Vlexique2.0: A rich lexicon of French verbal inflection with form-level frequencies

Sacha Beniamine, Maximin Coavoux, Olivier Bonami

Vlexique2.0: A rich lexicon of French verbal inflection with form-level frequencies

Sacha Beniamine[1], Maximin Coavoux[2], and Olivier Bonami[3]

[1]Surrey Morphology Group, University of Surrey, United Kingdom
[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[3]Laboratoire de Linguistique Formelle, Université Paris Cité, France

We present **Vlexique2.0**,[1] a lexicon of French Verbs based on Flexique1.1 (Bonami et al., 2014). It conforms to the Paralex standard (Beniamine et al., 2023) and includes phonemic and orthographic forms, as well as detailed form-level frequencies. It comprises 5273 lexemes inflected for 51 paradigm cells, and a total of 274855 forms.

Large inflected lexicons are essential for data-driven studies of inflection (Malouf et al., 2019), whether diachronic (Cathcart et al., 2022) or synchronic (Pellegrini, 2023), in morphological typology (Sims & Parker, 2016), or to obtain stimuli for psycholinguistic experiments (Copot & Bonami, accepted). For these applications, both orthography and transcription are valuable. Moreover, frequencies are critical to drawing a nuanced picture of inflectional structure. Yet, no previous French resource provided all three: Lexique (New et al., 2007, 64929 verbal forms) documents only partial paradigms. Flexique, derived from Lexique, provides full phonemic paradigms, but nothing else. Démonette2 (Namer et al., 2019) comprises full orthographic paradigms only.

To create Vlexique2.0 (figure 1), we updated Flexique, merged in forms from Démonette, added cell- and lexeme-level information, and frequency measurements. Frequencies were measured on the Open Subtitle Corpus (Lison & Tiedemann, 2016) and obtained by training a specialized morphological tagger.

Obtaining frequencies. Token frequencies cannot be reliably estimated from raw text due to syncretisms (ex 1) and homonymy both between part-of-speeches, and across distinct verbs (ex 2).[2] Unfortunately, most annotated corpora are either too small, or present skewed frequencies: for example, the first and second person are massively under-represented in journalistic corpora (eg. the French Treebank, Abeillé et al., 2003).

(1) a. Elle savait que vous **veniez**
she.3SG.F knew.IMP.3SG COMP you.2PL come.IND.IMP.2PL
'She knew that you were coming.'

b. Il faudrait que vous **veniez**
one.3SG.M must.COND.3SG COMP you.2PL come.SBJV.PRS.2PL
'You should come.'

(2) a. Je **comparais** devant une cour [...]
I.1SG appear.IND.PRS.1.SG before a court
'I appear before a court'

b. Je **comparais** juste les tailles !
I.1SG compare.IND.IMP.1.SG only the sizes !
'I was only comparing sizes' !

[1]Zenodo DOI: 10.5281/zenodo.10638682; website: https://sbeniamine.gitlab.io/vlexique/
[2]These examples were extracted from the Open Subtitle corpus.

To obtain frequencies, we needed to annotate a suitable corpus. The largest corpus of spoken French (CEFC-Orféo, 3.5M tokens, Benzitoun et al., 2016) being still too small, we resorted to the French section of the Open Subtitle corpus [400M tokens], a corpus of emulated spoken language. To disambiguate syncretic forms (ex 1), we trained a neural tagger based on `flaubert-base-cased` (Le et al., 2020) on the Universal Dependency `French_GSD` treebank version 2.6 (Guillaume et al., 2019), augmented with synthetic sentences to ensure ample data for every paradigm cell. We evaluated it on 2980 forms (1450 sentences), sampled from opensubtitles to include at least 5 occurrences of each paradigm cell. It identified verbs with 99.0 Fscore, and paradigm cells with 96.1 Fscore. We disambiguated homonym lexemes (ex 2) by using the Stanza lemmatizer (Qi et al., 2020). We then counted occurrences of inflected forms (triplets of cell, lemma and form).

Merging resources. We updated the verbal table from Flexique by adding 31 frequent lexemes, overabundant forms for 35 lexemes, and correcting 100 lexemes. We adjusted orthographic forms from Demonext by adding 13 verbs, correcting some forms, and normalizing conventions. We then added orthographic forms to Flexique, and adjusted overabundant forms using manual rules in order to keep only congruent phonemic/orthographic pairs.

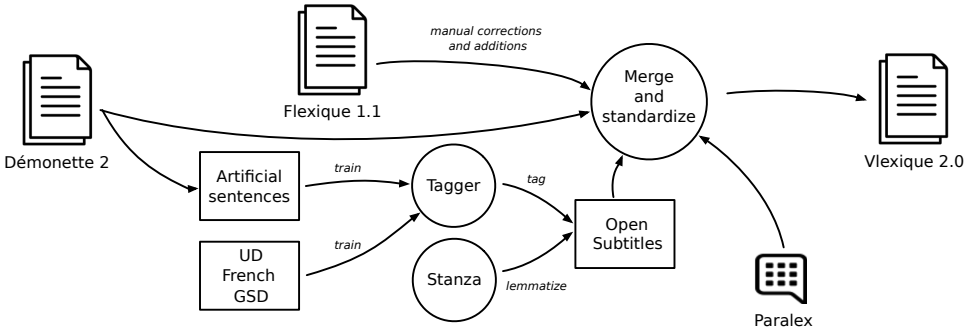

Figure 1: General pipeline

A paralex lexicon. The resulting lexicon is a relational database following the Paralex standard (Beniamine et al., 2023). The 'forms' table lists inflected forms. Each row is the combination of a paradigm cell, a lexeme, an orthographic form and a phonemic transcription. Overabundant forms constitute separate rows and are annotated with tags in order to facilitate data filtering. Token frequencies are given for each form. Cells are combinations of feature-values separated by dots (eg. 'ind.prs.1.sg'). To maximise inter-operability, the 'cells' table maps these to other schemes: GRACE (Adda et al., 1998), Flexique, Unimorph (Batsuren, 2022), Universal Dependencies (Zeman, 2023), French Treebank, and a semantic decomposition (Bonami & Boyé, 2005). Lexemes are documented in a separate table, associating the lexeme identifier to its citation form, a set of orthographic variants, and an inflection class marker. We report aggregated frequencies per lexeme and per cell in the relevant tables. The 'sounds' table describes each phoneme using distinctive features. A 'feature-value' table describes each value composing the cell identifiers. A 'tags' table documents the labels used to distinguish defective and overabundant forms.

Conclusion. The verbal subset of the Flexique lexicon has supported numerous morphological studies (among which Bonami & Beniamine, 2016; Malouf, 2017; Copot & Bonami, accepted). This updated and enhanced version will enable an even wider range of work. It is released openly (under CC BY-SA 4.0) and formatted according to the Paralex standard. We also release our pre-trained tagger.[3]

[3] https://doi.org/10.5281/zenodo.10697867

References

Abeillé, A., Clément, L., & Toussenel, F. (2003). *Building a Treebank for French*, (pp. 165–187). Dordrecht: Springer Netherlands.
URL https://doi.org/10.1007/978-94-010-0201-1_10

Adda, G., Mariani, J., Lecomte, J., Paroubek, P., & Rajman, M. (1998). The grace french part-of-speech tagging evaluation task. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, (pp. 433–441).

Batsuren, K. & al. (2022). UniMorph 4.0: Universal Morphology. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.) *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, (pp. 840–855). Marseille, France: European Language Resources Association.
URL https://aclanthology.org/2022.lrec-1.89

Beniamine, S., Anderson, C., Carroll, M., Naranjo, M. G., Herce, B., Pellegrini, M., Round, E., Sims-Williams, H., & Tresoldi, T. (2023). Paralex: a DeAR standard for rich lexicons of inflected forms. In *Presentation at International Symposium of Morphology*. https://www.paralex-standard.org.
URL https://ismo2023.ovh/fichiers/abstracts/4_ISMO_2023_Paralex.pdf

Benzitoun, C., Debaisieux, J.-M., & Deulofeu, H.-J. (2016). Le projet orféo: un corpus d'étude pour le français contemporain. *Corpus*, (15).

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

Bonami, O., & Boyé, G. (2005). French Pronominal Clitics and the Design of Paradigm Function Morphology. In G. Booij, L. Ducceschi, B. Fradin, A. Ralli, E. Guevara, , & S. Scalise (Eds.) *Fifth Mediterranean Morphology Meeting*, (pp. 291–322). Fréjus, France.
URL https://shs.hal.science/halshs-00276797

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).

Cathcart, C., Herce, B., & Bickel, B. (2022). Decoupling speed of change and long-term preference in language evolution: Insights from romance verb stem alternations. In *Proceedings of the Joint Conference on Language Evolution (JCoLE)*. JCoLE.
URL https://doi.org/10.5167/uzh-220700

Copot, M., & Bonami, O. (accepted). Behavioural evidence for implicative paradigmatic relations. *The Mental Lexicon*.

Guillaume, B., de Marneffe, M.-C., & Perrier, G. (2019). Conversion et améliorations de corpus du Français annotés en Universal Dependencies. *Traitement Automatique des Langues*, *60*(2), 71–95.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., & Schwab, D. (2020). FlauBERT: Unsupervised language model pre-training for French. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.) *Proceedings of the Twelfth Language Resources and Evaluation Conference*, (pp. 2479–2490). Marseille, France: European Language Resources Association.
URL https://aclanthology.org/2020.lrec-1.302

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, (pp. 923–929). Portorož, Slovenia: European Language Resources Association (ELRA). See http://www.opensubtitles.org/.
URL https://aclanthology.org/L16-1147

Malouf, R. (2017). Abstractive morphological learning with a recurrent neural network. *Morphology*, *27*(4), 431–458.

Malouf, R., Ackerman, F., & Semenuks, A. (2019). Lexical databases for computational analyses: A linguistic perspective. In *Proceedings of the Society for Computation in Linguistics*.

Namer, F., Barque, L., Bonami, O., Haas, P., Hathout, N., & Tribout, D. (2019). Demonette2 — Une base de données dérivationnelles du français à grande échelle : premiers résultats. In *Actes de TALN*. Toulouse, France.
URL https://halshs.archives-ouvertes.fr/halshs-02275652/document

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*, 661–677.

Pellegrini, M. (2023). *Paradigm Structure and Predictability in Latin Inflection: An Entropy-based Approach*. Springer Cham.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
URL https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Sims, A., & Parker, J. (2016). How inflection classes work: On the informativity of implicative structure. *Word Structure*, *9*(2), 215–239.

Zeman, D. & al. (2023). Universal dependencies 2.13. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
URL http://hdl.handle.net/11234/1-5287