionalization in English and German. *Linguistica Pragensia/Akademie Ved CR, Ústav pro Jazyk Ceskỳ* 1–13.

Mikolov, Tomas, Kai Chan, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*, Scottsdale.

Roché, Michel. 2004. Mot construit ? mot non construit ? quelques réflexions à partir des dérivés en *-ier*(*e*). *Verbum* 26(2). 459–480.

Prosody-morphology interactions in Mantauran Rukai

Benjamin Macaulay
The Graduate Center, City University of New York

1 Introduction

This paper presents asymmetries in how certain types of morphology interact with prosody and intonation in Mantauran Rukai. Rukai is a Formosan language, i.e. one of the sixteen indigenous Austronesian languages of Taiwan. Mantauran is the variety of Rukai spoken in the village of ’oponoho (Chinese: Wanshan 萬山), in Maolin District, Kaohsiung City. There is considerable variation among Rukai ‘dialects’, with Mantauran standing out as the only Formosan language with stress aligned to the left edge of the word. While Mantauran Rukai has seen a number of descriptions of its morphological/syntactic structures (Zeitoun, 2007, inter alia), including an in-depth analysis of the distribution and usage of morphemes in the language, the current study finds yet-undescribed complex interactions between morphological and prosodic structures, as well as typologically-rare patterns such as an alternation between 1st- and 3rd-syllable stress.[1] This paper will provide an updated stress assignment system based on evidence from intonation, an outline of how the system interacts with morphology, and a diachronic account of an asymmetry in the prosody-morphology interface.

2 Existing literature

While all other varieties of Rukai have either pen-/antepenultimate stress (Li, 1977), Mantauran has been described in the literature as having initial stress on all forms. Zeitoun (2007, 26) notes additionally that prefixes can bear stress, and that a secondary stress is available on the third syllable of words of a certain (unspecified) length, giving the example *vélevèle* ‘banana’, where the acute accent ´ marks primary stress, and the grave accent ` marks secondary stress.[2]

Another relevant structure is ‘echo vowels’, found in all Rukai varieties, as well as the nearby languages Saaroa, Kanakanavu, and Tsou (Li, 1977, 25). ‘Echo vowels’ are epenthesized at the end of (otherwise) consonant-final words. This vowel matches the preceding vowel in quality, unless the preceding vowel is /a/, in which case the echo vowel surfaces as [ə] in Rukai.

3 Revisiting stress assignment

While the existing literature describes Mantauran as having initial stress, the current study finds a more complex system, which is sensitive to word length.

3.1 Stress on 2- and 3-syllable words

In words of 2–3 syllables, stress is realized on the first syllable of the word. Examples include disyllabic *kóne* ‘DYN.SUBJ/eat’, and trisyllabic *tá’olro* ‘dog’.[3] Stress is realized as an F0 maximum,

[1] Data presented here elicited in Nov. 2019 from two speakers in ’oponoho village (73F and 60F), totalling one hour of audio recordings, as part of a larger survey of prosody and intonation in 15 Formosan languages/varieties.

[2] All symbols are IPA except: <c> /ts/; <dh> /ð/; <e> /ə/; <lr> /ɭ/; <ng> /ŋ/; <’> /ʔ/.

[3] Aside from Leipzig glosses, others are taken from Zeitoun (2007), including ‘DYN’ (dynamic verb); ‘STAT’ (stative verb); ‘(N)FIN’ (non-/finite); ‘HUM’ (human).

with 2/3-syllable words beginning with a high tone (H) and ending with a low (L).

3.2 Stress on 5+-syllable words

Words of 5+ syllables surface with two distinct intonational contours, both when produced in isolation, and when they form a prosodic phrase within a larger utterance. Both contours are available on all lexical items, and occur in free variation (at least for words in isolation). One contour, the 'peak' contour, starts with an initial rise, has a peak on the third syllable, and then falls in F0 to the end of the phrase: [L _ H ... L] (where '_' is a syllable unspecified for tone). An example can be seen in Figure 1, showing a pitch track of *a-valrovalro* 'pl-young.woman'.

The other contour is the 'plateau' contour, in which there is a high plateau for the first three syllables, followed by a fall spanning the rest of the word: [H H H ... L]. An example can be seen in Figure 2, showing a pitch track of *kangahatengate* 'palate' with 'plateau' intonation.

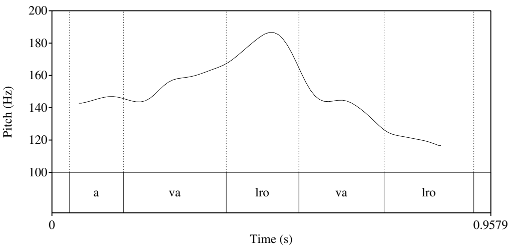


Figure 1: Pitch track of *a-valrovalro* 'pl-young.woman' with 'peak' intonation.

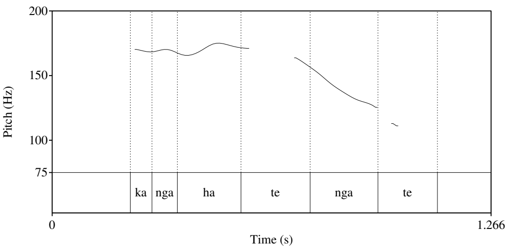


Figure 2: Pitch track of *kangahatengate* 'palate' with 'plateau' intonation.

What both 'peak' and 'plateau' contours have in common is a F0 maximum (H tone) on the third syllable of the word. I argue that this H is indicative of prominence on the third syllable of the word (*a-valróvalro*; *kangahátengate*), as it is in 2-/3-syllable words. The difference between the two contours is thus the initial L vs. H tone, which can be analyzed as boundary tones rather than pitch accents marking the prominent syllable. This same variability in initial %L vs. %H boundary tones preceding the prominent syllable was also noted in this study in the nearby languages Saaroa and Tsou. There is no acoustic evidence of secondary stress.

3.3 4-syllable words, and the stress alternation

Some 4-syllable words pattern with the stress-initial 2-/3-syllable words, while others pattern with the 5+-syllable words with 'peak' and 'plateau' intonation. The 4-syllable words that surfaced with initial stress in the current study were generally those that had echo vowels, the word-final epenthetic vowels of predictable quality. For example, *típitipi* 'slap' can be analyzed as underlyingly /tipitip/, with the surfacing final [i] epenthetic. Without this [i], the underlying form has one fewer syllable nucleus, and stress is thus assigned to the initial syllable, as it is for underlyingly trisyllabic words like *tá'olro* 'dog'.

This alternation can be seen most easily through the addition of suffixes: both *típitipi* 'slap' and *tipitíp-a* 'slap-IMP' surface with four syllables, but only *tipitíp-a* has four nuclei in the underlying form. With this in mind, stress assignment in Mantauran can be summarized as follows: stress falls on the first syllable if the domain of stress is less than 4 syllables, but on the third syllable if the domain of stress is 4+ syllables.

One way to account for this alternation metrically is to make the final syllable extrametrical, and build trochees from the left edge. Word-prominence is assigned to the first foot, but shifts right one foot when possible. The extrametrical final syllable is unavailable as a landing site. Examples can be seen in (1–2), in which the shift to third syllable stress is available in *kangahátengate* ‘palate’ but not the shorter *tá’olro* ‘dog’.

(1)	x						x	ω (Word)
	x				x			φ (Foot)
	x	x	<x>		x	x	<x>	σ (Syllable)
	ta	*’o*	*lro*	↛	*ta*	*’o*	*lro*	‘dog’

(2)	x									x					ω
	x		x		x			x		x		x			φ
	x	x	x	x	x	<x>		x	x	x	x	x	<x>		σ
	ka	*nga*	*ha*	*te*	*nga*	*te*	→	*ka*	*nga*	*ha*	*te*	*nga*	*te*		‘palate’

4 The interaction of prosody and morphology

Since stress assignment in Mantauran is different for domains of 2–3 vs. 4+ syllables, the two types of stress assignment serve as a diagnostic for what kind of morphology is included in the domain of stress assignment. As shown by the pair *típitipi* ‘slap’ vs. *tipitíp-a* ‘slap-IMP’, the imperative suffix *-a* is included in the domain of stress assignment, as it expands the domain past the 4-syllable boundary necessary for third syllable stress. The ‘echo vowels’ such as the final [i] in *típitipi* are not counted in this domain.

Many prefixes are also part of the domain of stress assignment, including the stative prefixes *ma-* ‘STAT.FIN’ and *ka-* ‘STAT.NFIN’, and the human plural prefix *a-*. However, two prefixes are not included in the domain of stress assignment: the dynamic finite verb marker *o(m)-*, and subjective nominalization prefix *ta-*. This can be seen in the examples below, where ‘[ ]$_{\text{Str}}$’ marks the domain of stress: (3a–c) show domains of third-syllable stress including *ma-*, *a-*, and *ka-*, but excluding *ta-*, while in (3d), the exclusion of *o(m)-* causes the smaller domain to surface with first syllable stress. (3e) shows an example where *o(m)-* surfaces as the prevocalic [om] allomorph, which is still excluded from the domain of stress.

(3)	a.	[*ma-somíkace*]$_{\text{Str}}$	‘STAT.FIN-healthy’	(*ma-* included)
	b.	[*a-valróvalro*]$_{\text{Str}}$	‘pl.HUM-healthy’	(*a-* included)
	c.	ta-[*ka-eáea*]$_{\text{Str}}$	‘SUBJNMZ-STAT.NFIN-one’	(*ta-* excluded; *ka-* included)
	d.	*o-*[*lrího’o*]$_{\text{Str}}$	‘DYN.FIN-know’	(*o(m)-* excluded)
	e.	*om-*[*íki*]$_{\text{Str}}$	‘DYN.FIN-exist’	(*o(m)-* excluded)

Of the morphemes that attach to the end of the word, only the imperative suffix *-a* was included in the domain of stress in the data elicited in this study. Two other types of morphemes were found attached word-finally, however. One are the weak pronouns, described by Zeitoun (2007) (and other authors) as enclitics. As most of the Formosan languages surveyed in the current study (except Paiwan) excluded clitics from the domain of stress assignment, this is unsurprising. However, the negator *ka*, described in these works as a suffix, is also excluded from the domain of stress, and on this basis, I analyze it too as an enclitic =*ka*. The exclusion of =*ka* and the pronominal clitics =*li* ‘1sg.GEN’ and =*i* ‘3sg.GEN’ can be seen in examples (4a–b), which can only be analyzed through the 1st/3rd-syllable stress paradigm if the domain of stress contains only the stem.

(4)	a.	*o-*[*lrího’o*]$_{\text{Str}}$=*ka*=*li*	‘DYN.FIN-know=NEG=1sg.GEN’
	b.	*om-*[*íki*]$_{\text{Str}}$=*ka*=*i*	‘DYN.FIN-exist=NEG=3sg.GEN’

The morphology included and excluded from the domain of stress in Mantauran Rukai is summarized in Table 1.

	Domain of Stress Assignment				
o(m)- 'Dyn.Fin' *ta-* 'SubjNmz'	*ma-* 'Stat.Fin' *ka-* 'Stat.NFin' *a-* 'pl.hum'	STEM	-*a* 'imp'	($\text{V}_{\text{epenthetic}}$) =*ka* 'neg'	Pron.

Table 1: Morphology included and excluded from the domain of primary stress assignment in Mantauran Rukai. 'Pron.' here stands for pronominal suffixes/enclitics.

5 Motivating the *o(m)-* vs. *ma-* asymmetry

Mantauran Rukai has a dynamic vs. stative verbal paradigm, with finite verbs marked with *o(m)-* if they are dynamic or *ma-* if they are static. However, while *ma-* is included in the domain of stress assignment, *o(m)-* is not. While unusual, this asymmetry has a likely historical origin. While both *o(m)-* and *ma-* in modern Mantauran are monosyllabic, *o(m)-* is only so in Mantauran. In other dialects like Budai, the cognate is *o-a-*; and Ross (2009, 312) reconstructs disyllabic **u-a-* to Proto-Rukai (vs. monosyllabic **m(a)-*, largely unchanged today).

When Proto-Rukai **u-a-* was reduced to modern Mantauran *o(m)-*, it is likely that the third-syllable stress that fell on the first syllable of the stem (*[*u-a-σ́σσ*]$_{\text{Str}}$) was reanalyzed as the first syllable of a stress domain excluding the prefixal material (*o(m)-*[σ́σσ]$_{\text{Str}}$). This reanalysis can be seen in (5a–b), showing the prosodic structures of *o-(a-)lriho'o* 'Dyn.Fin-know' and *ma-somikace* 'Stat.Fin-healthy' in pre-Mantauran and modern Mantauran. In both stages, the prominence in *o-(a-)lriho'o* falls on the syllable *lri*.

(5)																
			x			→		x					x			ω
	x		x					x			x		x			φ
	x	x	x	x	<x>		<x>	x	x	<x>	x	x	x	x	<x>	σ
	**o-*	*a-*	*lri*	*ho*	*'o*		*o-*	*lri*	*ho*	*'o*	*ma-*	*so*	*mi*	*ka*	*ce*	
	a. pre-Mantauran						b. modern Mantaruan				c. pre-/modern Mantauran					

This account of extrametrical *o(m)-* also places the shift from **u-a-* to modern *o(m)-* after the historical stage when Mantauran diverged from the right-edge stress elsewhere in Rukai.

6 Conclusion

Revisiting Mantauran Rukai with new evidence from intonation finds a number of typologically-uncommon patterns, including an alternation between 1st-/3rd-syllable stress, and a morphological paradigm in which some forms are part of the domain of stress but not others. The exclusion of some morphemes from the domain of stress may have a historical origin, where the foot structure of **u-a-* caused a reanalysis of the domain of stress assignment.

References

Li, Paul Jen-Kuei. 1977. The internal relationships of Rukai. *Bulletin of the Institute of History and Philology* 48(1). 1–92.

Ross, Malcolm D. 2009. Proto Austronesian verbal morphology: a reappraisal. In K. Alexander Adelaar & Andrew Pawley (eds.), *Austronesian historical linguistics and culture history: A festschrift for Robert Blust*, 295–326. Canberra, Australia: Pacific Linguistics.

Zeitoun, Elizabeth. 2007. *A grammar of Mantauran (Rukai)*, vol. A4-2 Language and Linguistics Monograph Series. Taipei: Academia Sinica.

What French eventive nominalizations without verbal bases tell us about the salience of paradigmatic networks

Alice Missud & Florence Villoing
Université Paris Nanterre - MoDyCo (CNRS)

1 Introduction

French nominalizations in *-ion*, *-age* and *-ment* can derive from constructed verbs(1).

(1) a. *créole*$_N$ ‘Creole’ → *créoliser*$_V$ ‘to creolize’ → *créolisation*$_N$ ‘creolization’
b. *plan*$_N$ ‘plan’ → *planifier*$_V$ ‘to plan’ → *planification*$_N$ ‘planning’
c. *jour*$_N$ ‘gap/chink’ → *ajourer*$_V$ ‘to perforate’ → *ajouration*$_N$ ‘perforation’
d. *cadre*$_N$ ‘frame’ → *encadrer*$_V$ ‘to frame’ → *encadrement*$_N$ ‘framing’
e. *goutte*$_N$ ‘drop’ → *égoutter*$_V$ ‘to drain’ → *égouttage*$_N$ ‘draining’
f. *frein*$_N$ ‘brake’ → *freiner*$_V$ ‘to brake’ → *freinage*$_N$ ‘braking’

In such cases, the successive derivation schemes constitute a derivational family of triplets (thus constituting a derivational paradigm in the sense of Bauer 1997 and Štekauer 2014) (2) that consists of a verb constructed on a nominal or adjectival base either by affixation (suffixation in *-iser* (1a), in *-ifier* (1b), prefixation in *a-* (1c), *en-* (1d), *é-* (1e), or by conversion (1f)), and of a resulting nominalization in *-ion*, *-age* or *-ment*. These successive formations do not constitute a case of double suffixation as previously discussed in cases of interfixation or affixal offset (Plénat 2005, Roché 2009) since each derivative carries a specific semantic value. They are not relevant to the bracketing paradox (Pesetsky 1985, Sproat 1992, Spencer 1988, Harley 2010) either, because the interpretations of the derivatives are in line with the morphological schemas they derive from. They also do not constitute a case of parasynthetic formation because missing verbs often eventually appear after the derived nominalization with the expected semantic interpretation.

(2) N / Adj → V → N-*ion* / N-*age* / N-*ment*

Although these formations are not a majority, they still represent a substantial proportion of derivatives: 26.4% of the nouns in *-ion*, *-age* and *-ment* in VerNom (Missud et al. 2020) are derived from a constructed verbal base (as assessed in Missud & Villoing 2020). Analyzing the data revealed a certain proportion of coinages in *-ion*, *-age* and *-ment* for which the constructed verbal base (denominal or deadjectival) does not appear in corpora or occurs at a very low frequency (3).

(3) a. *-ion*: *macbeth*$_N$ ‘Macbeth’ → *macbethisation*$_N$ ‘macbethization’ (°macbethiser$_V$ ‘to macbethize’), *translucide*$_A$ ‘translucent’ → *translucidation*$_N$ ‘translucentation’ (*translucider*$_V$ ‘to make translucent’), *disneyland*$_N$ ‘Disneyland’ → *disneylandification*$_N$ ‘disneylandification’ (*disneylandifier*$_V$ ‘to disneylandify’)
b. *-age*: *bandelette*$_N$ ‘strip’ → *bandelettage*$_N$ ‘strip application’ (*bandeletter*$_V$ ‘to apply strips’), *bestof*$_N$ ‘best-of’ → *bestofage*$_N$ ‘making a best-of’ (°*bestofer*$_V$ ‘to make a best-of’), *dofus*$_N$ ‘Dofus (video game)’ → *dofusage*$_N$ ‘playing Dofus’ (°*dofuser*$_V$ ‘to play Dofus’)
c. *-ment*: *stupide*$_A$ ‘stupid’ → *enstupidement*$_N$ ‘making stupid’ (°*enstupider*$_V$ ‘to make stupid’), *bruyère*$_N$ ‘heather’ → *embruyèrement*$_N$ ‘proliferation of heather’ (*embruyérer*$_V$ ‘to proliferate heather’)

These rare or nonexistent verbs constitute the focus of our study; we hypothesize that such nominalizations are formed directly on the base noun or adjective as the intermediate derivation in the family of triplets (verb formation) is not lexicalized. The successive derivations of triplets in (2) becomes (4).

(4) N / Adj (→ °V) → N-*ion* / N-*age* / N-*ment*

The degree symbol (°) indicates that the verb is possible but not attested in corpora, as previously used by French morphologists following D. Corbin. We analyze them as potential words: as widely discussed in morphology, they are the conceptual (and not actual) result of a productive rule (Booij 1977, Rainer 2012 for an extensive state-of-the-art). They are words that have not been lexicalized, or even attested, but nonetheless seem perfectly acceptable as they meet all the criteria that would make the rule derive them, while not interfering with an already existing form with the same meaning. Although such words have been identified (but not explained) by Roché (2007) and Lignon et al. (2014) in the case of *-ion* nominalizations, our data show that it also concerns *-age* and *-ment* suffixations.

Thus, despite its absence or low frequency, the verb is perfectly identifiable and can be easily interpreted. When looking at the corresponding *-ion*, *-age* or *-ment* nominalization, it is identifiable in i) its form as the nominalization reveals the phonological form of the verbalizing schema although the verb did not appear (3a), as well as in ii) its semantics - specifically in cases of conversion which do not show any affix (3b).

The purpose of our research is to identify the conditions that allow such coinages. We will show that a fundamental condition for these types of formations is that they correspond to a derivational network – a stable and identifiable formal and semantic relation between two or multiple lexemes, in the sense of Hathout (2011), Roché (2017), Bonami & Strnadová (2019), Fradin (2020), that is very salient in the French lexicon (i.e. occurring more frequently that expected in our data). The very high salience of such derivational network implies that some member of a triplet does not necessarily have to be coined, or can be coined with a very low frequency as previously shown (regarding other data) in Villoing & Namer (2012), Roché (2017). However, if the triangular relation in the paradigmatic network is not salient or frequent enough, then the direct derivation from noun/adjective to N-*ion*, N-*age* or N-*ment* is not possible and the verbalization step is necessary.

2 Collecting the data

The data we used were first extracted from frCOW (Schäfer & Bildhauer 2012, Schäfer 2015), a massive French web corpus consisting of 1.9 billion words. We collected every word tagged as a noun ending in either *-ion*, *-age* or *-ment*, as well as all nouns and adjectives. As we were looking for N-*ion*, N-*age* and N-*ment* for which no corresponding verb could be found, we first deleted all nouns that were already included in VerNom (Missud et al. 2020), a lexical database automatically constructed from frCOW consisting of verb-noun pairs and covering *-ion*, *-age* and *-ment* suffixations. We kept N-*ion*, N-*age*, N-*ment*, other nouns and adjectives that were not lemmatized in the distributed version of frCOW, hoping that we would find neologisms as we hypothesized that newly-coined derivatives would be more likely to lack a corresponding verb than lexicalized ones. The remaining nouns in *-ion*, *-age* and *-ment* were then matched automatically with the other nouns and adjectives using regular expressions. As for morphological matching, if the potential suffixed nouns differed from the base nouns and adjectives on the formal level, they were expected to exclusively show signs of an *-iser* or *-ifier* suffixation or an *en-*, *é-* or *a-* prefixation[1]. Otherwise, in the case of verb to noun conversion, only the last syllable of the base noun could differ from the stem of the suffixed noun. In view of the large number of

[1] *dé-* prefixation was not taken into account as the confusion between deverbal *dé-* (as in *boutonner* 'to button' → *déboutonner* 'to unbutton') and denominal *dé-* (as in *os* 'bone' -> *désosser* 'to bone') required a time-consuming step of manual annotation.

	Description of the N-N pairs	Number of items	Number of selected triplets
Sample 1	30% of the pairs containing a noun in *-ion* that shows signs of an *-iser* or *-ifier* suffixation, a conversion, or an *en-*, *é-* or *a-* prefixation	918	186
Sample 2	30% of the pairs containing a noun in *-age* that shows signs of an *-iser* or *-ifier* suffixation, a conversion, or an *en-*, *é-* or *a-* prefixation	593	80
Sample 3	30% of the pairs containing a noun in *-ment* that shows signs of an *-iser* or *-ifier* suffixation, a conversion, or an *en-*, *é-* or *a-* prefixation	667	8

Table 1: Table 1: Extracted samples and selected triplets

pairs that were collected, we created 3 random samples that consisted of 30% of the pairs we found for each of the nominalization schemas. The details are shown in Table 1.

For all the pairs in the samples above, as frCOW does not provide dates, we looked for the suffixed noun using the Google search engine (strictly, using quotation marks) in order to find their earliest date of attestation if the base noun or adjective was semantically related. Although the method relies heavily on the web pages that Google allows users to access, this appeared to be the most convenient way to look for attestation dates since it is still able to capture many book release dates, news articles, forums and dictionaries. Some nouns that did not exist in Google's database were also searched on Twitter.

As we were looking for coinages, suffixed nouns that got more than 10 pages of results and that appeared before 1950 were ignored. Otherwise, we looked for a corresponding verb (inflected or not). In some cases, we also looked for spelling variants (for example: *mic**k**aeljacksoniser* instead of *mic**h**aeljacksoniser*). If the verb's earliest attestation date was later than the noun's according to the search results, or if no verb could be found at all, we considered it as an unattested verb and collected the base noun, the unattested verb and the suffixed noun as triplets[2].

3 Results

The concrete evidence that these nominalizations lacking a corresponding attested verb belong to salient paradigmatic networks in the French lexicon is reflected by the fact that the unattested verbs that the nominalizations infer correspond to the schemas that *-ion*, *-age* and *-ment* prefer. As previously shown in Missud & Villoing (2020), *-ion*, *-age* and *-ment* all display salient distinct preferences when it comes to constructed base selection. *-ion* is by far the most specialized as it strongly favors *-iser* verbs and is the one that selects *-ifier* verbs the most. It can also select a great proportion of converted verbs. *-age* is less categorical although it strongly favors converted verbs, as well as a significant proportion of the *é-* verbs of the data. While *-ment* is the one that has the least salient preferences, it still shows a preference for *en-* verbs and converted verbs (although converted verbs are mostly selected by *-age*). The proportions in Table 2 show the preferences of *-ion*, *-age* and *-ment* when a verb is lacking (or attested later than the nominalization) in the paradigm by dividing the number of occurrences of each cell by the total number of occurrences of the data.

As shown in Table 2, nominalizations without verbs in *-ion* suppose the existence of an *-iser* derived verb, as in the general case. The same applies to nouns in *-age* with no verb that suppose a converted verb, and nouns in *-ment* that suppose a denominal verb in *en-*. What is striking is that the more salient the preference is in the general case, the more it shows in cases where

[2] Note that the number of selected triplets containing a noun in *-ment* is extremely low compared to the number of items initially collected: this is due to the fact that *-ment* coinages are rare, and most *-ment* nouns in the sample were misspelled or taken from various Ancient and Middle French dictionaries on the web, which calls into question the actual productivity of the suffix (as assessed in Missud et al. 2020).

	-iser	*-ifier*	*conversion*	*a-*	*en-*	*é-*
-ion	**0.51**	0.08	0.08	-	-	-
-age	-	-	**0.27**	-	0.01	-
-ment	-	-	<0.01	-	**0.02**	-

Table 2: Proportions of infrequent verbs according to their construction for each verbalization schema (division of each cell by the total)

no verb is attested. *-ion* is the only suffixation that can derive nouns in *-is(ation)* (51% of all data) and *-ifi(cation)* (8%), *-ment* suffixation constructs *en*-X-*ment* nouns in most cases; only one nominalization in *-ment* implying a converted verb was identified (*zizi*, *ziziement* → °zizier'), and the semantic relation between the members of this triplet is unclear (*zizi* can refer to an onomatopoeia or a willy). *-ion* derives a similar proportion of *-ification* nouns and nouns linked to an unattested converted verb (8% in both cases), which reflects a tendency for converted verbs that was already identified in the general case and that might have been extrapolated because of the higher proportion of unattested converted verbs that have been collected. Although *é*-prefixed verbs were mostly represented amongst *-age* triplets in the general case, *-age* preference for *en-* is reflected here as *en*-X-*age* represents 2% of the data. The strongest preference for converted verbs is apparent as 27% of the data consist of *-age* nouns implying a converted verb.

The proportion of constructed verbs that the nominalizations select is a good indicator of the salience of the derivational network, and subsequently of the potential nominalization that lacks a corresponding verbal base. Inversely, nominalizations without a verb correctly predict the preferred verbal bases that the nominalizations select in the general case. As a result, it could be that the salience of such preferences also partially ensures the productivity of *-ion*, *-age* and *-ment*: since the paradigmatic network is productive (e.g. N/A → V-*iser* → N-*ion*, or N/A → converted verb → N-*age*), shortcuts are allowed and nominalizations no longer need to wait for the availability of a verb that corresponds to their preferences to be constructed; they can directly select a noun or an adjective that fits morphologically and semantically.

References

Bauer, Laurie. 1997. Derivational paradigms. In *Yearbook of morphology 1996*, 243–256. Springer.

Bonami, Olivier & Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29(2). 167–197.

Booij, Geert Evert. 1977. *Dutch morphology: A study of word formation in generative grammar*. Dordrecht: Foris.

Fradin, Bernard. 2020. Characterizing derivational paradigms. In J. Fernández-Domínguez, A. Bagasheva & C. Lara-Clares (eds.), *Empirical approaches to linguistic theory*, vol. 16, 49–84. Leiden: Koninklijke Brill.

Harley, Heidi. 2010. Affixation and the mirror principle. *Interfaces in linguistics* 166. 186.

Hathout, Nabil. 2011. Une approche topologique de la construction des mots: propositions théoriques et application à la préfixation en anti. *Des unités morphologiques au lexique* 251–318.

Lignon, Stéphanie, Fiammetta Namer & Florence Villoing. 2014. De l'agglutination à la triangulation ou comment expliquer certaines séries morphologiques. In *Shs web of conferences*, vol. 8, 1813–1835. EDP Sciences.

Missud, Alice, Pascal Amsili & Florence Villoing. 2020. Vernom: une base de paires morphologiques acquise sur très gros corpus (vernom: a french derivational database acquired on a massive corpus). In *Actes de la 6e conférence conjointe journées d'études sur la parole (jep, 33e édition), traitement automatique des langues naturelles (taln, 27e édition), rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues (récital, 22e édition). volume 2: Traitement automatique des langues naturelles*, 305–313.

Missud, Alice & Florence Villoing. 2020. The morphology of rival –ion, –age and –ment selected verbal bases. In Dany Amiot & Delphine Tribout (eds.), *Lexique*, vol. 26, 29–52. Presses Universitaires de Lille.

Pesetsky, David. 1985. Morphology and logical form. *Linguistic inquiry* 16(2). 193–246.

Plénat, Marc. 2005. Rosinette, cousinette, putinette, starlinette, chipinette: décalage, infixation et épenthèse devant–ette. In Choï Jonin I., Bras M., Dagnac A. & Rouquier M. (eds.), *Questions de classification en linguistique : méthodes et descriptions. mélanges offerts au professeur christian molinier*, 275–298. Berne, Peter Lang.

Rainer, Franz. 2012. Morphological metaphysics: virtual, potential, and actual words. *Word Structure* 5(2). 165–182.

Roché, Michel. 2007. Logique lexicale et morphologie: la dérivation en-isme. *Selected proceedings of the 5th Décembrettes: Morphology in Toulouse* 45–58.

Roché, Michel. 2009. Un ou deux suffixes? une ou deux suffixations? In B. Fradin, F. Kerleroux & M. Plénat. (eds.), *Aperçus de morphologie du français*, 143–173. Presses Universitaires de Vincennes.

Roché, Michel. 2017. Les familles dérivationnelles : comment ça marche ?, Toulouse : Université Toulouse 2 Jean Jaurès.

Schäfer, Roland. 2015. Processing and querying large web corpora with the COW14 architecture. In Piotr Banski, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Längen & Andreas Witt (eds.), *Proceedings of challenges in the management of large corpora 3 (cmlc-3)*, UCREL Lancaster: IDS. `http://rolandschaefer.net/?p=749`.

Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet UÄŸur DoÄŸan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eight international conference on language resources and evaluation (lrec'12)*, 486–493. Istanbul, Turkey: European Language Resources Association (ELRA). `http://rolandschaefer.net/?p=70`.

Spencer, Andrew. 1988. Bracketing paradoxes and the english lexicon. *Language* 663–682.

Sproat, Richard. 1992. Unhappier is not a" bracketing paradox". *Linguistic Inquiry* 23(2). 347–352.

Štekauer, Pavol. 2014. Derivational paradigms. In R. Lieber & P. Štekauer (eds.), *The oxford handbook of derivational morphology*, 354–369. Oxford University Press Oxford.

Villoing, Florence & Fiammetta Namer. 2012. Composition néoclassique en-logue et en-logiste: Les noms en-logue sont-ils encore des noms de spécialistes? *La composition néoclassique, Verbum, Nancy: Presses Universitaires de Nancy. DOI* 10. 213–231.

Including the *Word Formation Latin* Resource in the LiLa Knowledge Base

Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini, Giovanni Moretti
Università Cattolica del Sacro Cuore, Milano

1 Background and Motivation

Nowadays, a continuously increasing quantity of resources (like corpora, dictionaries and lexica) and Natural Language Processing (NLP) tools is available for several languages. However, such resources and tools are often not able to interact with each other, making it difficult to search for pieces of information coming from different sources. To tackle this problem, in recent years there has been a trend towards applying techniques of the so-called Linked Data paradigm to linguistic data, creating a Linguistic Linked Data Cloud of interoperable resources (Cimiano et al., 2020).

The aim of the *LiLa* project[1] is to include Latin into this framework, by creating a Knowledge Base (KB) of interlinked resources for Latin using a common vocabulary for knowledge description. Here, we focus on the treatment of word formation in the LiLa KB, that already provides some derivational information taken from the Word Formation Latin (WFL) database (Litta & Passarotti, 2019). WFL on its part adopts a step-by-step, morphotactic approach where each lexeme is linked to the one from which it is directly derived by means of a specific word formation rule (WFR), thus providing a detailed, hierarchical information that is not currently encoded in the KB.

In this contribution, we describe a model designed to represent all the information contained in WFL in the LiLa KB, highlighting the theoretical principles underlying the differences in the current treatment of word formation in LiLa *vis-à-vis* the one of WFL. In Section 2, we describe the architecture of the LiLa KB on the one hand and of the WFL database on the other hand. We outline the model that we propose in order to include WFL in LiLa, showing how it interacts with other more general-purpose models developed by the Linked Data community, particularly the Morphology Module (Klimek et al., 2019) of the OntoLex-Lemon vocabulary for describing lexical resources (McCrae et al., 2017). In Section 3, we discuss some cases of research questions that are not easily answered with the information currently provided in the KB alone and require the use of WFL, and *vice versa*, exemplifying the benefit of having different pieces of information in a unified fashion, as it is allowed by the inclusion of WFL into LiLa.

2 LiLa and WFL

By adopting the data model of the Resource Description Framework (RDF), LiLa expresses information in terms of triples, that connect a subject – a labeled node in the graphical representations that follow – to an object – another labeled node, or a literal – by means of a property – a labeled edge. More specifically, the intuition behind the way in which LiLa achieves the desired interoperability between distributed resources is based on the central role of words: a pivotal role is played by the `Lemma`, defined as the canonical form of a lexical item, i.e. its citation form. The backbone of LiLa's architecture is a Lemma Bank that contains all the lemmas of the database of the morphological analyzer Lemlat (Passarotti et al., 2017). Among else, the Lemma Bank currently also provides some derivational information. Besides lemmas, two other classes

[1] `https://lila-erc.eu`.

of entities are involved in the treatment of word formation in LiLa, namely `Affixes` (in their turn divided into `Prefixes` and `Suffixes`) and `Bases`, defined simply as abstract connectors between lemmas that belong to the same family. Each lemma is linked to the base to which it is related by means of the property `hasBase`, and to the affixes it displays by means of the properties `hasPrefix` and `hasSuffix`. This results in a flat structure, as shown in Figure 1.

As for WFL, it is a derivational lexicon of Latin whose structure is devised according to the Item-and-Arrangement (I&A) morphology model (Hockett, 1954). Lexemes that are considered as deriving from one another are connected via WFRs of different kinds. More specifically, there are compounding rules and derivation rules; in turn, within derivation rules, affixation (divided into prefixation and suffixation) and conversion are distinguished. Furthermore, WFRs are classified according to the Part-of-Speech of the lexemes they take as input and output. This results in a hierarchical structure represented by a directed tree-graph, that takes root from the ancestor – the lexeme from which all the lexemes belonging to the same word formation family ultimately derive – and links it to all derivatives by means of the successive application of individual rules, as shown in Figure 2.

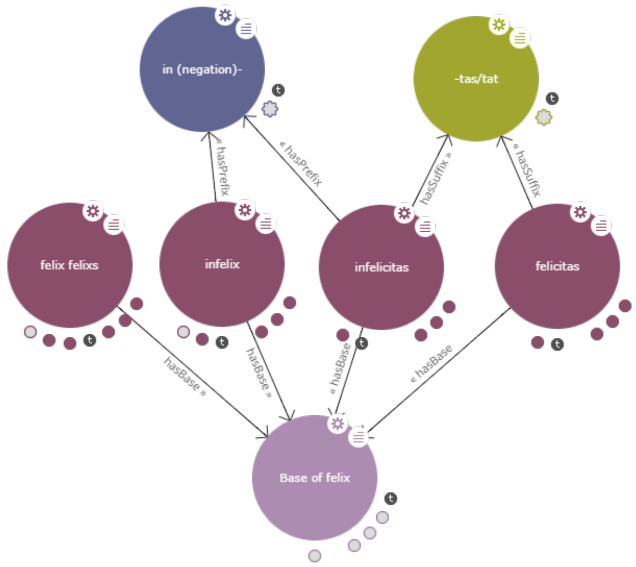


Figure 1: Word Formation in the Lemma Bank

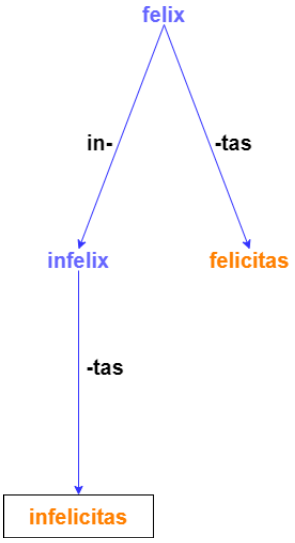


Figure 2: Word Formation in WFL

The flat organization of derivational information in the Lemma Bank was specifically envisaged to overcome some of the limits that have been observed regarding the treatment of word formation in WFL (Budassi & Litta, 2017). Indeed, the rigidly hierarchical structure of WFL forces it to make a choice about the directionality of conversion processes, even when there are doubts (e.g., does the noun ADVERSARIUS ‘opponent’ derive from the adjective ADVERSARIUS, or *vice versa*?), and to create fictional entries to account for cases where more than one affix appears to be simultaneously added to a base (e.g. EXAQUESCO ‘to become water’ from AQUA ‘water’, for which neither *AQUESCO nor *EXAQUO are attested as intermediate steps). As is argued by Litta et al. (2020), the flat approach adopted in LiLa allows for a more natural treatment of such cases, by providing a different modelling strategy compatible with Word-and-Paradigm (W&P) frameworks, and especially Construction Morphology (Booij, 2010). Nevertheless, this means that a lot of potentially useful information provided by WFL is not currently represented in LiLa. In what follows, we will describe the work that has been done in order to include such information within the architecture of LiLa, as summarized in the example in Figure 3.

In our model, the words of WFL that are derived from one another are treated as instances of the class `LexicalEntry` of OntoLex-Lemon, and they are seen as connected by an individual word

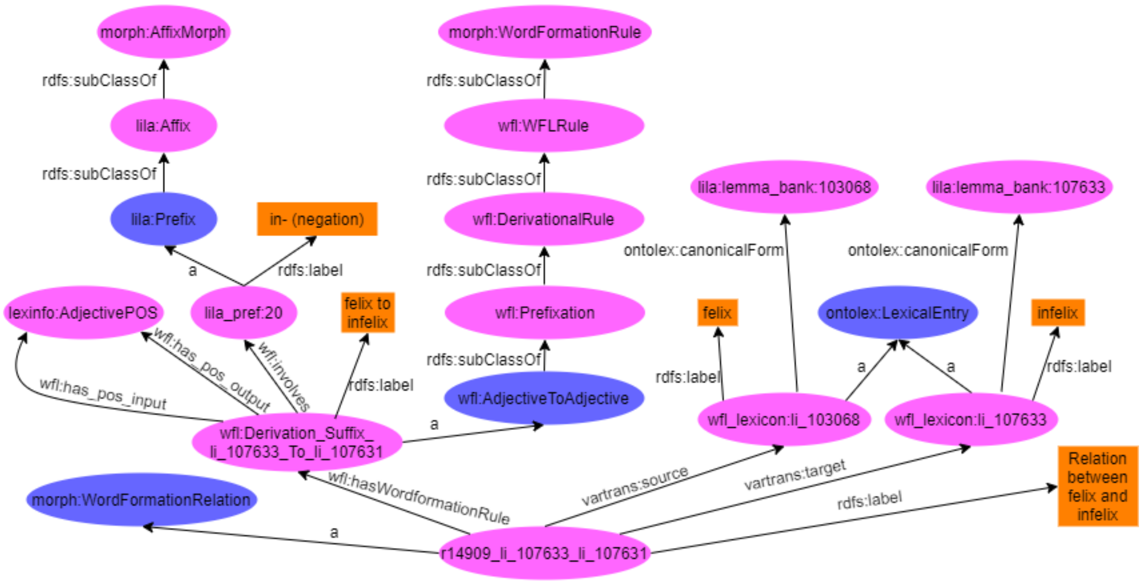


Figure 3: Example of prefixation in the WFL ontology

formation relation – i.e., an instance of the class `WordFormationRelation` of the Morphology Module.[2] More precisely, each relation is linked to the lexical entries of the input and output of the word formation process using properties taken from the Variation and Translation (vartrans) Module, namely `source` and `target`, respectively. Importantly, this allows to express the directionality of the word formation process as stated in WFL, thus ensuring that its hierarchical structure is preserved. Each relation is then linked to the WFR it instantiates according to WFL – in this case, an instance of the class of rules forming deadjectival adjectives – by means of the property `hasWordFormationRule`. The connection with the Morphology Module is achieved by establishing a sub-class relation between its rules (`WordFormationRule`) and the ones of WFL (`WFLRule`). Rules are classified in a way that accurately reflects the structure of WFL, as described above: firstly, there are two sub-classes `CompoundingRule` and `DerivationalRule`, with the latter in its turn displaying three sub-classes, `Suffixation`, `Prefixation` and `Conversion`; secondly, they are distinguished on the basis of the lexical category of the base and derivative, by providing a connection to the Parts-of-Speech of LexInfo (Cimiano et al., 2011) using the properties `has_pos_input` and `has_pos_output`, respectively. Lastly, each affixal rule is also linked to the prefix or suffix it displays, as expressed in the LiLa ontology, by means of the property `involves`, and again a sub-class relation is established between the affixes of LiLa and the ones of the Morphology Module to ensure interoperability.

3 Discussion and Conclusions

We have seen in Section 2 that the choice of a flat approach to word formation in the Lemma Bank of LiLa was motivated by the difficulties raised by the rigidly morphotactic approach of WFL in treating specific phenomena. However, it should be stressed that there are also relatively uncontroversial cases where the more detailed, hierarchical information provided by WFL on the order of application of different processes can prove helpful.

To give an example, it might be useful to be able to distinguish lexemes that are actually formed by means of an affix from the ones that simply display that affix because it has been

[2] Note that this module is still the object of discussion in the Linked Data community: our proposal reflects its current state, but some details might change in the future.

introduced in a previous step of their derivational history. This kind of information is not available in the Lemma Bank, that simply records the affixes present in each lemma. For instance, the adjectives INFRUCTUOSUS 'unfruitful' and INIURIOSUS 'injurious' have a similar structure on the surface, both displaying the prefix *in-* (negation) and the suffix *-os*. However, their derivational history is quite different, the former being undisputably formed by prefixation of *in-* to FRUCTUOSUS 'fruiful' (as *INFRUCTUS is not only not attested, but arguably not even possible as a Latin word), and the latter being clearly a result of the suffixation of *-os* to INIURIA 'injury' (*IURIOSUS). Therefore, the additional information provided by WFL is crucial.

Conversely, if the objective is to obtain all the lexemes that display a given affix, regardless of its position in the linear order of morphs and/or of its derivational history, this can be done with the information provided in the Lemma Bank. For instance, all the verbs containing the prefix *ob-* can trivially be extracted from the Lemma Bank, as they are all linked to that suffix by means of the property `hasPrefix`, while an analogous search in WFL would have trouble finding a verb like OBDURESCO 'to become hard', since it is not considered as formed directly by prefixation of *ob-*, but rather by suffixing *-sc* to the prefixed verb OBDURO 'to harden'.

One of the main advantages of the adoption of the Linked Data standards mentioned in Section 1 is exactly the possibility of not having to force a decision between the two approaches: both of them are made available within a unified framework, leaving up to scholars the choice of the one that is more compatible with their theoretical view, or that merely provides the kind of information more appropriate for the case at hand. This also allows to make the two approaches easily interact in case pieces of information from different sources are needed.

References

Booij, Geert. 2010. Construction morphology. *Language and linguistics compass* 4(7). 543–555.

Budassi, Marco & Eleonora Litta. 2017. In Trouble with the Rules. Theoretical Issues Raised by the Insertion of -sc- Verbs into Word Formation Latin. In *Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo)*, 15–26.

Cimiano, Philipp, Paul Buitelaar, John McCrae & Michael Sintek. 2011. LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Journal of Web Semantics* 9(1). 29–51.

Cimiano, Philipp, Christian Chiarcos, John McCrae & Jorge Gracia. 2020. *Linguistic Linked Data*. Springer.

Hockett, Charles F. 1954. Two models of grammatical description. *Word* 10. 210–234.

Klimek, Bettina, John McCrae, Julia Bosque-Gil, Maxim Ionov, James K. Tauber & Christian Chiarcos. 2019. Challenges for the Representation of Morphology in Ontology Lexicons. In *Proceedings of eLex*, 570–591.

Litta, Eleonora & Marco Passarotti. 2019. (When) inflection needs derivation: a word formation lexicon for Latin. In Nigel Holmes, Marijke Ottink, Josine Schrickx & Maria Selig (eds.), *Lemmata Linguistica Latina. Volume 1. Words and Sounds*, 224–239. Berlin, Boston: De Gruyter.

Litta, Eleonora, Marco Passarotti & Francesco Mambrini. 2020. Derivations and Connections: Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin. *The Prague Bulletin of Mathematical Linguistics* (115). 163–186.

McCrae, John, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar & Philipp Cimiano. 2017. The OntoLex-Lemon Model: Development and Applications. In *Proceedings of eLex*, 587–597.

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.

Regular and irregular noun plurals in German individuals with Down syndrome

Martina Penke

University of Cologne, Department of Special Education and Rehabilitation

1 Introduction

Individuals with Down syndrome (DS) display marked problems in the acquisition of inflectional morphology (e.g. Eadie et al., 2002). A typical finding in research on inflectional deficits is that regular and irregular inflected forms are affected differently. Dualistic approaches to inflection assume that this difference in vulnerability is due to the fact that the representations and mechanisms involved in the production of regular and irregular inflected forms rely on two independent modules of the human language faculty: a computational (i.e. grammar) component where regular affixation is carried out and regular inflected forms are produced, and a storage component – the mental lexicon – in which learned irregular inflected forms are stored and retrieved (Pinker, 1999). This dualistic view of inflection presupposes that language deficits should be found that selectively affect only one of these components sparing the other. Inflectional deficits in individuals with DS might constitute a case in point (Penke, 2019). If regular inflection requires morphological processing - assumed to be compromised in DS - regular inflection should be impaired in individuals with DS. In contrast, if irregular inflected forms are stored in the mental lexicon they should be less affected since the mental lexicon is typically developed according to or even exceeding mental-age expectations in these individuals (e.g. Næss et al., 2011). Here, I will present data on noun plural inflection in 57 German-speaking individuals (31 with DS, 26 typically-developing children) bearing on this issue.

Other than English, German displays overt regular and irregular inflectional endings on noun plurals. The German plural system consists of four different plural allomorphs. Plural nouns can be marked by /s/, by /e/, by /er/, by /n/ or they can remain unmarked. All German plural nouns other than /s/-inflected nouns are subject to a prosodic constraint that requires the plural form to end in a reduced syllable, i.e. an unstressed syllable with Schwa or a syllabic sonorant (Neef, 1998) (e.g. 1 Bär – 2 Bären 'bear(s)', 1 Tisch – 2 Tische 'table(s), 1 Kind – 2 Kinder 'child(ren)'). The plural ending /n/ is particularly suited to investigate selective deficits of regular or irregular inflected forms as it surfaces on regular as well as on irregular inflected nouns. On feminine nouns ending in /ə/ in the singular form (e.g. *[biːnə]* 'bee') the ending /n/ is completely predictable and considered to be regular (*[biːnən]* 'bees'). On masculine and neuter nouns that do not end in /ə/ in the singular form (e.g. *[bɛːɐ]* 'bear') the /n/ ending (e.g. *[bɛʀən]*) is considered to constitute an irregular ending since it is neither productive for masculine and neuter nouns nor predictable on the basis of the phonological shape of the nouns that take this ending (Wiese, 1999; Penke & Krause, 2002, Bartke et al., 2005). As the phonological complexity of regular and irregular /n/-inflected forms does not differ and as both types of /n/-plurals display a similar type frequency (Bartke et al., 2005), German /n/-plurals constitute an ideal test case to identify selective deficits of regular or irregular inflected forms.

2 Method

Noun plurals were elicited from 31 children and adolescents with DS (12 female) aged 4;07 to 19;02 years (*M* 14;05 years). For two of them the parents reported a mild hearing loss of less than 25 dB. For the remaining participants with DS no permanent hearing loss had been diagnosed. Nonverbal mental age (MA) was assessed using the SON-R 2.5-7 (Tellegen et al., 2007). It ranged from 2;11 to 6;05 years (*M* 4;05). Performance of the participants with DS was compared to a control group of 26 typically-developing (TD) children (13 female) matched in chronological age to the nonverbal mental age of the participants with DS (age TD group 3;04 - 5;07 years, *M* 4;05) (difference mental age DS group vs. chronological age TD group *p* = .63, ns.). All participants were monolingual speakers of German.

Elicitation of noun plurals followed the classical *wug*-test design (Berko, 1958). Participants were first presented with a picture displaying a single object named by the experimenter (e.g. *Look, this is a bee*). Then, a picture displaying three of these objects was presented and the participant was asked to produce a plural form (e.g. *Now there are some more. Now there are ___ ?*). In total, we elicited 40 noun plurals per participant. Here, I will focus on the 8 items eliciting regular /n/-plurals (henceforth n^{fem}-plurals) and the 8 items eliciting irregular /n/-plurals (henceforth n^{nonfem}-plurals). Regular and irregular /n/-items were matched for lemma and plural form frequency according to the CELEX database (Baayen et al. 1993) (n^{fem}-items: mean lemma frequency 24.4, mean plural frequency 12.6; n^{nonfem}-items: mean lemma frequency 37.4, mean plural frequency 17.1, difference p > .45 each). To tap into the productive abilities of the participants all tested items were of relatively low frequency.

Participants were tested individually after a short practice phase familiarizing them with the task. During testing, items were presented in the same previously randomized order for all participants and no feedback was given. All experimental sessions were video- and audiotaped. Participants' reactions were transcribed and transcripts were checked against the video files by a second independent researcher.

Produced forms were then evaluated for the correctness of the plural form of the 16 critical items. An inflectional error was counted if a wrong inflectional ending was used instead of the correct ending (e.g. **[biːnəs], *[bɛʁə]*) or if the inflectional ending was omitted (e.g. **[biːnə], *[bɛɐ]*). Based on these data, accuracy scores for n^{fem}- and n^{nonfem}-plurals were calculated for each participant and compared by a two-factorial, mixed *ANOVA*. The level of statistical significance was set at *p* < .05.

3 Results

Figure 1 presents the accuracy scores for the two groups of participants and the two types of /n/-plurals. For both types of /n/-plurals, the group of participants with DS achieved lower mean accuracy scores than the group of TD children. Whereas the group of TD children obtained a mean accuracy score of 91.3% for regular n^{fem}-plurals, the corresponding score for the group of participants with DS was at only 47%. For irregular n^{nonfem}-plurals the TD group obtained a mean accuracy score of 46.7%, the mean accuracy score of the participants with DS was 31.4%.

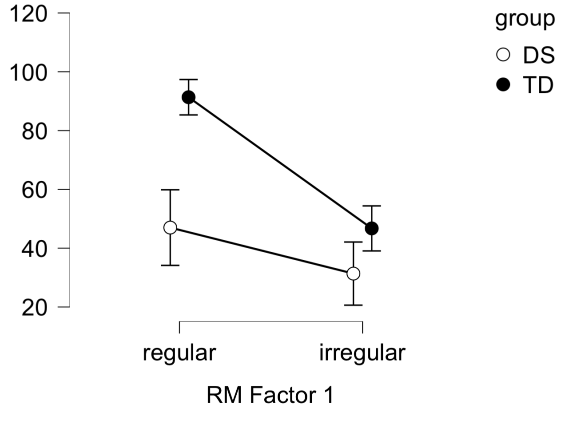


Figure 1: Accuracy scores for regular and irregular /n/-plurals obtained by the two participant groups.

A two-way factorial analysis of variance with SUBJECT GROUP (DS vs. TD) as between-subjects factor and REGULARITY (regular *n*^{fem}-plural vs. irregular *n*^{nonfem}-plural) as within-subject factor revealed a significant main effect of participant group ($F(1,55) = 21.2$, $p < .001$, $\eta p^2 = .28$) with the group of TD children achieving higher accuracy scores than the group of participants with DS. A significant main effect was also obtained for the factor REGULARITY ($F(1,55) = 119.2$, $p < .001$, $\eta p^2 = .68$), reflecting that accuracy scores for regular *n*^{fem}-plurals were significantly higher than accuracy scores for irregular *n*^{nonfem}-plurals in both groups of participants. In addition, the interaction of the factors SUBJECT GROUP X REGULARITY was also significant ($F(1,55) = 27.5$, $p < .001$, $\eta p^2 = .33$), indicating that group differences were more pronounced for regular compared to irregular plural forms. Indeed, post-hoc testing (*Bonferroni*) yielded no significant group difference for the accuracy scores obtained for irregular *n*^{nonfem}-plurals ($t = 2.2$, $p = .19$), whereas for regular *n*^{fem}-plurals the group of TD children significantly outperformed the group of participants with DS ($t = 6.3$, $p < .001$). Analyses correlating accuracy scores for both types of /n/-plurals with the chronological and the mental age of the participants with DS, yielded no significant relationships for *n*^{fem}-plurals ($p > .2$ each). For *n*^{nonfem}-plurals, accuracy scores displayed a tendency to increase with chronological age ($p = .061$).

Overall, 77.7% of the incorrect forms produced by the participants with DS were unmarked forms where the /n/-marking was missing. As mentioned above, a prosodic constraint requires all native German noun plurals to end in a reduced syllable. An error analysis was conducted to evaluate whether participants adhered to this constraint in their incorrectly produced noun plurals. Only nouns taking the *n*^{nonfem}-plural were evaluated in this analysis since nouns taking the *n*^{fem}-plural already end in a reduced Schwa-syllable in the singular form. In contrast, nouns taking the *n*^{nonfem}-plural typically end in a stressed syllable in the singular form. Leaving these forms unmarked, thus, results in a prosodically illicit plural form. This analysis yielded that a substantial proportion of the incorrectly produced plural forms for *n*^{nonfem}-nouns were left unmarked by the participants with DS (61.5%), thus violating the prosodic constraint on plural forms. This proportion was significantly higher than the proportion of unmarked forms produced by group of TD children (27.5%) ($t(54) = 2.95$, $p = .005$, $d = .79$). The high proportion of produced plural forms that do not adhere to the prosodic constraint on German plural nouns suggests that this prosodic constraint is not fully operative in the participants with DS. Interestingly, however, the huge majority of the unmarked forms produced by the participants with DS were accompanied by a quantifier (the numeral *three* or the quantifier *many*) (81.3%), suggesting that the concept of plural was expressed by the quantifier instead of the unavailable inflected plural form.

4 Discussion

The data indicate that noun plural inflection is impaired in German-speaking individuals with DS. This deficit selectively affects regular *n*^{fem}-plural formation. Whereas accuracy scores for regular *n*^{fem}-plurals were significantly lower compared to a group of TD children, accuracy scores for *n*^{nonfem}-plurals did not differ for the two groups of participants, suggesting that the production of irregular inflected /n/-plurals was at a level expected for the mental age attained by the participants with DS. This finding confirms previous findings on inflectional deficits in DS that have also found regular inflection (English past tense inflection and German past participle inflection) to be selectively affected in individuals with this syndrome (Eadie et al., 2002; Laws & Bishop, 2003; Penke, 2019).

Although regular and irregular /n/-inflected noun plurals display the same plural marker and the tested items were of similar frequency (lemma and word form frequency), participants' behaviour differed significantly with respect to regular and irregular /n/-plurals. TD children achieved a mean accuracy score for regular n*fem*-plurals of over 90%, indicating that they had acquired the regular n*fem*-plural marking. In contrast, accuracy scores for irregular /n/-plurals were significantly lower. The different development of regular and irregular inflected /n/-plurals in TD children and the finding that regular /n/-plurals were selectively affected in the group of participants with DS are in accordance with a dualistic view to inflection that states a qualitative difference between regular and irregular inflection.

The findings on incorrect, unmarked plural forms indicate that the observed language impairments in the participants with DS are not restricted to regular inflectional processes per se, but encompass prosodic constraints operating on the output of these processes. The observation that most unmarked nouns were produced with a preceding numeral suggests that participants with DS had already grasped the concept of plurality (see Clark & Nikitina, 2009) but had not yet acquired or could not access or produce the inflected plural form expressing this concept.

References

Baayen, H.; Piepenbrock, R. & H. van Rijn. 1993. *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistics Data Consortium, University of Pennsylvania.

Bartke, S.; Rösler, F.; Streb, J. & R. Wiese. 2005. An ERP study of German 'irregular' morphology. *Journal of Neurolinguistics* 18(1). 29-55.

Berko, J. 1958. The child's learning of English morphology. *Word* 14. 150-177.

Clark, E. & T. Nikitina. 2009. One vs. More than one: antecedents to plural marking in early language acquisition. *Linguistics* 47(1). 103-139.

Eadie, P.; Fey, M.; Douglas, J. & C. Parsons. 2002. Profiles of grammatical morphology and sentence imitation in children with specific language impairment and Down syndrome. *Journal of Speech, Language, and Hearing Research* 45(4). 720-732.

Laws, G. & D. Bishop. 2003. A comparison of language abilitites in adolescents with Down syndrome and children with specific language impairment. *Journal of Speech, Language, and Hearing Research* 46(6).1324-1339.

Næss, K.-A.; Lyster, S.-A.; Hulme, C. & M. Melby-Lervåg. 2011. Language and verbal short-term memory skills in children with Down syndrome: A meta-analytic review. *Research in Developmental Disabilities* 32(6). 2225-2234.

Neef, M. 1998. The reduced syllable in German. In R. Fabri, A. Ortmann & T. Parodi (eds.), *Models of inflection*, 244-265. Tübingen: Niemeyer.

Penke M. & M. Krause. 2002. German noun plurals: A challenge to the dual-mechanism model. *Brain and Language* 81. 303-311.

Penke, M. 2019. Regular and irregular inflection in Down syndrome – new evidence from German. *Cortex* 116. 192-208.

Pinker, S. 1999. *Words and rules*. New York: Basic Books.

Tellegen, P.; Laros, J. & F. Petermann. 2007. *SON-R 2½ -7. Non-verbaler intelligenztest*. Göttingen: Hogrefe.

Wiese, R. 1999. On default rules and other rules. *Behavioral and Brain Sciences* 22. 1043-1044.

The contribution of morphological skills to L2 reading comprehension

Serena Dal Maso
University of Verona

Sabrina Piccinin
University of Verona

1 Background

Over the last decades, psycholinguistic research has convincingly demonstrated the role of morphology as one of the organization criteria of the mental lexicon, both in adult speakers (see, e.g., Amenta & Crepaldi 2012 for review) and in developing readers (Burani et al. 2008; Colé et al. 2012). While such studies have mostly concentrated on the implicit unconscious processes underlying morphological organization, studies exploiting explicit metalinguistic measures of morphological skills have highlighted that speakers can also exhibit awareness of words' internal structure, and that such awareness is strongly associated with general reading comprehension abilities. Specifically, studies on L1 reading development have identified morphological awareness, defined as the ability of the speaker to perceive words' structure and to manipulate the smallest units of meaning in language (Carlisle 1995, 2000), as one of the strongest correlates of reading achievement. According to such studies, morphological skills can exert their influence in successful text comprehension, as demonstrated by the fact that children's knowledge about word structure emerges quite systematically as having a positive correlation with reading abilities and text comprehension skills. In other words, young readers with a high level of morphological awareness can better analyse meaning in morphologically complex words with cascading benefits to the understanding of the text as a whole (Carlisle 1995, 2000; Nagy et al. 2003; 2006). In this vein, Levesque, Kieffer & Deacon suggest that "[a]s a metalinguistic skill reflecting the synergy of sound and meaning, morphological awareness may be a foundational element of the linguistic system that works alongside other integration processes to build a mental model of the text while reading" (Levesque, Kieffer & Deacon 2017:18). Interest towards explicit metalinguistic abilities about morphology and how developing readers might put them to good use when reading a text have recently expanded to the field of L2 studies, with particular attention dedicated to children of immigrant families attending the L2 school system. A growing body of evidence has shown that morphological skills play a role in the reading skills of L2 and bilingual speakers too (e.g., Kieffer & Lesaux 2012; see Jeon & Yamashita 2014 for review), with findings indicating that the relationship between knowledge of morphology and reading comprehension becomes stronger between fourth and fifth grade, consistent with what has been observed in monolingual children.

2 Our study

So far, despite a growing attention to the issue of morphological knowledge in bilingual children in other languages (e.g., Vernice & Pagliarini 2018 on Italian, Fejzo 2020 on French), research on such population has mostly focused on English L2. Our study proposes to contribute to fill this gap, by focusing on Italian, for which studies focusing on children's morphological knowledge have largely made use of implicit on-line experimental techniques, focusing mainly on the contributions that morphology brings to L1 decoding skills, with the notable exception of the recent study by Vernice & Pagliarini (2018), which examined the relationship between morphological awareness and both decoding and reading comprehension skills in monolingual and bilingual pupils. Contrary to the literature on English L2, however, the study did not clearly find a correlation between reading comprehension skills in bilingual speakers (with Arabic L1) ranging from 3rd to 5th grade.

The goal of our study is twofold. First, we aim at further exploring the role of morphological knowledge in bilingual children's reading comprehension skills. By doing so, we will also reflect on the different dimensions underlying the construct of morphological awareness and on how they contribute to reading comprehension. It has indeed been acknowledged that the variety of tasks used in the literature may possibly tap different levels of morphological knowledge, not all of which may be relevant for text comprehension (McCutchen & Logan 2011). Recent literature (McCutchen & Logan 2011; Kuo & Anderson 2006; Deacon et al. 2017; Levesque et al. 2019; see also Carlisle 2000) tends to distinguish between morphological decoding, i.e., the ability to use morphemes to pronounce a word accurately, morphological structure awareness, i.e., awareness of the morphological structure of complex words, and morphological analysis, i.e., the ability to infer meaning from words' parts, identifying the latter as the crucial subcomponent of morphological knowledge involved in text understanding. On such premises, in order to disentangle the potential role of some of the subcomponents of morphological knowledge, we chose to assess the participants of our study through a combination of different tasks, focusing especially on the constructs of morphological (structure) awareness and morphological analysis, which are possibly more likely to impact on textual understanding.

2.1 Materials and Procedure

Participants were second-generation pupils with various language backgrounds attending 6th to 8th grade (n=47; mean age: 11,8) in three secondary schools located in Northern Italy. Preliminary tests were administered to ensure homogeneity of levels of the participants by using standardized assessment tools.

With regard to morphological skills, since no standardized measure is available for Italian, tests were specifically designed by the researchers, following some of the common proposals found in the literature. Specifically, we used the test of morphological structure (Carlisle 2000) to assess the subjects' sensitivity to the internal structure of the words, i.e., what is most commonly referred to as morphological (structure) awareness. In this task, subjects are required to identify either the derivative form or the base form of a word given as clue to be used in the context of a sentence provided, as in the following examples:

> [Decomposition] *Pescatore* ('fisherman'). *È severamente vietato* ____________ *in quel tratto di lago* ('It is strictly forbidden ____________ in that stretch of the lake')
> [Derivation] *Coltivare* ('cultivate'). *In Irlanda la* ______________ *di patate è molto diffusa.* ('In Ireland potatoes _________ is widespread').

The test was designed to assess both decomposition (i.e., identifying the base of a derivative word) and derivation skills (i.e., creating a derivative word starting from its given base).
The second task was a non-word suffix choice test (Tyler & Nagy 1989; Nagy et al. 2003): subjects were presented with sentences missing a word and were required to choose one among four given alternatives to fill in the gap. Crucially, such alternatives were all derived non-words, created through a legal combination of a non-existent base and an existent suffix, as exemplified below:

> *Dopo una lunga battaglia, i soldati infine si arresero a causa della* _______. ('After a long battle, the soldiers finally surrendered because of the________.')
> a) *ruvante;* b) *ruvabile;* c) *ruvezza;* d) *ruvista.*

This test assessed the subject's competence in the use of derivational suffixes, since it implicitly verified the students' ability to recognize the grammatical category of the word needed and to identify, among the given choices, the words that contained a suffix that was compatible with the needed grammatical category. While such a task is commonly supposed to assess morphological awareness, we believe it may tap into a more fine-grained aspect of morphological awareness, i.e., what Tyler & Nagy (1989) defined as syntactic knowledge about derivational morphology ("knowing that derivational suffixes mark words for syntactic category" Tyler & Nagy 1989:649). Finally, we administered a word knowledge test presenting morphologically complex words as target items (Deacon et al. 2017). The task consisted in a questionnaire in which subjects were presented with low-frequency words and were asked to indicate the meaning of such words by choosing among four given options. Crucially, in this test, the target words were composed of both a high-frequency base and a high-frequency suffix and were transparent from the point of view of the compositionality of meaning, as exemplified below:

Target word: *passivismo* ('the behaviour of someone who is passive')

a) *il comportamento di chi si crede superiore agli altri* ('the behavior of someone who believe her-/himselves superior to others')
b) *una persona che crede di essere superiore agli altri* ('a person who believes (s)he is superior to others')
c) *il comportamento di chi non prende l'iniziativa* ('the behavior of those who do not take the initiative')
d) *una persona che non prende mai l'iniziativa* ('a person who never takes the initiative')

The goal was indeed to encourage subjects to engage in meaning guessing strategies, relying on their knowledge of the meaning of word parts, rather than on their knowledge of the words as wholes. In other words, the test aimed at verifying pupils' awareness of suffixes' prototypical meanings and their ability to use such knowledge to infer the meaning of low-frequency semantically transparent complex words, i.e., what the literatures has defined as morphological analysis.

Finally, reading comprehension skills were measured through a standardized test specifically designed for Italian, *Prove di Lettura MT* (Cornoldi et al. 2017), in which participants are given a text and asked to answer a series of multiple-choice questions on its contents.

3 Results

Reading comprehension data show that, on average, subjects answered correctly to 7,9 questions out of 15 (52%), but almost half of the participants did not reach a sufficient level of performance in this test, according to the performance ranges set by the authors of the test. For what concerns morphological knowledge, taking into consideration the combined scores for the three tests, we observed that 74% of answers provided (globally) were correct, indicating overall a fairly good level of morphological skills. However, looking at the results of each specific test, there is an evident disproportion, in that the lexical knowledge test and the suffix choice test yielded respectively 59% and 65% of correct answers *versus* the 87% registered in the test of morphological structure. More specifically, in the derivation section of this test, 79% of the answers were correct, while in the decomposition section, the accuracy rate reached 94%.

Such results confirm the need to assess morphological knowledge on multiple levels: while pupils may have a generally well-developed sensitivity to the internal structure of words, this does not guarantee that they will be able to benefit from their knowledge of morphological structure. Indeed, recognizing word boundaries does not automatically entail being able to recognize affixes' prototypical meanings and syntactic functions. Crucially, since reading for understanding is a complex process implying a continuous integration of information in order to construct meaning, it is legitimate to expect that being able to use morphological information for understanding might be related to reading comprehension abilities. This hypothesis finds confirmation in our correlation analysis. Specifically, strongest correlations were found for the word knowledge test and the non-word suffix choice task (respectively, r=0.52, $p < .001$ and r=0.47, $p < .001$), while the correlation with the results of the derivational section of the test of morphological structure was weaker (r=0.35, p < 0.015) and no significant correlation with the results from the decomposition section was found. Ultimately, our study confirms the role of morphological skills in reading comprehension in second-generation bilingual pupils attending middle school in Italy, in line with the results found with other bilingual populations. At the same data, our data confirms the necessity of considering morphological knowledge as a multifaced construct, comprising different kinds of abilities which affect comprehension on multiple levels.

References

Amenta, Simona & Davide Crepaldi. 2012. Morphological processing as we know it: An analytical review of morphological effects in visual word identification. *Frontiers in psychology* 3:232.

Burani, Cristina, Stefania Marcolini, Maria De Luca & Pierluigi Zoccolotti. 2008. Morpheme-based reading aloud: evidence from dyslexic and skilled Italian readers. *Cognition* 108, 243–262.

Carlisle, Joanne F. (1995). Morphological awareness and early reading achievement, in Laurie B. Feldman (ed.), *Morphological Aspects of Language Processing*, 189–209. Hillsdale, NJ, England: Lawrence Erlbaum Associates.

Carlisle, Joanne F. 2000. Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing: An Interdisciplinary Journal* 12. 169-190.

Colé, Pascale, Sophie Bouton, Christel Leuwers, Severine Casalis & Liliane Sprenger-Charolles. 2012. Stem and derivational-suffix processing during reading by French second and third graders. *Applied Psycholinguistics 33*(1). 97-120.

Deacon, Hélène S., Xiuli Tong & Kathryn Francis. 2017. The relationship of morphological analysis and morphological decoding to reading comprehension. *Journal of Research in Reading* 40(1). 1-16.

Fejzo, Anila. 2020. The contribution of morphological awareness to vocabulary among L1 and L2 French-speaking 4th-graders. *Reading and Writing* 34. 659–679

Jeon, Eun Hee & Junko Yamashita. 2014. L2 Reading Comprehension and Its Correlates: A Meta-Analysis. *Language Learning* 64(1). 160-212.

Kieffer, Michael J. & Nonie K. Lesaux. 2012. Direct and indirect roles of morphological awareness in the English reading comprehension of native English, Spanish, Filipino, and Vietnamese speakers. *Language Learning* 62(4). 1170-1204.

Kuo, Li-jen & Richard C. Anderson. 2006. Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist* 41. 161–180.

Levesque, Kyle C., Michael J. Kieffer & Hélène S. Deacon. 2017. Morphological awareness and reading comprehension: Examining mediating factors. *Journal of Experimental Child Psychology* 160. 1-20.

Levesque, Kyle. C., Michael J. Kieffer & Hélène S. Deacon. 2019. Inferring meaning from meaningful parts: The contributions of morphological skills to the development of children's reading comprehension. *Reading Research Quarterly* 54(1). 63-80.

McCutchen, Deborah & Becky Logan. 2011. Inside incidental word learning: Children’s strategic use of morphological information to infer word meanings. *Reading Research Quarterly* 46(4). 334-349.

Nagy, William E., Virginia Berninger, Robert Abbott, Katherine Vaughan & Karin Vermeulen. 2003. Relationship of Morphology and Other Language Skills to Literacy Skills in At-Risk Second-Grade Readers and At-Risk Fourth-Grade Writers. *Journal of Educational Psychology* 95(4). 730-742.

Nagy, William E., Virginia Berninger & Robert Abbott. 2006. Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. *Journal of Educational Psychology* 98. 134-147.

Tyler, Andrea & William E. Nagy. 1989. The acquisition of English derivational morphology. *Journal of Memory and Language* 28*.* 649–667.

Vernice, Mirta & Elena Pagliarini. 2018. Morphological awareness a relevant predictor of reading fluency and comprehension? New evidence from Italian monolingual and Arabic-Italian bilingual children. *Frontiers in Communications* 3(11). 1-15.

Lexical strata in Japanese and Korean and the notion of lexeme

Clemens Poppe
Waseda University

1 Introduction: lexical strata and lexemes

Both Japanese and Korean are languages with three different lexical strata of an originally etymological nature. In both languages, the lexicon consists of so-called 'native' words, Chinese loan words, and more recent 'foreign' loan words (Shibatani 1990: Sohn 1999). In Japanese, for instance, both the free native element *hito* and the bound (°) Sino-Japanese elements °*nin* and °*jin* mean 'human, person' and are written with the same Chinese character. In Korean, the native element that expresses the meaning 'human, person' is *salam*, and the Sino-Korean element with the same meaning is °*in*.

In languages with neoclassical elements like French and English, we can find similar situations. Amiot and Dal (2007) discuss three different roots that refer to 'human': a native root *homme*, a Greek root, °*anthrop*, and a Latin root °*homin*, proposing that they belong to the same lexeme. A similar analysis is proposed for native and Sino-Japanese elements with corresponding meanings by Nagano & Shimada (2014), who also argue that *kanji* (Chinese characters) can be seen as representing lexemes.

The goal of this study is to critically discuss a number of issues in what we may call a 'shared lexeme' approach to native and Chinese loan elements in Japanese and Korean, and to look at the matter from the viewpoint of Construction Morphology (Booij 2010) and the closely related model of Relational Morphology (Jackendoff & Audring 2020).

2 Morphology, pragmatics, and orthography

For both Japanese and Korean, it has been pointed out that words belonging to the different strata have different ranges of meaning and stylistic values. Shibatani (1990: 144) writes that native words have broader meanings than Sino-Japanese and loan words, Sino-Japanese words have a more formal character and are used a lot in learned expressions, and foreign loan words have more modern and stylish connotations. A highly similar characterization of the three lexical strata in Korean is given by Sohn (1999: 88-89). In other words, native elements and Chinese elements have similar functions in the two languages, and what appear to be synonyms in reality are words with different pragmatic or sociolinguistic functions.

Given that such differences in abstractness and formality exist, the question arises why the elements that make up Sino-Japanese words should be analyzed as word-forms of the same lexeme as a corresponding native form with the same meaning. The main argument for such an analysis comes from the difference in morphological behavior between elements of the two strata: native elements are either free forms (in the case of nouns) or inflectional stems (in verbs and adjectives), whereas Sino-Japanese elements for which a corresponding native equivalent exists are bound forms. Based on this observation, Nagano & Shimada (2014) propose that nouns and verbs have two different stems: adopting the notion of 'stem space' (Montermini & Boyé 2012), they propose a distinction between a default stem and a compound stem. Their analysis is given in adapted form in (1a), where the default stem of nouns is called $\text{Stem}_{\text{Free}}$, a default stem which is in complementary distribution with a compound stem '$\text{Stem}_{\text{Comp}}$'. The forms in (1b) may serve as an example: the lexeme HITO ‘human, person’ has one free word-form *hito*, and two bound word-forms °*nin* and °*zin* (the distribution of which need not concern us here).

(1) a.

LEXEME	[− formal]	[+ formal]
[− bound]	$\text{Stem}_{\text{Free}}$	-
[+ bound]	-	$\text{Stem}_{\text{Comp}}$

b.

HITO	[− formal]	[+ formal]
[− bound]	hito	-
[+ bound]	-	°nin °zin

The tables in (1) give the impression that implicational relations exist between the two features. From the viewpoint of the language user, the [± formal] feature would seem to be the trigger to select a word based on bound or free elements. We would thus expect at least the implicational relations [+ formal]→[+ bound] and [− formal]→ [− bound]. Whether the relations also go in the opposite direction is a question that is more difficult to answer and depends on how we analyze the bound allomorphs of free forms which appear in compounds. There are two types of such allomorphs: *rendaku* stems and apophonic stems (see Labrune & Irwin 2021). To start with the latter type of allomorphy, the lexeme AME ‘rain’ is realized as *ame* in isolation, but either as *ame-* or *ama-* in compounds: *ame-huri* ‘rain-fall’ vs. *ama-gasa* ‘rain-umbrella’. Whether the free stem or the apophonic stem is used is only partially predictable, depending on a whole range of factors (Labrune & Irwin 2021). The phenomenon known as *rendaku* or ‘sequential voicing’ refers to cases in which a non-initial compound member that in isolation starts with one of the four voiceless obstruents /h/, /s/, /t/, or /k/ is realized with initial /b/, /z/, /d/, or /g/ in non-initial position in a compound, depending on several phonological, morphological, and lexical factors (Vance 2014; Irwin 2005). In the case of *hito* (‘human, person’), for instance, the *rendaku*-stem is °*-bito*, as in *mura-bito* (‘village-person = villager’), *turi-bito* (‘fishing-person = fisherman’), and *koi-bito* (‘love-person, i.e. lover’). It may be clear that if we analyze *rendaku* stems as [+ bound] forms, it is not possible to derive the value of the [± formal] from the [± bound] feature. To make things more complicated, sometimes even Sino-Japanese elements may undergo sequential voicing (Vance 1996). An example of such a word in *rin-goku* (‘next (to)-country = neighboring country’), which can also be pronounced as *rin-koku*, without *rendaku*. Nagano and Shimada (2014: 343, footnote 29) refer to *rendaku* as a “phonological voicing operation”, but this is a highly controversial characterization of the phenomenon; *rendaku* is far from regular and therefore should be viewed as morpho-phonological or lexical (van de Weijer et al. 2013; Vance 2014). As the marking of ‘compoundhood’ can be seen as the function of both rendaku and apophony (Labrune & Irwin 2021), it seems natural to treat them as compound stems. The consequences of such a view are shown in (2), where the *rendaku*-stem is given as Stem_{Ren} in (2a), and a concrete example in the form of the presumed structure of the lexeme KUNI ‘country’ in (2b). The following words are examples which contain the different allomorphs: *kuni* ‘country’, *shima-guni* ‘island country’, *koku-nai* ‘domestic’, *kan-koku* ‘South Korea’, and *chū-goku* ‘China’.

(2) a.

LEXEME	[− formal]	[+ formal]
[− bound]	$\text{Stem}_{\text{Free}}$	-
[+ bound]	Stem_{Ren}	Stem_{Ren} $\text{Stem}_{\text{Comp}}$

b.

KUNI	[− formal]	[+ formal]
[− bound]	kuni	-
[+ bound]	°-guni	°-goku °koku

Evidently, under the analysis in (2) we need the feature [± formal] to distinguish between native and Sino-Japanese bound stems. To deal with cases in which bound stems that are specified with the same value for the formality feature, we could assign a separate morphological feature [±rendaku] to *rendaku*-stems as in (3). This feature can be seen as an instance of the type of features used by Fradin (2003) to indicate that a stem is 'reserved' for a certain morphological construction (see also Amiot & Dal 2007). In other words, [±rendaku] could be formulated as [res: non-initial].

(3)

KUNI 'country'		[−formal]	[+formal]
[−bound]		kuni	-
[+bound]	[+rendaku]	°-guni	°-goku
	[−rendaku]		°koku

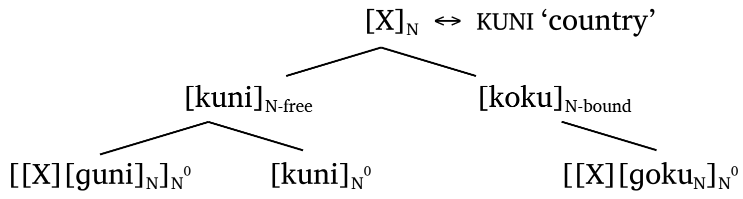
In Korean, a phenomenon similar to *rendaku* exists which involves obstruent tensification and nasal gemination and goes by the name of *sai-sios* (lit. 'between-s') (see Labrune 1999). The Korean name of this phenomenon refers to the letter *sios* ('s') of the native *hankul* script which is inserted between two members of a compound and written in the coda of the initial member of a compound, where it is optionally realized as /t/ if constraints on syllable structure are satisfied (Sohn 1999). The name 'compound tensification' refers to the tensification of the first consonant of the lax onset of the non-initial member of a compound. For instance, in the compound consisting of the native words *cho* ('candle') and *pul* ('fire'), the second member has an initial tense consonant /pp/ rather than the lax /p/ that appears in the isolation form: *cho(t)-ppul* ('candlelight'). The same tensification can be found in the often-cited compound *pom-ppi* 'spring rain', which consists of the lexemes *pom* 'spring' and *pi* 'rain'. The Sino-Korean element with the same meaning 'rain' is *u*, as in the word *u-chen* ('rain-sky = rainy weather'). As *sai-sios* related phenomena are not fully predictable, the *sai-sios* stems can also be thought to be lexically listed, as in the Japanese case in (3) above. The fact that Korean shows the same patterns as Japanese is important evidence for the idea that knowledge of Chinese characters is not necessary to acquire knowledge of the lexical relations that are interpreted as lexemic by Nagano & Shimada (2014). Korean is generally no longer written by means of a combination of Chinese characters (*hanca*), and more crucially, Chinese characters have no native Korean readings. Despite these facts, Song (1986: 493) explicitly states that "[a] Sino-Korean morpheme is free, if there does not exist a native word, denoting the same sense", and that "[i]f there is a native word, the SK morpheme is bound". The Korean case thus strengthens the case made for the morphological relatedness between native elements and Chinese loan elements in Japanese by Nagano & Shimada (2014).

From the above we may conclude that the complementary distribution in terms of the free vs. bound and informal vs. formal distinctions in both Japanese and Korean should be reflected in the morphological analysis. In the final part of this study, an alternative construction-based analysis is proposed in which a formal distinction is made between two types of bound stems: stems which are lexically marked as bound, and stems which are part of a schema that contains a lexical variable.

3 A view from Construction Morphology

In Construction Morphology (Booij 2010), the lexeme can be seen as a set of forms sharing particular semantic features, a morphosyntactic category, and often phonological properties. In a construction-based analysis of the Japanese and Korean data discussed above, native forms and Chinese loan forms can be thought to share their semantics and lexical class feature, but not their phonological form. This more abstract 'lexeme' node in the hierarchical network dominates two types of stems: as shown in (4), a native stem which is specified as a free stem, and a Chinese loan form which is specified as a bound stem. As a free stem, the native form [kuni] has a word form at the N^0 level. The *rendaku* allomorphs are dealt with by placing them one level lower than the more general schemas with a preceding variable 'X': $[[X][guni]_N]_{N^0}$ and $[[X][goku]_N]_{N^0}$. Although not shown in (4), these two constructions are at the same time instantiations of a more general *rendaku* schema, which captures the morpho-phonological nature of the allomorphy. (Note that both the $[kuni]_{N\text{-free}}$ and $[koku]_{N\text{-bound}}$ schemas also dominate further compound schemas which are not included in (4)) The distinction in formality, finally, can be captured by means of a default construction which expresses the implicational relation $[X]_{S\text{-bound}}$ ↔ [Pragmatics: formal], although possible alternative analyses will also be considered.

(4)



References

Amiot, Dany & Georgette Dal. 2007. Integrating neoclassical combining forms into a lexeme-based morphology. In Geert Booij, Luca Ducceschi, Bernard Fradin, Emiliano Guevera, Angela Ralli & Sergio Scalise (eds.), *On-line Proceedings of the Fifth Mediterranean Morphology Meeting*, 323–336. Bologna: Università di Bologna.

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

Fradin, Bernard. 2003. *Nouvelles approaches en morphologie*. Paris: Presses Universitaires de France.

Irwin, Mark. 2005. Rendaku-based lexical hierarchies in Japanese: The behavior of Sino-Japanese mononoms in hybrid noun compounds. *Journal of East Asian Linguistics* 14. 121–153.

Labrune, Laurence. 1999. Variation intra et inter-langue: morpho-phonologie du rendaku en japonais et du sai-sios en coréen. *Cahiers de Grammaire* 24. 117–152.

Labrune, Laurence & Mark Irwin. 2021. Japanese apophonic compounds. *Journal of Japanese Linguistics* 37(1). 25–67.

Montermini, Fabio & Gilles Boyé. 2012. Stem relations and inflectional class assignment in Italian. *Word Structure* 6. 69–87.

Nagano, Akiko & Masaharu Shimada. 2014. Morphological theory and orthography: *Kanji* as a representation of lexemes. *Journal of Linguistics* 50. 323–364.

Shibatani, Masayoshi. 1990. *The languages of Japan*. Cambridge: Cambridge University press.

Sohn, Ho-Min. 1999. *The Korean language*. Cambridge: Cambridge University Press.

Song, Ki Joong. 1986. Remarks on modern Sino-Korean. *Language Research* 22(4). 469–501.

Vance, Timothy J. 1996. Sequential voicing in Sino-Japanese. *The Journal of the Association of Teachers of Japanese* 30(1). 22–43.

Vance, Timothy J. 2014. If rendaku isn't a rule, what in the world is it? In Kaori Kabata & Tsuyoshi Ono (eds.). *Usage-based approaches to Japanese grammar: Towards the understanding of human language*, 137–152. Amsterdam & Philadelphia: John Benjamins.

van de Weijer, Jeroen, Clemens Poppe and Marjoleine Sloos (2013). Family matters: Lexical aspects of Japanese rendaku. In Jeroen van de Weijer & Tetsuo Nishihara (eds.). *Segmental variation in Japanese*, 129–148. Tokyo: Kaitakusha.

The role of attraction-repulsion dynamics in simulating the emergence of inflectional class systems

Erich R. Round
Surrey Morphology Group; Univ. of Queensland;
Max Planck Inst. for the Science of Human History

Sacha Beniamine
Surrey Morphology Group

Louise Esher
CNRS LLACAN

1 Introduction

Dynamic models of paradigm change can elucidate how the simplest of processes may lead to unexpected outcomes, and thereby can reveal new potential explanations for observed linguistic phenomena. Ackerman & Malouf (2015) present a model in which inflectional systems reduce in disorder through the action of an attraction-only dynamic, in which lexemes only ever grow more similar to one another over time. Here we emphasise that: (1) Attraction-only models cannot evolve the ***structured diversity*** which characterises true inflectional systems, because they inevitably remove all variation; and (2) Models with both attraction and repulsion enable the emergence of systems that are strikingly reminiscent of morphomic structure such as inflection classes. Thus, just one small ingredient — change based on dissimilarity — separates models that tend inexorably to uniformity, and which therefore are implausible for inflectional morphology, from those which evolve stable, morphome-like structure. These models have the potential to alter how we attempt to account for morphological complexity.

2 Structure in inflectional systems

Inflectional classes ('rhizomorphomes', Round, 2015) constitute groups of lexemes which share inflectional exponents; they are a type of 'morphomic', morphology-internal structure, mediating the mapping between content and form in inflection. Within complex inflectional systems in natural language, such structures are common, and are demonstrably both productive and psychologically real for speakers (Enger, 2014; Maiden, 2018); they are claimed to limit the complexity of the inflectional system by offering a systematic, recurrent and predictable means of distributing exponents (cf. Carstairs-McCarthy, 2010; Blevins, 2016).

A matter of ongoing debate is what kind of dynamics could potentially lead to such structure (Maiden, 2018; Carstairs-McCarthy, 2010). In this paper, we use computational iterated learning models to reveal for the first time some of the simplest conditions under which stable inflectional class systems can emerge. The insights afforded are of value both for our theoretical understanding of morphomic structure and for the formulation of explicit mathematical models of paradigm evolution, essential to tasks such as robust quantitative historical inference (Kelly & Nicholls, 2017).

3 Attraction-only models cannot evolve stable inflection classes

An early iterated learning model, implementing a simple paradigm cell filling task (Ackerman et al., 2009) in which a lexeme can change only by becoming more similar to another, is described in Ackerman & Malouf (2015). The initial input to the model consists of a lexicon in which paradigms are populated with randomly distributed exponents. At each cycle, the model must predict a held out value which we term the *focus*, at the intersection of a focal cell and lexeme. To predict the value of the focus, the model (i) picks a non-focal cell, which we call the *pivot*, (ii) selects all lexemes which share the exponent of the focal lexeme in the pivot cell, (iii) observes the exponents of these lexemes in the focal cell (exponents which jointly constitute *evidence*), and (iv) selects the most frequent exponent in the evidence to replace the held-out focus. Figure 1 illustrates one cycle. The result of one cycle is the input to the next.

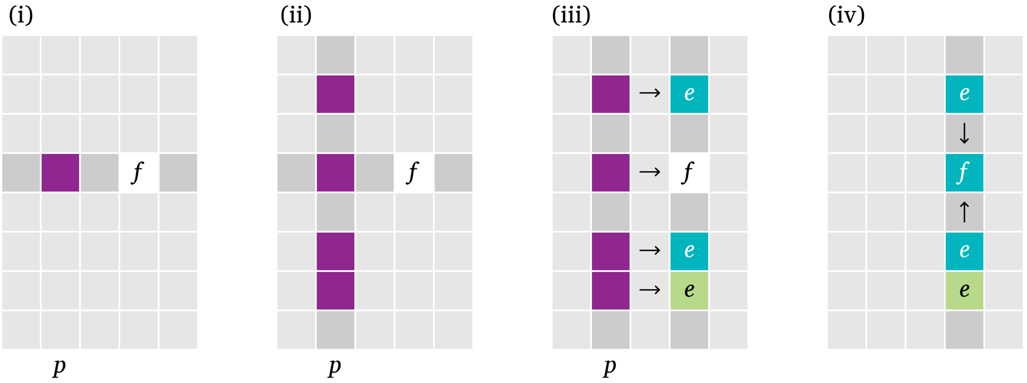


Figure 1: A cycle of Ackerman & Malouf's (2015) model. We label the focus f, the pivot cell p, and the evidence e.

This model is able to remove disorder from the system, because as lexemes change to be more like others, the dynamic is one of preferential attraction towards exponents that are already more frequent than their competitors. This rich-get-richer dynamic ensures that eventually, all lexemes converge on a single class (though discussion in Ackerman & Malouf (2015) focuses mainly on transitional states of the system just prior to ultimate uniformity). Ackerman & Malouf (2015) interpret this result as demonstrating the spontaneous emergence of self-organisational principles in morphological systems. We concur that the model exhibits self-organisation, but only of a radically homogenising kind. Here we investigate a family of minimally different dynamics (section 4); and their potential to generate outcomes which more closely resemble the morphological complexity of natural languages (section 5).

4 A family of dynamic, iterated learning models

As in Ackerman & Malouf's (2015) model, the initial input to our models consists of a lexicon in which paradigms are populated with randomly distributed exponents, and the basic task is again one of paradigm filling: at each cycle, the model must predict the *focus* based on *evidence* from multiple *pivots*, and the model's prediction is integrated into the lexicon input to the next cycle. All models in the family are similarly abstract and impoverished: the input provides indices for exponents rather than phonological forms; each cycle represents a general change in the system as a whole, with no method for capturing inter-speaker variation; and the models lack disruptive processes.

Our innovation is to alter the principles by which the paradigm filling task is accomplished, which we implement as modulable parameters to facilitate the controlled observation of their effects. We allow multiple pivot forms, reflecting the well-established observation that prediction based on multiple forms is more reliable than prediction based on pairings of forms or cells (Stump & Finkel, 2013; Bonami & Beniamine, 2016). In order to replicate the Zipfian frequency distribution observed for inflectional forms in natural language (Blevins et al., 2017), we introduce the option of frequency weighting in two ways: sampling foci in inverse proportion to their lexeme frequency, and sampling pivots and evidence proportionally to their frequency. Furthermore, we allow the process to be influenced by *negative evidence* (Voorspoels et al., 2015), which introduces an evolutionary *repulsion dynamic*, which at the right strength can enable structured diversity to emerge. To do this, rather than looking only at lexemes which have the same exponent as the focal lexeme in the pivot, we also observe lexemes with different exponents, and use these lexemes to provide evidence of exponents which are expected to be different from the focus. The relative proportion of negative and positive evidence can be varied, and the incorporation of at least some negative evidence proves crucial to the emergence of morphomic structure.

5 Results and implications of the attraction-repulsion dynamic

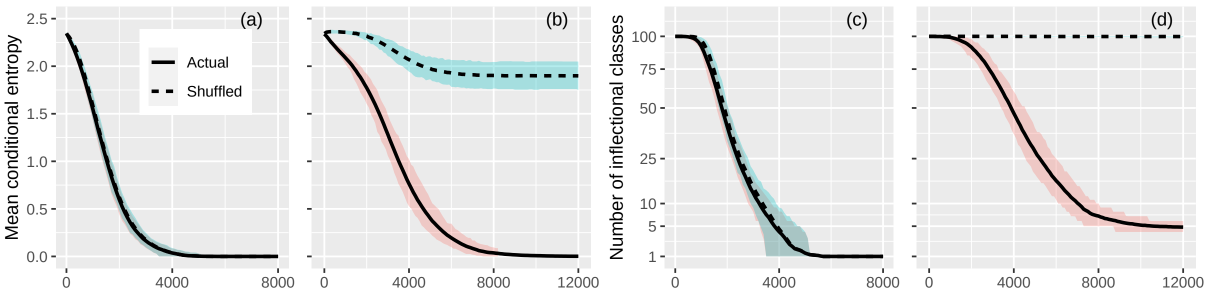


Figure 2: Evolution of mean conditional entropy (a,b) and number of inflectional classes (c,d) for Ackerman & Malouf's (2015) model (a,c) and an attraction-repulsion model (b,d), initial conditions: 100 lexemes, 8 cells, 6 exponents. Lines show means of 100 runs, shading shows 90% variation. The horizontal axis measures evolutionary cycles.

Figure 2 compares a replication of Ackerman & Malouf's (2015) model with a model which attends to negative evidence as well as positive (with a 30%-70% weighting). Both models were run 100 times and throughout their evolutions we track two measures: firstly, the complexity of the paradigm cell filling task measured as mean conditional entropy between cells (Ackerman et al., 2009; Ackerman & Malouf, 2015); and secondly, the number of inflection classes (unique rows in the lexicon). As Figure 2 shows, there is a dramatic reduction in conditional entropy over time in both models. However, we emphasise that this entropy metric has the undesirable property of conflating two kinds of systemic order: (1) order simply due to ***lack of variation***, and (2) order due to ***structured variation***. Our second metric, the number of inflection classes present, differentiates these two 'orderly' scenarios from one another, and shows that the attraction-repulsion model not only lowers entropy, but also does so while preserving distinct, stable inflectional classes, numbering 4.8 on average.

We also verify this finding by a second method. If low entropy, or a low number of inflectional classes, is due overwhelmingly to a *lack of variation* at any stage in the evolutionary process, then taking the exponents of any given cell and shuffling them among the language's lexemes should have little effect. This is what we see in the A&M model (Figure 2a,c), where the entropy and number of classes of this 'shuffled' version of the evolving system barely differ from the actual system itself. In contrast, if these properties are due to *structured variation*, then shuffling should cause both to be disrupted and both metrics will rise. This is precisely what we observe in the attraction-repulsion model (Figure 2b,d).

More generally, we found that a system's progression to uniformity could be slowed by incorporating additional parameters: reducing the amount of available pivots and evidence, or sampling these according to a Zipfian distribution to simulate frequency effects. However, stable inflectional structure still did not emerge in these enriched models, because the dynamic was still one of pure attraction. It is only once negative evidence is attended to, and thus a repulsion dynamic introduced into the system, that the models will consistently develop stable paradigmatic structure.

6 Conclusion

We have presented a new and fundamental mechanism for the spontaneous emergence of self-organising structure, which closely resembles what we find in natural language morphologies. Our findings have several topical ramifications. We clarify that morphome-like structure is able to emerge via a dynamic process consisting merely of piecemeal individual changes, within a system which does not explicitly represent morphomic structure (e.g. by means of morphomic indices). The simplicity of the developments involved indicates that, contrary to the prevalent characterisation of morphomic structure as 'unnatural' morphology, it is entirely plausible to view such structure as a *natural phenomenon* liable to arise spontaneously within inflectional systems.

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. 2015. The No Blur Principle Effects as an Emergent Property of Language Systems. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, vol. 41, doi:10.20354/B4414110014.

Blevins, James P. 2016. *Word and Paradigm Morphology*. Oxford: OUP.

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 & S. Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure* 9(2). 156–182. doi:10.3366/word.2016.0092.

Carstairs-McCarthy, Andrew. 2010. *The Evolution of Morphology*. Oxford: OUP.

Enger, Hans-Olav. 2014. Reinforcement in inflection classes: Two cues may be better than one. *Word Structure* 7(2). 153–181. doi:10.3366/word.2014.0064.

Kelly, Luke J. & Geoff K. Nicholls. 2017. Lateral Transfer in Stochastic Dollo Models. *The Annals of Applied Statistics* 11(2). 1146–1168. http://www.jstor.org/stable/26362220.

Maiden, Martin. 2018. *The Romance verb: Morphomic structure and diachrony*. Oxford: OUP.

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.

Stump, Gregory T. & Raphael Finkel. 2013. *Morphological Typology: From Word to Paradigm*. Cambridge: CUP.

Voorspoels, Wouter, Daniel J Navarro, Amy Perfors, Keith Ransom & Gert Storms. 2015. How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology* 81. 1–25. doi:10.1016/j.cogpsych.2015.07.001.

The Median Threshold Hypothesis : Measuring morphological productivity from frequency lists

Gauvain Schalchli
Université Bordeaux-Montaigne/CLLE

1 Context: A decline of frequency based studies on quantitative productivity because of methodology process limitations

Morphological productivity can be studied from a theorical point of view with attention focused on constraints on rules/processes/schemas application. However, another important point of view is the quantitative study of the extent of use of the morphological units. A first approach of quantitative studies is based on the type frequency of the morphological units. Specifically, this approach focused on new types in diachrony (Aronoff and Lindsay 2014; Berg 2020), neology (Cartier et al. 2018) or contemporary synchrony (Dal and Namer 2012, 2015; Dal et al. 2018). A different approach, inspired from corpus linguistics and psycholinguistics, is based on token frequency. In that approach, two aspects of productivity of morphological categories are captured: 1) the extent of new formations (the constitutive aspect) 2) lexicalized idiosyncratic items (the limitative aspect). Token frequency is estimated from the number of occurrences of the lexical units in a large and representative corpus.

The frequency-based quantitative approach of morphological productivity has been developed during the 90's from the works of Harald Baayen (Harald Baayen 1989, 1991, 1992a, 1992b, 1993, 1994, 1996, 2001, 2002; Harald Baayen and Lieber 1991; Chitashvili and Baayen 1993; Harald Baayen and Renouf 1996; Harald Baayen and Neijt 1997; Harald Baayen and Tweedie 1998; Plag, Dalton-Puffer, and Baayen 1999; Hay and Baayen 2002, 2003). He principally developed an index measure named potential productivity and defined, for one morphological process, the ratio between hapax number of the process and its cumulative frequency of occurrence (Baayen 2009; Gaeta and Ricca 2015; Dal and Namer 2016). This measure has been applied on a large scale for English (Baayen and Lieber 1991), Italian (Gaeta and Ricca 2003, 2006) and Dutch (Baayen 1989).

In French, the use of potential productivity index began in the first decade of the twenty-first century (Dal 2003). During that period, it has been applied on different special cases, like *-et/-ette* suffixation (Fradin, Hathout, and Meunier 2003), suffixes *-ité* and *-able* comparison (Grabar et al. 2006), comparison between *-able*, *-ité* et *-is(er)* suffixes (Namer 2003), denominal adjectival suffixes (Grabar and Zweigenbaum 2003), compound nouns (Voskovskaia 2009). However, the major study has concerned only 8 different processes (Dal et al. 2008).

At the same time of the first French studies, strong evidence has been produced of subtantial limitations of the Baayen's index. (Evert and Lüdeling 2001) shows that calculations based on automatic procedures are very different of those based on manual procedure. (Hay 2001) shows that relative frequency of bases impacts productivity. (Gaeta and Ricca 2003a, 2003b, 2006) show that: 1) the interaction between prefixation and suffixation in derivation cycles plays a role in productivity calculations; 2) productivity comparison between different processes with Baayen's index is only available for equal cumulative frequency.

Overall, these observations clarified and reinforced potential productivity, but also limited its scope and made its procedure more cumbersome. In correlation with that effect, we can observe an important slowdown in potential productivity studies. In Harald Baayen research, the last innovative paper on potential productivity dates from the beginning of the first decade of twenty-first century (Hay and Baayen 2003).[1] In French, the large scale project coordinated by Georgette Dal and initiated by the "morphological productivity" team of the GDR 2220 has been abandoned and the majority of works on productivity after 2008 deals with other quantitative strategies (Dal and Namer 2010; Koehl 2010, 2012; Cartier et al. 2018; Missud, Amsili, and Villoing 2020)[2]. We can observe the same tendency at the international level. From the two more recent reviews on productivity (Dal and Namer 2016; Gaeta and Ricca 2015), the last original works on potential productivity cited (Dal et al. 2008; Gaeta 2007) date from the first decade of twenty-first century (excepted some applied works like (Chmielik and Grabar 2011; Vendrell and Domínguez 2012; Wieling et al. 2014). In our own bibliographical exploration, we just found just three unreferenced works on potential productivity (Hennecke and Baayen 2017; Voskovskaia 2013, 2019). Moreover, the majority of research on productivity attempts to develop alternative strategies (Fernández-Domínguez 2010; Säily 2011, 2016; Berg 2020). The last but not least indication of that slowdown, the last international handbook on morphology (Audring and Masini 2019) contains no specific chapter on productivity, and only briefly cites potential productivity in the sub-chapter 3.5.2 "Productivity and blocking" (Lieber 2019).

2 Methodology: Facilitating the estimation of productivity from frequency lists by the Median Threshold Hypothesis

The fundamental intuition about frequency-based estimation of productivity is that the high part and the low part of the scale don't represent the same aspect of morphological knowledge (Fernández-Domínguez 2010). It is explicitly argued by Baayen (1992:110): "Any measure of morphological productivity [...] will have to satisfy a number of requirements. [...] such a measure should express "the statistically determinable readiness with which an element enters into new combinations." [...] taking into account these formations which are characterized by formally or semantically idiosyncratic properties should have the effect of lowering the value of the productivity measure." We propose to call the high frequency part of a lexicon his head and the low frequency part his tail.

Because it is inspired from biological probabilistic models (Chitashvili and Baayen 1993), Baayen chose the hapax count as an estimator for neologisms, however, not all hapaxes are neologisms and not all neologisms are hapaxes. On the other hand, the choice of cumulative frequency as estimator of high frequency items is another approximation. For example, (Baayen, Wurm, and Aycock 2007) use a threshold of 6 occurrences per million rather than the unique hapax rank for defined probable neologisms. Potential productivity estimate productivity with the minimal part of the lexicon tail and an extensive but confusing estimator of the lexicon head.

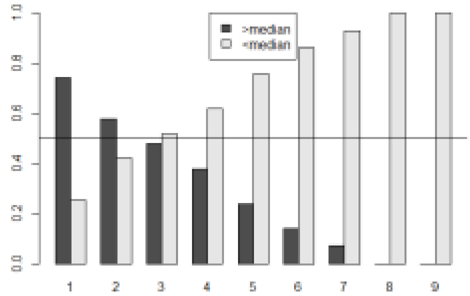
From a descriptive (VS inferential) statistical view on lexical frequency data, information about productivity extracted from high frequency items VS low frequency items could embrace the entire frequency scale. In order to avoid all groundless theoretical hypotheses, we propose to cut the lexicon and its frequency scale in two equal parts centered on the

[1] Further works of Harald Baayen did'nt answer the Gaeta & Ricca's discussion and did'nt apply the variable corpus approach (e.g. (Denistia and Baayen 2019; Shen and Baayen 2021))

[2] See also Dal & Namer (2012), Dal et al (2018)

median rank. From this point of departure, the estimation of productivity consists of comparing the number of instances of a morphological process in the head of the scale (over the median) with those present in the tail (under the median). The higher the number of the instances of the process in the tail, the more likely that the process is productive. Likewise, the higher the number of instances of the process in the head, the less likely that its productivity is strong.

In the following, we use the lexical frequency list of Lexique3 extracted from a 50 million words corpus of film subtitles and whose occurence counts are strongly correlated with lexical decision times (New et al. 2007). This list contains approximately 40,000 lexemes with phonological transcriptions. For that selection, the median rank frequency is 0.39 occurrences per million. 148 entries have exactly that frequency, 19,501 have a higher frequency than the median: this the head of the lexicon, while 20,382 have a lower frequency than the median: this is the tail of the lexicon. If we count the number of word forms more and less frequent than the median for each number of syllables by word, we find the barplot below:



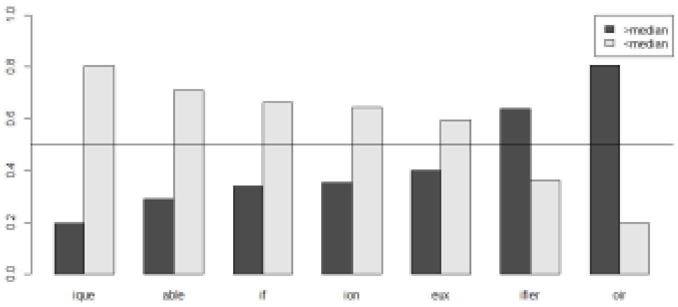
This barplot shows that short words belong mainly in the head and long words mainly in the tail. This observation deals with productivity, as it is well known that constructed words are typically longer than unconstructed words. We can express the relationship between the two sub-populations of a class of lexemes (e.g. one-syllable lexemes) by a ratio of T/H, similar to that of potential productivity, where T is the tail population and H is the head population. Applied on morphological categories, we can interpret the ratio in term of productivity. If the ratio is around 1, productivity level is medium. The more the ratio increases from one, the more the process is productive, and likewise unproductive in the contrary case. For example, in our sample, the ratios for one-syllable and two-syllable words which are rarely a result of productive morphological processes are approximately 0.3 and 0.7 whereas that of three-syllables and four-syllables syllable words which are more frequently morphologically constructed are approximately 1.2 and 1.75.

3 Results: Testing the hypothesis by discriminating French suffixal word endings from non-suffixal word endings

On the one hand, the median threshold hypothesis functions as a null hypothesis in statistical tests and allows us to differentiate between productive and non-productive processes or lexical properties: if the ratio of the numbers of instances of a process on either side of the median (T/H) is close to zero, then the index postulates that the process is not productive. If, on the contrary, the ratio is significantly greater than 0, then it must be postulated that the process is productive to some degree.

Moreover, the productivity index based on the median threshold hypothesis can give rise to different cases that logically split the productivity spectrum by acting as an index of the degree of productivity. If the tail of the vocabulary contains no instantiations of the process under study, the index is equal to 0, which is equivalent to null productivity. If the tail of the vocabulary contains as many instances of the process as the head, then the productivity is equal to 1, which corresponds to a significant productivity. Between these two first values, we can interpret the index in a gradual way: the closer the index is to zero, the lower the productivity and the closer it is to 1, the more significant it is. Finally, if the number of instances in the tail is higher than the number of instances in the head, then the productivity is high and the further away from 1 the higher it is.

Dal et al (2008) classify seven suffixes in three levels of productivity based on potential productivity index. From their counting, *-able* and *-ique* are highly productive, *-eux, -if, -ion* and *-ifier* are moderately productive and *-oir* has a low level of productivity. Applying our index to the data of Lexique3[3] shows a comparable classification of these seven suffixes:



-ique and *-able* have the higher tail proportion of attestations and *-oir* have the higher head proportion. About the ratios, *-ique* and *-able* are comparable to five-syllables words. *-if, -ion* and *-eux's* ratios are comparable to four-syllables, *-ifier* to two-syllable words and *-oir* to one-syllable words.
In order to validate the hypothesis on a large scale, we will present its application to different lexical categories and to different suffixes and morphological problems like allomorphy and competition from french data and from different corpora.

5 Conclusion

We propose a new quantitative estimation of productivity, comparable to Baayen's potential productivity but avoiding its shortcomings. The T/H ratio is based on frequency of occurrence, it is easy to compute from a list of frequencies, it is robust against corpus size variation and against automatic morphological analysis, it allows the comparability of all processes whatever their frequency of occurrence in the corpus, it takes into account in a consistent and interpretable way the effect of high frequencies on lexicalization and the relevance of low frequencies for morphology.

However, this proposal is simplistic and many refinements may be considered in order to advance the modeling of morphological productivity and to adapt it to different contexts or varieties. First of all, other thresholds can be easily tested. Second of all, it is consistent with

[3] For this example, we worked without manual validation of the morphological analysability and without category selection.

diachronic estimates and surveys from occasionalisms and non-conventional corpora such as the web. Finally, it allows us to imagine a dynamic interpretation of productivity as a function of the saturation phase of the derivational domain of the measured construction.

References

Aronoff, Mark, and Mark Lindsay. 2014. 'Productivity, Blocking, and Lexicalization'. *The Oxford Handbook of Derivational Morphology*. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199641642.001.0001/oxfordhb-9780199641642-e-005 (April 16, 2021).

Audring, and Masini, eds. 2019. The Oxford Handbook of Morphological Theory *The Oxford Handbook of Morphological Theory*. Oxford University Press. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199668984.001.0001/oxfordhb-9780199668984 (April 17, 2021).

Baayen, H., Lee H. Wurm, and Joanna Aycock. 2007. 'Lexical Dynamics for Low-Frequency Complex Words: A Regression Study across Tasks and Modalities'. *The Mental Lexicon* 2(3): 419–63.

Baayen, Harald. 1989. *A Corpus-Based Approach to Morphological Productivity: Statistical Analysis and Psycholinguistic Interpretation.*

———. 1991. 'A Stochastic Process for Word Frequency Distributions'. In *29th Annual Meeting of the Association for Computational Linguistics*, , 271–78.

———. 1992a. 'Quantitative Aspects of Morphological Productivity'. In *Yearbook of Morphology 1991*, Springer, 109–49.

———. 1992b. 'Statistical Models for Word Frequency Distributions: A Linguistic Evaluation'. *Computers and the Humanities* 26(5): 347–63.

———. 1993. 'On Frequency, Transparency and Productivity'. In *Yearbook of Morphology 1992*, Springer, 181–208.

———. 1994. 'Productivity in Language Production'. *Language and Cognitive Processes* 9(3): 447–69.

———. 1996. 'The Effects of Lexical Specialization on the Growth Curve of the Vocabulary'. *Computational Linguistics* 22(4): 455–80.

———. 2001. 'Word Frequency Distributions'. *Text, speech and language technology; 18.*

———. 2002. 'Affix Ordering and Productivity: A Blend of Phonotactics and Prosody, Frequency, and Lexical Strata’. In *Yearbook of Morphology 2001*, Yearbook of Morphology, eds. Geert Booij and Jaap Van Marle. Dordrecht: Springer Netherlands, 181–82. https://doi.org/10.1007/978-94-017-3726-5_6 (March 30, 2021).

———. 2009. ‘43. Corpus Linguistics in Morphology: Morphological Productivity’. *Corpus linguistics. An international handbook*: 900–919.

Baayen, Harald, and Rochelle Lieber. 1991. ‘Productivity and English Derivation: A Corpus-Based Study’. *Linguistics* 29(5): 801–44.

Baayen, Harald, and Anneke Neijt. 1997. ‘Productivity in Context: A Case Study of a Dutch Suffix’.

Baayen, Harald, and Antoinette Renouf. 1996. ‘Chronicling the Times: Productive Lexical Innovations in an English Newspaper’. *Language*: 69–96.

Baayen, Harald, and Fiona J. Tweedie. 1998. ‘Sample-Size Invariance of LNRE Model Parameters: Problems and Opportunities’. *Journal of Quantitative Linguistics* 5(3): 145–54.

Berg, Kristian. 2020. ‘Changes in the Productivity of Word-Formation Patterns: Some Methodological Remarks’. *Linguistics* 58(4): 1117–50.

Cartier, Emmanuel et al. 2018. ‘Détection automatique, description linguistique et suivi des néologismes en corpus : point d’étape sur les tendances du français contemporain’. *SHS Web of Conferences* 46: 08002.

Chitashvili, R. J, and R. H. Baayen. 1993. ‘Word Frequency Distributions’. In *Quantitative TextAnalysis*, Wissenschaftlicher Verlag Trier, eds. G. Altmann and L. Hrebicek. , 54–135.

Chmielik, Jolanta, and Natalia Grabar. 2011. ‘Détection de La Spécialisation Scientifique et Technique Des Documents Biomédicaux Grâce Aux Informations Morphologiques’. *TAL* 51(2): 151–79.

Dal, Georgette. 2003. *La Productivité En Questions et En Expérimentations*. Larousse.

———. 2008. ‘Quelques Préalables Au Calcul de La Productivité Des Règles Constructionnelles et Premiers Résultats.’ In *Congrès Mondial de Linguistique Française*, EDP Sciences, 142.

———. 2018. ‘Toile versus dictionnaires : Les nominalisations du français en-age et en-ment’. *SHS Web of Conferences* 46: 08003.

Dal, Georgette, and Fiammetta Namer. 2010. ‘Les Noms En-Ance/-Ence Du

Français: Quel (s) Patron (s) Constructionnel (s)?' *2ème Congrès Mondial de Linguistique Française*: 060.

———. 2012. 'Faut-il brûler les dictionnaires ? Ou comment les ressources numériques ont révolutionné les recherches en morphologie'. *SHS Web of Conferences* 1: 1261–76.

———. 2015. '133. Internet'. https://halshs.archives-ouvertes.fr/halshs-02275998 (April 16, 2021).

———. 2016. 'Productivity'. In *The Cambridge Handbook of Morphology*, ed. Andrew Hippisley & Gregory T. Stump. Cambridge University Press., 70–90. https://hal.archives-ouvertes.fr/hal-01303313 (December 18, 2020).

Denistia, Karlina, and R. Harald Baayen. 2019. 'The Indonesian Prefixes PE- and PEN-: A Study in Productivity and Allomorphy'. *Morphology* 3(29): 385–407.

Evert, Stefan, and Anke Lüdeling. 2001. 'Measuring Morphological Productivity: Is Automatic Preprocessing Sufficient'. In *Proceedings of the Corpus Linguistics 2001 Conference*, UCREL, 167–75.

Fernández-Domínguez, Jesús. 2010. 'Productivity vs. Lexicalization: Frequency-Based Hypotheses on Word-Formation'. *Poznań Studies in Contemporary Linguistics* 46(2): 193–219.

Fradin, Bernard, Nabil Hathout, and Fanny Meunier. 2003. 'La Suffixation En-ET et La Question de La Productivité'. *Langue française*: 56–78.

Gaeta, Livio. 2007. 'On the Double Nature of Productivity in Inflectional Morphology'. *Morphology* 17(2): 181–205.

Gaeta, Livio, and Davide Ricca. 2003a. 'Frequency and Productivity in Italian Derivation: A Comparison between Corpus-Based and Lexicographical Data'.

———. 2003b. 'Italian Prefixes and Productivity: A Quantitative Approach'. *Acta Linguistica Hungarica* 50(1–2): 93–112.

———. 2006. 'Productivity in Italian Word Formation: A Variable-Corpus Approach'.

———. 2015. 'Productivity'.

Grabar, Natalia et al. 2006. 'Productivité Quantitative Des Suffixations Par-Ité et-Able Dans Un Corpus Journalistique Moderne'. In *TALN*, , 167–75.

Grabar, Natalia, and Pierre Zweigenbaum. 2003. 'Productivité à Travers Domaines et Genres: Dérivés

Adjectivaux et Langue Médicale'. *Langue française*: 102–25.

Hay, Jennifer. 2001. 'Lexical Frequency in Morphology: Is Everything Relative?' *Linguistics*.

Hay, Jennifer, and Harald Baayen. 2002. 'Parsing and Productivity'. In *Yearbook of Morphology 2001*, Springer, 203–35.

———. 2003. 'Phonotactics, Parsing and Productivity'. *Italian Journal of Linguistics* 15: 99–130.

Hennecke, Inga, and Harald Baayen. 2017. 'A Quantitative Survey of N Prep N Constructions in Romance Languages and Prepositional Variability'. *Quaderns de filología. Estudis lingüístics* (22): 129–46.

Koehl, Aurore. 2010. 'Les Noms de Propriété Adjectivale En-Eur et-Esse: Un Modèle Évolutif Original'. *2ème Congrès Mondial de Linguistique Française*: 066.

———. 2012. 'La Construction Morphologique Des Noms Désadjectivaux Suffixés En Français'. PhD Thesis. Université de Lorraine.

Lieber, Rochelle. 2019. 'Theoretical Issues in Word Formation'. *The Oxford Handbook of Morphological Theory*. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199668984.001.0001/oxfordhb-9780199668984-e-3 (April 17, 2021).

Missud, Alice, Pascal Amsili, and Florence Villoing. 2020. 'VerNom: Une Base de Paires Morphologiques Acquise Sur Très Gros Corpus (VerNom: A French Derivational Database Acquired on a Massive Corpus)'. In *Actes de La 6e Conférence Conjointe Journées d'Études Sur La Parole (JEP, 33e Édition), Traitement Automatique Des Langues Naturelles (TALN, 27e Édition), Rencontre Des Étudiants Chercheurs En Informatique Pour Le Traitement Automatique Des Langues (RÉCITAL, 22e Édition). Volume 2: Traitement Automatique Des Langues Naturelles*, , 305–13.

Namer, Fiammetta. 2003. 'Productivité Morphologique, Représentativité et Complexité de La Base: Le Système MoQuête'. *Langue française*: 79–101.

New, Boris, Marc Brysbaert, Jean Veronis, and Christophe Pallier. 2007. 'The Use of Film Subtitles to Estimate Word Frequencies'. *APPLIED PSYCHOLINGUISTICS* 28(4): 661–77.

Plag, Ingo, Christiane Dalton-Puffer, and Harald Baayen. 1999. 'Morphological Productivity across Speech and Writing'. *English Language & Linguistics* 3(2): 209–28.

Säily, Tanja. 2011. 'Variation in Morphological Productivity in the BNC: Sociolinguistic and Methodological Considerations'. 7(1): 119–41.

———. 2016. 'Sociolinguistic Variation in Morphological Productivity in Eighteenth-Century English'. *Corpus Linguistics and Linguistic Theory* 12(1): 129–51.

Shen, Tian, and R. Harald Baayen. 2021. 'Adjective–Noun Compounds in Mandarin: A Study on Productivity'. *Corpus Linguistics and Linguistic Theory*. https://www.degruyter.com/document/doi/10.1515/cllt-2020-0059/html (June 25, 2021).

Vendrell, Mercedes Roldán, and Jesús Fernández Domínguez. 2012. 'Emergent Neologisms and Lexical Gaps in Specialised Languages'. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 18(1): 9–26.

Voskovskaia, Elena. 2009. 'Morphological Productivity and Family Size: Evidence from French Compound Nouns Garde-x and N-de-N'. In *Mediterranean Morphology Meetings*, , 123–33.

———. 2013. 'La Productivité Des Noms Composés En Français Du XVIIe Au Début Du XXe Siècle'. PhD Thesis.

———. 2019. 'Composés NN et NA Dans La Littérature Française Du 17e Au 20e Siècle: La Productivité Morphologique'. In Paris.: Université Paris Diderot.

Wieling, Martijn, Simonetta Montemagni, John Nerbonne, and R. Harald Baayen. 2014. *Lexical Differences Between Tuscan Dialects and Standard Italian: A Sociolinguistic Analysis Using Generalized Additive Mixed Modeling.*

When the *s* remains / does not remain an *s*: Further explorations in the acoustics of the English plural suffix

Marcel Schlechtweg & Greville G. Corbett
Carl von Ossietzky Universität Oldenburg, Germany & University of Surrey, UK

Whether or not specific variables trigger an acoustic distinction between phonologically identical forms has been a hotly debated territory. A key example is word-final *s* in English, and the question whether it systematically varies with different grammatical characteristics or functions. It has been demonstrated that affixal *s* (as in *laps*) differs from non-affixal word-final *s* (as in *lapse*) in duration (see, e.g., Plag et al. 2017; Seyfarth et al. 2018). Moreover, within affixal s, studies have revealed that several types differ in duration, such as plural (*cars*) and plural-genitive s (*cars'*) (see, e.g., Plag et al. 2020). These results put into question several established models of speech production, and the general understanding of the interplay of morphology, phonology, and phonetics (see, e.g., Bermúdez-Otero 2018; Fromkin 1971; Harley 1984; Kiparsky 1982; Levelt 1989; Levelt, Roelofs & Meyer 1999). In the spirit of such models, we should not expect acoustic variation since there is no direct connection between morphology and phonetics. Once the abstract and discrete phonological character of, say, the *s* has been formed, we should no longer expect duration differences, provided that the different conditions are comparable in terms of sentence position, frequency, and so on. The results reported above are thus more compatible with models that allow a flexible interplay between phonetics and other domains, such as morphology (see also, e.g., Pierrehumbert 2001, 2002). Here, we present two other studies to investigate the duration of the *s* in English, and to further examine the plausibility of different (psycho-) linguistic models.

The first experiment (Schlechtweg & Corbett 2021) investigated whether the *s* duration is distinct in regular plural (RPN, e.g., *toggles*) and pluralia tantum nouns (PTN, e.g., *goggles*). We conducted a reading experiment with Praat (Boersma & Weenink 2019) and tested the hypothesis that the *s* differs in duration. Two theoretical reasons for a possible duration difference are the informative and paradigmatic characteristics of the *s* in RPN and PTN (see also, e.g., Cohen 2014; Demuth 2011; Rose 2017). That is, first, the *s* is more informative in RPN than in PTN, since it distinguishes the plural from the singular form only in RPN (*toggles* vs. *toggle*) but not in PTN, which do not have a singular counterpart. Second, in terms of paradigmatic predictability, the *s* is more predictable in PTN than in RPN, since the former do not have a singular form. These two characteristics could in principle produce a difference in *s* duration.

One example of our test sentences is given in (1), and all 18 test items are given in Table 1.

(1) a. *The goggles appear to be broken and they're useless.*
b. *The toggles appear to be broken and they're useless.*

Table 1. Test items used.

PTN	RPN
shears	*beers*
trousers	*browsers*
earnings	*yearnings*
pliers	*fires*
tweezers	*freezers*
goggles	*toggles*
tongs	*gongs*
jeans	*screens*
odds	*pods*

We used nine test pairs with one PTN and a comparable singular-dominant RPN each. Potentially confounding variables were controlled for across the two conditions. The 18 test items ended in [z] and were inanimate. In each pair, identical sentences were used, the only difference being the target noun, which was either the PTN or the RPN. The PTN and the equivalent RPN had the same number of syllables, the same stress pattern, the same rhyme in the ultimate syllable, and at least two identical segments before the target segment [z]. Moreover, the frequencies of the two were not different (Mean RPN = 4.2 per million words (pmw); SD RPN = 4.7 pmw; Mean PTN = 4.2 pmw; SD PTN = 5.4 pmw; independent t test: t = 0.02, *p* = .988). The frequencies were gathered from the ukWaC corpus[1], a two-billion-words corpus containing materials from UK-based web pages. Also, the sequence "target noun + following word" had a frequency of 0 pmw in the ukWaC for all 18 test items and we therefore controlled for the syntagmatic probability (see, e.g., Cohen 2014). Each subject was tested on both items of a pair to exclude the influence of inter-subject variability. 36 filler sentences increased the distance between the two members of each pair. The two conditions were counterbalanced in the experiment. Analysis of the data of 40 native speakers of English and nine item pairs revealed no significant difference between the two conditions (for the descriptive statistics, see Table 2). Linear mixed effects models, conducted with the lme4 package (Bates, Maechler, Bolker & Walker 2015) in R (R Core Team 2021) and containing a random (intercepts for subject and item) and fixed effects structure (fixed effect of interest: noun type; several fixed effects control variables such as speech rate, frequency, and agreement) confirmed the finding.

Table 2. Descriptive statistics. Duration of *s* in seconds. Statistical outliers excluded. Total of 648 sound files.

Type of value	PTN	RPN
Mean	0.067	0.067
Standard deviation	0.017	0.017

Although we did not find a difference in *s* duration between PTN and RPN, another interesting effect was detected. Independently of the distinction between PTN and RPN, we found that the nouns in sentences with a past verb (e.g., *The odds / pods eventually dropped.*) contained a longer plural suffix than the nouns in sentences with a present tense verb (e.g., *The goggles / toggles appear to be broken and they're useless.*). Put differently, the *s* became longer if there was no overt number agreement between the noun and the verb in the sentence. One problem was, however, that the sentences did not only differ in terms of agreement but also in other respects, which might have caused the effect (e.g., *goggles / toggles* have two syllables, *odds / pods* have only one syllable). We decided to conduct another experiment to see whether the effect remains if different agreement conditions are entirely controlled for.

In the second study, we asked whether and how morphosyntactic agreement has an impact on the duration of word-final *s*. For this purpose, consider the examples in (2).

(2) a. *The blue cabs always break down.*
b. *The blue cabs always broke down.*
c. *These blue cabs always break down.*
d. *These blue cabs always broke down.*

We see four different situations. When we have a present tense verb, one finds an overt distinction in agreement between singular and plural (see (2a) *cabs break* (plural) in comparison to *cab breaks* (singular)). In (2b), however, there is no overt agreement of the past tense verb (see *cabs broke* (plural) in comparison to *cab broke* (singular)). In (2c), we observe agreement not only between the plural noun and the present tense verb, as in (2a), but also between the determiner and the plural noun (see *These cabs* (plural) in comparison to *This cab* (singular)). In

[1] http://corpus.leeds.ac.uk/itweb/htdocs/Query.html# (Hartley, Sharoff, Stephenson, Wilson, Babych & Thomas 2011).

(2d), there is no agreement between the plural noun and the past tense verb, as in (2b), but there is agreement between the determiner and the plural noun. We examined whether the plural *s* on the noun differs in duration across these conditions, with informative and syntagmatic reasons being potential candidates for acoustic differences (see, e.g., Rose 2017). For one, the *s* is more informative of plurality if there is no overt agreement, hence if there is no other plurality indication. Second, without overt agreement, the *s* is less predictable in the sentence.

We tested 12 native speakers of English, using 16 nouns in a reading study conducted with Praat (Boersma & Weenink 2020). Each person was exposed to all items in all conditions (16 items per person x 4 conditions per item = 64 experimental cases per person). As can be seen in the examples, we carefully controlled for potentially confounding variables by using the same sentences in all conditions and by exposing all subjects to all items in the four conditions. All target nouns are regular plurals, singular-dominant, monosyllabic, inanimate, and contain the voiced /z/ in the plural. All verbs are irregular and have the same number of syllables in the present and past tense. The order of the four conditions was counterbalanced both within and across subjects. In a very first analysis, based on the automatic segmentation with the MAUS tool (Kisler et al. 2017; Schiel 1999), we did not detect acoustic differences. However, in a subsequent analysis based on the important and more reliable and accurate manual segmentation (see, e.g., Schiel, Draxler & Harrington 2011; Schuppler, Grill, Menrath & Morales-Cordovilla 2014), we found that, using linear mixed effects models, the *s* was shorter if *these* occurred at the sentence beginning in comparison to if *the* was used.

In sum, after several studies showed that different types of the English *s* systematically differ in their duration, we did not find differences between PTN and RPN but between different agreement conditions. Our results provide some further and slight support for models permitting a flexible interaction between phonetics and higher-order levels such as morphology or morpho-syntax.

References

Bates, Douglas, Martin Maechler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. Version 1.1.26. *Journal of Statistical Software* 67 (1). 1–48.

Bermúdez-Otero, Ricardo. 2018. Stratal phonology. In S.J. Hannahs & Anna R.K. Bosch (eds.), *The Routledge handbook of phonological theory*, 100–134. New York, NY: Routledge.

Boersma, Paul & David Weenink. 2019/2020. *Praat: Doing phonetics by computer* [Computer program]. Retrieved from http://www.praat.org.

Cohen, Clara. 2014. Probabilistic reduction and probabilistic enhancement: contextual and paradigmatic effects on morpheme pronunciation. *Morphology* 24. 291–323.

Demuth, Katherine. 2011. The acquisition of phonology. In John Goldsmith, Jason Riggle & Alan C.L. Yu (eds.), *The handbook of phonological theory* (2nd ed.), 571–595. Malden, MA: Wiley Blackwell.

Fromkin, Victoria A. 1971. The non-anomalous nature of anomalous utterances. In Victoria A. Fromkin (ed.), *Speech errors as linguistic evidence* (1973, Janua Linguarum 77), 215–242. The Hague: Mouton (Reprinted from *Language* 47 (1), 27–52).

Harley, Trevor A. 1984. A critique of top-down independent levels models of speech production: evidence from non-plan-internal speech errors. *Cognitive Science* 8. 191–219.

Hartley, Tony, Serge Sharoff, Paul Stephenson, James Wilson, Bogdan Babych & Martin Thomas. 2011. *IntelliText*. http://corpus.leeds.ac.uk/itweb/htdocs/Query.html#.

Kiparsky, Paul. 1982. From cyclic phonology to lexical phonology (Part 1). In Harry van der Hulst & Norval Smith (eds.), *The structure of phonological representations*, 131–176. Dordrecht: Foris.

Kisler, Thomas, Uwe D. Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347.

Levelt, Willem J. M. 1989. *Speaking: from intention to articulation.* Cambridge, MA: The MIT Press.

Levelt, Willem J. M., Ardi Roelofs & Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22. 1−75.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: word frequency, lenition and contrast. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure* (Typological Studies in Language 45), 137−157. Amsterdam: John Benjamins.

Pierrehumbert, Janet B. 2002. Word-specific phonetics. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory phonology 7* (Phonology and Phonetics 4−1), 101−140. Berlin: Mouton de Gruyter.

Plag, Ingo, Julia Homann & Gero Kunter. 2017. Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics* 53. 181−216.

Plag, Ingo, Arne Lohmann, Sonia Ben Hedia & Julia Zimmermann. 2020. An <s> is an <s'>, or is it? Plural and genitive plural are not homophonous. In Lívia Körtvélyessy & Pavol Štekauer (eds.), *Complex words: Advances in morphology*, 260-292. Cambridge: Cambridge University Press.

R Core Team. 2021. *R: A language and environment for statistical computing*. R version 4.0.4. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Rose, Darcy Elizabeth. 2017. *Predicting plurality: an examination of the effects of morphological predictability on the learning and realization of bound morphemes.* Christchurch: University of Canterbury. (Doctoral dissertation)

Schiel, Florian. 1999. Automatic phonetic transcription of non-prompted speech. *Proceedings of the International Congress of the Phonetic Sciences (ICPhS)*. 607–610.

Schiel, Florian, Christoph Draxler & Jonathan Harrington. 2011. Phonemic segmentation and labelling using the MAUS technique. Workshop *New Tools and Methods for Very-Large-Scale Phonetics Research*, University of Pennsylvania, January 28-31, 2011.

Schlechtweg, Marcel & Greville G. Corbett. 2021. The duration of word-final *s* in English: A comparison of regular-plural and pluralia-tantum nouns. *Morphology.* (Online first)

Schuppler, Barbara, Sebastian Grill, André Menrath & Juan A. Morales-Cordovilla. 2014. Automatic phonetic transcription in two steps: forced alignment and burst deletion. In Laurent Besacier, Adrian-Horia Dediu & Carlos Martín-Vide (eds.), *Statistical language and speech processing: Second International Conference, SLP 2014 Grenoble, France, October 14-16, 2014 Proceedings*, 132−146. Cham: Springer.

Seyfarth, Scott, Marc Garellek, Gwendolyn Gillingham, Farrell Ackerman & Robert Malouf. 2018. Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience* 33 (1). 32−49.

An extensive analysis of blends in Contemporary Italian

M. Silvia Micheli
University of Milano – Bicocca

1 Introduction

Blending is generally considered a scarcely productive mechanism in Italian word-formation, mostly exploited for the creation of names of companies or associations (e.g., *Polfer* 'railway police' < *pol(izia)* + *fer(roviaria)*; see Thornton 1993: 148). More recently, Cacchiani (2016) has shown that the significant transfer of English blends has led to their gradual increase in productivity, though mostly in specific domains where creativity is widely exploited, such as children's literature and brand naming. Previous studies on Italian blends have been focused on both phonological and morphological properties shown by blend forms (cf. Thornton 1993, 2004; Bertinetto 2001). They have highlighted that Italian blends show a strong tendency to shorten only the first element, the second element remaining intact (e.g., *cantautore* < *cant(ante)* 'singer' + *autore* 'author'). In Thornton (1993: 148), they are not considered as prototypical blends, but rather as *partial* (or *peripheral*) *blends*, in that the second constituent does not undergo modification, contrary to what can be frequently observed in English, where both source words undergo a shortening, and the initial part of the first word combines with the final part of the second word (e.g. *vog* < *v(olcanic)* + *(f)og*, *fanzine* < *fan(atic)* + *(maga)zine)*. Although this tendency does not represent a strict rule: since the modification of the second constituent is also attested (e.g. *immigriano* 'the variety of Italian spoken by immigrants', *immigr(ato)* 'immigrant' + *(ital)iano* 'Italian', see Thornton 2004 for other examples), this implies that in Italian blends and neoclassical compounds containing a native combining form (henceforth, CF)[1] and an autonomous word frequently exhibit comparable formal features (e.g., *cinesaga* 'film saga', where *cine-* is a shortening of *cinema* 'id.'). The boundaries between blend's parts (or splinters) and other morphological elements (especially CFs and secreted affixes) have been thoroughly investigated in several studies (see, among others, Fradin 2000 and Fradin, Montermini & Plénat 2009 on French; Mattiello 2017, 2020 on English). According to the framework of Natural Morphology, blends are placed outside grammar due to their irregularity and unpredictability, which make them different from both grammatical word-formation mechanisms (such as compounding and derivation) and mechanisms placed at the boundaries between two subcomponents of morphology (belonging to so called "marginal morphology", see Dressler 2000; e.g., CFs), which both allow a prediction of a regular output. Other scholars (among others, Plag 2003) have argued that blending can be considered as a rule-governed (i.e., grammatical) phenomenon, by virtue of the (language-specific) phonological regularities that they show. A tendency towards regularity (and productivity) in blending has been recognized also in Mattiello (2013, 2017), which highlighted the role of analogy in conferring regularity (and predictability) to English blends. In particular, it has been shown that blends significantly frequent in use can serve as model for the creation of new words and produce series through "analogy via schema" (e.g. *-(a)holic* 'person addicted to' in *shopaholic*, *sportsaholic*, *chocoholic*), thus moving closer to regular morphological elements, such as CFs and true affixes (e.g. *-zilla* 'an overbearing person or an aggressive species' in *mumzilla*, *brandzilla*, *teenzilla*, from *Godzilla*). The goal of this study is twofold. On the one hand, it aims at deepening the boundaries between

[1] Since neoclassical CFs do not represent the outcome of a process of shortening, our dataset does not include compounds containing Greek/Latin CFs (e.g., *cardiologia* 'cardiology', *cardiochirurgia* 'cardiac surgery' < *cardio-* 'heart' + *chirurgia* 'surgery'). Similarly, we have left out words containing native CFs which have not undergone a shortening but just a modification (e.g., *mafiostruttura* 'mafia structure' where *mafio-* < *mafia*).

compounding with CFs and blending in Italian, identifying splinters that have acquired more regularity and morpheme status. On the other hand, we provide an updated description of Italian blending by analysing a sample of blends attested in the last two decades.

2 Methodology

This study is based on a sample of neologisms extracted from the Treccani Neologism Dictionary, which includes Italian new words (both nonce words/occasionalism and true neologisms) attested in a period ranging from around 2004 to March 2020.[2] The selection of words to be analysed has been carried out manually. The analysis provided in this paper consists of two parts. In the first part, we have extracted words where at least one of the two constituents has undergone a shortening and classified them into the categories illustrated in Table 1; the classification has been based on parameters identified by previous studies.[3] The first parameter deals with the presence of morphological series,[4] that is typical of CFs or secreted affixes, in contrast with blends which are generally type hapaxes (see Mattiello 2020; Fradin, Montermini & Plénat 2009); moreover, an increase in terms of type frequency can be considered as a cue that a splinter is gradually acquiring morpheme status. From the phonological point of view, previous studies have noted that prototypical CFs tend to show the structure of the minimal prosodic word (see Thornton 1996), while in blends the shortening occurs in many different patterns and can lead to a significant reduction of the source word (see the already mentioned case of *vog*). Moreover, it should be taken into account that overlap (both local and global) of constituents as well as a certain similarity/assonance between the source words are frequently found in blends, not in CFs. From the semantic standpoint, the emergence of a new, more abstract/specialized meaning can be considered as a clue that a splinter is gradually acquiring the status of secreted affix (being thus the word closer to derivation than compounding).

	Series	Significant reduction of the source word	Overlap	Semantic change
Words containing a secreted element	√	X	X	√
Compounds with shortened CFs	√	X	X	X
Blends	X	(√)	(√)	X

Table 1 Parameters of analysis

In the second part of the study, we focus on Italian blends and provide a description of their features according to the parameters proposed by Gries (2004: 646), i.e., shortening of source words, linearization, and overlap. The whole analysis is supported by a corpus investigation based on a corpus of Contemporary Italian, i.e., Timestamped JSI web corpus 2014-2020 (7.6 billion tokens), searched through the SketchEngine interface. This resource will provide quantitative data to verify the presence of morphological series and to distinguish nonce blends (i.e., occasionalisms) from blends accepted in the lexicon (i.e., neologisms).

3 Preliminary results

The sample extracted from the Treccani Neologism Dictionary consists of 743 words, including 67 adjectives (9%), 501 nouns (67.4%), 144 (agent) nouns that can also function as adjectives, 24 names (3.2%), and 6 verbs (0.8%). Figure 1 summarizes the results of our analysis.

[2] The collection of neologisms is available at the following link (accessed: 29/06/2021): https://www.treccani.it/magazine/lingua_italiana/neologismi/.

[3] Parameters in brackets are not strictly binary: they represent trends, e.g., the significant reduction of the source word is frequent in blending but does not necessarily occur in all blends, while morphological series represent a clear indication that a splinter has acquired a morpheme status.

[4] Conventionally, we consider a series as a set of at least 15 types attested within corpora.

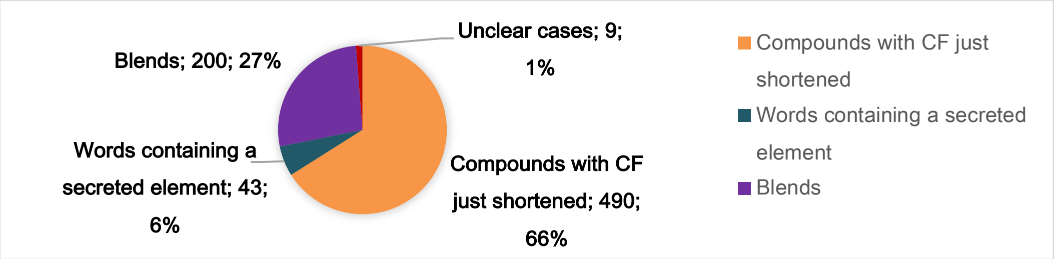


Figure 1 Classification of the dataset

Our dataset mostly includes compounds with CFs, while blends represent about one-third of the sample. More specifically, compounds with CFs just shortened represent the most attested category (i.e., 490 types), which includes both CFs already identified by previous studies (e.g. *cine-* < *cine(ma)* 'cinema', *catto-* < *catto(lico)* 'Catholic', etc.) and a set of morphological elements which have been previously considered as blend's parts, e.g., *risto-* (< *risto(rante)* 'restaurant'), *panta-* (< *panta(loni)* 'trousers'), *aperi-* (< *aperi(tivo)* 'happy hour'). They now occur in morphological series well attested in corpora and have acquired a certain degree of regularity; they mostly represent the leftmost constituent, but some cases where they are the rightmost element are attested, e.g., *-fonino* < *(tele)fonino* 'mobile'. From the semantic point of view, this kind of CFs do not show a semantic abstraction or specialization but reflect the original meaning of the source word. On the other hand, our dataset also contains a limited number of words (closer to derivation than compounding) containing secreted elements, i.e., elements that have undergone a semantic change. In particular, a process of abstraction can be found in *nazi-* (< *nazista* 'nazist'), when it refers to 'a person that takes radical positions, hard-liner' (e.g. *nazivegano* 'radical vegan'; cf. Eng./It. *grammarnazi*), and in *turbo-* (< *turbina* 'turbine'), which conveys the idea of speed (e.g., in *turbo-vacanza* 'short holiday') or intensification (e.g., *turbo-buonista* 'super feel-good') when it combines with adjectives. A case worthy of particular attention is represented by *-iota* (e.g., *destriota* 'typical member or sympathizer of right parties'; originally a splinter from *(id)iota* 'idiot'), in that it has developed a new more specific meaning (i.e., 'idiot' > 'typical member of a given group characterized by obtuseness'), which always entails an evaluative (i.e., pejorative) value: as we will show in more detail, the emergence of this new meaning is related to the dissemination of the word *italiota* 'typical average Italian' as used in political discourse. Finally, we have found some unclear cases, namely words made up of a CF as leftmost constituent and a segment of a word, e.g., *angloliano* 'mix of Italian and English' (< *anglo-* + *(ita)liano*), *cybertariato* 'proletariat of digital workers' (< *cyber-* + *(prole)tariato*).

As far as blends are concerned, our dataset contains 200 blends, which have been analysed according to the three parameters discussed in the previous Section (i.e., shortening, linearization, overlap). The quantitative results for each parameter are illustrated in the following Table and will be discussed in more details during the presentation.

Shortening	**Word1**	**Word2**	**Word1Word2**	**-**
	61	31	76	33
	cimitour 'cemetery tour' < *cimi(tero)* 'cemetery' + *tour*	*mielenoso* 'both sweet and toxic' < *miele* 'honey' + *(vele)noso* 'toxic'	*acqumba* 'Zumba in the water' < *acq(ua)* 'water' + *(Z)umba* 'id.'	*blogorroico* 'prolific blog writer' < *blog* + *logorroico* 'loquacious'
Linearization	**+**		**-**	
	195		6	
	genobiltà 'genetic aristocracy' < *gen(etica)* + *nobiltà* 'aristocracy'		*sprecheurare* 'to waste European founds' < *sprecare* 'to waste' + *euro* 'id.'	
Overlap	124		77	
	narcisindaco 'narcissist mayor' < *narcisi(sta)* 'narcissist' + *(si)ndaco* 'mayor'		*erogossip* 'erotic rumors' < *ero(tico)* 'erotic' + *gossip*	

Table 2 Italian Blends: Results

4 Discussion

The analysis has provided a classification of words where at least one constituent has suffered a shortening. It has been shown that elements such as *risto-*, *panta-*, *-fonino* are ascribable to the category of CFs (rather than to that of splinters), in that they occur in series and are well attested within corpora. We have also identified the case of *-iota* that demonstrates that a splinter can develop a new meaning, which makes it comparable with true affixes. On the other hand, the analysis of Italian blends has highlighted that all types of shortening identified by Gries (2004) are well-attested in Italian, including cases where the source words are not shortened. As far as linearization parameter is concerned, although most blends are made up of splinters arranged one after the other, some cases where a splinter is inserted within the other are attested. Finally, the analysis of overlap in Italian blending has confirmed that it represents a factor that favours blend formation, together with phonological resemblance.

References

Bertinetto, Pier Marco. 2001. Blends and syllabic structure: A four-fold comparison. In Mercé Lorente, Núria Alturo, Emili Boix, Maria-Rosa Lloret & Lluís Payrató (eds.), *La gramática i la semántica en l'estudi de la variació*, 59–112. Barcelona: Promociones y Publicaciones Universitarias S.A.

Cacchiani, Silvia. 2016. On Italian Lexical Blends. Borrowings, Hybridity, Adaptations and Native Word Formations. In Sebastian Knospe, Alexander Onysko & Maik Goth (eds.), *Crossing Languages to Play with Words*, 305–336. Berlin-Boston: De Gruyter.

Dressler, Wolfgang U. 2000. Extragrammatical versus marginal morphology. In Ursula Doleschal & Anna M. Thornton (eds.), *Extragrammatical and Marginal Morphology*, 1–9. Munich: LINCOM.

Fradin, Bernard. 2000. Combining forms, blends and related phenomena. In Doleschal, Ursula & Thornton, Anna M. (eds.), *Extragrammatical and marginal morphology*, 11-59. München: LINCOM.

Fradin, Bernard. 2015. Blend. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer, (eds.), *Word-Formation. An International Handbook of the language of Europe*, 386–413. Berlin-New York: De Gruyter.

Fradin, Bernard, Montermini, Fabio & Marc Plénat. 2009. Morphologie grammaticale et extragrammaticale. In Bernard Fradin, Françoise Kerleroux & Marc Plénat (eds.), *Aperçus de morphologie du français*, 21–45. Saint-Denis: Presses Universitaires de Vincennes.

Gries, Stephan. 2004. Shouldn't it be breakfunch? A quantitative analysis of blend structure in English. *Linguistics* 42(3). 639–667.

Mattiello, Elisa. 2017. Paradigmatic Morphology. Splinters, Combining Forms, and Secreted Affixes. *SKASE Journal of theoretical linguistics* 15(1). 2–22.

Mattiello, Elisa. 2020. A corpus-based analysis of English blends. *Lexis* 14(2). 1–28.

Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press.

Thornton, Anna M. 1993. Italian blends. In Livia Tonelli & Wolfgang U. Dressler (eds.), *Natural Morphology: Perspectives for the Nineties*, 143–155. Padova: Unipress.

Thornton, Anna M. 1996. On some phenomena of prosodic morphology in Italian. Accorciamenti, hypocoristics and prosodic delimitation. *Probus* 8. 81–112.

Thornton, Anna M. 2004. Parole macedonia. In Maria Grossmann & Franz Rainer (eds.), *La formazione delle parole in italiano*, 599–610. Tübingen: Niemeyer.

Exploring morphological connexions within the mental lexicon: evidence from speakers from diverse educational backgrounds

Despoina Stefanou,
Madeleine Voga et
Hélène Giraudo

1 Introduction

The role of morphology in word perception has been studied through various protocols and settings, among which lexical access protocols, often conducted with the masked priming technique. These studies demonstrate the existence and the role of morphological connections within the mental lexicon (e.g., Feldman, O'Connor & Moscoso del Prado Martin, 2009), while the theoretical specification of what is implied by the term "mental lexicon" is currently the subject of much debate. In what follows, we take the theoretical option of a mental lexicon in which individual words are represented and form connexions with each other based on their systematic common characteristics (form and/or meaning). The experimental evidence accumulated until now clearly shows that morphologically related words tend to facilitate processing of each other, in the same language (e.g., Drews & Zwitserlood, 1995) but also through languages, i.e., in cross-linguistic priming, giving rise to language co-activation effects (e.g., Mulder, Dijkstra, Schreuder & Baayen, 2014). Without going into detail here, we will admit that experimental research, despite some methodological criticisms, has the potential of enriching our understanding of how language in general, and morphology in particular, work (Baayen 2014). This is precisely why the role of some facts about language and its users should make the object of intensive research.

An important issue is about the fact that, contrary to the popular opinion relative to generative linguistics "idealized speaker", all speakers are not equivalent with respect to language use, and possibly to language representation. In the case of masked priming protocols tapping into morphological processes, the variable "speaker" is not often considered, or, to put it in another way, it is a special profile of speaker which is taken into account. The participants tested in most published literature tend to be highly educated students of which the majority is female, very often attending philological curricula. However, it is widely admitted that "differences in individual language users may lead to remarkably different use of the possibilities offered by the grammar of 'the language'" (Baayen 2014: 100). These can be sex differences (Kimura 2000, for a comparison between the verbal skills of men and women), or differences related to speakers' experience with language, leading to the study of variables such as the "vocabulary size" (e.g., Mainz, Shao, Brysbaert & Meyer, 2017) or the exposure to print. Differences between speakers can also be related to exposure to heritage language and its use.

2 Our study

Given the above, we sought to combine the questioning related to morphological processing and representation, to that related to speakers' background, i.e., speakers/readers with diverse educational backgrounds. This questioning arises as a result of previous findings that will be briefly presented below.

2.1 Previous findings on the *-isme/-iste* and *-isme/-ique* connexion

The protocol and results we briefly describe in this section were part of a study (Voga & Anastassiadis-Symeonidis, 2018) designed to compare bilingual and monolingual processing, in which bilingual participants (Exp. 1a) were the "prototypical student group", whereas monolinguals (Exp. 1b) were students in the public technical school of Thessaloniki. The experiment was designed to be "transposable" from a bilingual to a monolingual group. All the experiments presented here (§2.1 and §2.2) used the masked priming technique with a 48ms SOA (Stimulus Onset Asynchrony), which is a prime duration that prevents the participant from consciously processing the prime, and which generally leads to morphological priming effects (and identity ones[1]). The task was lexical decision (LD, YES, it is a word/NO, it is not). The stimuli tested in these experiments were selected to activate the morphological connexion between *-isme* and *-iste* (cf. table 1, a and b), two related morphemes that exist in Greek as well as in French, and to compare it with the *-isme/-ique* connexion (cf. table 1, c).

		Primes			
Targets		Translation/ Identity	Phon ovrl.	Morphological	Unrelated
a) Cognates 0-base *-iste*	*pluraliste* 10.3 lett. 1.81 occ./m.	*πλουραλιστής* /pluralistís/ 'pluralist'	95%	*πλουραλισμός* /pluralismós/ (75%)	*ξεχείλισμα* 'overflowing'
b) Noncognates Greek-base *-iste*	*individualiste* 10.22 lett. 2.06 occ./m.	*ατομικιστής* /atomikistís/ 'atomist'	–	*ατομικισμός* /atomikismós/ (–)	*αστεροσκοπείο* 'observatory'
c) Cognates Greek-base *-ique*	*monarchique* 10.4 lett. 4.5 occ./m.	*μοναρχικός* /monarhikós/ 'monarchist'	85%	*μοναρχισμός* /monarhismós/ (55%)	*αφαίρεση* 'substraction'

Table 1. Stimuli sample (number of letters and lexical frequency) and phonological overlap for the nine experimental conditions (3 priming conditions: translation, morphological, unrelated x three types of target).

The pattern of results found for the bilingual group (i.e., university students having Greek as L1 and having spent some years in France) and for the technical school group were quite different. The bilingual group (N = 29) exhibited translation and morphological priming effects (83ms and 49ms respectively) which occurred simultaneously for cognates Greek-base. This result extends the cognate effect to complex primes and targets and demonstrates that there is a cross-language connexion between *-isme* and *-ique*. Our bilingual speakers showed no effect for non-cognates, which is not surprising, given that in most cases the morphological effect is concomitant to the cognate (translation) effect. As for condition a), i.e., the 0-base cognates (in the L1 of our subjects, since *plural-* is not a stem in Greek), it induces morphological priming (85ms) but no translation priming. This result highlights two facts: i) that the contact with a whole word (lexical) entry is necessary to trigger translation effects: morphological segments such as πλουραλ- /plural-/ do not constitute entry units for the L1 lexicon of our subjects, and as such they cannot contact the corresponding lexical entry (Corbin, 1987: 457-459, 'ils ne sont les produits d'aucune Règle de Construction de Mot'). Psycho-linguistically speaking, 0-base cognates should have an intermediate status

[1] Please note that the identity effect under monolingual conditions corresponds to a translation effect under bilingual conditions, given that it is the same word which is tested; in Exp. 1a the priming direction is from L1 to L2 (cross-language cross-script priming, given that Greek and French have different alphabets), and in Exp. 1b the priming direction is from L1 to L1.

between the constructed and the non-constructed word; ii) that overlapping (phonological) form between prime and target does not suffice to induce cross-language effects, confirming that masked priming cross-language effects are not simply form priming effects.

The (unpublished) results of Exp. 1b (27 participants studying in the public technical school, age: 18-23) did not show any significant morphological priming, which is surprising, given that both primes and targets were in their L1. Only one condition, the identity condition of Cognates Greek-base *-ique* (Table 1, c) managed to yield a 52ms significant effect. Many questions arise from this finding: is it because our monolingual participants do not know these words, for instance the Cognates 0-base *-iste* words, which are not very frequent and cannot be connected to any morphological family of Greek words? Or would it be because they simply did not have the time to read the prime words, which were all quite lengthy (appr. 10 letters long)? If this were the case, how can we explain the fact that no priming effect is found for the morphological condition of Cognates Greek-base, despite the robust identity effect found for this type of word? In sum, while L2 speakers exhibit priming effects in most of cross-language conditions (Exp. 1a), monolingual speakers do not perceive morphological relations in their own L1, a fact that could imply qualitative differences within the processing system, depending on the type of speaker. Such an account recalls that claiming the inability of L2 learners to rely on the computational component (e.g., Clahsen, Felser, Neubauer, Sato & Silva, 2010) and their inclination to list forms in the lexicon rather than creating them with stems and affixes, as native speakers do.

2.2 Evidence from groups of speakers from diverse educational backgrounds

Given the above, it seemed crucial to us to repeat the experiment 1a with speakers of another language, we therefore ran the experiment with participants who are students in a "Second chance school" (age: 17-24) and have French as their "school and everyday language". To do so, we had to make the necessary adjustments, mainly the suppression of condition b (Table 1), given that these words (generally) do not exist in French. Another difference was that the morphological prime for the c) condition was the base, ex. *monarchie* for the target *monarchique*. Excepted these two differences, exp.2 was identical to Exp. 1a (§2.1), in its monolingual French version, i.e., both primes and targets were in French. Most of our participants in this experiment had French as their 'school language', but in most cases, French was not the "home language". We do not have the space here to provide a detailed description of this population, we wish however to underline that most participants had a "terminale" class level (i.e., the high school degree/ A level year), with some of them declaring a level equivalent to that of "seconde" (i.e., 11/10th grade, before integrating the school). Two groups were created, based on a double assessment of participants linguistic competence: her/his score in a French vocabulary test and the proportion of errors in the lexical decision protocol. Table 2 summarizes the results of the group that performed better (less than 22% error rate in the LD task). What our results show is that these participants do not process the *-ique* and the *-iste* derivations (targets) in the same way.

Words	Primes							
Targets	Identity (Id.)		Morph. (M)		Unrel. (U)		Net prim. effect	
	RT	Err.	RT	Err.	RT	Err.	U- Id.	U - M
-iste, ex. pluraliste	1060	14,7	1060	13,1	1056	14,4	0	-4
-ique, ex. monarchique	993	6,9	1023	9,7	1059	6,6	66*	36*

Table 2. Reaction times (in milliseconds) and percentages of errors for the lexical decisions to the two types of targets in the three priming conditions (identity, morphological and unrelated). Net priming

effects are given relative to the unrelated condition and statistically significant priming effects (identity and morphological) are marked with an asterisk.

In Exp. 2, the *-isme/-iste* induces no priming, whereas for *-ique* conditions, robust identity and morphological (base) priming are found, showing the strength of the connexion between the base and its derivation. In the discussion, these results will be compared to those of Exp. 1a and 1b, underlining the influence of the variable "type of speaker", as well as its effect in terms of strength of the connexions between words (and their parts). Our results will be interpreted with respect to what Mulder et al. (2014) call the "larger chain of morphological relations", including series effects (Dal Maso & Giraudo, 2019). With respect to the discussion related to multilingualism, they fit well the view of the mental lexicon as a unified lexico-semantic architecture (Schoonbaert, Duyck, Brysbaert & Hartsuiker, 2009; Voga, 2020).

References

Baayen, Harald. 2014. Experimental and psycholinguistic approaches. In Rachel Lieber & Pavol Štekauer (eds.), *The Oxford handbook of derivational morphology*, 95–117. Oxford: Oxford University Press.

Clahsen, H., C. Felser, K. Neubauer, M. Sato & R. Silva. 2010. Morphological structure in native and non-native language processing. *Language Learning* 60. 21-43.

Corbin, Danielle. 1987/1991. *Morphologie dérivationnelle et structuration du lexique* (2 vols.) Tübingen/Villeneuve d'Ascq: Max Niemeyer Verlag/Presses Universitaires de Lille.

Dal Maso, S. & H. Giraudo. 2019. On the interplay between family and series effects in morphological masked priming. *Morphology* 29. 293-315.

Drews, E. & P. Zwitserlood. 1995. Morphological and orthographic similarity in visual word recognition. *Journal of Experimental Psychology: Human Perception & Performance* 21. 1098-1116.

Feldman, L.B., P. O'Connor & F. Moscoso del Prado Martin. 2009. Early Morphological Processing is Morpho-semantic and not simply Morpho-orthographic: An exception to form-then-meaning accounts of word recognition. *Psychological Bulletin and Review* 16(4). 684-691.

Kimura, D. 2000. *Sex and Cognition*. The MIT Press, Cambridge, MA.

Mainz, N., Z. Shao, M. Brysbaert & A. Meyer. 2017. Vocabulary Knowledge Predicts Lexical Processing: Evidence from a Group of Participants with Diverse Educational Backgrounds. *Frontiers in Psychology* 8. 1164.

Mulder, K., T. Dijkstra, R. Schreuder & H. Baayen. 2014. Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Journal of Memory and Language* 72. 59-84.

Schoonbaert, S., W. Duyck, M. Brysbaert & R. Hartsuiker. 2009. Semantic and translation priming from a first language to a second and back: Making sense of the findings. *Memory & Cognition* 37(5). 569–586.

Voga, M., & A. Anastassiadis-Symeonidis. 2018. Connecting lexica in bilingual cross-script morphological processing: base and series effects in language co-activation. *Lexique* 23. 160-184.

Voga, M. 2020. Lexical co-activation with prefixed cognates and non-cognates: evidence from cross-script masked priming. In M. Schlechtweg (ed.), *The learnability of complex constructions from a cross-linguistic perspective*. Berlin: Mouton De Gruyter, 7-38.

A constructionist approach to the distinction between reduplication and repetition: A case study of Turkish

Yui Suzuki
The University of Tokyo

1 Background

In this paper I explore the similarities and differences between two iterative processes in Turkish: reduplication (a morphological process) and repetition (a syntactic process). More specifically, I conduct wordhood tests on deontic and emphatic iterations in Turkish from a Construction Morphology (CxM) perspective (Booij 2010; Booij 2018; Audring & Masini 2018) and argue that CxM successfully captures the similarities and differences between the two types of Turkish iteration discussed here.

1.1 Reduplication and repetition

Both reduplication and repetition refer to iteration of linguistic forms (Gil 2005; cf. Inkelas & Zoll 2005; Rubino 2005; Downing & Inkelas 2015; Inkelas & Downing 2015; Finkbeiner & Freywald 2018; Urdze 2018). However, they are different in terms of wordhood: whereas reduplication is a morphological process that yields a word, repetition is a syntactic process that produces a phrase.

1.2 Phenomena

As mentioned above, in this paper I compare two types of iteration in Turkish: **deontic iteration** and **emphatic iteration**. There is little research on these two types of iteration, though brief descriptions of deontic and emphatic iterations are found in Yaman (2017: 67) and Lewis (2000: 234), respectively.

In deontic iteration, the iteration of a past tense verb is used to express deontic mood, i.e., 'had better', as in (1). In contrast, in emphatic iteration, the iteration of an inflected verb fulfils an emphatic function, as in (2).

(1) *Ev-im-e* ***gel-dí-n*** ***gel-di-n*** (*yoksa* *yemek* *kal-ma-yacak*).
house-1SG-DAT come-PST-2SG come-PST-2SG otherwise meal remain-NEG-FUT
'You **had better come** to my house (nothing will be left of the meal).'

(2) *Ev-im-e* ***gel-dí-n*** ***gel-dí-n*** (*ama* *ben-i* *bul-ama-dı-n*).
house-1.SG-DAT come-PST-2SG come-PST-2SG but 1SG-ACC find-POSS-PST-2SG
'You **did come** to my house (but you couldn't find me).'

As can be seen in the examples above, a fully inflected verb may be used in both types of iteration. However, in Turkish deontic iteration, only a past tense verb can be used; in contrast, any tense can be used in emphatic iteration.

The two types of iteration are similar in that they have conventionalized meanings, deontic mood and emphasis, that cannot be derived from the constituents of the whole. For example, although the past tense suffix is used in deontic iteration, it does not describe a past event.

Importantly, some observations show that these two types of iteration have different wordhood statuses, suggesting that deontic iteration is an instance of reduplication and emphatic iteration an instance of repetition. For example, in cases of deontic iteration, the two verbs cannot be interrupted by another element (see (3)). In contrast, in cases of emphatic iteration, the two verbs can be interrupted by another element (see (6)). This indicates that deontic iteration results in the formation of a morphological word and is therefore a type of reduplication. In contrast, emphatic iteration results in the formation of two morphological words and is therefore a type of repetition. By investigating these two types of iteration, we can explore the similarities and differences between reduplication and repetition.

1.3 Theoretical framework

To capture the similarities and differences between Turkish reduplication and repetition, I adopt a constructionist approach (Goldberg 2006, Hilpert 2014) and specifically use the framework of Construction Morphology (CxM) (Booij 2010, Booij 2018, Audring & Masini 2018). CxM is SIGN-BASED and WORD-BASED (Audring & Masini 2018). In this approach, constructions are signs, that is, conventionalized pairings of form and meaning. The generalization of the form-meaning is represented by a schema, which may be abstract or partially/fully specified.

In addition, CxM is word-based rather than morpheme-based. Word-based approaches take the word, not the morpheme, as the smallest lexical entry. Constructions include both simplex words and complex words as conventionalized pairings of form and meaning on the word level.

There have been a number of CxM analyses of reduplication (Booij 2010; Booij 2018), and these have argued that the holistic meaning of reduplication is analyzed by a schema at a word level (e.g., total reduplication with a plural meaning in Malay (Booij 2018: 285)). However, few studies have been analyzed both reduplication and repetition in a unified way (cf. Nagaya 2020), a gap this paper hopes to begin to address.

2 Research questions

The present study addresses the following two research questions:

- **Descriptive question:** What are the similarities and differences between deontic iteration and emphatic iteration in terms of form and meaning?
- **Theoretical question:** How can CxM, a word-based constructionist theory of morphology, capture the similarities and differences between deontic iteration and emphatic iteration, or more generally reduplication and repetition?

3 Methodology

To answer the research questions above, I conducted wordhood tests on deontic and emphatic iterations in Turkish following Gil's (2005) framework for distinguishing between reduplication and repetition: stress, the size of a copy, the number of copies, interruptibility, and meaning.

4 Wordhood tests

In this section, I investigate the wordhood of Turkish deontic and emphatic iterations. CxM framework assumes that constructions have a tripartite parallel structure (Booij 2010: 6): phonological, morpho-syntactic, and semantic. Thus, I describe deontic and emphatic iterations in terms of these three structures.

4.1 Deontic iteration

- **Phonological properties:** Stress is assigned only to the final syllable of the first element (1).
- **Morpho-syntactic properties:** A fully inflected verb with the past tense suffix *-di* is used for the iteration. The two verbs cannot be separated by inserting another element, such as the particle *ya*, as in (3). In addition, the number of copies is limited to two. See (4).

(3)	**Beşiktaş-a*	***git-ti-m***	***ya***	***git-ti-m.***
	Beşiktaş-DAT	go-PST-1SG	PARTICLE	go-PST-1SG
	Intended: 'I **had better go** to Beşiktaş.'			
(4)	**Ödev-i*	***yap-tı-n***	***yap-tı-n***	***yap-tı-n.***
	homework-ACC	do-PST-2SG	do-PST-2SG	do-PST-2SG
	Intended: 'You **had better do** your homework.'			

- **Semantic properties:** Deontic iteration results in conventionalized and unpredictable meanings. It is productive and can be used with a variety of verbs.

4.2 Emphatic iteration

- **Phonological properties:** Stress is assigned to the final syllable of both elements (2).
- **Morpho-syntactic properties:** A fully inflected verb is iterated whether the verb is past tense or not. The two verbs can be separated by inserting another linguistic form, as in (5). The number of copies is not limited to two, as in (6).

(5)	*Beşiktaş-a*	***git-ti-m***	***ya***	***git-ti-m***.
	Beşiktaş-DAT	go-PST-1SG	PARTICLE	go-PST-1SG
	'I **did go** to Beşiktaş.'			
(6)	*Ödev-i*	***yap-tı-n***	***yap-tı-n***	***yap-tı-n***.
	homework-ACC	do-PST-2SG	do-PST-2SG	do-PST-2SG
	'You **did do** your homework.'			

- **Semantic properties:** The emphatic iteration fulfils an iconic yet conventionalized pragmatic task, i.e., emphasis. This iteration is very productive and can be used with a variety of verbs.

4.3 Summary

Table 1 summarizes the results of the wordhood tests I conducted to compare the two types of iteration discussed here.

Table 1: Results of wordhood tests

Level	Wordhood test	Deontic iteration	Emphatic iteration
phonology	stress	one	two or more
morpho-syntax	size of a copy	fully inflected verb	
	interruptibility	uninterruptible	interruptible
	number of copies	only two	two or more
semantics	meaning	conventionalized	
		deontic mood (non-iconic)	emphasis (iconic)

5 Analysis

On the phonological level, deontic iteration results in a single phonological word, while emphatic iteration results in a phrase. In Turkish, a phonological word has one and only one stress. As deontic iteration results in a form with only one stress, this supports the finding that this type of iteration produces a single phonological word. In contrast, in emphatic iteration, stress is assigned to each verb, indicating that this type of iteration results in a phrase composed of more than one phonological word.

As for morpho-syntactic structure, the wordhood tests described above indicate that Turkish deontic iteration results in the formation of a single morphological word that is uninterruptible (see (4)); in contrast, emphatic iteration results in the formation of a phrase that is interruptible (see (7)). Morphological words cannot be interrupted; syntactic phrases may be interrupted by another element. Additionally, Turkish deontic iteration results in forms that are restricted in terms of copies (see (3)), while Turkish emphatic iteration does not have a restriction on number copies (see (6)). Morphological processes are non-recursive; syntactic processes are recursive (Matthews 1991: 213).

Thus, both phonologically and morpho-syntactically, there is evidence to support the claim that deontic iteration is an instance of reduplication, and that emphatic iteration is an instance of repetition. Interestingly, in terms of semantic structure, both deontic and emphatic iterations result in conventionalized meanings which cannot be directly derived from the parts of the construction.

Based on these claims, I propose the schemas for Turkish deontic and emphatic iterations shown in (7) and (8), respectively. The phonological, morpho-syntactic, and semantic structures are represented from left to right.

(7) $<[\omega]_i \leftrightarrow [[\mathrm{V}_j\text{-PST-SBJ}]{\sim}[\mathrm{V}_j\text{-PST-SBJ}]]_i \leftrightarrow [\text{had better SEM}_j \text{ soon}]_i>$

(8) $<[\omega_i\ \ \omega_i\ (\omega_i)]_k \leftrightarrow [[\mathrm{V}_j\text{-TENSE-SBJ}]_i\ [\mathrm{V}_j\text{-TENSE-SBJ}]_i([\mathrm{V}_j\text{-TENSE-SBJ}]_i)]_k \leftrightarrow [\text{do SEM}_j]_k>$

6 Discussion

We are now in a position to answer the questions raised in Section 2. Descriptively, the schemas in (7) and (8) serve to describe the similarities and differences between the two types of iteration. Theoretically, a CxM approach maintains the distinction between reduplication and repetition, yet still captures the similarities between these types of iteration. Both phonologically and morphologically, deontic iteration forms one word and is thus a type of reduplication resulting in a morphological construction; emphatic iteration forms a phrase and is thus a type of repetition resulting in syntactic construction. Additionally, both types of iteration have conventionalized meanings as a whole and can therefore be analyzed as constructions with iterative forms showing different word properties. To conclude, CxM is useful for describing and analyzing reduplication and repetition in a unified way.

References

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

Booij, Geert (ed.). 2018. *The construction of words: advances in construction morphology*. Cham: Springer.

Downing, Laura J. & Sharon Inkelas. 2015. What is reduplication? Typology and analysis part 2/2: the analysis of reduplication. *Language and Linguistics Compass* 9(12). 516–528.

Finkbeiner, Rita & Ulrike Freywald (eds.). 2018. *Exact repetition in grammar and discourse*. Berlin: De Gruyter Mouton.

Gil, David. 2005. From repetition to reduplication in Riau Indonesian. In Bernhard Hurch (ed.), *Studies on reduplication*, 31–64. Berlin: De Gruyter Mouton.

Goldberg, Adele E. 2006. *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.

Göksel, Aslı & Celia Kerslake. 2005. *Turkish: a comprehensive grammar*. Oxon: Routledge.

Hilpert, Martin. 2019. *Construction grammar and its application to English*. 2nd ed. Edinburgh: Edinburgh University Press.

Inkelas, Sharon & Cheryl Zoll. 2005. *Reduplication: doubling in morphology*. Cambridge: Cambridge University Press.

Inkelas, Sharon & Laura J. Downing. 2015. What is reduplication? Typology and analysis part 1/2: the typology of reduplication. *Language and Linguistics Compass* 9(12). 502–515.

Lewis, Geoffrey. 2000. *Turkish grammar*. 2nd ed. Oxford: Oxford University Press.

Masini, Francesca & Jenny Audring. 2018. Construction morphology. In Jenny Audring & Francesca Masini (eds.), *The Oxford handbook of morphological theory*, 365–389. Oxford: Oxford University Press.

Matthews, P. Hugoe. 1991. *Morphology*. 2nd ed. Cambridge: Cambridge University Press.

Nagaya, Naonori. 2020. Reduplication and repetition from a constructionist perspective. *Belgian Journal of Linguistics* 34. 259–272.

Rubino, Carl. 2005. Reduplication: form, function and distribution. In Bernhard Hurch (ed.), *Studies on reduplication*, 11–29. Berlin: De Gruyter Mouton.

Urdze, Aide (ed.). 2018. *Non-prototypical reduplication*. Berlin: De Gruyter Mouton.

Yaman, Ertuğrul. 2017. *Türkiye türkçesinde zaman kaymaları*. Ankara: Türk Dil Kurumu Yayınları.

Social gender and derivational morphology: a distributional study of the gendered import of learned morphology in French

Marine Wauquier   *Olivier Bonami*
Université de Paris, LLF, CNRS

1 Introduction

French suffixes *-euse* and *-rice* are clearly morphological rivals for the formation of both feminine instrument nouns (*agrafeuse* 'stapler', *excavatrice* 'excavator') and agent nouns denoting women (*danseuse* 'female dancer', *rédactrice* 'female author'). However the literature on nouns designating women gives circumstantial evidence for differences in meaning: agent nouns in *-euse* are said to denote lower-level professions, such as *coiffeuse* 'hairdresser' or *serveuse* 'waitress' (Lenoble-Pinson, 2008), or nouns with a pejorative connotation, such as *entraîneuse* 'barmaid' or *allumeuse* 'tease' (Dawes, 2003), while *-rice* is favored for more socially valued positions (*directrice* 'female manager'). This has recently been confirmed quantitatively on the basis of distributional semantics (Wauquier et al., 2020a).

While it is plausible that the two suffixes have specialized to convey classes of meanings related to gender stereotypes, previous studies have not taken into account the fact that the two suffixes also differ in their place in the French morphological system. In parallel with other suffixes such as *-ion*, *-if*, etc., *-rice* originates in learned vocabulary borrowed from Latin from Middle French on (see Rainer & Buridant 2015 for an overview). While all these suffixes then became productive in their own right, their learned origin may have an influence on the types of concepts that they are used to designate. Crucially, the *-euse*/*-rice* pair is paralleled by a distinction between two processes using the same suffix *-eur* to form masculine agent and instrument nouns: learned *-eur* attaches to the same learned stems as *-ion*, *-if* or *-rice* (Bonami et al.'s (2009) 'hidden stem'),[1] while nonlearned *-eur* attaches to the same ordinary, nonlearned stems as *-euse* or other nonlearned suffixes such as *-age*.

Against this background, the present study attempts to assess to what extent the observed differences between *-euse* and *-rice* follow from their status as learned vs. nonlearned formations: if the differences in meaning between *-euse* and *-rice* follow from their learned status, we expect them to be paralleled by differences between learned and nonlearned masculine nouns in *-eur*, which are otherwise morphologically parallel, modulo gender. If the effect of learned vs. nonlearned is strong enough, we might even be able to document parallel effects for other morphosemantic types such as action nouns in *-ion* vs. *-age*.

2 Data

We built three datasets of deverbal feminine agent nouns (AGF), masculine agent nouns (AGM), and action nouns (ACT), with a contrast between a learned and a nonlearned alternative in each case. Feminine agent nouns and action nouns were extracted from Lexeur (Wauquier et al., 2020b), while masculine agent nouns were borrowed from the dataset documented in Huyghe & Wauquier (2020). All agent nouns were manually filtered so as to exclude polysemy with

[1] Most learned formations in *-eur* end in *-teur*, but there are exceptions in both directions: *professeur* 'professor' is learned, *acheteur* 'buyer' is not.

an instrument reading, and only nouns with a frequency of 50 or more in the FrCoW corpus (Schäfer, 2015; Schäfer & Bildhauer, 2012) were retained. The size of our final datasets are given in table 1.

	Learned	Nonlearned
Feminine agent nouns (*-rice* vs. *-euse*)	158	301
Masculine agent nouns	141	462
Action nouns (*-ion* vs. *-age*)	750	629

Table 1: Description of our dataset

To assess the semantic properties of these nouns, we used a distributional semantic model (DSM) obtained by applying the gensim (Řehůřek & Sojka, 2010) implementation of word2vec (Mikolov et al., 2013) to a tagged and lemmatized version of the FrCoW corpus.[2]

3 Quantitative assessment

We first assessed whether our DSM captures differences between learned and nonlearned derivatives in our three datasets. To this end, we trained classifiers to predict from the semantic representation of a lexeme whether it was formed using a learned or the corresponding nonlearned process. Specifically, we used gradient boosting (Friedman, 2001; Mason et al., 2000) applied to decision trees as our binary classification method.[3] To avoid differences in accuracy due to differences in dataset size, we randomly subsampled each of the subdatasets to 141 items, the size of the smallest of our 6 subdatasets. We report the aggregated accuracy of 10-fold cross-validation. The results of this first assessment is suprisingly good: despite a small training set, each of the three classifiers reaches an accuracy between 0.77 and 0.83, well above the 0.5 baseline. This clearly indicates that there are distributional cues separating learned and nonlearned nouns. Importantly, this holds across feminine agent nouns, masculine agent nouns, and action nouns.

The fact that all three morphosemantic types of nouns differ in their distribution does not entail that they differ in the same fashion. Further exploration however indicates that the relevant distributional properties overlap strongly. First, the analysis of dimension importance indicates that one and the same specific dimension has markedly more predictive power for all three models. Examination of the three datasets confirms in each case a highly significant contrast in values of that dimension between learned and nonlearned exemplars, although the distributions strongly overlap. Second, we used each of the three models to conduct *extrinsic prediction* on data from another semantic type: for instance, the model trained on feminine agent nouns is used to predict the constrast between learned and nonlearned masculine agent nouns. The results are shown in Table 2, with intrinsic prediction results on the diagonal.

The striking conclusion is that intrinsic and extrinsic prediction lead to very similar accuracy. All nine 95% confidence intervals overlap, so that one may not conclude that the learned vs. nonlearned distinction in one dataset is better predicted by vectors for the corresponding morphosemantic type or another one.

[2] We used the cbow variant of the algorithm with the following hyperparameters: 2 training epochs, 5 negative samples, window size 5, vector size 100. We used the tagging provided with the corpus and improved the lemmatization semi-automatically to correctly have separate lemmas for nouns of different grammatical genders but proper gender neutralizations of all non-nouns.

[3] We used the Python implementation of gradient boosting in the scikit-learn package (Pedregosa et al., 2011), with the following hyperparameters for all models: 500 estimators, max depth of 2, deviance loss function.

	Test data		
Training data	AGF	AGM	ACT
AGF	0.80	0.77	0.79
AGM	0.77	0.77	0.82
ACT	0.76	0.79	0.83

Table 2: Accuracy of the three classifiers applied to the three datasets

The evidence thus strongly suggests that there are general distributional differences between learned and nonlearned deverbal formations in French that are not limited to feminine agent nouns. A likely cause of these contrasts is the fact that learned formations entered the language in particular sociolinguistic circumstances, and that analogical extension of their use led to a partial specialization for some type of concepts.

4 Qualitative evaluation

While we have shown that the difference between *-rice* and *-euse* follows at least in part from their respective learned vs. nonlearned character, it remains to be seen whether there is a link between this difference and the observed difference in connotations. To assess this, we built, for each of the 6 processes under consideration, the *centroid* representing the average of their respective vector representations. Intuitively this should capture what the processes have in common, neutralizing individual lexical semantics. We identified their 100 nearest neighbors in the DSM, and examined qualitatively the semantic properties of these neighbors. Because we are interested in connotations linked to social gender, we focus on agent nouns.

Two main oppositions emerge from the comparison of the four lists of neighbors. The first involves the overall axiological valence of learned and nonlearned centroids, regardless of the targeted gender. Both feminine and masculine nonlearned centroids display a much higher proportion of negatively valued neighbors than their learned equivalent, for which neighbors are at best positively valued (*dirigeante* ‘female leader’, *chirurgienne* ‘female surgeon’, *avocate* ‘female lawyer’ for the feminine; *érudit* ‘scholar’, *académicien* ‘academician’, *orateur* ‘orator’ for the masculine), at worst neutral (*entrepeneure* ‘businesswoman’, *sculptrice* ‘female sculptor’, *collaboratrice* ‘female associate’ for the feminine; *exécutant* ‘subordinate’, *journaliste* ‘journalist’, *comptable* ‘accountant’ for the masculine). The second opposition concerns the types of axiological properties displayed by the neighbors of the nonlearned centroids with regard to gender. Neighbors of the feminine nonlearned centroid involve connotation with respect to sexuality (*nymphomane* ‘nymphomaniac’, *tapineuse* ‘prostitute’, *catin* ‘harlot’) and physical characterization (*laideron* ‘plain Jane’, *monstresse* ‘monstress’, midinette ‘starry-eyed girl’). On the other hand, the axiological valence of the masculine nonlearned centroid’s neighbors also involves sexuality (*dragueur* ‘womanizer’, *séducteur* ‘seducer’), but mainly builds on other domains such as criminal activities (*truand* ‘gangster’, *voleur* ‘thief’) or behavioral characterization (*tâcheron* ‘drudge’, *poivrot* ‘drunkard’).

These results indicate that the contrast between *-rice* and *-euse* in terms of connotations exists over and above the fact that they contrast in terms of learnedness. We submit that the basic contrast between learned vs. nonlearned formations is recruited to different purposes depending on the morphosemantic type. For action nouns, it implements a contrast between intellectual and technical domains of reference (Wauquier et al., 2020b). For agent nouns, it readily encodes gendered axiological judgements, which are different in the masculine and in the feminine as a consequence of gender stereotypes.

References

Bonami, Olivier, Gilles Boyé & Françoise Kerleroux. 2009. L'allomorphie radicale et la relation flexion-construction. In Bernard Fradin, Françoise Kerleroux & Marc Plénat (eds.), *Aperçus de morphologie du français*, 103–125. Saint-Denis: Presses de l'Université de Vincennes.

Dawes, Elizabeth. 2003. La féminisation des titres et fonctions dans la francophonie: de la morphologie à l'idéologie. *Ethnologies* 25(2). 195–213.

Friedman, Jerome H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.

Huyghe, Richard & Marine Wauquier. 2020. What's in an agent? a distributional semantics approach to agent nouns in french. *Morphology* 30. 185–218.

Lenoble-Pinson, Michèle. 2008. Mettre au féminin les noms de métier: résistances culturelles et sociolinguistiques. *Le français aujourd'hui* (4). 73–79.

Mason, Llew, Jonathan Baxter, Peter L Bartlett & Marcus R Frean. 2000. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*, 512–518.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. `http://arxiv.org/abs/1301.3781`.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Rainer, Franz & Claude Buridant. 2015. From old french to modern french. In P.O. Muller, I. Ohnheiser, S. Olsen & F. Rainer (eds.), *Word-formation: an international handbook of the languages of europe*, vol. 3, 1975–2000. Berlin/Boston: De Gruyter Mouton.

Řehůřek, Radim & Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA. `http://is.muni.cz/publication/884893/en`.

Schäfer, Roland. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of challenges in the management of large corpora*, 28–34.

Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the eighth international conference on language resources and evaluation*, 486–493.

Wauquier, Marine, Nabil Hathout & Cécile Fabre. 2020a. Contributions of distributional semantics to the semantic study of French morphologically derived agent nouns. In J. Audring, N. Koutsoukos & C. Manouilidou (eds.), *Rules, patterns, schemas and analogy, mmm12 online proceedings*, vol. 12, 111–121.

Wauquier, Marine, Nabil Hathout & Cécile Fabre. 2020b. Semantic discrimination of technicality in French nominalizations. *Zeitschrift für Wortbildung / Journal of Word Formation* 4(2). 100–119.

A diachronic approach to the formal idiosyncrasies of indexes

Tim Zingler

Affiliation TBD

1 Introduction

This work builds on the insight that indexes (i.e., argument-indexing agreement markers and/or pronouns) show a wider range of formal and distributional idiosyncrasies than the exponents of other inflectional categories (e.g., Julien 2002: ch. 5; Fuß 2005: 62-67). In order to support this claim, I will discuss indexes that are extrametrical with respect to reduplication and “mobile” affixes that can occur in different slots of otherwise identical words. In addition, I will illustrate indexes that can freely occur on either member of a phrase-level construction as well as indexes that behave like full-fledged affixes in one type of context but like clear-cut words in another. The claim that the range of these traits is unique to indexes is based on a larger project by the author, which in addition to indexes also investigates exponents of definiteness, case, and tense. The latter are all more homogeneous in their behavior than the indexes. The explanations for this discrepancy rest on the different diachronic pathways to which indexes are subject and which themselves constitute an important topic for further research. The overall database comprises 60 languages, which belong to 60 WALS genera and are evenly distributed across five macro-areas.

2 Data

For the purposes of this contribution, I will focus on the three phenomena that most clearly distinguish indexes from the other exponents: extrametricality, mobility, and duality. Each of these will be defined in the relevant sub-sections below. Section 3 will then argue that these data require diachronic explanations and will suggest such explanations for each of the patterns.

2.1 Extrametricality

The definition of extrametricality that I employ here is wider than usual. Whereas the concept is traditionally applied to strings that are irregular in that they fall outside a stress domain, I will also use this term here for strings that are irregular in that they fall outside a domain of reduplication. There are two indexes in my sample that fail to undergo reduplication processes that otherwise apply to all morphological items in the relevant position of the verb, whereas none of the definiteness, case, or tense markers show such behavior. This is illustrated below with data from Fwe (Atlantic-Congo; fwe; Gunnink 2018), where locative arguments are indexed on the verb via markers for the noun classes 16-18. The data of interest are given in (1) and (2).

(1) ndi-a-endí-end-i=ko
1SG-PST-RDP-go-PST=LOC$_{17}$
‘I kept going there.’

(2) ndì-ngòngòt-á=hò
1SG-knock-FV=LOC$_{16}$
‘I knock on it.’ (Gunnink 2018: 272)

Pluractional reduplication in Fwe targets the verb stem, which includes the root as well as all inflectional and derivational suffixes (Gunnink 2018: 199-200, 249). As can be gleaned from (1), then, the class 17 index is not a suffix for the purposes of this process because the reduplicant *endí* consists of the root and the past tense morph but excludes the following locative index *ko*. Meanwhile, in (2), the high tone that usually co-expresses present tense on the final mora is on the penultimate mora because the verb is in clause-final position (Gunnink 2018: 272). Since the final mora corresponds to the class 16 locative index *ho*, however, the latter must be part of the phonological word for the purpose of tone assignment. Note that while Gunnink (2018: 271-272) classifies the locative indexes as enclitics, they always take up the final slot in the verb template and thus have the syntagmatic distribution of affixes. Finally, while the Fwe example centers on postposed elements, most other extrametrical items in my sample are preposed. The relevance of this distributional fact will be addressed in Section 3.

2.2 Mobility

Indexes also show the ability to take up one of several slots in both morphological templates and phrase-level syntactic constructions without bringing about a semantic difference between the resulting alternatives. It is this behavior for which I suggest the umbrella term “mobility” here. There are two clear cases of “mobile affixes” among the indexes in my sample, twice as many as there are among the remaining data combined. In addition, there are also two indexes that show syntactic mobility, which I did not find for any exponent of the other three categories. (Note that *multiple* case marking across an NP is a kind of concord, not mobility as defined here.)

A mobile affix as defined here can be found in San Francisco del Mar Huave (Huavean; hue; Kim 2008). The second-person marker can occur on either side of a subordinate verb, and the two orderings are explicitly described as equally acceptable (Kim 2008: 346). These two options are contrasted in (3) and (4) below.

(3) m-e-chutu-r (4) chutu-m-ia-r
SB-2-sit-2.INTR sit-SB-2-2.INTR
‘that you (SG) sit’ ‘that you (SG) sit’ (Kim 2008: 347)

The segmental variation between the indexes follows from allophonic principles. That is, diphthongization affects vowels that precede a tautosyllabic plain consonant, as in (4), but not those in open syllables, as in (3); cf. Kim (2008: 53). The fact that the index is subject to such processes suggests that it is part of a larger phonological word. However, it constitutes an idiosyncrasy because it undermines the idea that phonological words map onto grammatical words whose constituents follow a rigid order (cf. Dixon & Aikhenvald 2003; Haspelmath 2011).

Mobility at (roughly) the level of a phrase can be seen in Lillooet (Salishan; lil; van Eijk 1997), where the third-person plural marker *wit* can occur either on the auxiliary or on the lexical verb of an otherwise identical construction. These possibilities are juxtaposed in (5) and (6).

(5) waˀ-wit-ás=mał=ƛ’uˀ ˀíƛ’əm
AUX-3PL-SBJV=HORT=well sing
‘Let them sing/they might as well sing.’ (van Eijk 1997: 153)

(6) waˀ-as=máɬ=ƛ’uˀ ˀíƛ’əm-wit
AUX-SBJV=HORT=well sing-3PL
‘Let them sing/they might as well sing.’ (van Eijk 1997: 153)

Here too, the index is not simply a free word because it crucially falls within a larger domain in terms of stress assignment. Primary stress (marked by an acute accent) usually falls on the first syllable whose nucleus is not a schwa, but it can also be marked on the third syllable provided this is not the last syllable of the phonological word (van Eijk 1997: 14, 17). Hence, the primary stress on the third syllable in (5) can only be explained if the preceding index accounts for the second syllable. In contrast to the San Francisco del Mar Huave item, however, the Lillooet index is an even more drastic idiosyncrasy because its freedom extends to the syntactic level. Note also that neither the Huave nor the Lillooet marker is adequately classified as a clitic. This is because neither element is limited to second position in the clause or to a specific position with respect to a phrase, whereas both indexes *are* limited in terms of their possible hosts/stems.

2.3 Duality

By “duality,” I refer to the fact that some indexes behave like full words in some contexts but like prototypical affixes in others. As such, they clearly differ from clitics understood as “syntactic affixes” (cf. Anderson 2005), which always interact phonologically with a phrasal host and thus show the same behavior in all contexts. Note that while duality is also found among markers of the other categories, languages often have multiple paradigms of indexes (e.g., Cardinaletti & Starke 1999). This increases the likelihood that duality is more common in indexation than in other grammatical domains. Yet, further research on all aspects of this issue is required.

One index that clearly shows duality comes from the Umari Norte variety of Hup (Naduhup; jup; Epps 2008). The third-person singular marker typically occurs before the verb, but since it can also appear in other syntagmatic contexts it is not simply a prefix (cf. Epps 2008: 285, 755-756). The relevant contrast is illustrated in (7) and (8), which show the unmarked preverbal and the marked clause-final position of the index, respectively.

(7) “hɨ̃́t tã=hám-ãʔ?” tɔ̃=nɔ-máh-ãh
where 3SG=go-Q 3SG=say-REP-DYN
‘“Where did he go?” he said.’ (Epps 2008: 135)

(8) maŋgă táʔ-ay hɨ́d-ăn yamhidɔʔ-nɨ́h tɨ́h?
Margarita RI-INCH 3PL-OBJ sing-NEG 3SG
‘What about Margarita, didn’t she sing to them?’ (Epps 2008: 172)

In (7), both tokens of the index undergo consonant cluster reduction, due to which they lose their final /h/, and vowel harmony, due to which their vowel qualities are assimilated to those of the following vowels. Both of these processes are limited to the phonological word (cf. Epps 2008: 103-104), which lends credence to the affix analysis of these variants. By contrast, the underlying form seen in (8) is not subject to any phonological process and can be freely placed within the clause in a way that is typically taken to define grammatical words. Hence, the behavior of the third-person index differs on both the phonological and the syntactic dimension depending on its syntagmatic context.

3 Explanation and discussion

What unites all the indexes analyzed above is that they behave like bound elements on some set of criteria and/or in some contexts but like free words on another set of criteria and/or in other contexts. These mismatches constitute the idiosyncrasies of interest here, and it will be assumed in this contribution that they ultimately come about because the indexes at issue are in the process of grammaticalizing from free pronouns to pronominal/agreement affixes (cf. Bybee 2015: 152-153). Yet, the different types of idiosyncrasies illustrated here must nevertheless have resulted from different diachronic trajectories, and these will be sketched in this section.

With respect to indexes that are extrametrical in terms of reduplication, it is important to bear in mind that free pronouns are typically emphatic in nature. Since this emphatic status is defined by segmental weight and the ability to bear prominence, pronouns will first have to lose the properties they share with phonological words before they can begin to integrate into another word domain (i.e., the verb). Meanwhile, markers of case and tense often derive from adpositions and auxiliaries, respectively, and both of these form classes are typically already phonologically reduced. Hence, indexes apparently have to go through more diachronic stages to become full-fledged affixes than do exponents of other categories, and this makes indexes more likely to show idiosyncratic behavior at any given point in time.

Most extrametrical indexes in my database are preposed, and this tendency holds for both reduplication and other (supra)segmental processes. Crucially, a preposed position is likely to limit the degree of formal integration in that preposed elements fuse with following items less easily than do postposed ones with preceding items (e.g., Himmelmann 2014). Given that indexes are preposed much more frequently than exponents of most other grammatical categories (cf. Siewierska 2004: 165), indexes thus face unique obstacles on their path toward full affix-hood. Finally, reduplication processes typically include a word edge as part of their target domain, and since indexes predominantly occur at word edges (Bybee 1985), they are simply more likely to fall within a domain of reduplication. This, in turn, is a logical prerequisite for being considered extrametrical with regard to such a domain. In sum, then, the interaction of reduplication, extrametricality, and indexes can be derived from general morphological and diachronic facts.

Word-internal mobility plausibly derives from “exuberant agreement” (cf. Harris 2008), in which the same index is marked in multiple locations of a single verb form. At that stage, language users might reanalyze the templatic position of the index as flexible. If so, this would pave the way for a period during which a given index can interchangeably be marked in any of several slots. While the nature of word-external mobility is more obscure, it might be explained by the fact that agreement ultimately references a cognitive entity rather than a morphosyntactic unit (cf. Kibrik 2019). Once this basic assumption is granted, the grammatical properties of a given referent would be compatible with any member of the predicate because each member of the predicate can be tied to the referent in some form. That indexes are usually only marked once, on the verb, might then be due to economy considerations as well as to the fact that verbs are the most frequently available collocate and ultimately reanalyzed as the only possible one.

Finally, the phenomenon of duality as defined here is unremarkable on the recognition that the diachronic development from a syntactic construction to a morphologically complex string does not necessarily involve an intermediate clitic stage (cf. Lehmann 2020: 226). That is, to the extent that the fusion of formerly independent elements is conditioned by the token frequency of their collocation (e.g., Bybee 2002), frequent combinations as in (7) will show fusion whereas infrequent ones such as in (8) will not. It follows that, once duality is clearly distinguished from clitic-hood, the former might turn out to be cross-linguistically frequent in its own right.

In sum, the idiosyncratic behavior of indexes largely seems to derive from the fact that indexation is a more important communicative function than that expressed by other inflectional categories. As such, the grammatical properties of referents are often expressed multiple times per clause and/or emphasized, and the relevant constructions may then lead to extrametrical or mobile affixes, etc. This idea obviously needs to be fleshed out in more detail, but it promises to reveal an interesting correspondence between form and meaning in that a functionally more complex category might then also be expressed by a more complex set of exponents.

References

Anderson, Stephen. 2005. *Aspects of the theory of clitics*. Oxford: Oxford University Press.

Bybee, Joan. 1985. *Morphology*. Amsterdam: Benjamins.

Bybee, Joan. 2002. Sequentiality as the basis of constituent structure. In Talmy Givón & Bertrand Malle (eds.), *The evolution of language out of pre-language*, 107–132. Amsterdam: Benjamins.

Bybee, Joan. 2015. *Language change*. Cambridge: Cambridge University Press.

Cardinaletti, Anna & Michal Starke. 1999. The typology of structural deficiency: A case study of the three classes of pronouns. In Henk van Riemsdijk (ed.), *Clitics in the languages of Europe*, 145–233. Berlin: De Gruyter.

Dixon, R. M. W. & Alexandra Aikhenvald. 2003. Word: A typological framework. In R. M. W. Dixon & Alexandra Aikhenvald (eds.), *Word*, 1–41. Cambridge: Cambridge University Press.

Epps, Patience. 2008. *A grammar of Hup*. Berlin: De Gruyter.

Fuß, Eric. 2005. *The rise of agreement*. Amsterdam: Benjamins.

Gunnink, Hilde. 2018. *A grammar of Fwe*. PhD thesis, University of Gent.

Harris, Alice. 2008. Explaining exuberant agreement. In Thorhallur Eythórsson (ed.), *Grammatical change and linguistic theory*, 265–283. Amsterdam: Benjamins.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45. 31–80.

Himmelmann, Nikolaus. 2014. Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. *Language* 90. 927–960.

Julien, Marit. 2002. *Syntactic heads and word formation*. Oxford: Oxford University Press.

Kibrik, Andrej. 2019. Rethinking agreement: Cognition-to-form mapping. *Cognitive Linguistics* 30. 37–83.

Kim, Yuni. 2008. *Topics in the phonology and morphology of San Francisco del Mar Huave*. PhD thesis, University of California, Berkeley.

Lehmann, Christian. 2020. Univerbation. *Folia Linguistica Historica* 41. 205–252.

Siewierska, Anna. 2004. *Person*. Cambridge: Cambridge University Press.

Van Eijk, Jan. 1997. *The Lillooet language*. Vancouver: University of British Columbia Press.

Index