# Computational methods for descriptive and theoretical morphology: a brief introduction

Olivier Bonami, Benoît Sagot

Morphology
DOI 10.1007/s11525-017-9317-8

# Computational methods for descriptive and theoretical morphology: a brief introduction

**Olivier Bonami**[1] · **Benoît Sagot**[2]



## 1 Setting the scene

While computational morphology is a well-established sub-field of computational linguistics with important applications in Natural Language Processing (NLP), it has long been somewhat isolated from descriptive and theoretical morphology. Historically, computational approaches to morphological analysis predominantly relied on finite-state transducers, either as a framework for manually developing morphological analysers and generators (Vauquois et al. 1965; Koskenniemi 1984; Beesley and Karttunen 2003)—also for generating large-scale lexica—or for the unsupervised acquisition of morphological analyses (Creutz and Lagus 2005; Monson et al. 2008).[1] Yet none of these lines of research have managed to arouse much interest among theoretical morphologists.

Discussing the theoretical morphologists' lack of interest in manually developed finite-state descriptions, Karttunen (2003) puzzles, somewhat bitterly, over its possible cause. We suggest that it results from a mutual misunderstanding. Because morphology is comparatively simple in terms of formal complexity, computational morphologists have tended to focus on the large-scale applicability of well-understood

[1] See Roark and Sproat (2007) for an accessible and linguistically motivated introduction to finite-state approaches to morphology.

✉ O. Bonami
olivier.bonami@linguist.univ-paris-diderot.fr

B. Sagot
benoit.sagot@inria.fr

1 Université Paris–Diderot, Laboratoire de linguistique formelle, Paris, France

2 Inria, Paris, France

Published online: 31 October 2017

analytic schemes. This presented little interest for descriptive and theoretical morphologists, either because the descriptive gains of a computational approach were not obvious,[2] or because the high-level questions regarding the structure and typology of morphological systems were simply not addressed in computational work.

Computational approaches to the unsupervised acquisition of morphological structure have also failed to convince theoretical and descriptive morphologists despite the early adoption of information-theoretic approaches (Goldsmith 2001). In this case, this lack of interest can probably be analysed as a result of often morphologically unrealistic but computationally useful approximations, such as the use of strictly concatenative approaches, and of the insufficient quality of the resulting analyses.

The situation has changed radically over the last decade. Three main research strands have emerged that have made computational methods directly relevant to the practising morphologist. First, implemented morphological fragments designed to test the validity of analyses have become more commonplace among morphologists. The framework of Network Morphology (Corbett and Fraser 1993; Brown and Hippisley 2012) is at the centre of the stage in this area: from the start, Network Morphology has paired a formal theory of morphology with an implementation scheme based on the DATR language (Evans and Gazdar 1996), and network morphologists have favoured the practice of testing in detail the consequences of their analyses through large-scale implementations; however other efforts are notable that attempt to implement existing morphological theories (Stump and Finkel)[3] or design new implemented theories (Walther 2013; Sagot and Walther 2013).

Second, large-scale, morphologically-annotated lexica and other lexical databases are being developed for a growing number of languages, often using computational methods based on previous lexicographic work (Romary et al. 2004) or on corpus-based methods (Oliver et al. 2003; Sagot 2005; Zanchetta and Baroni 2005; Forsberg et al. 2006)[4] to accelerate development. These lexica have commonly been designed with different applications in mind, especially psycholinguistics (e.g. CELEX for English, German and Dutch (Burnage 1990; Baayen et al. 1993); Lexique for French (New et al. 2004)) or NLP (e.g. DELA (Courtois 1990) and Le*fff* (Sagot 2010) among many others). Yet such resources constitute a useful testing ground for morphological theories and ideas, by enabling experiments on whole systems rather than focusing on small samples of interesting cases. This area is important enough that the development of databases that document in detail lexical features of prime importance to morphologists, such as phonological representations (Bonami et al. 2014; Hathout et al. 2014) or derivational relations (Hathout and Namer 2014; Žabokrtský et al. 2016), has become a lively field of research. It is also notable that the free availability of an increasing number of resources, including all resources cited in this paragraph published in this century, favours their use for descriptive purposes.

[2]This is despite the existence of large-scale efforts to develop morphological resources not only for major languages but also for minority or less-described languages. In this regard, see for instance the work carried out in the Apertium and Giellatekno projects (respectively https://svn.code.sf.net/p/apertium/svn/languages and https://victorio.uit.no/langtech/trunk/langs).

[3]http://www.cs.uky.edu/~raphael/linguistics/claw.html.

[4]With a different point of view, see also Bender et al. (2014), Zamaraeva (2016).

Finally, morphologists are increasingly relying on computational quantitative methods to infer automatically analyses of morphological systems from raw or annotated data.[5] Two notable lines of research concern implicative structure and inflectional classification. Work on implicative structure has employed set-theoretic (Stump and Finkel 2013) or information-theoretic methods (Ackerman et al. 2009; Ackerman and Malouf 2013; Bonami and Beniamine 2016) to examine the predictability of forms in inflectional paradigms. Work on inflectional classification has mostly used information-theoretic criteria to construct optimal classifications of inflectional systems (Walther and Sagot 2011; Brown and Evans 2012; Lee 2014; Beniamine et al. 2017). These lines of research present a promising convergence with an emerging interest within computational linguistics for the inference of full inflectional paradigms from partial information (Durrett and DeNero 2013; Ahlberg et al. 2014; Nicolai et al. 2015). This issue has recently been the focus of a shared task (Cotterell et al. 2017), whereby the performance of competing systems was assessed on a single collection of datasets. Although they were not the only competitors, approaches based on deep neural networks dominated the competition. Such approaches have the potential of providing new ways of assessing the structure of inflection systems, as the paper by Malouf in the present issue demonstrates.

## 2 The present special issue

This special issue grew out of a workshop with the same title organised as part of the 17th International Morphology Meeting (Vienna, Austria, February 18–21, 2016). It shares with that workshop the goal of showcasing recent advances in descriptive and theoretical morphology that make crucial use of computational methods.

The first three papers each address an important morphological problem through computational modelling.

The main goal of Rob Malouf's paper is to establish that inflectional systems can be learned without any explicit information on the shape of representations to be learned. To this end, Malouf designs a recurrent neural network intended to learn a paradigm function in the sense of Stump (2001): the network input is a pairing of a lexeme identifier and a paradigm cell, and the output is a phonological form. The paper documents models for datasets in 7 different languages, attaining an accuracy varying between 86% (Irish nouns) and 99.9% (French verbs) depending on the dataset. This in itself is quite an important result, showing that conventional knowledge on the adequacy of neural networks for the modelling of morphology needs to be reassessed in light of advances in deep learning. Perhaps the most interesting (if tentative) part of the paper is an analysis of network weights for a network trained on French verbs. Malouf shows that the network learned to classify phonemes in natural classes, to group paradigm cells by stem alternants, and to segment words into stems

[5] This is part of a more general trend of using experimental and quantitative methods in morphology, that is well represented in the pages of this journal. Here we focus on work that presents specific computational challenges, rather than the mere use of off-the-shelf statistical analysis software. We also focus on work whose empirical basis is linguistic rather than psycholinguistic (see for instance, among many others, Keuleers 2008; Pirrelli et al. 2011; O'Donnell 2015).

and exponents. Importantly, the network was trained on basically unstructured data: the input contains no phonological or semantic information, and the output consists of raw sequences of unanalysed phonemes. Thus, the network does seem to have learned from scratch all crucial components of an analysis of the French system.

The paper by Sebastian Banks presents an attempt at evaluating typological claims on the distribution of portmanteau person markers in conjugation paradigms. Bank considers two similar but distinct claims: (i) that paradigm cells relating first and second person arguments are most likely to be realised as portmanteaux, and (ii) that paradigm cells where the subjects outranks the object in the person hierarchy are most likely to give rise to portmanteaus. As Banks emphasises, an immediate concern when trying to assess such a claim is that available morphological descriptions are often not commensurate: there is no universally agreed upon method to establish what features a morph realises; hence existing descriptions can't be trusted to have made commensurate decisions on what is a portmanteau and what isn't. Hence, the heart of the paper is the design of a segmentation algorithm that starts from a raw conjugation paradigm and, on the basis of intuitively reasonable heuristics, infers morphs and assigns content to them. Banks then goes on to apply his algorithms to datasets from 26 languages with subject and object agreement, and concludes that there is support for claim (i) but not for claim (ii).

The paper by Roland Mühlenbernd and Dankmar Enke addresses how the direction of morphosemantic change can be modelled using tools from Evolutionary Game Theory. The specific goal is to determine what factors can be hypothesised to explain tendencies in directions of evolution of imperfective aspect marker systems. Building on previous work by Deo (2015), they define a signalling game of exchange of aspectual information between speakers, and go on to run simulations where mixed populations of learning and mature agents exchange information over time and make the system evolve until it reaches a stable state. The simulations show that a combination of three conjectures produces a distribution of outcomes that approaches the observed distribution in the languages of the world: (i) hearers do not always have access to contextual cues indicating whether the speaker reports a habitual or progressive situation; (ii) using multiple markers when a single one would do comes with a cost; (iii) children are exposed to a comparatively higher proportion of reports of progressive situations.

The remaining three papers pertain to the field of (morphological) computational grammar engineering, but have very different scope.

The paper by Berthold Crysmann shows how a detailed analysis of various reduplication patterns can be integrated into a large-scale, detailed computational HPSG grammar of Hausa running on the LKB system (Copestake 2002). Hausa inflection uses intricate patterns of both partial and total reduplication. Crysmann argues that the two should be kept strictly separate, but develops a detailed analysis that highlights both what they have in common (sharing of structure) and where they differ (the phonological or morphological nature of the shared structure). Two features of the analysis should be highlighted: First, the analysis captures in a fully declarative fashion the intricate diversity of patterns of similarity and difference between base and reduplicant; hence there is no argument in favour of a processual view of morphology to be drawn from reduplication. Second, the paper shows how total redupli-

cation can be fully integrated in the grammar by postulating a morphosyntactic word with no autonomous interpretation.

The paper by Atticus Harrigan and co-authors also focuses on the design and implementation of a computational grammar. The system of interest is Plains Cree conjugation, and the grammar relies exclusively on finite-state methods. Starting from a descriptive overview of the system, the authors describe and motivate a detailed finite state analysis that takes into account in particular multiple patterns of partial reduplication, circumfixal markers, and the direct/inverse organisation of the person marking system. The system is evaluated on a manually-annotated corpus, exhibiting a performance comparable to that of similar endeavours. Of particular linguistic interest is the final section where statistics on the morphological profile of the language are derived from the corpus, providing an insightful characterisation of the enumerative complexity (in the sense of Ackerman and Malouf 2013) of a polysynthetic language.

Finally, the paper by Jane Chandlee takes a step back on the development of morphological grammars by investigating the formal complexity of morphological mechanisms. She investigates the formal expressive power necessary to encode "morphological maps," i.e. mappings from an underlying representation of an inflected form to the surface form itself, based on a large inventory of realisational mechanisms in multiple typologically distinct languages. Although several realisational mechanisms raise interesting and not fully solved issues, and a few others are not covered by this study, the conclusion of the author is that morphology generally requires a limited expressive power that is a strict subset of regular finite-state transducers. If expressive power is used as a way to assess complexity, this would show that morphology is significantly less computationally complex than syntax.

**Acknowledgements** We thank our handling editor Adam Albright for his support. This work, as well as the 2016 workshop it is a follow-up to, was partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083).

## References

Ackerman, F., & Malouf, R. (2013). Morphological organization: the low conditional entropy conjecture. *Language*, *89*, 429–464.

Ackerman, F., Blevins, J. P., & Malouf, R. (2009). Parts and wholes: implicative patterns in inflectional paradigms. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar* (pp. 54–82). Oxford: Oxford University Press.

Ahlberg, M., Forsberg, M., & Hulden, M. (2014). Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, EACL '14.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1993). *The CELEX lexical data base on CD-ROM*.

Beesley, K. R., & Karttunen, L. (2003). *Finite state morphology. Studies in computational linguistics*. Stanford: CSLI Publications.

Bender, E. M., Crowgey, J., Goodman, M. W., & Xia, F. (2014). Learning grammar specifications from IGT: a case study of Chintang. In *Proceedings of the 2014 Workshop on the use of computational methods in the study of Endangered languages*, ACL 2014.

Beniamine, S., Bonami, O., & Sagot, B. (2017, in press). Inferring inflection classes with description length. *Journal of Language Modelling*.

Bonami, O., & Beniamine, S. (2016). Joint predictiveness in inflectional paradigms. *Word Structure*, *9*(2), 156–182.

Bonami, O., Caron, G., & Plancq, C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer, & S. Prévost (Eds.), *Actes du quatrième Congrès Mondial de Linguistique Française* (pp. 2583–2596).

Brown, D., & Evans, R. (2012). Morphological complexity and unsupervised learning: validating Russian inflectional classes using high frequency data. In F. Kiefer, M. Ladányi, & P. Siptár (Eds.), *Current Issues in Morphological Theory: (Ir)regularity, analogy and frequency* (pp. 135–162). Amsterdam: John Benjamins.

Brown, D., & Hippisley, A. (2012). *Network Morphology: a defaults based theory of word structure*. Cambridge: Cambridge University Press.

Burnage, G. (1990). *Celex: a guide for users* (Tech. rep.). University of Nijmegen, Center for Lexical Information.

Copestake, A. (2002). *Implementing typed feature structure grammars*. Stanford: CSLI Publications.

Corbett, G. G., & Fraser, N. M. (1993). Network morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics*, *29*, 113–142.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J., & Hulden, M. (2017). CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages. In *Proceedings of the CoNLL SIGMORPHON 2017 shared task: universal morphological reinflection*, Vancouver, Canada (pp. 1–30).

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, *87*(1), 11–22.

Creutz, M., & Lagus, K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the international and interdisciplinary conference on adaptive knowledge representation and reasoning*, Espoo, Finland (pp. 106–113).

Deo, A. (2015). The semantic and pragmatic underpinnings of grammaticalization paths: the progressive to imperfective shift. *Semantics and Pragmatics*, *8*(14), 1–52.

Durrett, G., & DeNero, J. (2013). Supervised learning of complete morphological paradigms. In *Proceedings of the North American chapter of the association for computational linguistics*, NAACL '13, Atlanta, Georgia, United States (pp. 1185–1195).

Evans, R., & Gazdar, G. (1996). DATR: a language for lexical knowledge representation. *Computational Linguistics*, *22*, 167–216.

Forsberg, M., Hammarström, H., & Ranta, A. (2006). Morphological lexicon extraction from raw text data. In *LNAI: Vol. 4139*. *Proceedings of FinTAL 2006*, Turku, Finland (pp. 488–499). Berlin: Springer.

Goldsmith, J. A. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, *27*, 153–189.

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

Hathout, N., Sajous, F., & Calderone, B. (2014). GLÀFF, a large versatile French lexicon. In *Proceedings of LREC 2014*.

Karttunen, L. (2003). Computing with realizational morphology. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 205–216). Heidelberg: Springer.

Keuleers, E. (2008). *Memory-based learning of inflectional morphology*. PhD thesis, University of Antwerp.

Koskenniemi, K. (1984). A general computational model for word-form recognition and production. In *Proceedings of the 22nd annual meeting of the Association for Computational Linguistics*, ACL '84, Stanford, California, United States, (pp. 178–181).

Lee, J. (2014). *Automatic morphological alignment and clustering* (Tech. rep.). University of Chicago.

Monson, C., Carbonell, J., Lavie, A., & Levin, L. (2008). Paramor: finding paradigms across morphology. In C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, A. Peñas, V. Petras, & D. Santos (Eds.), *Lecture notes in computer science (LNCS): Vol. 5152*. *Advances in multilingual and multimodal information retrieval, revised selected papers from the 8th workshop of the cross-language evaluation forum (CLEF 2007)*, Budapest, Hungary (pp. 900–907). Berlin: Springer.

New, B., Brysbaert, M., Segui, J., Ferrand, L., & Rastle, K. (2004). The processing of singular and plural nouns in French and English. *Journal of Memory and Language*, *51*, 568–585.

Nicolai, G., Cherry, C., & Kondrak, G. (2015). Inflection Generation as Discriminative String Transduction. In *Proceedings of the 2015 conference of the North American chapter of the association*

*for computational linguistics: human language technologies*, NAACL-HLT '15, Denver, Colorado, United States (pp. 922–931).

O'Donnell, T. J. (2015). *Productivity and reuse in language*. Cambridge: MIT Press.

Oliver, A., Castellón, I., & Màrquez, L. (2003). Use of Internet for augmenting coverage in a lexical acquisition system from raw corpora: application to Russian. In *Proceedings of the RANLP'03 international workshop on information extraction for Slavonic and other Central and Eastern European languages (IESL'03)*, Borovets, Bulgaria.

Pirrelli, V., Ferro, M., & Calderone, B. (2011). Learning paradigms in time and space. computational evidence from romance languages. Morphological autonomy: Perspectives from Romance inflectional morphology (pp. 135–157).

Roark, B., & Sproat, R. (2007). *Computational approaches to morphology and syntax*. Oxford: Oxford University Press.

Romary, L., Salmont-Alt, S., & Francopoulo, G. (2004). Standards going concrete: from Lmf to Morphalou. In *Proceedings of coling 2014*.

Sagot, B. (2005). Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture notes in artificial intelligence: Vol. 3658. Proceedings of TSD'05*, Karlovy Vary, Czech Republic (pp. 156–163). Berlin: Springer.

Sagot, B. (2010). The *Lefff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of LREC 2010*.

Sagot, B., & Walther, G. (2013). Implementing a formal model of inflectional morphology. In *Proceedings of systems and frameworks in computational morphology* (pp. 115–134).

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

Stump, G. T., & Finkel, R. (2013). *Morphological typology: from word to paradigm*. Cambridge: Cambridge University Press.

Vauquois, B., Veillon, G., & Veyrunes, J. (1965). Application des grammaires formelles aux modèles linguistiques en traduction automatique. *Kybernetika*, *1*(3), 281–289.

Walther, G. (2013). *De la canonicité en morphologie: perspective empirique, théorique et computationnelle*. PhD thesis, Université Paris Diderot.

Walther, G., & Sagot, B. (2011). Modélisation et implémentation de phénomènes flexionnels non-canoniques. *Traitement Automatique des Langues*, *52*(2), 91–122.

Žabokrtský, Z., Ševčíková, M., Straka, M., Vidra, J., & Limburská, A. (2016). Merging data resources for inflectional and derivational morphology in Czech. In *Proceedings of the 10th international conference on language resources and evaluation* (pp. 1307–1314).

Zamaraeva, O. (2016). Inferring morphotactics from interlinear glossed text: combining clustering and precision grammars. In *Proceedings of the 14th SIGMORPHON workshop*, Berlin.

Zanchetta, E., & Baroni, M. (2005). Morph-it! a free corpus-based morphological resource for the Italian language. In *Proceedings of the corpus linguistics conference*, Birmingham, United Kingdom (pp. 1–12).