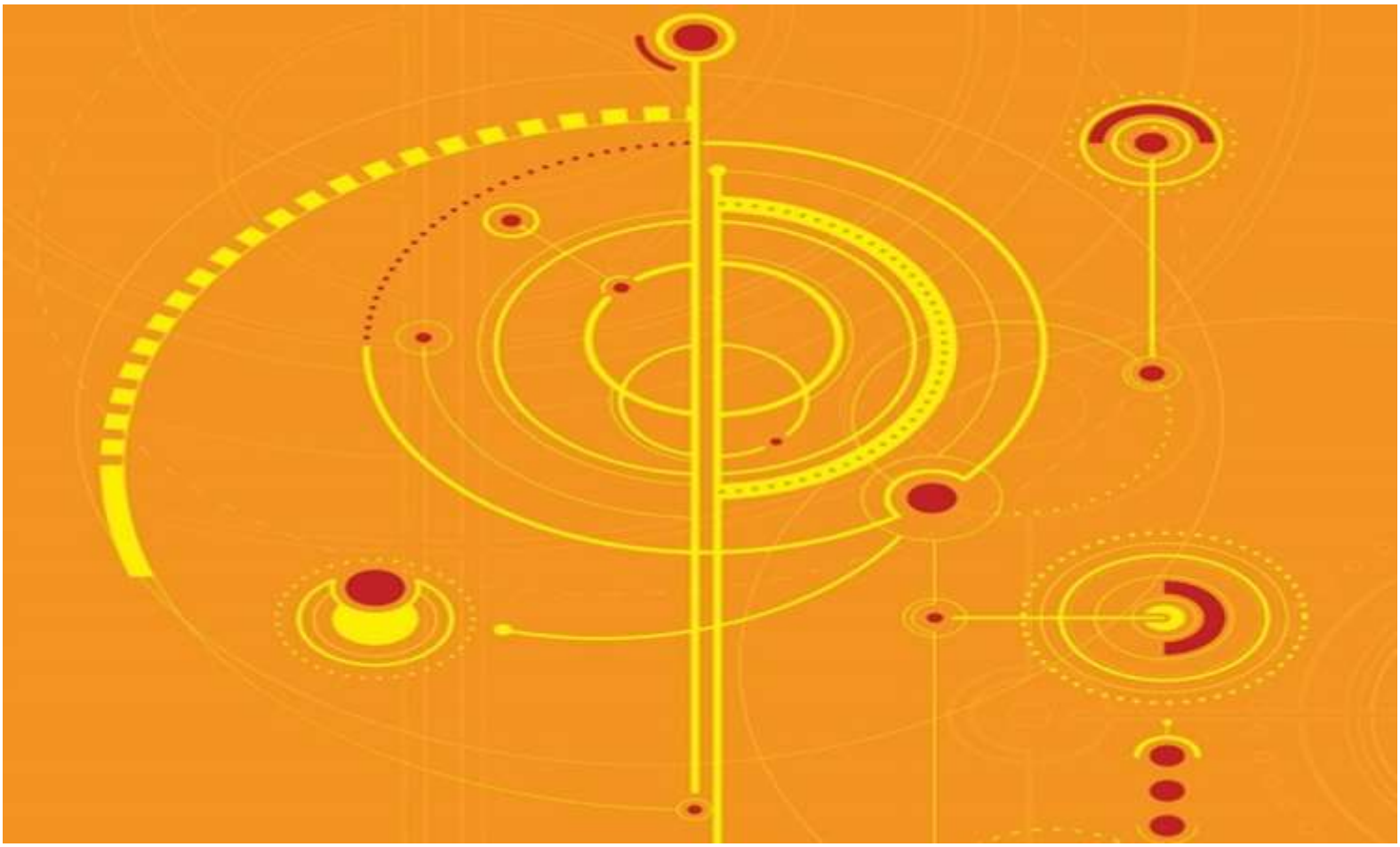
Second Workshop on Paradigmatic Word Formation Modelling

ParadigMo II

Program and Abstracts



Bordeaux, France

3-4 June 2021

Thursday, 3 June 2021

08:45 – 09:15	*Registration*	
09:15 – 09:30	Welcome	
09:30 – 10:30	**Andrew Spencer** (University of Essex)	*Guest talk 1* Paradigms, features, and the lexeme
10:30 – 11:00	*Coffee break*	
11:00 – 11:30	**Nabil Hathout** (CLLE, CNRS & U. Toulouse Jean Jaurès)**,** **Fiammetta Namer** (U. Lorraine & ATILF, CNRS) **Michel Roché** (U. Toulouse Jean Jaurès)	French ethnonyms, toponyms, demonyms and their paradigmatic organization)
11:30 – 12:00	**Petr Kos** (U. of South Bohemia)	Paradigmatic nature of Dokulil's onomasiological theory of word-formation)
12:00 – 13:30	*Lunch*	
13:30 – 14:00	**Livio Gaeta** (U. Turin)	Paradigmatic relations across morphology and syntax: Particle verbs in two Walser German varieties
14:00 – 14:30	**Camiel Hamans** (Amsterdam U./Adam Mickiewicz U. Poznan)	Paradigms in non-morphemic word formation
14:30 – 15:00	*Coffee break*	
15:00 – 15:30	**Alexandra Soares Rodrigues** (ESE – I. P. Bragança, CELGA-ILTEC – U. Coimbra) **Pedro João Rodrigues** (CeDRI, ESTiG – I. P. de Bragança)	Modelling word-formation paradigms: networks visually representing their multidimensionality, complexity and theoretical infiniteness
15:30 – 16:00	**Lior Laks** (Bar-Ilan U.) **Fiammetta Namer** (U. Lorraine & ATILF)**,**	Démonette meets Semitic morphology: A paradigm-based model for the derivational resources in French and Hebrew
16:00 – 16:30	**Michael Bilynsky** (Ivan Franko Lviv National U.)	The English de-verbal lexis as a problem of suffix-sensitive paradigm discovery

Friday, 4 June 2021

09:15 – 09:30	Welcome	
~~09:30 – 10:30~~	**Jesus Fernandez-Dominguez** (U. of Granada)	*Guest talk 2* CANCELLED
~~10:30 – 11:00~~	*Coffee break* *CANCELLED*	
10:00 - 10:30	**Fabio Montermini** (CLLE, CNRS & U. Toulouse Jean Jaurès)**,** **Delphine Tribout** (U. Lille & STL)	Measuring morphological series membership. The example of feminine *-eur* nouns in French
10:30 – 11:00	**Gauvain Schalchli** (U. Bordeaux-Montaigne & CLLE)	Derivational Paradigms and The Frequency Factor: The French *-ion* Nouns Allomorphy Problem
11:00 – 11:30	**Magda Ševčíková** **Lukáš Kyjánek** **Barbora Vidová Hladká** (Charles U., Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics)	Agent noun formation in Czech: An empirical study on suffix rivalry
11:30 – 13:00	*Lunch*	
13:00 – 13:30	**Matías Guzmán Naranjo** (U. Tübingen) **Olivier Bonami** (LLF, U. Paris)	Comparing derivational processes with distributional semantics
13:30 – 14:00	**Maria Copot** (LLF, U. Paris) **Timothee Mickus** (ATILF, CNRS/U. Lorraine) **Olivier Bonami** (LLF, U. Paris)	Striking out on one's own: idiosyncratic frequency as a measure of derivation vs inflection
14:00 – 14:30	*Coffee break*	
14:30 – 15:00	**Bożena Cetnarowska** (U. Silesia, Katowice (Poland))	Derivational paradigms or paradigms of function? Competition between Polish affixal formations, morphological compounds and phrasal lexemes
15:00 – 15:30	**Lior Laks** (Bar-Ilan U.)	Towards uniformity within derivational paradigms: Evidence from Hebrew

Booklet of abstracts

Summary

PLENARY TALKS

Paradigms, features, and the lexeme

Andrew Spencer
University of Essex

According to one tradition, inflection interfaces with syntax/phrasal semantics, defining the (intralexemic) forms of a lexeme compatible with insertion into syntactic representations, while derivation interfaces with the lexicon, defining interlexemic relations. However, much recent work on paradigms emphasizes similarities between inflectional paradigms (Π_i) and derivational paradigms (Π_δ), even sometimes denying the distinction (and hence effectively rejecting the lexeme concept). I argue for an architectural distinction between Π_is and Π_δs, centering on the notion of lexeme, and defining a Π_i as an Orthogonal Multi-dimensional Feature Structure (OMDFS).

Bonami & Strnadová (2018) argue that all the standard non-canonical properties of Π_is are replicated in Π_δs. To their list I add transpositions: Π_δs can include transpositional lexemes such as Romance/Slavic/Greek/… Relational Adjectives. However, when we 'reverse engineer' some of the Π_δ correlates we find that the comparisons stand up less well, in part because Π_δs are not, in practice, definable as a non-trivial OMDFS. Following Stump (2016) I assume a content vs form/realized paradigm for Π_is, and, following Sadler & Spencer (2001), a related distinction between m-features (morphomic) and s-features (syntactico-semantic). It seems that neither of these distinctions can easily be reproduced in Π_δs.

The differences between Π_is and Π_δs can be made to flow from the assumption that Π_is realize *obligatory* contrasts in OMDFSs (Jakobson's Principle: inflection is the set of distinctions the language *must* express), while Π_δs define networks (*réseaux*—Fradin) of related lexemes which *permit* the naming of a Thing, Property, Situation, …, but which don't require it. This puts the onus on solving the problem of lexemic individuation, which is in any case a sine qua non for any lexeme-based model.

COMMUNICATIONS

The English de-verbal lexis as a problem of suffix-sensitive *paradigm discovery*

Michael Bilynsky
Ivan Franko Lviv National University
Mykhaylo.bilynskyy@lnu.edu.ua

1. Introductory remarks

The proposal consists in presenting an electronic framework for the study of verb-related lexemes in English.

We proceed from the assumption that "a morphological family is a tuple *F = (w1, wn)* such that any member *wi* of the family is morphologically related to any other member *wj*. A morphological family *F* is complete if there exists no larger morphological family that contains all members of F. A morphological family is partial if it is not complete" (Bonami, Strnadová 2018: 169).

2. Derivation: adjacent notions and disputable cases

A word family entails connections of word-relatedness on the premise of sameness of the root for all its members. Part of these ties are **derivational**. Conversely, by the scope of application to lexemic facts word-formedness is broader than derivability ('genuine', or 'live', derivation) of coinages. In lexicology, *derivation* as a term is adjacent to *word building* (rarely used), *word formation* (WF), almost parallel, but more often applicable to diachronic issues, and *transposition*, an older term which presupposes a change of the part-of-speech (POS) status of the verb. The substance (nouns) and quality (adjectives and participles) as verb-related suffixed lexemes (sometimes referred to as complexes) are steps, also known as zones or branches of de-verbal families.

Derivation is effected by the attachment of a formative, in our case a suffix, to the base. Tentatively, we equate the notions of the stem, the root or even the whole verb to which the suffix is attachable so as to produce de-verbal coinages.

Sometimes, a de-verbal coinage is of a different status than supposed, or we think a lexeme to be a derivative when it is actually non-derived, i.e. A-morphous, or seeming, coinages with the imagined verb; "sham" derivatives, which are non-derived borrowings, cf. interesting details in (ten Hacken 2020: 8-9); "shadow" borrowings that look like loans or, "copies", but, in fact, are coinages from borrowed verbs with missing morphemic counterparts in the source language. There are also delusionary psycholinguistic effects in derivation.

In word structure, the 'seams' between the suffix and the stem may not be fully transparent, which is an issue of natural morphology. They upset derivation, but comply with word-relatedness.

3. What is the factual evidence ?

Over 17,700 verbs have been found related to at least one documented coinage. The total number of words that reveal suffix arrangements around the respective common-root verbs in English de-verbal lexis proves to surpass 30,000 lexemes. Some additions, including nonce coinages, found in corpora or other sources, as well as recent, and, sometimes, previous specialist literature on de-verbal words, can be added to the framework.

All de-verbal family constituents are provided with dated reference to the earliest Oxford English Dictionary (OED) quotation.

For the sake of an introductory illustration we take a random, but quite well-represented example of a de-verbal family, which has only three categories missing from the complete set:

amuse (1480); 1: amusing (1603); 2: amusement (1611); 3: amuser (1583); 4: amusee (1838); 5: amusive (1728); 6: amusing (1597); 7: amusable (1832); 8: amused (1600); 9: amusively (1776); 10: amusiveness (1805); 11: amusingly (1812); 12: amusingness (1823); 15: amusedly (1844); 17: amusement (1673).

Such crowded families are certainly rare. Derivational constraints will appear part and parcel of derivation. Archaic lexemes are marked by the asterisks. Variant suffixes are called blur suffixes. Blur suffixes, as in de-verbal adjectives in the example below, are attributed respective dated textual prototypes:

adopt (1548); 1: adopting (1591); 2: adoption (1387); 3: adopter (1572); 4: adoptee (1892); 5: adoptive (1430; (-ive, 1430; -ant, 1671); 6: adopting (1717); 7: adoptable (1843); 8: adopted (1590); 9: adoptively (1844); 14: adoptability (1843); 15: adoptedly (1603); 17: adoption (1382); 18: adoptional (1861)

Verbs that are derived from other verbs do not fall under the notion of transposition. Derived verbs and their transposed derivatives, if they have them, are taken for separate de-verbal families. Verbs with no derivatives have been disregarded.

4. **The Paradigm Cell Filling and Finding Principles**

In derivation, there is the notion of a derivation paradigm (cf. Bauer 1998). Derivation paradigms are exemplified by WF-pairs. They are also known as derivation sets (Bauer 2019: 160). The second element of such a pair can figure as a parent for the subsequent pair. Such derivation rests on Dokulil's notion of a chain, or a later conception of suffix ordering.

A paradigm may be manifested by a single de-verbal item of a verb-derivative pair. This is, however, contradictory to derivation. Yet a one-member paradigm potentially is 'paradigm-able' too. According to Štekauer "the essential features of derivational paradigms is the availability of slots (filled with potential words) that are more important for the paradigm than the forms which fill them" (Štekauer 2014: 369; cf., Bauer 2019: 159).

As suggested by Boyé and Schalchli (2019: 245; cf., also, 2016) **the cell cultures** of inflectional morphology are to be brought to derivation and provide the tools for an extension of the Paradigm Cell Filling Principle (PCFP). Though originally, and still continuously (Ackerman, Malouf 2015), applied to the study of inflectional forms, this method can also reveal the paradigmatic binding of derivatives.

The idea of the growing lexicon, "le lexique tel qu'il naît" (Roché 2011) is reflected in the phenomenon of *paradigm discovery* (cf. Erdman et al. 2020), also known as paradigm *cell* discovery (Elsner et al. 2019: 152) or a *morphological forest* (Luo et al. 2017).

In the discussed framework, English verbs with their shared-root lexemes were keyed into an equal number of 'open-to-editing' e-grids. The developed software allows multiple queries. They show and analyze the filled cells as well as present, and re-fittingly (with changing parameters) visualize data on word-formedness and derivation.

In the meaningful description of de-verbal coinages, it is expedient to distinguish between such notions as onomasiological features, bases and links (or relaters). The first two reveal the POS affiliations, or, correspondingly, sub-paradigms, of de-verbal words. The relater is of syntactic (propositional) relevance in the sense of opposing diatheses (active voice *vs.* passive voice) of the paraphrases of de-verbal elements. The relaters are treated as grammar-oriented categorizers. They retain the syntactic categories of the verb and motivate the respective proposition-oriented sub-paradigms inside de-verbal families. Basing on derivational chains there are also suffix-order sub-paradigms. The singled out sub-paradigms of de-verbal coinages may involve the same lexical material and be partially or fully overlapping.

In the e-grid, recognizable categories in a default suffix were awarded separate cells. Notionally, this term combines both the cognitive features of a slot and the modelling characteristics of a square.

If a category has no default suffix, we fill the cell with a derivative in one of the singly attested blur suffixes. This situation resembles the No Blur Principle for inflectional morphology (Blevins 2016: 196). However, there could be several blur suffixes for one position in some types of derivatives from numerous verbs. They admit of the Enumeration of Cells (E-Cell) operation just as the cells in the default or single blur suffix. Blur suffixes can take up the position of a single blur suffix in the place where other categories have default suffixes when attested.

All the coinages contained in the framework were taken from the Oxford English Dictionary (updated to OED 3, where possible) which possesses the most complete inventory of de-verbal lexis.

In the downloadable examples from the framework upon a given query, the verb will be noted as zero and the de-verbal coinages as the ordinal taxonomy positions of D n (1-20). None of the known works on de-verbal morphology operates with such an inventory of suffixes.

The agreed categories are successively tagged. Here, omitting the logic of this succession, we stress that the numbering proves beyond a mere formality. The framework is sifting and classificatory, with extensions to textual evidence. The found morphological data are marked by the corresponding number label given in brackets within the E-Cell calculus of de-verbal types in the default or blur (single, precedent or traded-off) suffix.

We distinguish ACTION NOUNS (D1) in -ING, -AGE, -AL, (-A/-E)NCE, -MENT, (-T/-S)ION and (T/-S)URE that are free of same-word factitive lexicalization. There are also ACTION NOUNS (D2) with the same suffixes that admit of factitive lexicalization. De-verbal families have AGENTS (D3) in -ANT, -ER, -IVE, -OR and PATIENTS or/and OBJECTS (D4) in -ANT, -EE and -ER.

Among the non-nouns, we distinguish ADJECTIVES (D5) in (-A/-E)NT, -FUL, -IVE, -ORY, -OUS AND -Y, active diathesis PRESENT PARTICIPLES (D6) in -ING, MODAL ADJECTIVES (D7) in (-A/-I)BLE and passive diathesis PAST PARTICIPLES (D8) in -ED. Second-order coinages from adjectives and participles are ADVERBS (D9, D11, D13, D15) in -LY and NOUNS (D10, D12, D14, D16) in -NESS and, selectively (not after participles), in -ITY.

RESULT substantives (D17) are mostly lexicalizations, coincident suffix wise (but not always) with ACTION NOUNS (D2). Hence, in a variant notation they are given as D2'. Sporadic de-nominal adjectival derivatives (D18), their third order adverbs (D19) and, eventually, sporadic second-order de-nominal substantive coinages (D20) conclude the E-Cell calculus.

The framework admits the integration of cells (I-Cell) and the appropriate data rebuilding procedures. With this purpose, it can present ACTION NOUNS jointly (D1 and/or D2). Also, it can disregard lexicalization and record each noun in the said suffixes by its earliest attestation. Similarly, ADJECTIVES (D5) and active diathesis PRESENT PARTICIPLES (D6) as well as their second-order derivatives (D9 and D11 for adverbs and, respectively, D10 and D12 for nouns) can be cell-integrated too. Here, taking place would be the overlap of the two PCF patterns principles (of Filling and Finding) and the correspondingly merged data (PCF(x2)P) interplays.

Even though there are suffix exponents for each slot, cells which overlap in cases of coincidental (shared) suffixes pose a PC2ndFP problem. Possibly, involved in this procedure are PRESENT PARTICIPLES in -ING and ACTION NOUN and/or RESULT NOUN in -ING. This kind of cell economy (cf. Blevins 2016, 184-189), or syncretism (suffix and, eventually, coinage homonymy) also occurs in AGENT NOUNS AND ADJECTIVES IN -IVE and -ANT, AGENT NOUN AND PATIENT NOUN in -ER and -ANT (note the 3-times occurrence of -ANT in the slots of de-verbal derivation).

The data takes into account the whole of the OED, which is in line with the requirement that in constructing extensive derivational morphology "more data is better data, exhaustive data is the best" (Boyé, Schalchli 2019: 245).

However, the suggested framework has certain drawbacks. Some of the derivatives and their verbs are not in active use. In about one fifth of the drawn families, both a verb and (a) coinage(s) or, eventually, either a verb or (a) coinage(s), are recognized by the OED as archaic. In the cells of the e-grids, we mark them by a preceding asterisk.

There is also a problem with homonymous verbs. Usually, they are two, but could be three and even four verbs. In cases of coincident derivation there arises homonymy in some cells as well.

We assume that a verb and a derivative are related through the suffix and a shared root not only formally but also *semantically*. However, the framework is extendable into by-families on the principle of recognizable meanings of the verb (polysemy). Then, some coinages are taken as related just to specific meanings of verbs or, possibly, their valence and prepositional government. With this purpose, the grid for the earliest uses of semantically unspecified verbs and shared-root derivatives yield sub-grid(s) which may be somewhat diversified in terms of composition. If the blur suffixes are cell-coincident as regards pair or multiple cell doublets with plausible Gaussian ecological rivalry (cf. Aronoff 2016), an arbitrary derivative may fill the grid. If the E-Cell principle obliges, the blur suffixes are filled in by-cells.

The contents of a given by-cell, both the blur suffix and the corresponding earliest OED citation date marker, are exchangeable with those of the corresponding main cell. Such a swap covers all the families where it applies, when required. The whole framework with this upgraded part of the evidence then becomes accessible to all queries, once it has been rebuilt. In such cases of refitted cells, the downloaded lists of examples contain derivatives with "stocked suffixes" notation. This is demonstrative of the exchange and could be corrected manually. In a slightly more morphological vision of this situation, suffix exponents are doublets and the respective cells prove re-suffixed.

5. Layered construal of paradigmatic relations

The shared-root verb-related words are made storable and searchable (by the PCF(x2)P) in the **procedure of layered construal** of paradigmatic relations. We introduce the **2-storey cell (re-)entry/recovery principle (2SCE/RP)** for extracting shared-root derivatives from those stored in e-grids by their filled-in positions. It is possible to find exclusive and inclusive (respectively, "on their own" and "contained elsewhere", cf. the elsewhere condition) sample sets. Numerically, the sets that are drawn from the framework, reflect the tokenization of (sub-)paradigms. These are groups of cells, or recurring partials of paradigms (cf. Elsner et al. 2019: 137).

The data contained in the above-described de-verbal family grids flows into the layered-sifting grid. Sifting takes place at the second, "finding" or "discovery" (cf. Elsner et al. 2019: 130-131, 148-152), stage of the process. The upper level presents the main sifted data, and the lower level does the subsidiary sifted data.

The main sifting procedure focuses on the ***island*** *paradigmatic contouring* of the filled cells. In it, we are interested in the exhaustive ("all-examples-counted-in") tokenization of the composition of shared-root de-verbal coinages. All possible variants of the make-up of de-verbal paradigms can be checked against the filled cells. They will include strictly derivational, purely word-related and mixed paradigms.

The exclusive query reveals the individual realized WF potential of a verb. It can be taken as a verb's morphological **WF class identifier**. The realization strengths of such classes can be established. The inclusive query gets more informative with the growth of gaps in shared-rood relatedness as gaps can refer to the predictability of coinages.

The lower sifters, in their turn, can contour specified sub-categories as well as the amount of deficiency and mutual co-occurrence of cells inside them. This type of queries strategy is more likely to reveal gold sub-paradigms in the 'hedged setting' whereas the upper level sifters are more fit for revealing paradigm hapaxes.

The de-verbal paradigms are scaled numerically and assessed as realized frequency. Output trajectories range from gold to hapax paradigms, which would be unrecoverable non-electronically without the Paradigm Cell Filling and Finding Principles (PCF(x2)P). It is possible to check all the theoretical combinations of cells and get the list of matching verb-stem sets for each configured combination.

We introduce the Zipfian function of realized paradigm tokens into Štekauer's Predictability Rate (cf. Štekauer 2005: 58) as a measure of mutual occurrence by a filled/gapped cell or combined cells in the present-day de-verbal derivation framework.

The prediction of the composition of configured shared-root coinages rests on the computable mutual occurrence forecasts. They are calculable as uneven ratio probabilities. There are multiple combinability inferences predicting that, if a family possesses a given category, then in a certain percentage of cases it has got a comparable category. This can shed light on the issues of the tightness of links between the cells/nodes of the paradigm (Fradin 2020: 84) by using the philosophy of joint predictiveness (Bonami, Strnadová 2018: 195). Also, cf. an idea about the "degrees of paradigm coherence and applicability" in (Bauer 1998:248). Hypothetically, cells are more mutually predictable within sub-paradigms than between them.

The likelihood of adverbs and nouns derived from adjectives/participle as two-link single/multiple chains is numerically correlative with the attested substantive branch of a de-verbal family. This proves the overall relatedness of the width and depth of derivation processes in shared-root families. We call it the De-verbal Family Branching Principle.

Incorporated through the e-sifters into alternatively rebuilt frameworks are multiple thematic (cf., e.g., Fernández-Alcaina, Čermák 2018; Bagasheva 2017), etymological and chronological parameters of verbs. Then, one over-all De-verbal Morphology will produce split and variant morphologies.

6. Comparable approaches and concluding remarks

The suggested approach appears to be in line with the recent version of realization theory of Paradigm Function Morphology (Stump 2019) and the philosophy of "giving second life to lexical resources" (Hathout, Namer 2016; cf., also, Hathout 2009). The sub-paradigms of English de-verbal word formation seem comparable and worth comparing with the data on French de-verbal morphology (cf. Hathout, Namer 2019, Namer, Hathout 2020).

The developed framework has shed some light on multiple parametric descriptions of de-verbal families in English. Just as verbs reveal relationships of sense adjacency, so can categories of coinages forge relations in a reflection of synonymy in derivation that can be describable in terms of yet a further extension of the cells culture in present-day morphology.

References

Ackerman, F. and Malouf, R. (2015). The no blur principle effects as an emergent property of language systems. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, 41, pages 1-14.

Aronoff, M. (2016). Competition and the lexicon. In E. Annibale, C. Iacobini, and M. Voghera, editors, *Livelli di Analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della società di linguistica Italiana*, pages 39–52. Roma. Bulzoni Editore.

Bagasheva, A. (2017). Comparative semantic concepts in affixation. In J. Santana-Lario and S. Valera, editors, *Competing Patterns in English Affixation*, pages 33–66. Bern. Peter Lang.

Bauer, L. (1997). Derivational paradigms. In G. Booij and J. van Marle, editors, *Yearbook of Morphology* 1996, pages 243–256. Springer.

Bauer, L. (2019). Notions of paradigm and their value in word-formation. *Word Structure,* 12 (2), pages153–175.

Blevins, J. P. (2016). *Word and Paradigm Morphology*. Oxford. Oxford University Press.

Bonami, O. and Strnadová, J. (2019). Paradigm structure and predictability in derivational morphology. *Morphology,* 29, pages 167–197.

Boyé, G. and Schalchli, G. (2016). The status of paradigms. In A. Hippisley and G. Stump, editors, *The Cambridge Handbook of Morphology*, pages 206–234, Cambridge. Cambridge University Press.

Boyé, G. and Schalchli, G. (2019). Realistic data and paradigm: the paradigm cell finding problem. *Morphology*, 29, pages 199–248.

Elsner, M., Sims, A. D., Erdmann, A., Hernandez, A., Jaffe, E., Jin, L., Johnson, M. B., Karim, S., King, D. L., Lamberti Nunes, L., Oh, B.-D., Rasmussen, N., Shain, C., Antetomaso, S., Dickinson, K. V., Diewald, N., McKenzie, M., and Stevens-Guille, S. (2019). Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modelling*, 7(1), pages 53–98.

Erdmann, A., Elsner, M., Wu, Sh., Cotterell, R., and Habash, N. (2020). The paradigm discovery problem. *Proceedings of the 58th annual Meeting of the Association of Computational Linguistics*, pages 7778–7790. https://www.aclweb.org/anthology/2020.acl-main.695.pdf

Fernández-Alcaina, C. and Čermák, J. (2018) Derivational paradigms and competition in English: A diachronic study on competing causative verbs and their derivatives. *SKASE Journal of Theoretical Linguistics*, 15(3), pages 69–97.

Fradin, B. (2020). Characterizing derivational paradigms. In J. Fernández-Domínguez, A. Bagasheva, and C. Lara Clares, editors, *Paradigmatic Relations in Word Formation*, pages 49–84. Leiden-Boston. Brill.

ten Hacken, P., and Panocová, R. (2020). Word formation, borrowing and their interaction. In P. ten Hacken, R. Panocová, editors, *The Interaction of Borrowing and Word Formation*, pages 3–14. Edinburgh. Edinburgh University Press.

Hathout, N. (2009). Acquisition of morphological families and derivational series from a machine readable dictionary. In F. Montermini, G. Boyé, and J. Tseng, editors, *Selected Proceedings of the 6th Décembrettes*: Morphology in Bordeaux, pages166–180. Somerville, MA Cascadilla Proceedings Project.

Hathout, N. and Namer, F. (2016). Giving lexical resources a second life: Démonette,a multi-sourced morpho-semantic network for French. *Language Resources and Evaluation Conference*, May 2016, Portoroz, Slovenia. hal 02054275.

Hathout, N. and Namer, F. (2019). Paradigms in word formation: what are we up to. *Morphology*, 29, pages 153–165.

Luo, J., Narasimhan, K., and Barzilay, R. (2017). Unsupervised learning of morphological forests. *Transactions of the Association of Computational Linguistics*, 5, pages 354–364.

Namer, F. and Hathout, N. (2020). ParaDis and Démonette – from theory to resources for derivational paradigms. *The Prague Bulletin of Mathematical Linguistics*, De Gruyter, 114 (1), pages 5–34.

Proffitt, M. (2018) The Oxford English Dictionary. (http://www.oed.com)

Roché, M. (2011). Quelle morphologie ? In M. Roché, G. Boyé, N. Hathout, S. Lignon, and M. Plénat, editors, *Des unités morphologiques au lexique, langues et syntaxe*, pages 15–39. Paris. Lavoisier.

Štekauer, P.(2005). *Meaning Predictability in Word Formation*. Amsterdam/Philadelphia. John Benjamins.

Štekauer, P. (2014). Derivational paradigms. In Lieber, R. and Štekauer, P., editors, *The Oxford Handbook of Derivational Morphology*, pages 354–369. Oxford. Oxford University Press.

Stump, G. (2019). Paradigm Function Morphology. In J. Audring and F. Masini, editors, *The Oxford Handbook of Morphological Theory*, pages 285–304. Oxford. Oxford University Press.

Derivational paradigms or paradigms of function? Competition between Polish affixal formations, morphological compounds and phrasal lexemes

Bożena Cetnarowska
University of Silesia, Katowice (Poland)

1 Introduction

This paper provides support for the broadening of the notion of derivational paradigms – i.e. of paradigms of function (Bauer 2019), understood as sets of derivational families whose members realise the same cognitive categories - to the notion of word-formation paradigms (Gaeta and Angster 2019, Bagasheva 2020).

Word-formation paradigms can be employed to capture the competition between affixal derivatives, morphological compounds and phrasal lexemes. Phrasal lexemes are defined as multi-word expressions (MWEs) which show phrasal internal complexity but which have the naming function and show lexical integrity (Booij 2010, Masini 2009). Their formation can be regarded as a case of "periphrastic word-formation" (Booij 2002).

2 Derivational paradigms and feminine occupation terms

The concept of a derivational paradigm – resembling an inflectional paradigm and consisting of a table with cells (Hathout and Namer 2019) - has been applied felicitously to the discussion of selected derivational categories, such as names of young animals (Manova et al. 2019). Other derivational and conceptual categories may turn out to be less amenable to the analysis in terms of paradigms with cells to be filled by affixal formations. For instance, a derivational paradigm consisting of suffixal feminine nouns in English is defective and incomplete, with most cells being unoccupied (except for those filled by a couple of institutionalized *-ess*, *-trix,* or *-ette* nouns, e.g. *actress, aviatrix*). A word-formation paradigm for feminine occupation terms which would contain both affixal derivatives and compounds related to a given base, could indicate that some cells are filled by compounds with gender-specific lexemes *woman* and *lady* (as in 1).

(1) a. *writer* → **writeress*, **writerette*, *woman writer*
b. *president* → **presidentess*, **presidentrix*, *woman president*, *lady president*

Derivation of feminine forms is productive in Slavonic languages (see Čmejrková 2003 on Czech). However, it would be useful to conceive of word-formation paradigms for feminine forms as including both derivatives, compounds and compound-like phrasal lexemes. Polish suffixal feminine occupation terms compete with multi-word units which do not meet the criteria of morphological compounds. The morphologically well-formed derivatives with the feminine suffix *-k(a)* listed in (2a, 3a) are rejected by normative grammarians (see Łaziński 2006 and Szymanek 2010 for more discussion). Speakers of Polish tend to avoid such forms and replace them by multi-word units (2b, 3b, 2c, 3c).

(2) a. (*)*prezydent-k-a* (president + FEM + NOM.SG) 'female president'
b. *kobieta prezydent* (woman + NOM.SG president.NOM.SG) 'female president'

c. *pan-i prezydent* (lady + NOM.SG president.NOM.SG) 'lady president'

(3) a. (*)*kancler-k-a* (chancellor + FEM + NOM.SG) 'female chancellor'
b. *kobiet-a kanclerz* (woman + NOM.SG chancellor.NOM.SG) 'female chancellor'
c. *pan-i kanclerz* (lady + NOM.SG chancellor.NOM.SG) 'lady chancellor'

While (2) and (3) testify to the coexistence of competing phrasal nouns, in (4) the cooccurrence of phrasal nouns with a suffixal derivative is exemplified. If the lexical units in (4a–c) are regarded as being placed in the same cell of a word-formation paradigm, note should be taken that they are not exact synonyms (e.g. they differ in emotional colouring, degree of politeness, stylistic value or range of usage). This would explain why the forms in (4a–c) do not block each other (see Aronoff 1976, Plag 1999 on synonymy blocking).

(4) a. *profesor-k-a* (professor + FEM + NOM.SG)'female professor or secondary school teacher'
b. *kobiet-a profesor* (woman + NOM.SG professor.NOM.SG) 'female professor'
c. *pan-i profesor* (lady + NOM.SG professor.NOM.SG) 'lady professor'

3 Morphological condensation

The coexistence of suffixal forms and multi-word expressions (MWEs) in Polish is exemplified further in (5–6). The suffixal formations are regarded as products of the process of morphological condensation of MWEs, i.e. the process of univerbation (Martincová 2015, Szymanek 2010). The univerbated forms are generally perceived as more colloquial than MWEs (as shown in 5), but some of them are stylistically neutral and thus fully synonymous with suffixal derivatives (as in 6). Booij and Masini (2015) propose second-order construction schemas (postulated within the theory of Construction Morphology) to model paradigmatic relations between morphological and phrasal schemas (the latter generalizing over sets of A + N or N + A phrasal nouns).

(5) a. *szkoł-a podstaw-ow-a* (school + NOM.SG base + ADJZ + NOM.SG) 'primary school'
b. *podstaw-ów-k-a* (base + ADJZ + NMLZ + NOM.SG) '(colloq.) primary school'

(6) a. *statek kabl-ow-y* (ship.NOM.SG cable + ADJZ + NOM.SG) 'cable-laying ship'
b. *kabl-owi-ec* (cable + ADJZ + NMLZ.NOM.SG) 'cable-laying ship;

4 Coexistence of morphological compounds and MWEs

Numerous pairs can be found of Polish morphological compounds coexisting with MWEs built of the same stems (cf. Masini 2019 for a discussion of competition between compounds and MWEs in Italian). Three types of situations when blocking seems to be suspended are discussed below.

Firstly, in some of those instances exemplified in (7–8), blocking does not operate due to the lack of synonymy between morphological compounds and MWEs. This can be treated as competition between patterns (i.e. type blocking, cf. van Marle 1986, Rainer 2005) rather than as token blocking. There is a "division of labour" between the two types of composite expressions in Polish. A + N compounds proper in (7a, 8a), which are characterized by the occurrence of a linking vowel (LV) between the two stems, are attributive exocentric compounds whereas N + A multi-word expressions (7b, 8b) require an endocentric interpretation.

(7) a. *równoległ-o-bok* (parallel + LV + side.NOM.SG) ‘parallelogram’
b. *bok równoległ-y* (side.NOM.SG parallel + NOM.SG) ‘parallel side’
(8) a. *doln-o-płat* (low + LV + wing.NOM.SG) ‘low-wing plane’
b. *płat doln-y* (wing.NOM.SG low + NOM.SG) ‘low wing’

Secondly, the coexistence of multi-word expressions side by side with synonymous compounds may be indicative of a change in progress. The N + N MWE with coordinate multifunctional interpretation in (9a) sounds dated and is usually replaced by the compound proper given as (9b). Both the compound in (10b) and the N + N MWE in (10a) are attested, yet a Google search for the compound brings fewer results that the search for the multi-word unit.

(9) a. ?*spódnic-a spodni-e* (skirt + NOM.SG trouser + NOM.PL) ‘culottes’
b. *spódnic-o-spodni-e* (skirt + LV + trouser + NOM.PL) ‘culottes’
(10) a. *piekarni-a cukierni-a* (bakery + NOM.SG café + NOM.SG) ‘bakery and café’
b. *piekarni-o-cukierni-a* (bakery + LV + café + NOM.SG) ‘bakery and café’

Thirdly, some morphological compounds are not institutionalized (e.g. 11a, 12a) and are normally replaced by appropriate multi-word units (11b, 12b). However, such compounds can occur as attention-seeking devices (Lipka 1987, Konieczna 2012) in journalese or in texts posted on blogs.

(11) a. ?*prezydent-o-bójc-a* (president + LV + killer + NOM.SG) ‘presidential assassin’
b. *zabójc-a prezydent-a* (killer + NOM.SG president + GEN.SG) ‘presidential assasin’
(12) a. ?*krwi-o-dawani-e* (blood + LV + giving + NOM.SG) ‘blood donation’
b. *oddawani-e krw-i* (donating + NOM.SG blood + GEN.SG) ‘blood donation’

5 Conclusion

The modelling of competition between affixal derivatives, morphological compounds and phrasal lexemes requires the broadening of the notion of the “derivational paradigm” and the inspection of the products of “periphrastic word-formation”. This also necessitates a reconsideration of what kind of (and what degree of) differentiation can be exhibited between derivatives, compounds and compound-like expressions which can be treated as filling the same cell in the paradigm.

References

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge MA: MIT Press.
Bagasheva, Alexandra. 2020. Paradigmaticity in compounding. In Jesús Fernández-Domínguez, Alexandra Bagasheva & Cristina Lara-Clares (eds.), *Paradigmatic relations in word-formation*, 21–48. Leiden: Brill.
Bauer, Lauer. 2019. Notions of paradigm and their value in word-formation. *Word Structure* 12(2). 153–175.
Booij, Geert. 2002. Separable complex verbs in Dutch: A case of periphrastic word formation. In Nicole Dehé, Ray Jackendoff, Andrew McIntyre & Silke Urban (eds.), *Verb-particle explorations*, 21–42. Berlin & New York: De Gruyter.
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, Livia Kőrtvélyessy & Pavol Štekauer (eds.), *Semantics of complex words*, 47–66. Cham: Springer.

Čmejrková, Svetla. 2003. Communicating gender in Czech. In Marlis Hellinger & Hadumod Bussmann (eds.), *Gender across languages. The linguistic representation of women and men*, vol. 3, 27–58. Amsterdam: John Benjamins.

Gaeta, Livio & Marco Angster. 2019. Stripping paradigmatic relations out of the syntax. *Morphology* 29. 249–270.

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

Konieczna, Ewa. 2012. Analogical modelling and paradigmatic word formation as attention-seeking devices. In Angela Ralli, Geert Booij, Sergio Scalise & Athanasios Karasimos (eds.), *Morphology and the architecture of grammar. Online proceedings of the Eight Mediterranean Morphology Meeting MMM8 (Cagliari) 14–17 September 2011*, 168–191. Patras: University of Patras.

Lipka, Leonhard. 1987. Word-formation and text in English and German. In Brigitte Asbach-Schnitker & Johannes Roggenhofer (eds.), *Neuere Forschungen zur Wortbildung und Historiographie der Linguistik*, 59–67. Tübingen: Gunter Narr.

Łaziński, Marek. 2006. *O panach i paniach. Polskie rzeczowniki tytularne i ich asymetria rodzajowo-płciowa*. Warszawa: Wydawnictwo PWN.

Manova, Stela, Dmitri Sitchinava & Maria Shvedova. 2019. Derivational paradigms: Rules, patterns or neural networks? Paper presented at the SLS Annual Meeting, Potsdam, September 2019.

Martincová, Olga. 2015. Multi-word expressions and univerbation in Slavic. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-formation: An international handbook of the languages of Europe, vol.1*, 742–757. Berlin: Mouton de Gruyter.

Masini, Francesca. 2009. Phrasal lexemes, compounds and phrases: A constructionist perspective. *Word Structure* 2(2). 254–271.

Masini, Francesca. 2019. Competition between morphological words and multiword expressions. In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler & Hans Christian Luschützky (eds.), *Competition in inflection and word-formation*, 281–305. Cham: Springer.

van Marle, Jaap. 1986. The Domain Hypothesis: The study of rival morphological processes. *Linguistics* 24. 601–621.

Plag, Ingo. 1999. *Morphological productivity: Structural constraints in English derivation.* Berlin & New York: Mouton de Gruyter.

Rainer, Franz. 2005. Constraints on productivity. In Pavol Štekauer & Rochelle Lieber (eds.), *Handbook of word-formation*, 335–352. Dordrecht: Springer.

Szymanek, Bogdan. 2010. *A panorama of Polish word-formation*. Lublin: Wydawnictwo KUL.

Striking out on one's own: idiosyncratic frequency as a measure of derivation vs inflection

Maria Copot — LLF, Université de Paris
Timothee Mickus — ATILF, CNRS/Université de Lorraine
Olivier Bonami — LLF, Université de Paris

1 Motivation

A growing body of work argues that derivation mirrors the paradigmatic organization of inflection (among many others, (Bauer, 1997; Blevins, 2001; Stump, 2005; Stekauer, 2014; **?**; Bonami & Strnadová, 2019)), suggesting that differences between the two domains be reassessed as gradient rather than categorical. For instance, it is generally agreed that what differentiates inflection and derivation is the semantic predictability of their output: inflection is employed to talk about the same concept in different contexts (*I read* vs *she reads*), while the role of derivation is to create words for new concepts (*to read* vs *readable*), which more easily undergo semantic shifts. Bonami & Paperno (2018) provide quantitative evidence for that conclusion, using methods from distributional semantics (see e.g. Boleda (2019)). Here we address a related but separate possible contrast between inflection and derivation. Because derived lexemes are distinct lexical entities, we expect their frequency to vary independently from that of their base. As an example, French *friperie* 'thrift shop' is twice as frequent as its base *fripe* 'used garment' while *dentellerie* 'lace shop' is 1000 times less frequent than its base *dentelle* 'lace'. This example highlights that frequency differentials in derivation may be independent of base semantics: the relative rarity of *dentellerie* is due to the shops selling lace typically being designated by another name; and *fripe* is just dropping out of usage independently of its derivative. By contrast, we expect frequency variation within inflectional paradigms to be more limited, and predictable from lexeme semantics (e.g. *eye* is more likely to be used in the plural than *nose*). In this presentation we provide quantitative evidence for the reality of that differential, using distributional semantics to assess semantic predictability.

2 Materials

We used the FrCoW-16X web-crawled corpus of French (Schäfer & Bildhauer, 2012; Schäfer, 2015) as our primary source of data. Whenever lemma annotations were missing, we converted the token into the most appropriate lemma given the POS tag using Levenshtein distance. We also computed word type frequency, as well as word2vec representations of both lexemes and word types (100D) - the former treats donkey and donkeys as instances of the same thing, while the second has separate vectors for each. We established inflectional paradigms using the GLÀFF lexicon (Sajous et al., 2013) and identified derivational families based on Démonette (Hathout & Namer, 2014). Due to the large amounts of data necessary, and the need to avoid systematic homophony between cells, we had to exclude several processes, and focus on 21 inflectional cells (20 from the verbal paradigm, and noun pluralisation), and two derivational cells (agent and action nouns).

3 Methodology

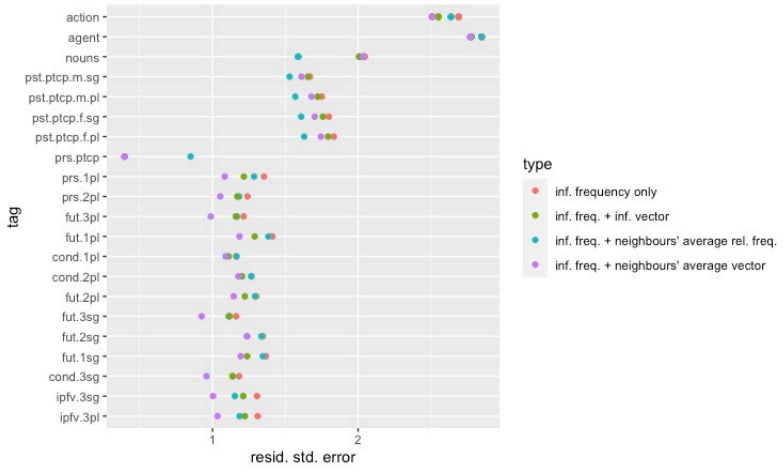
For each of the cells thus identified, we trained four types of linear models (frequency data is log-transformed): they all try to predict the frequency of an inflected or derived word based on the frequency of a reference form (for this purpose, we chose the citation form for inflected words, and the base form for derived words), alone or in conjunction with various types of semantic information. We consider four types of models:

1. Prediction based on frequency of the reference form alone, serves as a baseline.
2. 1 + reference form semantics. Adds an 8D vector[1] of the reference form to the predictors, which is the most basic type of semantic information, and the one we expect to perform worst.
3. 1 + neighbour frequency. Adds the average relative frequency of the semantic neighbours that underwent the same morphological process.[2]
4. 1 + neighbour semantics. Includes neighbour semantics more directly, by adding as a second predictor the weighted average 8D vector of the word's semantic neighbours that underwent the same process.[3]

For each of the models, we take residual standard error (RSE) as a measure of unpredictability. The higher the RSE, the more unpredictable the frequency (and relatedly, usage and meaning) of items in that cell. Since RSE is a continuous metric, it's a good candidate to capture the continuous nature of the inflection-derivation distinction.

4 Results

Within each cell, the four models give similar RSEs.



Residual standard error for the different processes selected, by model type

[1]The 8D vectors are obtained by applying a Truncated SVD to the 100D word type vectors.

[2]Semantic neighbours are lexemes that have a cosine similarity to the reference lexeme higher than 0.7 (based on the 100D lexeme vectors). The relative frequency is the frequency of the neighbour's form in the cell of interest divided by the frequency of the neighbour's reference form.

[3]Neighbours are selected in the same way as for Model type 3, and the 8D vectors are vectors for the neighbour's form in the cell of interest.

The cells with the highest RSE by far are traditionally derivational ones (agent and action nouns, ∼2.75), followed by noun pluralisation and the past participles (∼1.75), followed by the rest of the verbal inflectional cells included in the study (∼1.25). This is in line with the literature's placement of these cells along the inflection-derivation continuum, and certainly in line with intuitions about the predictability of their semantics. As expected, the model that most directly includes neighbour semantic information performs best in almost all cases. Exceptions to this are the past participles and nominal pluralisation, for which the best fit is the model based on average neighbour relative frequency. This reflects the inherent semantic ambiguity of these cells: past participles often simultaneously behave as participles and adjectives, and noun plurals often denote slightly different concepts than their singular form (as an extreme case, glass vs glasses). The semantic variance inherent in the output of these cells is likely what makes neighbour frequency a slightly better predictor than neighbour semantics.

5 Conclusion

The RSEs of models predicting the frequency of a form seem to successfully capture intuitions about the different nature of inflection and derivation, as well as its gradient character. The measure taps directly into the observation that inflection yields ways to talk about the same concept in different contexts, while derivation produces words for new concepts, a distinction which manifests itself in frequency predictability. The frequency properties of inflectional cells tend to be more systematic, in that they can easily be deduced from restricted samples of related forms, and/or from their base. On the other hand, derivation yields new lexemes with a lower degree of interdependence, and thus is more prone to intraparadigmatic semantic shifts resulting in changes of semantic neighbourhood and frequency profile, thereby increasing variation in this domain.

References

Bauer, Laurie. 1997. *Derivational paradigms* 243–256. Dordrecht: Springer Netherlands. doi: 10.1007/978-94-017-3718-0_13. https://doi.org/10.1007/978-94-017-3718-0_13.

Blevins, James. 2001. Paradigmatic derivation. *Transactions of the Philological Society* 99. 211–222. doi:10.1111/1467-968X.00080.

Boleda, Gemma. 2019. Distributional semantics and linguistic theory. *CoRR* abs/1905.01896. http://arxiv.org/abs/1905.01896.

Bonami, Olivier & Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio* 17. 173–195.

Bonami, Olivier & Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 28. 167–197. doi:https://doi.org/10.1007/s11525-018-9322-6. http://rdcu.be/Io3H.

Hathout, Nabil & Fiammetta Namer. 2014. Demonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5). 125–168. https://halshs.archives-ouvertes.fr/halshs-01110404.

Sajous, Franck, Nabil Hathout & Basilio Calderone. 2013. GLÁFF, un Gros Lexique Á tout Faire du Français. In *Actes de la 20e conférence sur le traitement automatique des langues naturelles (taln'2013)*, 285–298. Les Sables d'Olonne, France.

Schäfer, Roland. 2015. Processing and querying large web corpora with the cow14 architecture, .

Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new

efficient tool chain. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* .

Stekauer, Pavol. 2014. *Štekauer, p. 2014. 'derivational paradigms.' in: Lieber, r. – Štekauer, p. (eds.) the oxford handbook of derivational morphology. oxford: Oxford university press, 354-369* 354–369.

Stump, Gregory T. 2005. *Referrals and morphomes in sora verb inflection* 227–251. Dordrecht: Springer Netherlands. doi:10.1007/1-4020-4066-0_8. https://doi.org/10.1007/1-4020-4066-0_8.

Paradigmatic relations across morphology and syntax: Particle verbs in two Walser German varieties

Livio Gaeta
University of Turin

1 Introduction: Particle verbs in German

German particle verbs are a well-known case where the boundaries between components are difficult to draw as they have been treated with good arguments in syntax as well as in morphology (cf. Dehé et al. 2002, McIntyre 2015 for a general discussion).

One argument in favor of a morphological analysis comes from the paradigmatic relations that the particle verbs entertain with bona fide prefixed verbs like *aufladen* 'to load' vs. *entladen* 'to unload', etc. On the other hand, they are difficult to analyze straightforwardly as morphological objects because they display the well-known syntactic separation in contrast to bona fide prefixes as shown by *Hans lädt das Heu auf* 'Hans loads the hay' vs. *Hans entlädt das Heu* 'Hans unloads the hay'. Such a syntactic separation is crucially combined with the displacement of the verbal complex insofar as its finite and the non-finite pieces are placed in two different sentence positions, respectively the second and the final position: *Hans hat das Heu aufgeladen / entladen* 'Hans has loaded / unloaded the hay'.

2 Particle verbs in two Walser German varieties

In the paper, these two properties – namely paradigmatic relations through the lexicon with other complex words and syntactic displacement – will be discussed on the basis of two Walser German varieties spoken in linguistic islands found in northern Italy, namely in Gressoney and Issime.

Because of isolation and of long-standing contacts with Romance varieties, they display an interesting development insofar as in Titsch (1) spoken in Gressoney syntactic displacement is partially preserved, while in Töitschu (2) spoken in Issime it is completely lost (cf. Gaeta & Angster 2020):

(1) *Noa där Mäsch heintsch d'Lammiene uf d'Gheimnesse verkantot*
after the mass have.they the=lambs on the=mysteries auctioned.PSTPTCP
'After the mass they have auctioned the lambs for the Mysteries'.

(2) *hentsch k'offurut as lamji däm heilege Chin*
have.they offered.PSTPTCP a.N lamb[N].DIM the.N.SG.DAT holy child[N]
'they offered a little lamb to the holy child'.

In concomitance with the loss of syntactic displacement, in Töitschu (3) particle verbs have disappeared in favor of phrasal verbs, while they are well preserved in Titsch (4):

(3) *z bruat hescht gleit i / *igleit sua*
the bread have.2SG put.PSTPTCP in / *in.put.PSTPTCP so
'You have put the bread inside in this way'.

(4) *heintsch demnoa Heilége mét dem water zéemegleit / *gleit zéeme*

have.3PL hence saints with the weather together.put.PSTPCP / *put.PSTPCP together
'Hence they have combined the Saints with the weather'.

3 Particle and phrasal verbs along the Germanic/Romance edge

The reanalysis of particle verbs as phrasal verbs consisting of a verb immediately followed by a locative adverb is a generalized feature throughout the Töitschu lexicon, which stands in neat contrast with the conservative behavior of Titsch that resembles the rest of German varieties including Standard German. On the other hand, phrasal verbs are commonly found in Piedmontese and more in general in the Northern Italian contact varieties, as exemplified in the following table in which the paradigmatic relations centering on the base verb put are reported:

German	Titsch	Töitschu	Piedmontese	Italian	
einlegen	*élecke*	*lécken dri*	*büté 'ndrinta*	*mettere dentro*	('to put inside')
niederlegen	*embrélecke*	–	*büté giü*	*mettere giù*	('to put down')
auflegen	*uflecke*	*lécken ouf*	*büté sü*	*mettere su*	('to put up')
auslegen	*uslecke*	*lécken ous*	*büté fora*	*metter fuori*	('to put out')
vorlegen	*vorlecke*	*lécken vür*	–	–	('to put forward')
zulegen	*zuelecke*	*lécken zu*	–	–	('to put to')
zusammenlegen	*zéemelecke*	*lécken zseeme*	*büté 'nsema*	*mettere insieme*	('to put together')
Tab. 1: Correspondence patterns of particle verbs					

4 Particle verbs and paradigmatic force

On the other hand, in Töitschu the model of the particle verbs did not completely disappear, as shown by pairs of verbs in which both possibilities are found: *brechen ous* 'to escape, overflow' vs. *ousbrechen* 'to pierce (a wall)', etc. Moreover, a number of true prefixed verbs is still attested in Töitschu like *ischissen* 'to bake', *ubergien* 'to take over, overflow', *ubersprinnhen* 'to climb over', etc. which are similar to their cognates found in Titsch: *ésschiesse* 'to bake', *òbergé* 'to take over', *òbersprénge* 'to climb over, omit', etc.

Finally, in spite of the diffusion of the phrasal verbs, paradigmatic relations between verbs and corresponding abstract nouns still survive in Töitschu as shown by pairs *brechen ab* 'to dissuade, discourage, break off' / *abpruch* 'debris', *voan an* 'to begin' / *anvanh* 'beginn', which reflect the similar correspondences found in Titsch: *afoa* 'to begin' / *afang* 'begin', etc.

It has to be added that Titsch did not simply reflect a conservative system similar to the Standard German variety. In fact, particle verbs in Titsch have also lost at least partially properties like the morphological separation (6), whereby the subordinating particle *zu* has to be inserted between the particle and the base verb in Standard German (7):

(6) *òn fer di häscht nit khät de förcht z'vorwerz goa*
and for those have.2SG not had.PSTPTCP the fear to = forward go

‘and for those you were not scared of going ahead’.

(7) *und für die hast du keine Angst gehabt, vorwärts-zu-gehen / *zu vorwärtsgehen*
and for those have.2SG you NEG fear had.PSTPTCP forward-to-go / to forward.go
‘and for those you were not scared of going ahead’.

In the paper, both syntactic and morphological properties of particle verbs in Titsch and Töitsch will systematically be investigated showing similarities and differences with regard to each other as well as to Standard German, by paying particular attention to the impact of paradigmatic relations on their peculiar development.

Bibliography

Dehé, Nicole, Ray Jackendoff, Andrew McIntyre & Silke Urban (eds.) 2002. *Verb particle explorations*. Berlin / New York: Mouton de Gruyter.

Gaeta, Livio & Marco Angster. 2020. Loanword Formation in Minority Languages: Lexical Strata in Titsch and Töitschu. In Pius ten Hacken & Renáta Panocová (eds.), *The Interaction of Borrowing and Word Formation*. Edinburgh: Edinburgh University Press, 215-236.

McIntyre, Andrew. 2015. Particle-verb formation. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-Formation. An International Handbook of the Languages of Europe*. Berlin / New York: Mouton de Gruyter, vol. 2, 434-449.

Comparing derivational processes with distributional semantics

Matías Guzmán Naranjo Universität Tübingen
Olivier Bonami Université de Paris

1 Motivation

A main challenge for the study of word formation is the elusive nature of the semantic relations between derivationally-related words. While broad characterizations of these relations are readily available, applying them on a large scale proves to be both extremely time consuming and error-prone. This is due to a variety of well-identified factors, including but not limited to the polysemy of derivational processes, the unpredictable polysemy of both bases and derivatives, and the sheer amount of lexical data that needs to be examined. That situation however is something of an embarassment for paradigmatic approaches to derivation, which typically rely on the identification of morphosemantic relations to organize paradigms (Štekauer, 2014).

In this context it is tempting to rely on distributional methods to assess the semantics of morphological processes. A basic tenet of distributional semantics is that the distribution of a word is informative on its meaning: semantically similar words are expected to have similar distributions (Lenci, 2008), where the distribution of a word can be modeled as a vector in a high-dimensional vector space. In that context, the semantics of a morphological process can be seen as a function that maps input vectors to output vectors (Marelli & Baroni, 2015). The goal of the present research is to assess empirically the adequacy of that idea on the basis of large scale morphological data. Our research hypothesis is the following: if functions mapping input vectors to output vectors capture the semantics of derivational processes, then semantically similar processes should lead to similar functions—e.g. we expect quasi-synomnymous processes such as English *-age* and *-ion* nouns to be represented by very similar functions, but that these functions should be dissimilar to the function for *-er* agent nouns. We assess similarity between functions representing morphological relations using a cross-prediction task (Mickus et al., 2019): e.g. we expect that the function inferred from observing the relation between *-age* nouns and their base will also be good at predicting from a verb what the meaning of the corresponding *-ion* noun is.

2 Methods

Our dataset combines information on derivational relatedness in French compiled from various sources: Hathout & Namer (2014) for relations between verbs, action nouns, agent/ instrument nouns; Tribout (2010) for nouns and verbs related by conversion; Koehl (2012) for deadjectival nouns; Strnadová (2014) for derived adjectives; and Bonami & Thuilier (2019) for derived verbs in *-iser* and *-ifier*.

We use 100-dimensional distributional vectors to represent the meaning of the lexemes under study. These were obtained using using the Gensim (Řehůřek, 2010) implementation of word2vec (Mikolov et al., 2013)[1] with a version of the FrCoW corpus (Schäfer, 2015; Schäfer & Bildhauer, 2012) where each word is replaced with a tagged lemma; hence these vectors represent the distribution of lexemes in the context of other lexemes, rather than words in the context of other words.

[1]We used the skipgram algorithm with the following hyperparameters: 2 training epochs, 5 negative samples, window size 5, vector size 100.

We excluded all lexemes in the dataset for which there were no corresponding vectors in the distributional space. Additionally, we removed derivation pairs related by a derivational process with a type frequency of less than 50. The final dataset contained 21,990 pairs exemplifying 35 distinct processes. While that dataset does not give a full picture of the French derivation system (in particular denominal nouns are absent), it is diverse enough that different degrees of similarity are instantiated. The list of processes and number of types per process are given in Table 1.

Process	Pairs	Process	Pairs	Process	Pairs
CONVERSION:N > V	2372	*-aire*:N > A	439	*-if*:V > A	135
CONVERSION:V > N	2364	*-eux*:N > A	407	*-ien*:N > A	102
-ion:V > N	1960	*-iser*:A > V	383	*-erie*:A > N	101
-ique:N > A	1790	*-if*:N > A	375	*-ance*:V > N	95
-age:V > N	1644	*-eur*:V > A	356	*-erie*:V > N	87
-eur:V > N	1605	*-Vble*:V > A	324	*-té*:A > N	82
-ment:V > N	1299	PST.PART:V > A	317	*-ure*:V > N	75
-é:V > A	1272	*-iser*:N > V	286	*-ième*:N > A	68
-ité:A > N	1097	*-el*:N > A	281	*-ée*:V > N	66
-ant:V > A	915	*-ier*:N > A	204	*-itude*:A > N	62
-euse:V > N	536	*-rice*:V > N	198	*-ifier*:A > V	51
-al:N > A	458	CONVERSION:N > A	184		

Table 1: Processes in the dataset

For each process under consideration, we computed the average difference vector between the derived vector and the base vector. Average difference vectors approximate a function taking the base vector as input and giving the derived vector as output (Marelli & Baroni, 2015; Bonami & Paperno, 2018). The similarity between two processes can then be operationalized as the cosine similarity between their difference vectors. We use agglomerative clustering with an unweighted average linking function (Sokal & Michener, 1958) to deduce a dendrogram of similarities among processes.

3 Results

The result is shown in Figure 1. The groupings shown in Figure 1 are remarkably close to expectations. With only two exceptions, the models groups together broad classes of processes, such as deadjectival nouns, denominal adjectives, denominal verbs, and deverbal adjectives. It also identifies more fine-grained groupings, discriminating e.g. Angent/Instrument from Action deverbal nouns, and identifying the close proximity between *-age*, *-ment* and *-ion* among deverbal nominalizations.

4 Evaluation

In order to have an independent evaluation of the quality of our assessment of similarity, we collected expert opinion. 7 professional French morphologists not familiar with the present study provided hierarchical classifications of the 35 processes under investigation based on their similarity. We then compared the trees provided by experts among themselves and with the tree obtained from the vectors, using the proportion of shared clusters as a measure of

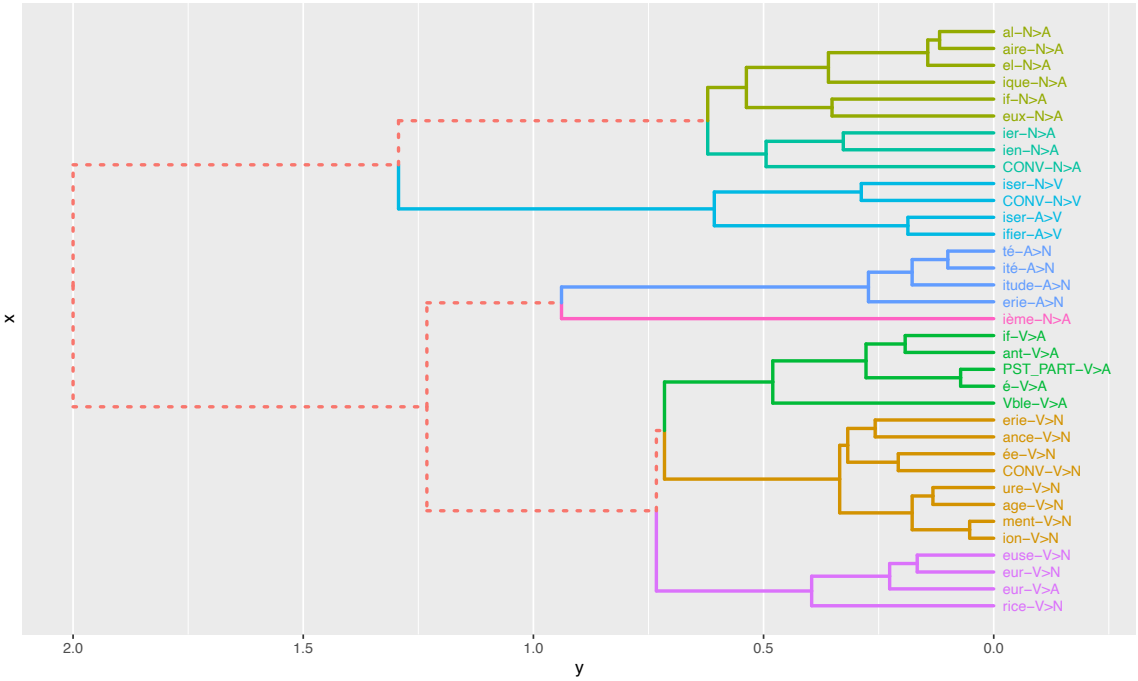


Figure 1: Hierarchical clustering

similarity between two trees. This is defined by the following formula, where clusters is a function from trees to the set of clusters it contains, i.e., the set of leaf node labels that are the yield of a non-leaf node in the tree.

$$\text{sim}(T, T') = \frac{2 \times |\text{clusters}(T) \cap \text{clusters}(T')|}{|\text{clusters}(T)| + |\text{clusters}(T')|}$$

Table 2 summarizes the comparison. Agreement across experts on the exact classification is quite variable, with some experts being more consensual than others. Importantly for our purposes, the vector-based classification does not stand out among the other classification.

Source	E1	E2	E3	E4	E5	E6	E7	**vectors**
Similarity	0.636	0.683	0.597	0.693	0.699	0.567	0.584	**0.616**

Table 2: Average similarity between each classivication tree and all other trees

5 Conclusions

The overall conclusion then is that comparisons of average difference vectors do an excellent job of capturing similarities and differences between derivational processes. It is worth emphasizing that this is done without any information on the forms of words being related, nor explicit information about part of speech: distributional information is all they have. We argue that this provides strong support for the view that difference vectors can reliably be used to explore morphosemantic relations in derivational paradigms.

References

Bonami, Olivier & Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio* 17(2). 173–195.

Bonami, Olivier & Juliette Thuilier. 2019. A statistical approach to affix rivalry: French *-iser* and *-ifier*. *Word Structure* 12(1). 4–41.

Hathout, Nabil & Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5). 125–168.

Koehl, Aurore. 2012. *La construction morphologique des noms désadjectivaux suffixés en français*: Université de Lorraine dissertation.

Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica* 20. 1–31.

Marelli, Marco & Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review* 122. 485–515.

Mickus, Timothee, Olivier Bonami & Denis Paperno. 2019. Distributional effects of gender contrasts across categories. In *Proceedings of the 2nd meeting of the society for computation in linguistics*, .

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`.

Řehůřek, Radim. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the lrec 2010 workshop on new challenges for nlp frameworks*, 45–50.

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.

Sokal, Robert R. & Charles D. Michener. 1958. A statistical method for evaluating systematic relationships. *The University of Kansas Scientific Bulletin* 38(1409–1438).

Strnadová, Jana. 2014. *Les réseaux adjectivaux: Sur la grammaire des adjectifs dénominaux en français*: Université Paris Diderot et Univerzita Karlova V Praze dissertation.

Tribout, Delphine. 2010. *Les conversions de nom à verbe et de verbe à nom en français*: Université Paris Diderot dissertation.

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

Paradigms in non-morphemic word formation

Camiel Hamans
Amsterdam University/Adam Mickiewicz
University Poznań

1 Introduction

In this paper it will be shown how essential paradigms are for non-morphemic word formation. Following Van Marle (1994), the term paradigm will be used here in an informal way referring to series of linguistic forms that are related. The relationship in the series presented here is of a formal nature. However, they appear to have a certain psychological reality (cf. Fernández-Domínguez et al. 2020).

Four processes of non-morphemic word formation will be discussed: at the one hand hypocoristic formation and embellished clipping, on the other libfixing and blending as the basis for paradigmatic productivity. In all four processes the notion morpheme does not play a role or a minor role only, therefore these processes are called non-morphemic.
The data that will be presented are taken from Dutch and English.

2 Hypocoristics

In (1) examples are presented of hypocoristic formation in English and Dutch.

(1a) English		(1b) Dutch	
Andy <	Andrew	Gerrie <	Gerard (Ger)
Debbie <	Deborah	Japie <	Jacob (Jaap)
Monty <	Montgomery	Jannie <	Johanna (Jan?)

In all these examples clipping occurs first; in English standard back clipping (cf. Lappe 2007) , in Dutch a more complicated form of clipping may occur, combining the first part of the original name and the final consonant or the final segments of the full name (*Johannes* > *Johan* > *Jan*).
In Dutch most clipped forms also occur as first names. It is remarkable that next to the disyllabic female first name *Jannie* only the monosyllabic male form *Jan* can be found. *Gerrie* can be a male and a female first name, however, the clipped form tout court *Ger* refers to males only.
The word formation process of the hypocoristic forms in (1) are the result of clipping followed by suffixation. Clipping is the non-morphemic word formation process here, whereas suffixation makes use of a hypocoristic suffix *-y* or *-ie* respectively, which of course is a morpheme. So far, the notion paradigm does not yet play a role. That changes when we look at the next process in sections 3 and 4.

3 Embellished clippings

Embellished clippings (Bauer and Huddleston 2002: 1632) are words formed of a clipped word followed by a suffix and thus resemble hypocoristics structurally. For examples see (2)

(2a) English Embellished clippings with an attested independent clipped form			(2b) English Embellished clippings without an attested clipped form		
sissy	< sister	(sis)	granny	< grandmother	(*gran)
ciggy	< cigarette	(cig)	(n)uncie/nunky	< uncle	(*unc)
bevvy	< beverage	(bev)	hanky	< handkerchief	(*hank)

The data under (2a) may be explained as a special form of diminutive formation. If so, one can equate this form of an embellished clipping with hypocoristic formation under (1). In addition,

one has to accept clipped forms such *sig*, *cig*, *bev* and *bro* as independent lexical items. However, this is not possible with the examples under (2b).

In Dutch one finds a similar process in informal registers (see below, 3). However, there are hardly independent clipped lexical items which could function as a base form for embellishment, which makes an explanation in terms of a quasi-diminutive or hypocoristic formation process very unlikely, even tough the suffix *-ie* also is the diminutive suffix in this informal register of Dutch.

(3) Dutch (informal register)
Embellished clippings

jochie	< jongen	'boy'	(joch)
makkie	< gemakkelijk	'something which can easily be done'	(*mak)
gympie	< gymschoen	'gym shoe, sneakers'	(? gym; different meaning)

Both English and Dutch also exhibit a process of pseudo-embellished clippings (4).

(4a) English

chappie	< chap
blokie	< bloke
foody	< food

(4b) Dutch

jonkie	<jong	'young person'
nakie	< naakt	'standing naked'
zopie	< sope/soep	'beverage'

It is remarkable that all the examples presented so far, thus (1) - (4), are disyllabic and trochaic. It seems to be the syllabic and prosodic structure that determines the outcome of the processes of hypocoristic formation and of embellishment.

The role of the paradigm is still not very clear. However, one may guess that series of diminutives, especially in Dutch where diminutives are highly frequent, promoted the use of the diminutive suffix as a hypocoristic suffix and subsequently also as a suffix for embellishment and pseudo-embellishment.

4 The suffix *-o*

Another embellishment suffix, however, demonstrates clearly how important a series or paradigm of related forms is for the origin of the suffix (cf. Hamans 2012, 2018 & 2020). This suffix is *-o* as in (5) and (6):

(5a) English
Embellished clippings

afro	< African
lesbo	< lesbian
relo	< relative (N)

(5b) Dutch
Embellished clippings

alto	< alternatief	'alternative person'
depro	< depressief	'depressed person'
sago	< chagrijnig	'cantankerous person'

(6a) English
Pseudo-embellished clippings

sicko	< sick
kiddo	< kid
creepo	< creep

(6b) Dutch
Pseudo-embellished clippings

lullo	< lul ('prick')	'dumb person'
duffo	< duf ('dull')	'dull person'
jazzo	< jazz ('jazz')	'fan of old style jazz'

Here, no diminutive suffix or another suffix can be found that promotes *-o*. This new suffix originates as final ending in series of neoclassical clipped forms, such as:

(7)a English

psycho	< psychopath
homo	< homosexual
dipso	< dipsomaniac

(7b) Dutch

aso	< asocial	'antisocial person'
impo	< impotent	'impotent man'
pedo	< pedofiel	'pedophile'

The series in (7) form paradigms with forms ending in final *-o* which originates from the long base forms and which gets assigned a psychologically real connotation [final part of a clipped, disyllabic, trochaic form, informal and with a negative meaning]. The naïve language user subsequently uses this final *-o* as a suffix as in (5) and (6).

5 Libfixing

Another word formation process in which the morpheme does not play a role is libfixing. Again the role of the paradigm is essential here when it comes to productivity. The term 'libfix' goes back to Zwicky (2010). Libfixes are parts of words that operate as if they are affixes. They are highly productive and very frequent in current English (Norde and Sippach 2019). An example is the word part *-preneur* as in (8) which starts with the existing form *entrepreneur*

(8) entrepreneur
ecopreneur
biopreneur
soloopreneur
etc.

This part *-preneur* could be 'liberated' since the naïve language user recognized the part *entre* in other French loanwords in English as in the paradigm in (9)

(9) entrepot
entremets
entresol
entrecote

Thus, the naïve native speaker of English reanalyzed *entrepreneur* as consisting of two parts *entre* and *preneur*, of which the last was given a sort of affix status.

6 Reanalysis

Libfixing does not necessarily arise from paradigmatic comparison as in (9), it may also be the result of simple reanalysis of opaque forms, as in (10)

(10) Armageddon
snowmageddon
carmageddon
heatmageddon
Obamageddon

It will be clear that a libfix such as *-mageddon* only can become productive when it is used in a series of related forms, thus in a paradigm.

7 Blends

Just as libfixes can form the basis for a productive non-morphemic process of word formation, blends can. This is not the place to discuss blend formation itself. It suffices here to claim that the second, right hand, source word of blends usually provides the head of the blend and consequently determines the outcome in terms of syllabic structure and prosody (see for an overview Renner et al 2012 and Bauer et al. 2013). What counts here is what may happen after the blend is coined. The moment a blend is used to produce a second similar form, the final part

of the blend, the head, may become as productive as the libfixes discussed above, see (11) and (12)

(11) stay + vacation →	staycation	(12) mock + documentary →	mockumentary
	daycation		shockumnetary
	gaycation		socumentary
	kidcation		dogumentary

The blends *staycation* and *mockumentary*, or similar forms, must have been reanalyzed with the result that the parts *-cation* and *-umentary* could be used in the same way as the libfixes, discussed before. The difference between these two processes is that in the case of libfixing the starting point can be paradigmatic comparison followed by reanalysis or reanalysis only, whereas in the case of blends first blend formation has to occur before reanalysis may take place. As will be clear, there must be a certain mass of corresponding or related forms before the final part of a blend can develop into a productive affix-like phenomenon. That mass is a paradigm and this mass can lead to paradigmatic productivity.

References

Bauer, Laurie and Rodney Huddleston. (2002). Lexical word-formation. Rodney Huddleston and Geoffrey K. Pullum (eds.). *The Cambridge Grammar of the English Language.* Cambridge: CUP: 1621-1722.

Bauer, Laurie, Rochelle Lieber and Ingo Plag (2013). *The Oxford Refence Guide to English Morphology.* Oxford: OUP.

Fernández-Domínguez, Jesús, Alexandra Bagasheva and Christina Lara-Clares (2020). What Paradigms and What For? Jesús Fernández-Domínguez, Alexandra Bagasheva and Christina Lara-Clares (eds.) *Paradigmatic Relations in Word Formation.* Leiden/Boston: Brill.

Hamans, Camiel (2012). From Prof to Provo: Some observations on Dutch Clippings. Bert Botma and Roland Noske (eds.) *Phonological Explorations: Empirical, Theoretical and Diachronic Issues.* Berlin/Boston: de Gruyter.

Hamans Camiel (2018). Between *Abi* and *Propjes:* Some remarks about Clipping in English, German, Dutch and Swedish. *SKASE Journal of Theoretical Linguistics* 15, 2: 24-59.

Hamans, Camiel (2020). How an 'Italian' suffix became productive in Germanic languages. Pius ten Hacken and Renáta Panocová (eds.). *The interaction of borrowing and word formation.* Edinburgh: Edinburgh University Press.

Lappe, Sabine (2007). *English Prosodic Morphology.* Dordrecht: Springer.

Marle, Jaap van (1994). Paradigms. Ronald E. Asher (ed.) *Encyclopedia of language and linguistics.* Oxford: Pergamon, 6: 2927-2930.

Norde, Muriel and Sarah Sippach (2019). Nerdalicious scientainment: A network analysis of English libfixes. *Word Structure* 12: 353-384.

Renner, Vincent, François Maniez and Pierre J.L. Arnaud (eds) (2012). *Cross-Disciplinary Perspectives on Lexical Blending.* Berlin/Boston: De Gruyter Mouton.

Zwicky, Arnold (2010) Libfixes https://arnoldzwicky.org/2010/01/23/libfixes/

French ethnonyms, toponyms, demonyms and their paradigmatic organization

Nabil Hathout
CLLE, CNRS &
Université Toulouse Jean Jaurès

Fiammetta Namer
Université de Lorraine &
ATILF, CNRS

Michel Roché
Université Toulouse Jean Jaurès

Word formation of **demonyms**, i.e. names of people from a particular geographical entity, e.g. *parisien*$_{\text{DEMON}}$ 'Parisian', *français*$_{\text{DEMON}}$ 'French person') has been the subject of many studies (for French, cf. Roché (2008); Molinier (2018)) and is a central issue in derivational morphology, especially in works that highlight the concept of paradigm (Booij, 1997, 2010; Booij & Masini, 2015; Boyé & Schalchli, 2019). In line with these works, we propose a communication whose main contribution is to consider globally the **ethnonyms** (i.e. names of humans who belong to a people or an ethnic group, regardless of any territorial anchorage, e.g. *Touareg*$_{\text{ETHN}}$), the **toponyms** (i.e. names of towns, e.g. *Paris*$_{\text{TOWN}}$ and names of countries, e.g. *France*$_{\text{COUNTRY}}$), their **demonyms**, and the **relational adjectives** (RA) that correspond to all these nouns in French (*parisien*$_{\text{RA-TOWN/RA-DEMON}}$, *français*$_{\text{RA-COUNTRY/RA-DEMON}}$, *touareg*$_{\text{RA-ETHN}}$), to the extent that they belong to the same derivational family. The result is a theoretical network of 10 cells (1XX), if we count as one the masculine and feminine forms of ethnonyms and demonyms, even if it is actually very unusual for the same family to include all the five names:

(1)

	Ethnic	**Country**	**Town**
Toponym		Toponym: name of country	Toponym: name of town
Toponym relational adjective		RA-country	RA-town
Demonym	Ethnonym: Ethnic group name	Demonym: country inhabitant	Demonym: town inhabitant
Demonym relational adjective	RA-ethnonym	RA-country-demonym	RA-town-demonym

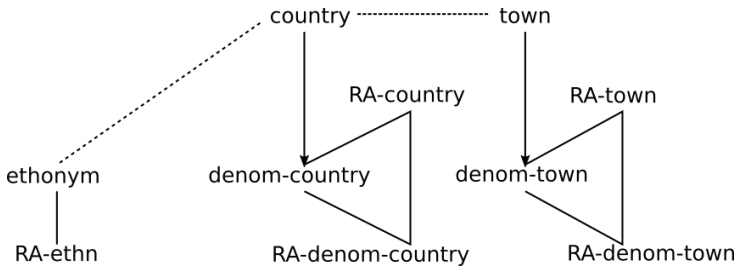
We show that these lexemes form a very coherent group semantically and respond to the same onomasiological needs even if they are very diverse at the morphological (i.e. formal) level, for reasons both that are linguistic (multiple inheritance and borrowing) and referential (historical contingencies, political structures). For example, in the *Kurde*$_{\text{ETHN}}$ 'Kurdish' *Kurdistan*$_{\text{COUNTRY}}$ family, the fact that Kurdistan does not exist as a state is manifested by the absence of a demonym (*kurdistanais* is the RA of the country name) whereas in the family of *Bolivie*$_{\text{COUNTRY}}$, which is a state, the demonym is *Bolivien* 'Bolivian' while the population ethnicity is expressed by other words (e.g. *Quechua*, *Aymara*...). More generally, the eymological relations between these words are diverse: a country name (*Algérie* 'Algeria') can be coined on a town name (*Alger* 'Algiers'), a town name (*Brasilia*) on a country name (*Brésil* 'Brazil'), or both can be identical (*Luxembourg*) or closely related (*Angers*$_{\text{TOWN}}$ / *Anjou*$_{\text{COUNTRY}}$). Historically, ethnonyms usually come first and country names are often coined on the ethnonym (*Turc*$_{\text{ETHN}}$ 'Turk' and *Turquie*$_{\text{COUNTRY}}$ 'Turkey', *Thaï*$_{\text{ETHN}}$ and *Thaïlande*$_{\text{COUNTRY}}$ 'Thailand'). Demonyms (*Thaïlandais* 'inhabitant of Thailand', *Algérois*, 'inhabitant of Algiers') are always present with the country names (*Thaïlande*) and city names (*Alger*), which they logically derive from since conceptually, they only exist in relation to them. The configuration represented by *Thaï*$_{\text{ETHN}}$, *Thaïlande*$_{\text{COUNTRY}}$, *Thaïlandais*$_{\text{DEMON}}$, where the three elements have different forms, is actually quite rare. More often, when there is already an ethnonym in the family, it is also used as demonym

(*Turc* 'Turk', *Russe* 'Russian'), and, in the absence of a specific ethnic denomination, it is the demonym that takes its place (so the demonym *Français* serves as an ethnonym and means "pure French" as opposed to ethnic minorities).

Some of these name and adjective reuses are systematic. The RA corresponding to an ethnonym almost always has the same form as the ethnonym itself, e.g. in *(masque) hopi* 'Hopy (mask)'. The RA corresponding to the demonym usually has the same form as the demonym itself (*français*$_{\text{RA-DEMON}}$ 'of French people', *Français*$_{\text{DEMON}}$) and the RA corresponding to the name of country also has the same form as the demonym, even when the name of country is itself derived by suffixation (*Russie* 'Russia', *Russe*$_{\text{DEMON}}$, *russe*$_{\text{RA-COUNTRY / RA-DEMON}}$ 'of Russia, of Russians'). In families yielded by a name of town, the reuse even can go further as with *beaujolais* and *Beaujolais*, respectively RA and demonym of the town of *Beaujeu*, which are reused as RA and demonym corresponding to the country of *Beaujolais*.

Diagram (2) summarizes these regularities (the bold lines mark a formal identity, the arrows a systematic dependence and the dotted lines a possible motivation). Many atypical configurations further complicate the morphological analysis, e.g. *Corse*$_{\text{DEMON}}$ 'Corsican' / *Corse*$_{\text{COUNTRY}}$ 'Corsica', *Hongrois*$_{\text{DEMON}}$ 'Hungarian' / *Hongrie*$_{\text{COUNTRY}}$ 'Hungary', *Argentin*$_{\text{DEMON}}$ 'Argentinean' / *Argentine*$_{\text{COUNTRY}}$ 'Argentina', *Poitevin*$_{\text{DEMON}}$ as demonym and RA of both *Poitiers*$_{\text{TOWN}}$ and *Poitou*$_{\text{COUNTRY}}$, etc. As it has been already demonstrated (e.g. in Booij (2010)), the analysis of these uncommon families is hardly possible if each element is considered in isolation in a one-to-one base-derivative relationship, whatever the adopted framework (morpheme segmentation or lexeme-based approach). Things become clear when the various elements are considered in the network they form with the other members of their derivational families.

(2)



These examples show that the semantic and formal organizations of these networks belong to independent dimensions and must be considered separately, since it is not possible to predict neither the presence nor the form of one element from another element of the family. The data we present illustrates several of the fundamental principles of the paradigmatic organization of the derivational lexicon (cf. Van Marle (1985); Bochner (1993); Stump (1991); Bauer (1997); Booij (2008), i.a.; for an overview, Štekauer (2014)): the structure of the derivational paradigms is defined by semantic regularities (determined by onomasiological needs, see Štekauer (2005)); there are natural sub-paradigms within these paradigms which bring together, for example, country names, their demonyms and their RAs (on this point, see Bochner (1993)); the paradigmatic structure defined by semantic properties allows the superposition of formally heterogeneous derivational families.

In order to capture both the semantic homogeneity of these families and their formal diversity, we need a model able to grasp paradigmatic regularities even when blurred by form-meaning discrepancies. In order to achieve this goal, the ethnic families are represented in a framework that combines the main contribution of classical lexeme-based morphology (independent formal, categorical and semantic description levels) and the paradigmatic organization of the lexicon (where all the members of a derivational family are accessible). This framework, called ParaDis "Paradigms vs Discrepancies" (Hathout & Namer, 2018) is made up of autonomous paradigmatic networks (at least one formal, and one semantic) connected to a corresponding morphological paradigm: thanks to the independence of these different paradigms,

local constraints and regularities (for instance, the formal identity between toponyms and the corresponding RAs) are expressed locally, within the appropriate paradigm, regardless of what happens elsewhere in the system.

In ParaDis, the families of French ethnonyms, toponyms, demonyms and their respective related RA are represented by several formal networks connected to a single semantic paradigm. In this way, ParaDis enables the description of what these paradigmatic organizations have in common and of where they differ.

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.

Booij, Geert. 2008. Paradigmatic morphology. In Bernard Fradin (ed.), *La raison morphologique. hommage à la mémoire de danielle corbin*, vol. 27, 29–38. Amsterdam / Philadelphia: John Benjamins.

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.

Booij, Geert E. 1997. Autonomous morphology and paradigmatic relations. In Geert E. Booij & Jaap van Marle (eds.), *Yearbook of morphology 1996*, 35–53. Dordrecht: Kluwer.

Boyé, Gilles & Gauvain Schalchli. 2019. Realistic data and paradigms: The Paradigms Cell Finding Problem. *Morphology* 29(2). 199–248.

Hathout, Nabil & Fiammetta Namer. 2018. La parasynthèse à travers les modèles : des RCL au ParaDis. In Olivier Bonami, Gilles Boyé, Georgette Dal, Hélène Giraudo & Fiammetta Namer (eds.), *The lexeme in descriptive and theorical morphology*, 365–399. Langage Sciences Press.

Molinier, Christian. 2018. Éthniques et gentilés. formes et propriétés respectives. *Le Français Moderne* 2018(2).

Roché, Michel. 2008. Structuration du lexique et principe d'économie : Le cas des ethniques. In Jacques Durand, Benoît Habert & Bernard Laks (eds.), *Actes du congrès mondial de linguistique française (cmlf-2008)*, 1571–1585. Paris: ILF.

Štekauer, Pavol. 2005. Onomasiological approach to word-formation. In Pavol Štekauer & Rochelle Lieber (eds.), *Handbook of word-formation*, 207–232. Dordrecht: Springer.

Stump, Gregory T. 1991. A paradigm-based theory of morphosemantic mismatches. *Language* 675–725.

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

Paradigmatic nature of Dokulil's onomasiological theory of word-formation

Petr Kos
University of South Bohemia
Czech Republic

The interest in paradigmatic approaches to word-formation dates back to the publication of Van Marle (1984). However, as early as in 1962, Miloš Dokulil published a theory of word-formation (Dokulil 1962), which clearly exhibits the paradigmatic nature of the process of coining new lexemes.

The aim of this contribution is to introduce Dokulil's onomasiological theory, demonstrate its paradigmatic nature, show how he set the paradigmatic approach into a wider conception of formation of lexemes within an onomasiological theory, and present the implications that appear relevant to the current discussions on the role of paradigms in word-formation.

Dokulil's theory is anchored in the functional-structural conception of the Prague school. He focuses on the elaboration of a method of synchronic description of a word-formation system, the aim of which is to reveal the mutual relations and connections of individual components and elements of this system in contemporary language.

In contrast to the traditional, genetic, concept of word formation of that time, there is a new emphasis on language as a system and on revealing the systemic nature of linguistic phenomena. So, he not only focuses on the dynamic (genetic) aspect of word-formation but also on the static aspect of 'word-formedness', the latter referring to the existing lexicon of complex words in the sense that it has a decisive impact on the former aspect. So, he aligns the genetic aspect with the aspect of the structure of the system and shows how the two aspects are complementary and firmly intertwined. "The mutual relationship of results and processes, which themselves become conditions for new processes, constitutes the specific character of the theory of word formation. Only a unification of these two aspects can account for the dialectic relation between word-formative processes and the functioning of word-formative structures" (Dokulil 1994: 130).

The dynamic process of word-formation is not only understood as an act of forming new lexemes (i.e., production, which occurs rather rarely) but also the reproduction of these processes in speech, together with their interpretation, which Dokulil sees as a ubiquitous phenomenon. It is thus in the interest of a synchronic theory to focus on those units which are still felt to be instantiations of the word-formative schemas of the language and can thus be repeatedly reproduced and interpreted in speech. If a word is used only rarely and speakers cannot find it in their memory, they have to create it on the basis of the relevant word-formation matrices. "Analogically, in perception one often does not connect such a [n unknown] word immediately with its (instantaneous) discourse meaning, but one deciphers it on the basis of its word-formative pattern - thus one attains the structural (word-formative) meaning of the word, starting from which one identifies its instantaneous lexical meaning on the basis of the situation and of the context" (Dokulil 1994: 131).

The key to the paradigmatic description of word-formation is the mutual relationship between the lexical and structural meanings. The identity and structure of cells in word-formation paradigms are given by the structural meanings, which are abstractions over the lexical meanings of the existing lexemes. The creation of lexical meanings, nevertheless, begins in the very process of naming by mapping a specific onomasiological structure on some

of the possible structural meanings, which are more general. Consequently, the lexical meaning should not be understood as a secondary idiosyncratic shift of the structural meaning, but it is a direct reflection of the onomasiological structure. Moreover, the existing lexical meanings are a source from which the structural meaning is abstracted.

Another key distinction made by Dokulil is the one between the three major onomasiological categories, namely the modificational (adding a modifying element to the contents of the given concept, such as diminutiveness or change of gender), the transpositional (the change of word class with no change of meaning), and the mutational (naming in the narrowest sense, providing names for (new) concepts in the extra-linguistic reality). In terms of their paradigmatic nature, it is the former two that resemble most the inflectional paradigms (see Kos 2020).

As has been mentioned above, the genetic aspect of word-formation is complementary to the static system of motivated complex words. The genesis of a word starts with the conceptualization of the extra-linguistic reality, so speakers do not name the reality itself but rather its reflection in their minds. The named concept is first classified into an existing category, which becomes the onomasiological base. The salient feature that distinguishes the named concept from other members of the category becomes the onomasiological mark. Together the onomasiological base and the onomasiological mark comprise the onomasiological structure, which can be seen as an interface between the conceptualization and the actual process of naming. The linguistic coding stage starts with the linguistic expression of the onomasiological mark, and this form is matched with a suitable word-formation type. The word-formation type is a result of the abstraction over a series of words with a homogeneous internal structure which have concrete lexical meanings. It is thus a generalization of the semantics of this series and of the mutual relationship of their components. The word-formation type is "regarded as a unity of onomasiological structure of a series of words (i.e. a unity of the structural meaning regarded as a whole and a unity of the mutual relation of the component parts of this structure), a unity of the lexico-grammatical category of the derivational base and a unity of the formative element" (Dokulil 1994: 139). In other words, in derivation, which is the predominant word-formation process in Czech, the word-formation type is the unity of the semantic relationship between the components, of the grammatical category of the derivational base, and of the form of the suffix. However, as Dokulil points out, this abstraction may also occur on other levels, which are either more specific or more general than the word-formation type. The formation of a word as a new lexical unit is concluded by the grammatical formation of this word, i.e., by matching it with a specific inflectional paradigm within the given word-class. Also, the newly formed lexeme becomes part of the system and contributes to the formation of other lexical units.

Dokulil's conception of the word-formation type is thus highly reminiscent of Booij's (2010) conception of constructions within Construction Morphology. While Booij's approach is more elaborate with the formal expression of the constructions, Dokulil's approach is more comprehensive in setting the paradigmatic approach within a broader conception of coining lexical units.

The general categories as "agent N" or "instrument N", which typically appear in discussions on derivational paradigms (see, among many others, Bonami & Strnadová 2019, Fradin 2020), are seen as instances of a superordinate level to the word-formation type, that of word-formation category. As a more abstract level, lacking the unity of the suffix and some specific semantic properties arising from the abstraction of the paradigmatic series, it can be seen as a sum of a number of word-formation types sharing the same, e.g., agentive, meaning.

In Dokulil's conception the word-formation category, however, rarely serves as a model for new words.

Another aspect that deserves some discussion is Dokulil's notion of parallel motivation. This is the situation when a word is motivated by more than one other word, e.g., the Czech *kovárna* 'forge' as a workshop for blacksmiths is motivated both by the activity *kovat* 'to forge' and the typical user *kovář* 'blacksmith', thus forming a paradigmatic system (in current terminology) in which all members are related through motivation.

In summary, the presentation will mostly deal with the implications that arise from the features briefly suggested above, i.e., the setting of paradigms within a broader conception of the process of coining new words, the relation of paradigms to conceptualization, the notion of the word-formation type, the distinction between the main onomasiological categories, the distinction between the structural and lexical meanings, and the role of parallel motivation. The presentation will be accompanied by numerous examples taken from Czech.

References

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

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

Dokulil, Miloš. 1962. *Tvoření slov v češtině*. Praha: Nakladatelství ČAV.

Dokulil, Miloš. 1994. The Prague School's Theoretical and Methodological Contribution to "Word Formation" (Derivology). In Philip A. Luelsdorff (ed.), *The Prague School of Structural and Functional Linguistics*. 123–161. Amsterdam/Philadelphia: John Benjamins.

Fradin, Bernard. 2020. Characterizing Derivational Paradigms. In Jesús Fernández-Domínguez, Alexandra Bagasheva & Cristina Lara Clares (eds.), *Paradigmatic relations in word formation*. 49–84. Leiden: Brill.

Kos, Petr. 2020. The level of paradigmaticity within derivational networks. In Jesús Fernández-Domínguez, Alexandra Bagasheva & Cristina Lara Clares (eds.), *Paradigmatic relations in word formation*. 85–99. Leiden: Brill.

Van Marle, Jaap. 1984. *On the paradigmatic dimension of morphological creativity*. Dordrecht: Foris.

Towards uniformity within derivational paradigms: Evidence from Hebrew

Lior Laks
Bar-Ilan University

This study examines derivational relations in Hebrew. It addresses two cases of doublet formation (see Kroch 1989; Thornton 2012; Fradin 2016; Aronoff 2017, among others), arguing that language change is motivated by paradigm uniformity in derivation.

The Hebrew verbal system consists of configurations called *patterns*. The patterns indicate the prosodic structure of verbs, their vocalic patterns and their affixes (if any). For example, the verbs *siper* 'tell' and *hidpis* 'print' are formed in *CiCeC*, and the verbs *hitraxec* 'wash oneslef' and *hitragel* 'get used to' are in *hitCaCeC*. The phonological shape of a verb is essential for determining the shape of other forms in the inflectional paradigm (Berman 1978; Schwarzwald 1981; Bolozky 1978; Ravid 1990; Bat-El 1994, Aronoff 1994). The formation of the same root in different patterns results in two (or more) different verbs, where the semantic relations between them can be of different degrees of transparency. The relations between verbs in different patterns are derivational and are manifested mainly in transitivity and voice alternations. Verbs that are formed in certain patterns share some typical semantic and syntactic features. For example, verbs in *CiCeC* are usually transitive verbs (*ximem* 'warmed X'), and *hitCaCeC* verbs are mostly intransitive (*hitxamem* 'warmed up'). It is crucial to note that such distinctions reflect strong tendencies rather than a clear cut distribution. The patterns system is considered a system of derivational relations, where each patetrn has its own inflectional classes (see Aronoff 1994).

The paradigmatic approach has been gaining a growing in derivation, in addition to its role in inflection. Many studies demonstrate the importance of paradigms in word formation (see Bauer 1997; Pounder 2000; Beecher 2004; Booij & Lieber 2004; Hathouth & Namer 2014; Štekauer 2014; Blevins 2016; Bonami& Strnadová, among others). The current study provides further evidence to the role of derivational paradigms in word formation and in morphological change. Specifically, I will show that doublet formation is triggered by structural and semantic transparency within paradigms. This will be demonstrated with respect to two case studies. Data collection in both cases studies is based on online web-searches, to be discussed in more details in the talk.

1 Doublet formation of adjective-related verbs

Some verbs that are semantically related to adjectives have doublets (1).

(1) a. **hitbayaši** ledaber ita panim mul panim (https://www.askpeople.co.il/question/32679)
'I was embarassed/shy to speak to her face to face'

b. bexol-ofen **hitbayšanti** ledaber ita (https://www.fxp.co.il/showthread.php?t=12544723)
'anyway I was embarrassed/shy to speak to her'

hitbayašti (2a) and *hitbayšanti* (2b) are the 1st person sg. forms of the verbs *hitbayeš* and *hitbayšen*. Both verbs share the meaning 'be embarrassed', and can be used in similar contexts. The verbs are also phonologically related; they are both formed in *hitCaCeC* and share the root consonants *b-y-š*. However, *hitbayšen* has a quadrilateral root *b-y-š-n*, as it is derived directly from the adjective *bayšan* 'shy'. This adjective is formed in the *CaCCan* pattern, which is typical for adjective formation, e.g. *xamdan* 'greedy'. The consonant *n*, which is not part of the root *b-y-š*, becomes part of a new root in the formarion of *hitbayšen*. Verbs like *hitbayšen* are not accepted by all Hebrew speakers, and most of them are not documented in dictionaries. However, web searches reveal that their formation becomes more and more productive. Similarly, the verbs *hitʕacel* (2a) and *hitʕaclen* (2b) (infinitive forms), are both related to *ʕaclan* 'lazy', while only *hitʕaclen* is derived directly from it.

(2) a. hexlateti **lehitʕacel** ve-pašut lecatet mi-wikipedya
'I decided to be lazy and simply quote from Wikipedia'
(https://hwzone.co.il/community/topic/275566-%D7%9C%D7%9E%D7%94-%D7%91%D7%97%D7%95%D7%A8%D7%A3-%D7%A7%D7%A8/)

b. hexlateti **lehitʕaclen** ve-lehišaer ba-taxana ha-krova
'I decided to be lazy and stay in the next station'
(http://israblog.nana10.co.il/blogread.asp?blog=64230&blogcode=12711065)

Why are such doublets formed? I argue that this is motivated by structural transparency between items that are part of a derivational paradigm. The change from *hitbayeš* to *hitbayšen* results in more structural transparent relation between the verb 'become embarrassed' and the adjective *bayšan* 'shy' that is related to it. A paradigm like *bayšan-hitbayšen* is structurally more transparent than a paradigm like *bayšan-hitbayeš*, as the transition between the two words in the former paradigm maintains all the consonants of the adjective regardless of whether they are part of the original root. The morphological mechanism aims at maintaining as much elements as possible, and as a result the related forms are more faithful to each other (McCarthy & Prince 1990; Bat-El 1994, 2017; Ussishkin 2005). Accordingly, the paradigm *bayšan-hitbayeš* is less transparent. Such cases also lend support to a word-based approach of word formation (Aronoff 1976, 2007; Blevins 2006, 2016), according to which, the lexicon consists of existing words and word formation relies on the relation between words.

2 The transitive ~ intransitive alternation: unifying the system

The derivational relations between verbs is manifested mainly transitivity alternations. One such alternations is the causative-inchoative alternation (Haspelmath 1987, 1993, among many others). In most cases, causative and inchoative verbs are morphologically distinct. In (3a), the transitive is formed in *CiCeC* and the intransitive is in *hitCaCeC*. In (3b) and (3c), the transitive verbs are in *hiCCiC* and the inchoative ones are in *hitCeCaC* (3b) and *niCCaC* (3c).

Some patterns are more typical of transitive verbs (*CiCeC*, *hiCCiC*), while others are more typical of intransitive verbs (*hitCaCeC*, *niCCaC*).

(3) Hebrew transitive ~ intransitive alternations

Transitive verb	**Pattern**	**Intransitive verb**	**Pattern**
yibeš 'dry X'	CiCeC	hityabeš 'become dry'	hitCaCeC
hirgiz 'make X upset'	hiCCiC	hitragez 'become upset'	hitCaCeC
hirdim 'put X to sleep'	hiCCiC	nirdam 'fall asleep'	niCCaC

In contrast, there is a group of labile verbs in *hiCCiC* that are ambiguous with respect to transitivity (Rosen 1956, Borer 1991, Lev 2016). For example, *hivri* denotes both 'make healthy' and 'become healthy'. The formation of intransitive verbs in *hiCCiC* is not productive and is considered irregular. Their existence seems to stand in contradiction to the morphological features of Hebrew transitive-intransitive paradigms. Indeed, there is a tendency to mend this irregularity via doublet formation. Some *hiCCiC* intransitive verbs have doublets in *hitCaCeC*. For example, *hexvir* denotes both 'make X pale' (4a) and 'become pale' (4b), while *hitxaver* (4c) is only intransitive. The intransitive doublets *hexvir* and *hitxaver* (1st person) surface in similar contexts. The change into *hitCaCeC* never occurs for the transitive meaning, as *hiCCiC* is typical for the formation of transitive verbs (Borer 1991).

(4) a. ha-mar?e ha-xadaš **hexvir** et paneha (http://10tv.nana10.co.il/Article/?ArticleID=790264)
'the new look made her face pale'

b. **hexvarti** ve-hitalafti, nlkaxti le-miyun
'I became pale and fainted, I was taken to ER' (https://www.tapuz.co.il/forums/viewmsg/1493/102140101)

c. **hitxavarti** ve-hitalafti le-kama šniyot (https://www.doctors.co.il/forum-1469/message-124838/)
'I became pale and fainted for a few seconds'

Similarly to the case in 1, the morphological change brings about uniformity within derivational paradigms. In contrast to the case in 1, which is motivated by **structural transparency** between derivationally related forms, the change here is motivated by **semantic transparency**. The morphological mechanism takes verbs, which are 'misbehaving' with respect to transitivity, and forms doublets in a pattern, which is more typical for their transitivity value. As a result, derivational paradigms become more uniform in the sense that more verbs are marked consistently with respect to transitivity.

The two cases demonstrate the central role of paradigms in derivation. In both cases, the morphological mechanism aims at creating more transparent and regular form-meaning relations within derivational paradigms. In 1, doublet formation takes place with respect to the root, while the patterns remains the same. In 2, the root remains the same and the pattern changes. In both cases, the morphological mechanism has to examine not only the word that undergoes change, but also its structural and semantic relations with other words within the relevant derivational paradigm. This highlights the necessity of paradigm accessibility in derivation and its role in morphological variation and change.

References

Aronoff, Mark. (1976). *Word Formation in Generative Grammar.* Cambridge: MIT Press.

Aronoff, Mark. (1994). *Morphology by Itself*. Cambridge: MIT Press.

Aronoff, Mark. (2007). In the beginning was the word. *Language* 83. 803-830.

Aronoff, Mark. (2017) (submitted). Competitors and alternants. To appear in a volume of *selected papers from IMM 17*, Vienna, 2016.

Bat-El, Outi. (1994). Stem modification and cluster transfer in Modern Hebrew. *Natural Language and Linguistic Theory* 12: 572-596.

Bat-El, Outi. (2017). Word-based items-and processes (WoBIP): Evidence from Hebrew morphology. In C. Bowern, L. Horn, & R. Zanuttini (eds.), *On Looking into Words (and beyond)*, 115-135. Berlin: Language Sciences Press.

Bauer, Laurie. (1997). Derivational paradigms. In G. Booij & J. van Marle (eds.), *Yearbook of Morphology 1996*. 243-56. Dordrecht: Kluwer.

Berman, Ruth A. (1978). *Modern Hebrew Structure*. Tel-Aviv: University Publishing Projects.

Blevins, James P. (2006). Word-based morphology. *Journal of Linguistics* 42(3). 531-573.

Blevins, James P. (2016). *Word and Paradigm Morphology*. Oxford: Oxford University Press.

Beecher, Henry. (2004). Derivational Paradigm in Word Formation.
http://pdfcast.org/pdf/research-paper-i-derivational-paradigm-in-word-formation

Bolozky, Shmuel. (1978). Word formation strategies in Modern Hebrew verb system: denominative Verbs. *Afroasiatic Linguistics* 5: 1-26.

Bonami, Olivier & Jana Strnadová. (2019). *Paradigm structure and predictability in derivational morphology*. *Morphology* 28.2, 167-197.

Booij, Geert E. and Rochelle Lieber. (2004). On the paradigmatic nature of affixal semantics in English and Dutch. *Linguistics* 42 (2), 327-357.

Borer Hagit. (1991). The causative-inchoative alternation: A case study in parallel morphology. The Linguistic Review 8, 119-158.

Fradin, Bernard. (2016). Competition in derivation: what can we learn from doublets? Paper read at *International Morphology Meeting*, Vienna.
/sites/llf.cnrs.fr/files/u48/IMM17-Diapos-2.pdf

Haspelmath, Martin. (1987). *Transitivity alternations of the anticausative type.* (Arbeitspapiere, N.F., Nr. 4), Institut für Sprachwissenschaft der Universität zu Köln, Cologne.

Haspelmath, Martin. (1993). More on the typology of inchoative/causative verb alternations. In: Comrie, B. & Polinsky, M. (Eds.), *Causatives and transitivity*. (Studies in Language Companion Series, 23. Benjamins, Amsterdam, 87-120.

Hathout, Nabil and Fiammetta Namer (2014). Démonette, a French derivational morpho-

semantic network. Linguistic *Issues in Language Technology* 11(5), 125-168.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1, 199-244.

Lev, Shaul. (2016). Hebrew Labile Alternations. MA thesis, Tel-Aviv University.

McCarthy, John & Alan Prince. (1990). Foot and word in Prosodic Morphology: the Arabic broken plural. *Natural Language and Linguistic Theory* 8: 209-283.

Pounder, Amanda. (2000). *Processes and Paradigms in Word-Formation Morphology*. Berlin: Gruyter.

Ravid, Dorit. (1990). Internal structure constraints on new-word formation devices in Modern Hebrew. *Folia Linguistica* 24. 289-347.

Schwarzwald, Ora R. (1981). *Grammar and reality in the Hebrew verb.* Ramat Gan: Bar Ilan University Press. (in Hebrew)

Štekauer, Pavol. (2014). Derivational paradigms. In R. Lieber & P. Štekauer (eds.), *The Oxford Handbook of Derivational Morphology,* 354-369. Oxford: Oxford University Press.

Thornton, Anna M. (2012) .Reduction and maintenance of overabundance: A case study on Italian verb paradigms. *Word Structure* 5: 183-207.

Ussishkin, Adam. (2005). A fixed prosodic theory of nonconcatenative templatic morphology. *Natural Language and Linguistic Theory* 23, 169-218.

Démonette meets Semitic morphology: A paradigm-based model for the derivational resources in French and Hebrew

Lior Laks
Bar-Ilan University
Fiammetta Namer
UMR 7118 ATILF, CNRS & Université de Lorraine, Nancy, France

1 Introduction

We present **Hebrewnette**, a database model for the representation of the derivational relations in the Hebrew lexicon. The model aims at revealing the nature of Hebrew paradigmatic organisation of words including different degrees of regularity. The model is based on a Word-based approach (Aronoff 1976, 2007; Blevins 2006, 2016), according to which the lexicon consists of existing words. The entries are coded in terms of both structural and semantic relations between words. We rely on an existing architecture initially designed for concatenative morphology, and currently being implemented on French: the Démonette database (Hathout & Namer 2014, Namer & Hathout 2020)[1]. Several features make Démonette an appropriate system to account for the paradigmatic organization of the lexicon, including form-meaning mismatches:

(i) Each entry describes a relation between two derivationally related words.

(ii) Both the words and their relation are identified by a set of morphological, phonological and semantic features. Some of the features on which we focus in this talk include directionality of the relation with respect to both structural and semantic dimensions, and the changes that are involved in the transition from one word to the other.

This talk will address the question of the adaptability of a database architecture designed for concatenative morphology to the non-concatenative nature of word formation in Semitic morphology. Specifically, are the features of French Démonette compatible with Hebrew derivational morphology? We first examine the differences between the paradigmatic behaviour of the two systems. We then add features that are unique to non-concatenative morphology, while making sure that such modifications do not compromise the principles of the existing architecture.

2 Non-concatenative morphology

Semitic word formation relies highly on non-concatenative morphology, *i.e.* the formation via root and pattern (Berman 1978, Bolozky 1978, Schwarzwald 1981, Ravid 1990, Aronoff 1994). Hebrew verbs are formed only in patterns, which indicate their prosodic structure, vocalic melody and affixes (if any) (Bat-El 1994, 2017). For example, *kivec* 'shrink' is formed in *CiCeC*, and *hitkavec* 'become shrunk' in *hitCaCeC*. The semantic relations between patterns are manifested mainly in transitivity alternations. Nouns and adjectives can be formed in patterns as well as by other strategies. In this study we focus on non-concatenative formation.

The current coverage of the Hebrewnette prototype contains relations among lexemes of 29 families of various size (between 3 and 16 members), and includes word formation processes with different degrees of productivity. Each relation is encoded with respect to the

[1] This work benefited from the support of the project DEMONEXT ANR-17-CE23-0005 of the French National Research Agency (ANR) and from the Chateaubriand Fellowship Program of the French Embassy in Israel.

Démonette's guidelines with the addition of Hebrew specific features like roots, melodic structures, and the distinction between the formal and semantic orientation of a relation.

3 Case studies

Applying the principles of Démonette on Hebrew provides insights on the morphological and the semantic relations between Hebrew words. We show it in two case studies.

3.1 Form/meaning mismatches : Cases of apparent morpho-semantic mismatches are reflected in causative/inchoative alternations (see Haspelmath 1987, 1993, Borer 1991, Doron 2003, Horvath & Siloni 2010, Rappaport Hovav and Levin 2012, among many others), where semantic and morphological directions seem to collide. In each pair of verbs in (1), the semantic relation is similar, where the transitive verbs denote causation of change in X's mental state, and the intransitive verbs denote the change in the mental state that X undergoes. However, the structural relations are different. In (1a), the relation seems *ascending* from the transitive to the intransitive verb, as the former is formed in an affixless pattern (*CiCeC*), while the latter is formed in a pattern with a prefix (*hitCaCeC*). In contrast, the structural relation in (1b) seems *descending* from the transitive to the intransitive. The transitive verb is formed in a pattern with a prefix (*hiCCiC*), while the intransitive verb is formed in *CaCaC*, which has no affixes. We assume that the transition from an affixless form to a form with an affix is ascending.

(1)

Transitive verb		Intransitive verb	
a. yiʕeš	'W make X desperate'	hityaʕeš	'X become desperate'
b. hitsis	‘W make X agitated’	tasas	'X become agitated ‘

The fact that Démonette encodes separately semantic and structural information about the direction of the derivational relation enables an accurate representation of such mismatches.

3.2 Faithfulness constraints and competing patterns: Hebrew has cases of doublet formation (see Bolozky 2003, Laks 2013), where two words share the same meaning and root consonants, but are formed in different patterns. One of the doublets is preferred due to faithfulness (Hammond 1988, McCarthy & Prince 1990, Bat-El 1994, 2017, Ussishkin 2005, Kihm 2011) to the base from which it is derived. This is demonstrated for instrument nouns that are assumed to derive from verbs. Both verbs in (2) have derived instrument nouns in two competing patterns.

(2)

Verb		**Verb pattern**	**Instrument noun**		**Nominal pattern**
a. sinen	'filter'	CiCeC	(i) masnen	'a filter'	maCCeC
			(ii) mesanen		meCaCeC
b. hidgiš	'emphasize'	hiCCiC	(i) madgeš	'a marker'	maCCeC
			(ii) madgiš		maCCiC

In both cases the form in (ii) is preferred over the form in (i) (Bolozky 1999, 2003). This is because in (2a) the formation of *mesanen* is more faithful to *sinen* as it involves only prefixation and changing the vowels, while the prosodic structure remains intact. In contrast, the formation of *masnen* changes the prosodic structure of the base, as it creates the consonant cluster /sn/ that does not exist in the verb *sinen*. In (2b), the formation of both *madgeš* and *madgiš* does not change the verbal prosodic structure, but *madgiš* is more faithful because its second vowel /i/ is identical to the second vowel of the related verb *hidgiš*. The formation of

both instrument nouns in (ii) involves fewer changes with respect to the verb, and as a result there is greater structural transparency between the verb and the instrument noun. The degree of faithfulness can be represented in Hebrewnette in terms of Levenshtein measure and thus can be predicted systematically.

4 Conclusions

These case studies pave the way to the implementation of the methodology of Démonette on Semitic morphology. The overall architecture of both tools can be based on the same principles. The design of Démonette's annotation system makes its features, initially designed for French, suitable for capturing both morphological and semantic relations between Hebrew words, regardless of the type of morphology (i.e. concatenative or non-concatenative). Other morphological tools and resources for Hebrew and other Semitic languages exist (*e.g.* Wintner 2004, Itai & Wintner 2008, Singh & Habash 2012, Nir et al. 2013, Klimek et al. 2016). However, they rely mostly on the consonantal root as the central entity used as a base for Word Formation, which implies that family networks are oriented tree-shaped graphs, where only ancestor-descendant relations are represented, and not paradigmatic relations between words. As a result, such systems do not allow a separation between structural and semantic properties, in the examination of such paradigmatic relations.

Specifically, we have shown that the proposed design of Hebrewnette allows the representation and analysis of cases of form-meaning mismatches and the role of faithfulness in selection between competing patterns (Aronoff 2016). The results also point out to the central role of paradigms in derivational relations, as has been shown in previous studies (see for example Bauer 1997, Pounder 2000, Beecher 2004, Booij & Lieber 2004, Štekauer 2014, Bonami & Strnadová 2019, Blevins 2016).

References

Aronoff, Mark. (1976). Word Formation in Generative Grammar. Cambridge: MIT Press.

Aronoff, Mark. (1994). Morphology by Itself. Cambridge: MIT Press.

Aronoff, Mark. (2007). In the beginning was the word. Language 83. 803-830.

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

Bat-El, Outi. (1994). Stem modification and cluster transfer in Modern Hebrew. Natural Language and Linguistic Theory 12: 572-596.

Bat-El, Outi. (2017). Word-based items-and processes (WoBIP): Evidence from Hebrew morphology. In C. Bowern, L. Horn, & R. Zanuttini (eds.), On Looking into Words (and beyond), 115-135. Berlin: Language Sciences Press.

Bauer, Laurie. (1997). Derivational paradigms. In G. Booij & J. van Marle (eds.), Yearbook of Morphology 1996. 243-56. Dordrecht: Kluwer.

Berman, Ruth A. (1978). Modern Hebrew Structure. Tel-Aviv: University Publishing Projects.

Blevins, James P. (2006). Word-based morphology. Journal of Linguistics 42(3). 531-573.

Blevins, James. P. (2016). Word and Paradigm Morphology. Oxford: Oxford University Press.

Beecher, Henry. (2004). Derivational Paradigm in Word Formation.
http://pdfcast.org/pdf/research-paper-i-derivational-paradigm-in-word-formation

Bolozky, Shmuel. (1978). Word formation strategies in Modern Hebrew verb system: denominative Verbs. Afroasiatic Linguistics 5: 1-26.

Bolozky, Shmuel. 1999. Measuring productivity in word formation: the case of Israeli Hebrew. Leiden: Brill.

Bolozky, Shmuel. 2003. Phonological and morphological variations in spoken Hebrew. In Benjamin H. Hary (ed.), Corpus Linguistics and Modern Hebrew. Tel Aviv: Rosenberg

Bonami, Olivier & Jana Strnadová. (2019). Paradigm structure and predictability in derivational morphology. Morphology 28.2, 167-197.

Booij, Gert E. & Rochelle Lieber. (2004). On the paradigmatic nature of affixal semantics in English and Dutch. Linguistics 42 (2), 327-357.

Borer Hagit. (1991). The causative-inchoative alternation: A case study in parallel morphology. The Linguistic Review 8, 119-158.

Doron, Edit. (2003). Agency and voice: The semantics of the Semitic templates. Natural Language Semantics 11, 1-67.

Hammond, Michael. (1988). Templatic transfer in Arabic broken plurals. Natural Language and Linguistic Theory 6, 247-270.

Haspelmath, Martin. (1987). Transitivity alternations of the anticausative type. (Arbeitspapiere, N.F., Nr. 4), Institut für Sprachwissenschaft der Universität zu Köln, Cologne.

Haspelmath, Martin. (1993). More on the typology of inchoative/causative verb alternations. In: Comrie, B. & Polinsky, M. (Eds.), Causatives and transitivity. (Studies in Language Companion Series, 23. Benjamins, Amsterdam, 87-120.

Hathout, Nabil & Fiammetta Namer (2014). Démonette, a French derivational morpho-semantic network." Linguistic Issues in Language Technology 11(5), 125-168.

Horvath, Julia & Tal Siloni. (2010). Lexicon versus syntax: Evidence from Morphological Causatives. In: Doron, E., Sichel, I., Hovav-Rappaport, M., (Eds.), Syntax, Lexical Semantics, and Event Structure. Oxford University Press, Oxford. 153-176.

Itai, Alon & Shuly Wintner. (2008). Language resources for Hebrew. Language Resources and Evaluation, 42, 75–98.

Kihm, Alain. 2011. Plural formation in Nubi and Arabic: A comparative study and a word-based
approach. Brill's Annual of Afroasiatic Languages and Linguistics 3. 1-21.

Klimek, Bettina, Natanel Arndt, Sebastian Krause & Timotheus Arndt. (2016). Creating Linked Data Morphological Language Resources with MMoOn. The Hebrew Morpheme Inventory. The 10th edition of the Language Resources and Evaluation Conference, 23-28 May 2016, Slovenia, Portorož.

Laks, Lior. (2013). Why and how do Hebrew verbs change their form? A morpho-thematic account. Morphology 23(3), 351-383.

McCarthy, John & Alan Prince. (1990). Foot and word in Prosodic Morphology: the Arabic broken plural. Natural Language and Linguistic Theory 8. 209-283.

Namer, Fiammetta & Nabil Hathout. (2020). ParaDis and Démonette –From Theory to Resources for Derivational Paradigms. The Prague Bulletin of Mathematical Linguistics 114. 5-33.

Nir Bracha, Brian MacWhinney & Shuly Wintner. (2013). The Hebrew CHILDES corpus: transcription and morphological analysis. Language Resources and Evaluation 47 (4), 973-1005.

Pounder, Amanda. (2000). Processes and Paradigms in Word-Formation Morphology. Berlin: Gruyter.

Rappaport Hovav, Malka & Beth Levin. (2012). Lexicon uniformity and the causative alternation. In: Everaert, M., Marelj, M., Siloni, T. (Eds.),The Theta System: Argument Structure at the Interface. University Press, Oxford, 150-176.

Ravid, Dorit. (1990). Internal structure constraints on new-word formation devices in Modern Hebrew. Folia Linguistica 24. 289-347.
Schwarzwald,Ora R. (1981). Grammar and reality in the Hebrew verb. Ramat Gan: Bar Ilan University Press. (in Hebrew)
Singh, Nimesh & Nizar Habash. (2012). Hebrew Morphological Preprocessing for Statistical Machine Translation. Proceedings of the 16th EAMT Conference, Trento, Italy
Štekauer, Pavol. (2014). Derivational paradigms. In R. Lieber & P. Štekauer (eds.), The Oxford Handbook of Derivational Morphology, 354-369. Oxford: Oxford University Press.
Ussishkin, Adam. (2005). A fixed prosodic theory of nonconcatenative templatic morphology. Natural Language and Linguistic Theory 23, 169-218.
Wintner, Shuly. (2004). Hebrew computational linguistics: Past and future. Artificial Intelligence Review 21(2), 113-138.

Measuring morphological series membership
The example of feminine *-eur* nouns in French

Fabio Montermini
CLLE, CNRS &
Université de Toulouse Jean Jaurès

Delphine Tribout
Université de Lille & STL

1 Introduction

Several recent works in morphology, especially on French, have proposed that the form of constructed words is determined by the interaction of different, possibly conflicting, constraints (see Hathout 2009; Plénat & Roché 2014; Roché & Plénat 2014, among others). If some of these constraints correspond to classic, and possibly universal, faithfulness and well-formedness constraints, others are intended to formalize an emergent and exemplar-based model of morphology, in which productively derived lexemes are modelled on the existing lexicon speakers have access to. The main function of constraints is to guarantee coherence and predictability within the lexicon, in particular by inserting newly constructed lexemes into existing lexical networks. The notions 'Morphological family' and 'Morphological series' play a major role in these models; a word-formation pattern is thus viewed as a strategy to insert a lexeme in a lexical network, at the intersection between a family and a series. Among the constraints proposed by the above-mentioned authors there are specific 'Family' and 'Series constraints', aiming at maintaining maximal homogeneity within families and series (cf. e.g. Roché 2011: 87; Plénat & Roché 2014: 72). The notions in question may be easily characterized intuitively. However, although they play a key role in constraint-based morphology, they are trickier to define in a rigorously formal way, especially when one deals with large sets of data. Core cases are often clearly identifiable, but dealing with the peripheral ones, and drawing clear borders for families and series is generally harder. In recent years, an important work has been done in the direction of a precise characterization of morphological families (see e.g. Roché 2017; Fradin 2018; Bonami & Strnadová 2019). Morphological series, on the other hand, have not yet received much attention in this respect.

Our proposal constitutes a first attempt to fill this gap. In particular, we propose a set of parameters that can be considered reliable indicators of the belonging of a particular lexeme to a morphological series. In the model we propose, series are viewed as prototypical spaces to which individual items (lexemes) may be compared. We claim that the distance of a specific lexeme from a series' prototype may be measured by means of explicit criteria. An advantage of this view is that it allows defining the Series constraint in a more explicit way. At a very general level, it entails that each new lexeme that is inserted into the series is as close as possible (formally and semantically) to the prototype (see also Roché 2011: 87). This global constraint may also be decomposed into smaller, possibly conflicting, ones, each of which is responsible for the fact that a specific lexeme tends to be close to the prototype according to a specific parameter. Our definition of 'series' is based on Hathout (2009: 35-36), who distinguishes between morphological and lexical series. Morphological series are sets of lexemes connected by systematic form / meaning relations (e.g. DÉRIVATION 'derivation', PRODUCTION 'production', both synchronically linked to a verb); lexical series include lexemes that can be grouped together only on the basis of their form or meaning (e.g. CONFECTION 'elaboration' or LOCOMOTION 'locomotion', for which no base verb can be identified in synchrony). If the two coexist, a morphological series is necessarily a subset of a larger lexi-

cal series. We consider that the two-level structure proposed by Hathout fits well with a model in which series membership may be measured in terms of prototypicality.

2 Data: feminine *-eur* nouns in French

For an assessment of the parameters that may be taken as relevant, and possibly correlating, in order to define a series' prototype, we chose to focus on feminine nouns ending with the sequence *-eur* in French. We created a database by extracting all relevant nouns from the *TLFi* dictionary and the Lexique.org corpus[1] (83 overall). Apart from its compactness, an advantage of this class of nouns is that it allows identifying a clear morphological series (cf. DOUCEUR ← DOUX 'sweet'; PROFONDEUR ← PROFOND 'deep')[2] and a larger lexical series (cf. RANCŒUR 'resentment'; VIGUEUR 'vigour'). It also includes lexemes which belong less obviously to the one or to the other category (cf. SŒUR 'sister', CHANDELEUR 'Candlemas', etc.).

3 Analysis

All nouns in the database were coded according to several parameters, which can eventually be correlated with each other in order to determine the properties contributing to define the series' prototype. The parameters chosen correspond to canonical phonological and semantic properties, but also to properties accounting for morphological and lexical relations.

First, data were coded according to their morphological structure, as detailed in Table 1 (derived / underived, autonomous / non autonomous base, category of the base), and to their size in phonemes and syllables.

Type		**Example**	**Number**
derived	deadjectival / autonomous base	DOUCEUR (← DOUX 'sweet')	37
	deadjectival / non-autonomous base	PUDEUR (← PUDIQUE 'modest')	11
	deverbal	VALEUR (← VALOIR 'be worth')	6
underived		VIGUEUR 'vigor'	29

Table 1: Distribution of feminine *-eur* nouns in the database

According to these criteria, the most prototypical lexemes have an adjectival autonomous base (37/83) which is underived and monosyllabic (34/83) such as MAIGREUR ← MAIGRE ('thin').

As a second step, *-eur* nouns were coded to the size and form of their morphological families (also based on Lexique.org). In this case, we distinguished between the 'parallel' family (i.e. the set of lexemes constructed on the same base adjective, e.g. DOUCEMENT, DOUCEÂTRE, ADOUCIR… for DOUCEUR) and the 'descending' family (i.e. the set of lexemes derived, directly or indirectly, from the *-eur* noun itself, e.g. VALEUREUX, VALEUREUSEMENT, VALORISER… for VALEUR). With respect to this parameter, prototypicality is measured according to the rate of

1 We only excluded feminine forms of agent nouns in *-eur*, such as AUTEURE ('author$_F$'), INGÉNIEURE ('engineer$_F$'), etc.

2 When both the *-eur* noun and the base adjective are mentioned, only the latter is glossed.

overlapping of morphological families. From this point of view, an observation we can make is that the highest rated lexemes all denote physical, visual or sensible, properties (e.g. DOUCEUR 'sweetness', MAIGREUR 'thinness', GRANDEUR 'size', BLANCHEUR 'whiteness'…). A clear distinction also emerges between derived *-eur* nouns, which tend to belong to large parallel families (average size 4.14 lexemes), but block further derivation, and the underived ones, which may be the roots of descending families (average size 2.48 lexemes). Moreover, when they serve as derivational bases the latter, but not the former, display specific stem allomorphies (cf. /œʁ/~/ɔʁ/, TUMEUR 'tumor' → TUMORAL; /œʁ/~/uʁ/, DOULEUR 'pain' → DOULOUREUX vs. CHAUD 'hot' → CHALEUR → CHALEUREUX).

Another parameter considered is the ratio between the frequency of the derivative and the frequency of the base (also from Lexique.org), a measure which has been taken in the literature as a good indicator of a derivative's parsability (cf. Plag & Baayen 2009; Sims & Parker 2015). Interestingly, derived *-eur* nouns displaying the lowest degree of parsability according to this parameter are also less transparent in other respects, e.g. are derived from non-autonomous adjectives or from verbs. In this case too, the lexemes which appear to be the most prototypical are those which denote physical properties (NOIRCEUR 'blackness', ROSEUR 'pinkness', VERDEUR 'greenness', GROSSEUR 'size'…).

The last criterion used in order to determine a prototype for feminine *-eur* nouns is semantic in nature. Nouns were coded according to their semantic type. It appears that most nouns (53/83) denote a physical property, either visual (BLANCHEUR, LARGEUR 'width') or sensible (DOUCEUR, CHALEUR). When a noun denotes a psychological property, it is most often underived (PEUR 'fear') or derived from a non-autonomous base (PUDEUR). Derived nouns may also have a psychological meaning, but it always coexists with a concrete reading (e.g. NOIRCEUR 'blackness / darkness'). Finally, all *-eur* nouns derived from adjectives denote individual-level predicates (Carlson 1977), except for those derived from colour adjectives. Some of them may also denote stage-level predicates. Interestingly, the few nouns allowing only a stage-level predicate interpretation have no base or no autonomous base (cf. HORREUR 'horror', STUPEUR 'astonishment', TERREUR 'terror').

4 Conclusion

Above we presented some measures which appear to be significant in defining a prototypical core for the derivational series of feminine *-eur* nouns in French, which in turn constitutes the centre of a larger lexical series. Some of these properties correspond to those already established by Koehl (2012), in particular concerning the phonology and the semantics of the base. Other possibly significant criteria to be taken into account include for instance the formal complexity of the base and the stem selected in the derivative, or purely phonological factors, such as the segment preceding *-eur*. It is likely that a complex model, in which all the criteria considered, or a part of them, are correlated would give even clearer results in order to determine the prototypical core of the series.

Moreover, the model we propose is intended to be largely applicable to various morphological and lexical series. Planned future work includes the testing of the model's robustness on larger sets of data and its implementation in a constraint-based model of morphological derivation.

References

Bonami, Olivier & Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29.2. 167-197.

Carlson, Gregory N. 1977. *Reference to Kinds in English*. PhD Thesis. University of Massachusetts.

Fradin, Bernard. 2018. The variety of derivational paradigms. Paper presented at *Word Formation III. Typology and Universals in Word Formation IV. Košice, Slovakia, 27-30 June 2018*.

Hathout, Nabil 2009. *Contributions à la description de la structure morphologique du lexique et à l'approche extensive en morphologie*. Habilitation à diriger des recherches. Université de Toulouse – le Mirail.

Koehl, Aurore. 2012. *La construction morphologique des noms désadjectivaux en français*. PhD Thesis. Université de Lorraine.

Lexique.org. *Lexique*. www.lexique.org.

Plag, Ingo & Harald Baayen. 2009. Suffix ordering and morphological processing. *Language* 85.1. 109-152.

Plénat, Marc & Michel Roché. 2014. La suffixation dénominale en *-at* et la loi des (sous-)séries. In Sophie David, Sarah Leroy & Florence Villoing (eds.), *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*, 47-74. Nanterre: Presses Universitaires de Paris Ouest.

Roché, Michel. 2011. Quel traitement unifié pour les dérivations en *-isme* et en *-iste* ?. In Michel Roché, Gilles Boyé, Nabil Hathout, Stéphanie Lignon & Marc Plénat, *Des unités morphologiques au lexique*, 69-143. Paris: Hermès-Lavoisier.

Roché, Michel. 2017. Les familles dérivationnelles : comment ça marche ?. Manuscript.

Roché, Michel & Marc Plénat. 2014. Le jeu des contraintes dans la sélection du thème présuffixal. In Franck Neveu, Peter Blumenthal, Linda Hriba, Annette Gerstenberger, Judith Meinschaefer & Sophie Prévost (eds.), *Actes du 4e Congrès Mondial de Linguistique Française. Berlin, Allemagne, 19-23 juillet 2014*, 1863-1878. Paris: Institut de Linguistique Française.

Sims, Andrea D. & Jeff Parker. 2015. Lexical processing and affix ordering: cross-linguistic predictions. *Morphology* 25.2. 143-182.

TLFi. *Trésor de la Langue Française informatisé*. http://atilf.atilf.fr/.

Derivational Paradigms and The Frequency Factor : The French *-ion* Nouns Allomorphy Problem

Gauvain Schalchli
Bordeaux-Montaigne University/CLLE

1 Introduction

The classical descriptive method for demonstrating the availability of derivational processes (for example an affix), is to show a series of word pairs with the same morphological difference as analysable by a proportional analogy (Dell 1970, 1979; Corbin 2012). That process doesn't take token frequency of lexemes into account, neither diachronic newness. In order to argue the importance of updating the derivational method of analysing data with these two types of properties, we illustrate the methodological issue by a well known case of derivational allomorphy in French illustrated by the following example from Bonami, Boyé & Kerleroux (2009:8, Table 6):

Classe	Description	Exemple		Effectif
1	**Rad3** ⊕ asjɔ̃	*vexation*	/vɛksasjɔ̃/	1093
2	**Rad3** ⊕ kasjɔ̃	*modification*	/modifikasjɔ̃/	95
3	**Rad3** ⊕ jɔ̃	*dispersion*	/dispɛrsjɔ̃/	86
4	**Rad3** ⊕ isjɔ̃	*composition*	/kɔ̃posisjɔ̃/	33
5	**Rad3** ⊕ sjɔ̃	*pollution*	/polysjɔ̃/	50
6	**X** ⊕ jɔ̃	*abstraction*	/abstraksjɔ̃/	277
7	**Pas de base autonome**	*compétition*	/kɔ̃petisjɔ̃/	474

Table 1 - Surface classification of -ion nouns

In addition to token frequency and diachronic datation, another shortcoming of the classical method is the simplification of derivational data by taking the lexical abstract unit (citation form, lemme or principle parts) as an approximation of lexemes, which is inadequate for inflectional languages. This approximation is possibly without bias for languages where the majority of inflected derivational bases have one and unique stem, but it is considered problematic in the case of systematic inflectional stem allomorphy (Zwicky 1992; Aronoff 1993; Brown 1998; Pirrelli & Battista 2000; Boyé 2000; Stump 2001).

The works of Boyé and Bonami on verbal inflection in French (Bonami et Boyé 2003, 2007, 2014; Boyé 2011) concerns initially many-indexed-bases verbal inflection in contemporary French. In line with Aronoff (1993)'s suggestions for Latin conjugation, their model describes French verbal paradigms by applying inflectional rules on a finite list of stems regularly distributed between the morpho-syntactic cells ("hidden stem"). But this model was also extended to derivation by allowing special stems, named "hidden stems", for derivational processes inside the stem list of each verbal lexeme used as input of derivational rules.

This inclusion of hidden stems to deverbal derivation as initially argued by Bonami, Boyé & Kerleroux (2009) has been applied to conversion by (Tribout 2010, 2012), and generalized to all inflectional categories and all derivational processes by (Roché 2010).

Though, this idea is counter-intuitive, because a stem only realised in derivational context (output) is attributed to the inflectional properties (derivational input). This choice preserves the unicity of derivational input of morphophonological processes and simplifies the description of derivational allomorphy.

However, in recent years, after the abstractive revolution (information-theoretic turn) of the word and-paradigms approaches (Blevins 2006, 2016; Ackerman et al 2009), the stem space model was abandoned by Bonami & Boyé for the benefit of a many-to-many relations between forms system (Bonami & Boyé 2014 ; Bonami 2014; Boyé 2016, 2017, 2019, 2020)

Moreover, psycholinguistic studies on lexical processing showed that the family-size of the phonological neighbours and the frequency estimations of lexical stimuli have an important effect on lexical decision times and other experimental tasks.[1]

The move from the stem spaces to the many-to-many relation nullifies the derivational analyses based on stem indexing. Without stems, works on derivational morphology like (Bonami, Boyé & Kerleroux (2009) ; Bonami & Boyé, 2005 ; Kerleroux, 2007 ; Boyé & Plénat, sous presse ; Plénat, 2008 ; Roché, 2010 ; Tribout, 2010, sous presse)[2] do not work. In this paper, we try to avoid this problem and integrate frequency to reformulate the classical derivational method in line with inflectional complexity and paradigmatic approaches of word formation.

For this discussion, we compare the proposition of a special stem by Bonami et al (2009) with data attested in a French lexical database validated by psycholinguistic experiments. This comparison illustrates the impact of taking account or not token frequency and diachronic datation on morphological analysis.

2 Verbal stem allomorphy: the empirical evidence

In the first step, we will discuss the empirical evidence of the "stem space" model of the French verbal allomorphy. The model attributes 12 "abstract stems" to each concrete verb of the French lexicon. However, for many verbs, several "abstract stems" have the same phonological content. For example, following Boyé (2011:56), LAVER presents only four different phonological stems and FINIR only two. The more irregular verb "ETRE" presents probably seven or eight distinct stems and CONCLURE probably just one (Bonami 2014:49,51,57).

In this section, we propose to evaluate quantitatively the empirical reality of the verbal stem allomorphy in French. In order to establish this evaluation, we compare *Flexique*[3], the reference lexical resource for inflection in French, with *Lexique3*[4], the reference lexical resource for psycholinguistic experiments in French.

After eliminating thematic vowels and vowel alternations, *Flexique*'s 5178 complete non-defective verbal paradigms (with 51 forms) contain only 6% cases of radical allomorphy. The *Lexique3* database contains 6399 verbal lexemes attested in a corpus of film subtitles or in a corpus of recent novels, but the paradigms are not complete because not all forms of each verb are necessarily attested. Table 2 illustrates this lack of paradigm coverage for extreme frequency ranges.

[1] For example see (AMBRIDGE et al. 2015) for a general discussion based on acquisition.

[2] Works cited by Bonami & Boyé (2014)

[3] http://www.llf.cnrs.fr/fr/flexique-fr.php

[4] http://www.lexique.org

The five highest cells			The five lowest cells		
CELL	FREQ	% of 6399 lexemes	CELL	FREQ	% of 6399 lexemes
inf	5294	83%	ind:pas:2p	14	0,2%
par:pas	5139	80%	sub:imp:2s	14	0,2%
ind:pre:3s	4265	67%	sub:imp:1p	12	0,2%
ind:imp:3s	3783	59%	sub:imp:2p	7	0,1%
par:pre	3370	53%	imp:pre:3s	1	0,02%

Table 2 – Headcounts of the five most frequently attested and the five lowest frequently attested cells of French conjugation

Overall, the coverage of the verbal lexicon in number of inflectional forms is about 25%. The same analysis as above results in a significant decrease of verbs presenting a surface stem allomorphy to 1.89% (94 allomorphic verbs and 5042 non-allomorphic verbs) of all attested verbal lexemes.

In the following section, we examine the part of these allomorphic verbs that is correlated to an *-ion* noun and the correspondence between the form of the *-ion* noun and that of the verbal stem when it is unique.

3 *-ion* nouns data: lexeme pairing and stem matching

In this section, we examine the following questions:

1. Do polyradical verbs correspond to an *-ion* noun and if so, is this relationship relevant to the study of *-ion* noun construction?
2. Do *-ion* derivatives of non-allomorphic verbs have a consistent formal relationship with the stem of these verbs?

The *Lexique3* database contains 1293 constructed deverbal feminine nouns by suffixation in *-ion*. Among these nouns, 994 (about 77%) do not have any allomorphic relation to their verbal base. 299 nouns present an allomorphy with respect to their verbal base. Among them, 252 (84%) are directly borrowed from Latin at an early date. The remaining 47 nouns correspond either to a prefix or to the influence of an older member of the derivational family or are attested prior to the 16th century.

Consequently, it seems that the hypothesis of a suppletive hidden stem in the verbal base could be used to "simplify" the reconstruction of about 4% of the nouns studied, but that only 14 of these nouns are not attested before the 20th century and that 13 of these are as well as explainable by a denominal prefix.

We therefore propose to abandon the hidden radical hypothesis and we explore in the following section the hypothesis of a suffixal allomorphy based on the influence of the lexical token frequency.

4 *-ion* nouns data: the frequency factor

Most likely, the *-ion* nouns derived from one of the handful of attested verbal lexemes with surface allomorphy are well lexicalized lexemes and are not the primary focus of derivational theory which deals primarily with productivity (Dal & Namer 2016).

If morphological units have to be attested with significant data, as assume Word-and-paradigm approach and empirical principles, the preceding of verbal allomorphy provides very weak evidence for the hidden verbal stems hypothesis of the construction of French *-ion* nouns as proposed by Bonami, Boyé & Kerleroux (2009). From the preceding description, attested data for French verbs generally seem to offer only one surface stem by verb for deverbal derivation. The allomorphy problem of *-ion* nouns has to be resolved with a different explanation than the result of the verbal allormophy of the base.

In the course of their analysis, Bonami, Boyé & Kerleroux (2009) argue about the productivity of *-ion* nouns. Following their observation, only two subclasses of *-ion* nouns from the seven distinguished (*-ation* and *-cation* vs. *-ion*, *-tion*, *-ition* and unanalysability) are productive (between the five other classes four are unproductive and one is morphologically unanalysable). We compared the frequency distribution of the two open classes with that of the five other classes by counting the respective lexemes inside each quartile's division of the frequency scale. The result of the comparison is presented in Table 3.

This comparison shows that lexemes of the productive classes are more frequent at both extremes of the scale and that the unproductive classes are concentrated in the middle

Inverse frequency order	PRODUCTIVE CLASSES	UNPRODUCTIVE CLASSES	COMPARISON
1st quartile	25,6%	21,7%	- 4,9%
2nd quartile	23,6%	28,1%	+ 4,5%
3rd quartile	23,3%	24,9%	+ 1,6%
last quartile	27,4%	15,9%	- 11,5%

Table 3 - Productive and unproductive classes comparison of -ion nouns derivation

frequencies. These statistics argue for a suffixal nature of the *-ion* polymorphy. In fact, the population of the productive classes is sufficient enough in the high range of frequencies to allow for the production of *-ation* and *-cation* low frequency nouns.

Table 4 presents the correlation between high and low frequency levels of *-ion* nouns for the variants of the *-ion* suffix. This confirms that *-ation* and *-cation* are the only productive variants of *-ion* deverbal construction and that all variants have enough frequent exemplars in the lexicon to lead new formations without any need of a complex mechanism from verbal indexed stem allomorphy.

ENDINGS	SUFFIX HEADCOUNTS		RATIOS
	high frequency range	low frequency range	
/asj§/ (*-ation*)	297	386	1,2996633
/j§/ (*-ion*)	26	10	0,38461538
/sj§/ (*-tion*)	26	8	0,30769231
/kasj§/ (*-cation*)	18	39	2,16666667
/isj§/ (*-ition*)	8	6	0,75

Table 4 - Quantitative realized productivity of -ion suffix variants correlated with levels of frequency

5 Conclusion

The data for French *-ion* nouns derivation provide evidence for an analogical, abstractive, usage-based approach of word formation which is particularly apt at capturing paradigmatic derivational phenomena. More specifically, our analysis shows that the discovery procedure for derivational processes such as suffixal constructions has to be refined in relation to the empirical definition of derivational paradigms. The series of attested derivational pairs allowing to hypothesize a morphological rule, like that which forms deverbal nouns in *-ion* for example, need to be divided in two parts. The high frequency part of the series serves as exemplars of the rules. These nouns are not actually constructed by the rule but actually they serve to construct the rule. The low frequency part of the series is the product of the rule. They display the actual productivity of the rule. The matching of derivatives and bases inside the high part of the frequency scale of attested data supports the producibility of the low frequency part.

This approach is in line with the exponent-constraints model of (Montermini 2018) and is an example of the series effect or the sub-series law described by (Plénat & Roché 2014). It is the same mechanism which leads to coalescence phenomena (Roché 2009; Amiot 2020) and triangulation (Lignon, Namer et Villoing 2014; Dal & Namer 2015).

More generally, it is an adaptation to derivational morphology of the Blevins' abstractive approach developed initially to inflection (Blevins 2006,2016). This hypothesis respects the word-and-paradigm principle that each morphological part must be supported by several morphological wholes (Ackerman, Blevins & Malouf 2009), that is: a paradigm. In other words, we can summarize by saying that morphology, as the generative face of redundancy (Jackendoff 1975, Aronoff 1976, Aronoff & Anshen 1998, Jackendoff & Audring 2020) is more likely just a complex (multidimensional) frequency effect.

References

Ackerman, Farrell, James P. Blevins, et Robert Malouf. 2009 Parts and Wholes:

Implicative Patterns in Inflectional Paradigms. In Analogy in Grammar Oxford University Press.

Ambridge, Ben, Evan Kidd, Caroline F. Rowland, et Anna L. Thaekston. 2015. « The ubiquity of frequency effects in first language acquisition ». *Journal of Child Language* 42(2): 239-73.

Amiot, Dany. 2020. Chapitre 48. Procédés morphologiques de création lexicale. In *Grande Grammaire Historique du Français (GGHF)*. De Gruyter Mouton.

Aronoff, Mark. 1993. Morphology by Itself: Stems and Inflectional Classes. Cambridge, Mass: The MIT Press.

Aronoff, M. (1976). Word formation in generative grammar. Linguistic Inquiry Monographs Cambridge, Mass, (1), 1-134.

Aronoff, M., & Anshen, F. (1998). Morphology and the lexicon: Lexicalization and productivity. Spencer & Swicky (Eds.) The handbook of morphology, 237-247.

Bonami, Olivier, et Gilles Boyé. 2003. « Supplétion et classes flexionnelles ». *Langages* 37(152): 102-26.

———. 2007. « Remarques sur les bases de la conjugaison ». In *Des sons et des sens (données et modèles en phonologie et en morphologie)*, Langues et syntaxe, éd. Elisabeth Delais Roussarie et Laurence Labrune. Hermès Sciences, 77-90.

———. 2014. « De formes en thèmes ». In *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*, éd. Florence Villoing, Sarah Leroy, et Sophie David. Presses Universitaires de Paris-Ouest, 17-45.

Bonami, Olivier, Gilles Boyé, et Françoise Kerleroux. 2009. « L'allomorphie radicale et la relation flexion-construction ». In *Aperçus de morphologie*, éd. Fradin, Bernard; Kerleroux, Françoise; Plénat, et Marc; Presses de l'Université de Vincennes, 103-25.

Boyé, Gilles. 2000. *Problèmes de morpho-phonologie verbale en français, en espagnol et en italien*. (phd thesis). Université Paris-Diderot - Paris VII.

———. 2011. Régularités et classes flexionnelles dans la conjugaison du français. In *Des unités morphologiques au lexique*, Langues et Syntaxe, éd. Roché et al. Hermes Science Publishing/Lavoisier, 41-68.

Brown, Dunstan. 1998. Stem Indexing and Morphonological Selection in the Russian Verb: A Network Morphology Account. In Models of Inflection. De Gruyter.

Corbin, Danielle. 2012. *Morphologie dérivationnelle et structuration du lexique*. De Gruyter Mouton.

Dal, Georgette, et Fiammetta Namer. 2015. La fréquence en morphologie : pour quels usages ? *Langages* N° 197(1): 47-68.

———. 2016. Productivity. In *The Cambridge Handbook of Morphology*, éd. Andrew Hippisley & Gregory T. Stump. Cambridge University Press., 70-90.

Dell, François Les règles phonologiques tardives et la morphologie dérivationnelle du français, 1970, M.I.T. Dissertation.

Dell, François « La morphologie dérivationnelle du français et l'organisation de la composante lexicale en grammaire generative », Revue Romane, XIV, 1979, 185-216.

Haspelmath, Martin, et Andrea D Sims. 2010. *Understanding Morphology*.

Jackendoff, Ray & Jenny Audring. 2020. *The texture of the lexicon*. Oxford: Oxford University Press.

Lignon, Stéphanie, Fiammetta Namer, et Florence Villoing. 2014. De l'agglutination à la triangulation ou

comment expliquer certaines séries morphologiques. SHS Web of Conferences 8: 1813-35.

Montermini, Fabio. 2018. Les affixes dérivationnels ont-ils des allomorphes ? Pour une modélisation de la variation des exposants dans une morphologie à contraintes. In The lexeme in descriptive and theoretical morphology. Ed. Bonami et al.

Pirrelli, Vito, et M. Battista. 2000. The paradigmatic dimension of stem allomorphy in Italian verb inflection. *Italian Journal of Linguistics* 12: 307-80.

Plénat, Marc, et Michel Roché. 2014. La suffixation dénominale en -at et la loi des (sous-) séries. In *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*, éd. Villoing F., S. David & S. Leroy. Nanterre, France, 47-74.

Roché, Michel. 2009. Un ou deux suffixes ? Une ou deux suffixations ?In Aperçus de morphologie du français, éd. Fradin : 143-73.

———. 2010. Base, thème, radical. *Recherches linguistiques de Vincennes* (39): 95-134.

Stump, Gregory T. 2001. *Inflectional Morphology: A Theory of Paradigm Structure.* Cambridge:Cambridge University Press.

Tribout, Delphine. 2010. How Many Conversions from Verb to Noun Are There in French? In CSLI Publications.

———. 2012. Verbal Stem Space and Verb to Noun Conversion in French. *Word Structure* 5(1): 109-28.

Zwicky, Arnold. 1992. Some choices in the theory of morphology. In *Formal Grammar; Theory and Implementation*, éd. Robert Levine. Oxford, 327-71.

Agent noun formation in Czech: An empirical study on suffix rivalry

Magda Ševčíková *Lukáš Kyjánek* *Barbora Vidová Hladká*
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

1 Introduction

Concurring with Bonami & Strnadová's (2019) account of paradigms in derivational morphology, agent nouns can be seen as paradigm cells encountered in derivational paradigms of verbs across European languages and beyond (Rainer 2015, Štekauer et al. 2012). The present paper deals with derivation of agent nouns in Czech. They are coined using a wide range of suffixes, the full list of which contains up to 35 items (Daneš et al. 1967). Following up on the recent research into rivalry in derivation (see Wauquier et al. 2020, Fernandéz-Alcaina & Čermák 2018, Santana-Lario & Valera 2017, Strnadová 2015, and references given there), the present study focuses on eight most frequent suffixes that compete for the agent noun cell within the derivational paradigms of Czech verbs; cf. (1).

(1) a. *uč-i-**tel*** 'teacher' < *uč-i-t* 'to teach'
b. *řid-i-**č*** 'driver' < *říd-i-t* 'to drive'
c. *řez-**ník*** 'butcher' < *řez-a-t* 'to cut'
d. *kov-**ář*** 'blacksmith' < *kov-a-t* 'to forge'
e. *soud-**ce*** 'judge' < *soud-i-t* 'to judge'
f. *kuř-**ák*** 'smoker' < *kouř-i-t* 'to smoke'
g. *kup-**ec*** 'buyer' < *koup-i-t* 'to buy'
h. *mluv-**čí*** 'speaker' < *mluv-i-t* 'to speak'

For the sake of this study, a corpus-based sample of nearly 1,200 agent nouns carrying one of these suffixes is assigned 30 features that are assumed to be relevant for exploiting the suffix rivalry. These features are used to feed machine-learning models in order to investigate which of the features are at play when choosing one of the competing items for a particular verb.

2 Rivalry in agent noun derivation in Czech

Derivation in Czech, as in other languages, is characterized by complex form–meaning relationships. A particular meaning is usually expressed by more than one affix (as in (1)), and many affixes convey more than one semantic category (cf. the suffix *-ák* in the agent noun in (1f), in an instrument noun in (2a), and in an augmentative noun in (2b)).

(2) a. *pad-**ák*** 'parachute' < *pad-a-t* 'to fall' b. *auť-**ák*** 'car.augment' < *auto* 'car'

In the literature on Czech derivation, deverbal agent nouns (nomina agentis; (1)) are distinguished from nouns with agentive meanings that are alleged to be motivated by nouns (nomina actoris, cf. (3); Daneš et al. 1967, p. 13ff., Grepl et al. 2000, p. 140ff.). Actual language data, though, paint a more complicated picture. In the morphological families of many agent nouns, both a verb and a noun are attested that can be considered as motivating items (cf. the verb *zvon-i-t* for (3b)). For agent nouns that attach the suffix directly to the verbal root without retaining the verbal theme (all in (1) except for (a) and (b)), one has to take into account not just one single verb, but two verbs with a common root and different themes (cf. the imperfective verb *kup-ova-t* 'to buy' in addition to the perfective *koup-i-t* 'to buy' in (1g)).

(3) a. *ryb-**ář*** 'fisher' < *ryba* 'fish' b. *zvon-**ík*** 'bell-ringer' < *zvon* 'bell'

3 Data-based modeling of agent suffix rivalry

3.1 Compilation of the data set to analyse

All nouns containing one of the eight agent suffixes chosen for the analysis were extracted from a 100-million-word corpus of written Czech (Křen et al. 2015). For each of the nouns, potential base words were identified. Those nouns for which no motivating verb was attested (as in (3a)) were excluded from the data set. The resulting list of 1,178 agent nouns, each with all possible motivating lexemes, was provided with a total of 30 features, ranging from very basic ones (e.g. grapheme strings of the agent suffix, of the agent noun and the motivating lexemes, of the verb theme; absolute corpus frequency of the lexemes) to more sophisticated. The focus was on phonological, inflectional and derivational properties of the agent nouns and their morphological families: For instance, the string shared by the agent noun and the motivating lexeme(s) (i.e., root or stem) was characterized as for the number of syllables and final vowel/consonant; the word class of the motivating lexemes was registered, as well as whether they are derived or unmotivated; specifically for the motivating verb, conjugation class and grammatical aspect were listed; it was also noted whether or not the verb has an aspectual counterpart and how it is formed.

The analysed suffixes and their absolute frequencies in the data set used for the machine learning experiments are summarized in Table 1. For the experiments, the data set was divided into training, evaluation, and hold-out subsets (in 60:20:20 splits).

Table 1: Absolute frequency of individual agent suffixes in the data set

Suffix	*-tel*	*-č*	*-ník*|*-ík*	*-ář*|*-ař*	*-ce*	*-ák*	*-ec*	*-čí*	TOTAL
Freq	426	388	106	96	66	50	32	14	**1,178**

3.2 Machine-learning experiments

The features assigned were used as predictors in machine-learning experiments. The agent noun suffix which was encoded as another feature with eight different values (*-tel, -č, -ce, -ák, -ník*|*-ík, -ec, -ář*|*-ař, -čí*) was used as the target class in the experiments. To deal with the imbalanced distribution of the suffixes, individual target classes were weighted proportionally to their frequencies in the data during the training phase. To predict the target class, we experimented with two machine-learning methods assumed to be linguistically interpretable, specifically logistic regression and decision trees. The two methods differ in how they model the impact of the given features on the target class of competing suffixes. While logistic regression estimates dependencies among the given features, decision trees propose a set of decisions over the features such that their disorder (entropy) is minimized. The abstract is limited to the former method; the most relevant results of experiments using both methods will be compared in the presentation.

As preparatory steps, the best hyper-parameters were tuned using grid search on the training and evaluation subsets of instances assigned with all 30 features. Large-scale bound-constrained optimization and l2 regularization penalty were applied.

The first experiment was based on the entire set of features. All suffixes were predicted with an accuracy of 69 %. The fact that only slightly more than two thirds of the agent nouns have been predicted correctly suggests that other factors than those represented by the features are at play. A more detailed insight into the results is provided in Table 2; see the 1st row. The best

results were achieved for the nouns in *-ník|-ík*, followed by *-tel* and *-č*. The model has failed to predict the suffixes *-ec* and *-čí*.

As the next step, the features were divided into five subsets according to the type of information they encode:

- (A) features encoding formal characteristics of the agent noun and related lexemes (the grapheme strings of the lexemes and of the verbal theme, its final character, etc.);
- (B) features capturing phonological characteristics of the base (syllable count, articulatory features of the end of the base);
- (C) inflectional features of the motivating verb(s) (grammatical aspect, conjugation class, etc.);
- (D) information about the members of individual morphological families (if a single motivating lexeme or multiple such lexemes are available, which word class they belong to, if a homonymous inanimate noun exists, if the verb has an aspectual counterpart, etc.);
- (E) quantitative characteristics of the agent noun and related lexemes (the absolute corpus frequency of each of the items).

Each subset was used as an input to a separate experiment which was run with the same hyper-parameters as applied within the first experiment with all 30 features.

When evaluated on the hold-out data containing all agent suffixes, none of the models based on the feature subsets (A) to (E) outperformed the original model (the (A)-based model with the accuracy of 62 % was the best, the (D)-based model performing at 43 % the worst one). The results of the individual models for individual suffixes (calculated, again, as F-scores) are listed in Table 2. The models that were trained on either formal features (A) or verbal inflectional features (C) achieved overall good results. Interestingly, the models that were limited to frequency features (E) performed slightly better than models exploiting phonological features (B) and rather complex information about the make-up of the morphological families (D). The last mentioned feature set, though, was sufficient to train the best-scoring model for the nouns in *-ník|ík* while, in contrast, the results based on the formal features (A) and verbal inflection (C) seem to be much less relevant here.

In the presentation, the linguistic interpretation will be the subject of focus. The results of the logistic regression experiments will be compared to an analogous set of experiments using decision trees.

Table 2: F-scores of models predicting individual agent suffixes based on all features vs. on feature subsets (on hold-out data; in percent)

Model/Suffix	*-tel*	*-č*	*-ník|-ík*	*-ář|-ař*	*-ce*	*-ák*	*-ec*	*-čí*
all features	**77**	**74**	83	59	**40**	47	0	0
subset (A)	69	68	39	56	36	48	**63**	**44**
subset (B)	59	44	80	40	32	**57**	40	16
subset (C)	72	64	39	46	25	14	43	22
subset (D)	57	19	**87**	59	35	12	42	0
subset (E)	44	62	76	**61**	33	0	24	0

Acknowledgements

This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation, the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101), and by the Grant No. START/HUM/010 of Grant schemes at Charles University (Reg. No. CZ.02.2.69/0.0/0.0/19_073/0016935). It was using language resources developed, stored, and distributed by the LINDAT/CLARIAH-CZ project.

References

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

Daneš, František et al. 1967. *Tvoření slov v češtině 2: Odvozování podstatných jmen*. Praha: Nakl. ČSAV.

Fernandéz-Alcaina, Cristina & Jan Čermák. 2018. Derivational paradigms and competition in English: a diachronic study on competing causative verbs and their derivatives. *SKASE Journal of Theoretical Linguistics* 15. 69–97.

Grepl, Miroslav et al. (eds.). 2000. *Příruční mluvnice češtiny*. Praha: NLN.

Křen, Michal et al. 2015. *SYN2015: A Representative Corpus of Written Czech*. Prague, Institute of the Czech National Corpus, Faculty of Arts, Charles University; `http://www.korpus.cz`.

Rainer, Franz. 2015. Agent and instrument nouns. In Peter O. Müller et al. (eds.), *Word-Formation. An International Handbook of the Languages of Europe*, vol. 2, 1304–1316. Berlin: de Gruyter.

Santana-Lario, Juan & Salvador Valera (eds.). 2017. *Competing patterns in English affixation*. Bern: Peter Lang.

Štekauer, Pavol et al. (eds.). 2012. *Word-Formation in the World's Languages*. Cambridge: CUP.

Strnadová, Jana. 2015. Multiple Derivation in French Denominal Adjectives. In *Carnets de Grammaire 22*, 327–346. Toulouse: CLLE-ERSS.

Wauquier, Marine et al. 2020. Contributions of distributional semantics to the semantic study of French morphologically derived agent nouns. In J. Audring et al. (eds.), *Online Proceedings of the 12th Mediterranean Morphology Meeting (MMM12)*, vol. 2, 111–122. Patras: Pasithee.

Modelling word-formation paradigms: networks visually representing their multidimensionality, complexity and theoretical infiniteness

Alexandra Soares Rodrigues
ESE – Instituto Politécnico de Bragança
CELGA-ILTEC – Universidade de Coimbra

Pedro João Rodrigues
CeDRI, ESTiG – Instituto Politécnico de Bragança

Abstract

The growing importance of a paradigmatic approach to word formation has been evident at scientific meetings on morphology in recent years, such as the 12th Mediterranean Morphology Meeting 2019, ParadigMo 2017, two workshops at SLE 2015, and in volumes such as *Lingue e Linguaggio* XVII(2), 2018, and *Morphology* 29(2), 2019.

Our work focuses on modelling word-formation paradigms. We propose that networks (Newman 2010) provide the means to model and visually represent word-formation paradigms. Networks enable us to represent paradigms at both the large scale and the small scale, bringing visual and conceptual evidence to the multidimensional relationships that shape paradigms (Štekauer (2014) and to the dynamics of the mental lexicon (Libben 2015, Elman 2011).

The relationships between the items of a paradigm can be founded on different features (Pounder 2000, van Marle 1985, Štekauer 2014), such as word class, semantic rules or formal features (Pounder labelled these features *lexical paradigms*). Štekauer (2014: 359) refers to semantic structures (AGENT, INSTRUMENT, ACTION) and to the formal realisation of these categories (suffixation in *-ation*, *-ment*, etc.). The feature that is responsible for cohesion among items of the paradigm is called the *axis of the paradigm* by Rodrigues & Rodrigues (2018). Bonami & Strnadová (2019: 170) use the term *paradigmatic system* to refer to relationships between pairs based on content (which includes syntactic and semantic categories). The term *series* is reserved for the relationships between pairs based on the share of a derivational affix (Hathout 2009).

Bearing these aspects in mind, we consider that a network model serves as the basis to describe and visualise the multiple and complex relationships built within and by derivational paradigms.

Our study is based on the analysis of a corpus comprising 8414 Portuguese deverbal nouns and their relationships with derivative verbs (Rodrigues 2008). Of those 8414 deverbal nouns, 4917 are deverbal event and state nouns (ACTION, PROCESS, STATE, etc.), and 3497 are individual deverbal nouns (AGENT, INSTRUMENT, PLACE, etc.). The analysis of the relationships between deverbal nouns and verbs yields the following aspects, which may be conceived as organised into networks:

a) the constraints between the morphological structures of the verbs and the morphological structures of the nouns (which nominaliser suffixes (do not) correlate with which morphological structures of the verbs). E.g., there is no paradigmatic series constituted by deverbal nouns with the suffix *-ção* in a relationship with verbs constructed with the suffix *-ec-* (*esclarecer* *: **esclareceção*), whereas the series constituted by deverbal nouns with the suffix *-mento* in a relationship with those verbs is a dense one (*esclarecer* : *esclarecimento*);

b) the constraints between the syntactic-semantic structures of the verbs and the morphological structures of the nouns (which nominaliser suffixes (do not) correlate with

which syntactic-semantic structures of the verbs). E.g., there is no paradigmatic series comprising unergative verbs of sound emission in a relationship with event nouns with the suffix *-agem* (*gritar* *: **gritagem*), whereas there is a series correlating this type of verb with deverbal nouns with the suffix *-aria* (*gritar* : *gritaria*);

c) the constraints between the syntactic-semantic structures of the verbs, the morphological structures of the nouns and the semantic structures of the nouns (which syntactic-semantic structures of the verbs (do not) correlate with which morphological structures of the nouns and with which semantic structures of the nouns). E.g., there is no relationship between unergative verbs of sound emission and nouns with the suffix *-aria* and the meaning of PLACE (*gritar* 'to roar' * : *gritaria* *'place' vs. *gritar* 'to roar' : *gritaria* 'uproar'); whereas there is a relationship between causative verbs with nouns with the suffix *-aria* and the meaning of PLACE (*barbear* 'to shave' : *barbearia* 'barbershop');

d) multi-suffixation, that is, the different paradigmatic series the same verb may belong to, when correlated with nouns bearing suffixes working in the same lexical paradigm (e.g., *refinar* 'to refine' */refinamento /refinadura /refinagem*, in which the four nouns are deverbal event nouns).

Bearing in mind the complexity, the multidimensionality and the theoretically infinite character of word-formation paradigms, networks present advantages over other representations, since they can show:

(We use "vertex" as a correspondent of "word" and "network" as a correspondent of "paradigm".)

- the different axes (Rodrigues & Rodrigues 2018) or features underlying different paradigms, whether they are organised around semantic features or formal features;

- series and families and the correlations that a base can establish within different series (verb : event noun (*refinar* : *refinamento/ refinação /refinagem*)) and within different families (verb : event noun / verb : agent noun (*refinar* : *refinação / refinar* : *refinador*)), that is, the degree (number of edges attached to the vertex) of the vertices of the network(s);

- the hubs, that is, the vertices with a higher degree (e.g., the bases that have more correlations with more derived words);

- morphological competition among paradigmatic series, measuring the size and density of the networks;

- niches (Lindsay & Aronoff 2013) inside lexical paradigms, based on the semantic specialisations of paradigms;

- the potentiality of paradigms (expansion of the network), by measuring the degree of frequency, predictability and productivity of the series (cf. Hawkins & Blakeslee 2004, Plag & Baayen 2009, Bell & Schäfer 2013; 2016);

- the correlation between the morphological complexity of the paradigm (bearing in mind the geodesic distance between vertices) and its regularity and saturation (Körtvélyessy 2015);

- cross-paradigms (Rodrigues & Rodrigues 2018), that is, "paradigms that interface with one another, in a structured network, by means of a feature that is shared by the several paradigms involved" (Rodrigues & Rodrigues 2019), and their new developments (expansion of the network) (e.g., Rodrigues and Rodrigues (2019) analyse the case of nouns with the suffix *-ção* which has come to acquire a new meaning of 'intensity/iteration' in Brazilian Portuguese').

References

Bell, Melanie J. & Martin Schäfer. 2013. Semantic transparency: challenges for distributional semantics. In Aurelie Herbelot, Roberto Zamparelli & Gemma Boleda (eds.), *Proceedings of*

the IWCS 2013 workshop: Towards a formal distributional semantics, 1–10. Potsdam: Association for Computational Linguistics.
Bell, Melanie. J. & Martin Schäfer. 2016. Modelling semantic transparency. *Morphology* 26(2). 157–199.
Bonami, Olivier & Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29(2). 167–197.
Elman, Jeffrey. 2011. Lexical knowledge without a lexicon? *The Mental Lexicon* 6(1). 1–33.
Gaeta, Livio & Marco Angster. 2019. Stripping paradigmatic relations out of the syntax. *Morphology* 29(2). 249–270.
Körtvélyessy, Lívia. 2015. *Evaluative morphology from a cross-linguistic perspective*. Newcastle: Cambridge Scholars Publishing.
Hathout, Nabil. 2009. Acquisition of morphological families and derivational series from a machine readable dictionary. *Décembrettes* 6. 166–180.
Hathout, Nabil & Fiammetta Namer. 2018. Defining paradigms in word formation: concepts, data and experiments. *Lingue e Linguaggio* XVII (2). 151–154.
Hathout, Nabil & Fiammetta Namer. 2019. Paradigms in word formation: what are we up to? *Morphology* 29 (2). 153–165.
Hawkins, Jeff & Sandra Blakeslee. 2004. *On intelligence*. New York: Henry Holt and Company.
Libben, Gary. 2015. Word-formation in psycholinguistics and neurocognitive research. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-formation. An international handbook of the languages of Europe*, vol. 1, 203–217. Berlin: Mouton de Gruyter.
Lindsay, Mark & Mark Aronoff. 2013. Natural selection in self-organizing morphological systems. In Nabil Hathout, Fabio Montermini, & Jesse Tseng (eds.), *Morphology in Toulouse: Selected proceedings of Décembrettes 7*, 133–153, München: Lincom.
Newman, Mark. 2010. *Networks. An introduction.* Oxford: Oxford University Press.
Plag, Ingo & Harald Baayen. 2009. Suffix ordering and morphological processing. *Language* 85(1), 109–152.
Pounder, Amanda. 2000. *Process and paradigms in word-formation morphology*. Berlin & New York: Mouton de Gruyter.
Rodrigues, Alexandra Soares. 2008. *Formação de substantivos deverbais sufixados em português.* München: Lincom.
Rodrigues, Alexandra Soares & Pedro João Rodrigues. 2018. Cross-paradigms or the interfaces of word-formation patterns: evidence from Portuguese. *Lingue e Linguaggio* XVII(2). 273–288.
Rodrigues, Alexandra Soares & Pedro João Rodrigues. 2019. Plasticity of morphological paradigms. In Jenny Audring, Nikos Koutsoukos, Christina Manouilidou (eds.), *Mediterranean morphology meetings: rules, patterns, schemas and analogy*, 98–110. Patras: University of Patras.
Štekauer, Pavol. 2014. Derivational paradigms. In Rochelle Lieber & Pavol Štekauer (eds.), *The Oxford handbook of derivational morphology*, 354–369. Oxford: Oxford University Press.
van Marle, Jaap. 1985. *On the paradigmatic dimension of morphological creativity*. Dordrecht: Foris.

Committees

Organisation and Reviewing

Organisation committee

Julien Antunes (Université Bordeaux-Montaigne, CLLE-ERSSàB)
Marie Armentia (Université Bordeaux-Montaigne, IKER)
Gilles Boyé (Université Bordeaux-Montaigne, CLLE-ERSSàB)
Helline Havet (Université Bordeaux-Montaigne, CLLE-ERSSàB)
Gauvain Schalchli (Université Bordeaux-Montaigne, CLLE-ERSSàB)

Program committee

Giorgio Francesco Arcodia (Università di Milano-Bicocca)
Matthew Baerman (University of Surrey)
Alexandra Bagasheva (SU "Kliment Ohridski")
Laurie Bauer (Victoria University of Wellington)
Olivier Bonami (Université Paris Diderot)
Geert Booij (University of Leiden)
Gilles Boyé (Université Bordeaux-Montaigne)
Berthold Crysmann (LLF, CNRS & U. Paris-Diderot)
Jesús Fernández Domínguez (University of Granada)
Bernard Fradin (LLF, CNRS & U. Paris-Diderot)
Livio Gaeta (Università di Torino)
Francesco Gardani (University of Zurich)
Hélène Giraudo (CLLE, CNRS, Toulouse)
Nabil Hathout (CLLE, CNRS, Toulouse)
Fabiola Henri (University of Kentucky)
Martin Hummel (Karl-Franzens-Universitaet, Graz)
Marianne Kilani-Schoch (University of Lausanne)
Jean-Pierre Koenig (University at Buffalo, The State University of New York)
Lior Laks (Tel-Aviv University)
Fabio Montermini (CLLE, CNRS, Toulouse)
Fiammetta Namer (UMR 7118 ATILF & Université de Lorraine)
Vito Pirrelli (ILC-CNR)
Ingo Plag (Heinrich-Heine-Universität Düsseldorf)
Franz Rainer (WU Vienna)
Gauvain Schalchli (Université Bordeaux-Montaigne)
Magda Sevcikova (Charles University, Prague)
Andrew Spencer (University of Essex)
Pavel Stichauer (Charles University, Prague)
Gregory Stump (University of Kentucky)
Anna Maria Thornton (Università dell'Aquila)
Delphine Tribout (Université de Lille)
Géraldine Walther (University of Zurich)
Jaap van Marle (Open Universiteit)