Adding Glawinette into Démonette: pratical consequences and theoretical questions

Nabil Hathout
CLLE, CNRS
Université de Toulouse

Fiammetta Namer
Université de Lorraine
ATILF, CNRS

Abstract

Glawinette is a derivational lexicon of French made up of morphological families and morphological series. It has been acquired automatically from GLAWI, a large machine readable dictionary and contains about 100 000 pairs of morphologically related lexemes. In this paper, we present Glawinette and discuss how we plan to include this new resource into the Démonette derivational database, what changes this may bring to the architecture of this database and how this inclusion will raise several theoretical questions regarding the content of Démonette and the nature of derivational paradigms.

1 Introduction

Glawinette (Hathout et al., 2020) is a newly created resource which provides a description of derivational morphology of French on a large scale. In this paper, we discuss its inclusion into the Démonette database (Hathout and Namer, 2014, 2016; Namer et al., 2019; Namer and Hathout, 2020). This will increase the size of Démonette and test the robustness of the principles underlying the structure of Démonette and its description formats by confronting them with the diversity of derivational relations contained in Glawinette. In addition, it will involve a manual revision of Glawinette. Verification and inclusion into Démonette will be done in batches, starting with the most reliable lexeme pairs and lexeme clusters. Batch processing will also make it easier to complement the descriptions of Glawinette and to fill in the semantic fields of Démonette.

2 Glawinette

Glawinette is a derivational morphological lexicon of French built from the GLAWI machine readable dictionary (Sajous and Hathout, 2015; Hathout and Sajous, 2016). Like Démonette, and before it Morphonette (Hathout, 2011a), Glawinette is a lexicon of derivational relations which enables a smooth and easy integration into Démonette. Morphological relations (i.e. pairs of morphologically related lexemes) are acquired from the definitions of GLAWI and the morphological sections of this dictionary. Specifically, these relations are extracted from the so-called morphological definitions (Martin, 1983), i.e., definitions where the *definiendum* is a complex lexeme whose meaning is described by a *definiens* that contains a member of its morphological family, as in (1) that link *glaçon* 'ice cube' to *glace* 'ice' and *développement* 'development' to *développer* 'to develop'. In these examples, the morphological relations are direct (base → derivative), but this is not always the case as in (2) where *conservation* 'conservation' is not the base for *conservateur* 'preservative'.

(1) a. *glaçon = morceau de glace* 'piece of ice'

b. *développement = action de développer, de se développer ou résultat de cette action, au propre et au figuré* 'act of developing or result of this action, literally and figuratively'

(2) *conservateur = substance chimique minérale ou organique, ajoutée aux aliments afin d'améliorer leur conservation* 'chemical substance, mineral or organic, added to food to improve its preservation'

Glawinette proposes a description of morphological relations within two fundamental structures in the paradigmatic organization of derivational morphology (Bochner, 1993; Van Marle, 1985; Bauer, 1997; Štekauer, 2014; Hathout and Namer, 2018, 2019; Bonami and Strnadová, 2019; Namer and Hathout, 2020): morphological families and morphological series (Roché, 2009; Hathout, 2011b; Fradin, 2018). In Glawinette, morphological families are related graphs of derivational relations like (3) which presents the family of the noun *prince* 'prince'. In addition, every relation (i.e. lexeme pairs) is part of a morphological series as in (4) which presents a part of the series that connects agent nouns in *-eur* and action nouns in *-ion*. The series are labeled by patterns consisting of two regular expressions that contain the same number of sequences `(.+)` and where these sequences represent the same strings.

(3) prince=N:princesse=N 'princess' prince=N:princier=A 'princely' prince=N:princillon=N 'petty prince' prince=N:princiser=V 'make become a prince' princesse=N:prince=N princier=A:prince=N princier=A:princièrement=R 'princely' princillon=N:prince=N princiser=V:prince=N princièrement=R:princier=A

(4)

`^(.+)eur$=N`	`^(.+)ion$=N`		
acteur	action	'actor'	'action'
animateur	animation	'animator'	'animation'
classificateur	classification	'classifier'	'classification'
formateur	formation	'trainer'	'training'

On the one hand, Glawinette takes advantage of the fact that lexeme pairs that enter into regular morphological relations form formal analogies (Lepage, 1998, 2004; Stroppa and Yvon, 2005; Hathout, 2008; Langlais and Yvon, 2008; Arndt-Lappe, 2015; Fam and Lepage, 2018, 2021), for example, *acteur=N:action=N::animateur=N:animation=N*. These analogies are directly acquired from the morphological definitions and morphological sections of GLAWI. On the other had, sets of relations such as (4) are made up of two sets of lexemes (the left and right columns) that exhibit morphologically relevant regularities. For example, all words in the left column of (4) contain a final sequence `eur`, all words in the right one contain a final sequence `ion`. Moreover, these sequences are morphologically relevant because the stem of the lexeme pairs in each line are identical (for example, we have the same stem `animat` in the two lexemes of the second line). Glawinette is also distinguished by its ability to describe the morphological series by means of "natural" patterns that are very similar to the ones used by linguists to characterize complex lexemes. For example, the relation *activiste=N:activisme=N* will be characterized by the pattern `^(.+)iste$=N/^(.+)isme$=N` and not `^(.+)te$=N/^(.+)me$=N` nor `^(.+)t(.+)$=N/^(.+)m(.+)$=N`. Glawinette contains 97 293 lexemes connected by 47 712 relations. These relations are divided into 15 904 morphological families and 5 400 series. Note that some of the relations described in Glawinette are already present in Démonette. This intersection will be used to complement the morphological descriptions of the relations in Glawinette.

3 Some "practical" consequences of the inclusion of Glawinette in Démonette

Glawinette will provide Démonette with more complete and varied morphological families. The families of Glawinette contain a large number of relations not yet covered in Démonette. This integration will test the capacity of the database architecture to describe a representative fragment of French morphological relations, which potentially are more complex than the ones currently described in Démonette. For example, they contain derivationally distant pairs such as *déformer=V:indéformable=A* 'to deform:undeformable' where *déformer* is a second level ascendant of *indéformable* (*déformer* → *déformable* 'deformable' → *indéformable*). This type of relation is hardly present in the current version of Démonette. Their inclusion will test the robustness of the tagsets used in Démonette.

The other interesting feature of the Glawinette relations is that they are semantically relevant as they are directly derived from definitions (and morphological sections). However, the relations of Glawinette, like the ones from other resources used to create Démonette, do not contain semantic characterization. Completion of these descriptions will be the main challenge in integrating Glawinette into

Démonette. Several paths will be explored for these descriptions. On the one hand, there will be a semi-automatic shallow completion at the level of the series of specific relations. For example, we can specify that the *-ion* derivatives in (4) are action nouns and propose for the pair *formateur=N:formation=N* 'trainer:training' a gloss such as 'a trainer carries out a training' which could be later completed in 'a trainer carries out a training of people to whom he teaches new skills'. Another strategy will take advantage of clusters of relations within the families to leverage the semantic descriptions of some of them, e.g., to predict the gloss of an indirect relation, or cross-formation (Becker, 1993), or that of a complex relation (e.g. *déformer:indéformable*) from existing, base-to-derivative glosses, defining the lexemes involved in these indirect and complex relations. For example, the pair *déformer:indéformable* can be glossed as 'something undeformable cannot be deformed' by superposition and adaptation of the glosses of the direct relations *déformable:indéformable* 'what is undeformable is not deformable' and *déformer:déformable* 'something deformable can be deformed'. The integration of the pairs from Glawinette will also involve a revision of the exponents of the morphological processes. For example, the pattern `^re(.+)er$=V/^(.+)er$=V` will be replaced by the pattern `^re(.+)$=V/^(.+)$=V` which is a more appropriate level of generalization as prefixation in *re-* is not limited to verbs of the first conjugation (with infinitives ending in *-er*). Series is the right level of granularity to make this kind of decision because it gathers homogeneous sets of similar relations. Moreover, families give a more complete view of all the specific derivational relations that hold between its lexemes.

4 Feeding Démonette with relations from Glawinette

The series of Glawinette will be integrated into Démonette one by one. These series are characterized by their yield (that is, the number of lexeme pairs they contain) and by the properties of the patterns that define them: the (cumulative) length of the patterns, the specificity of the exponents (i.e. the ratio of the number of words that match a pattern in the whole lexicon to the number of pairs contained in the series of relations, (Bybee, 1988)), and their versatility (i.e. the overall number of connections of the lexemes identified by the pattern). These features enable us to estimate the quality of the pairs contained in a series, to process the most reliable ones first and to devote more resources for the ones that are likely to contain errors. For example, the series `^(.+)er$=V/^(.+)age$=N` contains 1465 pairs that normally contains no errors. Conversely, the series `^(.+)tte$=N/^(.+)lle$=N` contains only the erroneous pair *batte=N:balle=N* 'bat:ball'. The very small size of the stem (`ba` contains only 2 characters) is an additional clue to this error. However, not all series that contain few pairs are incorrect, especially the ones with sufficiently long patterns like `^(.+)anisme$=N/^(.+)éen$=A` which only contains *européanisme=N:européen=A* 'europeanism:european'. By combining a number of such criteria, we can quickly identify potentially erroneous pairs and series that are most cost-effective to include in Démonette.

5 New theoretical questions

The inclusion of Glawinette in Démonette also contributes to the debate on several current theoretical issues in morphology. The families and series of Glawinette are the source material from which morphological paradigms can be built. The creation of these paradigms from the morphological series remains an open question that Glawinette will help clarify. They will lead to complement the architecture of Démonette with additional tables that will represent this paradigmatic organization (morphological families, morphological paradigms). This is not a trivial evolution because these structures are defined on top of multiple, redundant and unconstrained relational descriptions. At first, we will only include the derivational relations from Glawinette.

Various future decisions regarding the relations encoded in Démonette will be reconsidered with respect to the content of Glawinette. First, the relations in Démonette are symmetrical by design, whether direct base-to-derivative, complex ancestor-to-descendent or indirect between siblings. Whenever Démonette contains an entry (word1, word2), it also includes the corresponding (word2, word1) entry described by means of feature values that are symmetrical to the ones of (word1, word2). However, we observe that the morphological relations originating from the GLAWI dictionary are not symmetrical, and this will lead us to rethink the conditions of the systematic symmetrization of the entries in Démonette.

Second, the presence in Glawinette of lexeme pairs that are in complex relations like `^(.+)er$=V/^in(.+)able$=A` confirms the relevance of this type of relations and validates their description in Démonette. Moreover, these pairs empirically validate the intuition of speakers who unconsciously reanalyze these sequences as affixes in their own right (see Stump (2017, 2019) for a theoretical account of this phenomenon he calls "rule conflation").

On the other hand, the observation of indirect relations in Glawinette questions the systematic description of all indirect relations in Démonette. For example, the series `^(.+)eur$=N/^(.+)ion$=N` contains only 285 pairs in Glawinette when the series `^(.+)er$=V/^(.+)ation$=N` contains 1322 ones. Yet when a verb is the base of an action noun in -ation, then it should also be the base of an agent noun in -ateur: therefore, we would have expected similar figures for the two series. The integration of Glawinette thus leads to two questions: (*i*) explain the shift; (*ii*) account for it in Démonette, for example by completing the graphs (i.e. the families) on the fly according to users' wishes.

Finally, Glawinette may call into question theoretical certainties about the indentity of rule exponents. For instance, Glawinette contains 122 `^(.+)er$=V/^(.+)ion$=N` pairs compared to the 1322 `^(.+)er$=V/^(.+)ation$=N` series above; this calls into question the common conception that *-ation* is an allomorphic variant of *-ion* where the sequence /at/ is part of the verb stem(Bonami et al., 2009). In view of these numbers, it seems legitimate to consider *-ation* as an exponent in its own right and to adapt the description of these derivatives in Démonette accordingly. Conversely, the relations in Glawinette are essentially determined by the formal regularities that exist in the lexicon. Their inclusion in Démonette will impose to dissociate their formal, categorical and semantic components and will highlight the multiplicity of the possible generalizations.

6 Perspective

In the short term, we plan to integrate most of the relations of Glawinette into Démonette, which will significantly increase the number of entries in the database and the diversity of indirect and complex relations. This extension will provide additional material to conduct experimental and quantitative morphology experiments. The next step will be to exploit the definitions in GLAWI to generate glosses for the lexeme pairs in Glawinette. These glosses will then be used to feed the semantic section of Démonette. Finally, we plan to build a phonological version of Glawinette by combining the phonological transcriptions in GLAWI and in the lexeme table of Démonette in order to the characterize phonological operations and provide phonological patterns that will be used to complement the phonological fields of Démonette.

References

Sabine Arndt-Lappe. 2015. Word-formation and analogy. In Ingeborg Müller, Peter O.and Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation. An international handbook of the languages of Europe*, de Gruyter Mouton, Berlin/Boston, pages 822–841.

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

Thomas Becker. 1993. Back-formation, cross-formation, and 'bracketing paradoxes' in paradigmatic morphology. *Yearbook of Morphology* 1992:1–25.

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

Olivier Bonami, Gilles Boyé, and Françoise Kerleroux. 2009. L'allomorphie radicale et la relation flexion-construction. In Bernard Fradin, Françoise Kerleroux, and Marc Plénat, editors, *Aperçus de morphologie du français*, Presses universitaires de Vincennes, Saint-Denis, pages 103–125.

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

Joan L. Bybee. 1988. Morphology as lexical organization. In Michael Hammond and Michael Noonan, editors, *Theoretical Morphology. Approaches in Modern Linguistics*, Academic Press, San Diego, CA, pages 119–141.

Rashel Fam and Yves Lepage. 2018. Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan, page 1060–1066.

Rashel Fam and Yves Lepage. 2021. A study of analogical density in various corpora at various granularity. *Information* 12(8).

Bernard Fradin. 2018. Paradigms and the role of series in derivational morphology. *Lingue e Linguaggio* 2/2018:155–172.

Nabil Hathout. 2008. Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *Proceedings of the Coling workshop Textgraphs-3*. ACL, Manchester, pages 1–8.

Nabil Hathout. 2011a. Morphonette: a paradigm-based morphological network. *Lingue e linguaggio* 2011(2):243–262.

Nabil Hathout. 2011b. Une approche topologique de la construction des mots : propositions théoriques et application à la préfixation en *anti-*. In *Des unités morphologiques au lexique*, Hermès Science-Lavoisier, Paris, pages 251–318.

Nabil Hathout and Fiammetta Namer. 2014. La base lexicale Démonette : entre sémantique constructionnelle et morphologie dérivationnelle. In *Actes de la 21e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2014)*. ATALA, Marseille, pages 208–219.

Nabil Hathout and Fiammetta Namer. 2016. Giving lexical resources a second life: Démonette, a multi-sourced morpho-semantic network for French. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia.

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

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

Nabil Hathout and Franck Sajous. 2016. Wiktionnaire's Wikicode GLAWIfied: a workable French machine-readable dictionary. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia.

Nabil Hathout, Franck Sajous, Basilio Calderone, and Fiammetta Namer. 2020. Glawinette: a linguistically motivated derivational description of French acquired from GLAWI. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, pages 3870–3878.

Philippe Langlais and François Yvon. 2008. Scaling up analogical learning. In *Proceedings of the 22nd international conference on Computational Linguistics (COLING 2008)*. Manchester, UK, pages 51–54.

Yves Lepage. 1998. Solving analogies on words: An algorithm. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics*. Montréal, volume 2, pages 728–735.

Yves Lepage. 2004. Lower and higher estimates of the number of true analogies between sentences contained in a large multilingual corpus. In *Proceedings of the 20th international conference on Computational Linguistics (COLING-2004)*. Genève, pages 736–742.

Robert Martin. 1983. *Pour une logique du sens*. Linguistique nouvelle. Presses universitaires de France, Paris.

Fiammetta Namer, Lucie Barque, Olivier Bonami, Pauline Haas, Nabil Hathout, and Delphine Tribout. 2019. Demonette2 - une base de données dérivationnelles du français à grande échelle : premiers résultats. In *Actes de la 26^{e} conférence annuelle sur le traitement automatique des langues naturelles (TALN 2019)*. Toulouse, pages 233–243.

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

Michel Roché. 2009. Pour une morphologie *lexicale*. In *La morphologie lexicale est-elle possible ?*, Éditions Peeters, Leuven, volume 17 of *Mémoires de la Société de Linguistique, Nouvelle Série*, pages 65–87.

Franck Sajous and Nabil Hathout. 2015. GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*. Herstmonceux, England, pages 405–426.

Nicolas Stroppa and François Yvon. 2005. An analogical learner for morphological analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*. ACL, Ann Arbor, MI, pages 120–127.

Gregory Stump. 2017. Rule conflation in an inferential-realizational theory of morphotactics. *Acta Linguistica Academica* 64:79–124.

Gregory T. Stump. 2019. Some sources of apparent gaps in derivational paradigms. *Morphology* 29(2):271–292.

Jaap Van Marle. 1985. *On the Paradigmatic Dimension of Morphological Creativity*. Foris, Dordrecht.

Pavol Štekauer. 2014. Derivational paradigms. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, Oxford, Oxford University Press, Oxford, pages 354–369.