Flexique: an inflectional lexicon for spoken French

Olivier Bonami   Gauthier Caron   Clément Plancq
U. Paris-Sorbonne, IUF   U. de la Réunion   CNRS
Laboratoire de Linguistique Formelle (U. Paris Diderot & CNRS)

Version 0.91 — June 2013

Flexique was designed as a tool for studying the structure of the French inflection system.[1] In its current form it consists of three tables of French nouns, adjectives and verbs:

POS	lexemes	words
nouns	31,002	65,111
adjectives	11,252	45,008
verbs	4,987	253,174
total	50,461	384,608

Flexique derives from Lexique,[2] a lexical database of phonetic, lexical, morphosyntactic, and frequency information on French. Like its predecessor, Flexique is distributed as an open source resource, under a Creative Commons Attribution-NonCommercial-ShareAlike license.[3]

1 Design features

Flexique was derived from Lexique version 3.70, a database collecting information of various kinds on 142694 French words, and distributed as an open source resource. Lexique is an extremely useful resource, but can be frustrating for the investigation of inflection, for a number of reasons:

- Lexique only collects entries for word forms that have occurrences in one of two corpora, post-1950 Frantext[4] and the French Subtitles Corpus[5]. As a result, inflectional paradigms are not complete; in particular there are very few verbs whose full paradigm is documented.
- Because Lexique is word-centric rather than lexeme-centric, sometimes forms of the same lexeme are not described coherently.
- The phonetic transcriptions of Lexique are a bit too surfacy for many purposes. In particular, there is no explicit representation of schwa optionality or mid vowel neutralization.

[1] We wish to thank Boris New and Christophe Pallier, whose work on Lexique was an indispensable precondition for the constitution of Flexique, as well as an inspiration for attempting to build it. Boris New has been extremely helpful in sharing his expertise. Gilles Boyé and Delphine Tribout provided crucial help at various points.

Flexique was designed at the *Laboratoire de Linguistique Formelle* and funded by Olivier Bonami's IUF grant.

[2] New B., Pallier C., Ferrand L., Matos R. (2001) 'Une base de données lexicales du français contemporain sur internet: LEXIQUE', *L'Année Psychologique*, 101, 447-462. http://www.lexique.org

[3] http://creativecommons.org/licenses/by-nc-sa/3.0/

[4] http://www.frantext.fr

[5] http://www.lexique.org/projets/Diphones/

- Although Lexique is being constantly improved, it has never been thoroughly validated by hand. Thus many scattered errors remain, both in transcriptions and in morphosyntactic annotations.

Flexique was designed with the goal of complementing Lexique in these domains. In particular:

- Flexique is organized by lexemes rather than words; it provides full paradigms for all adjectives, nouns and verbs one of whose forms is documented in Lexique.
- The phonetic transcriptions strike a balance between surface correctness and generality; the idea is to have for each word a unique phonological representation from which all phonologically predictable phonetic variants of a word can be deduced. This entails having systematic information on the location of possible schwas, even when these are only very seldom realized. This also entails providing notations for neutralized vowels.

2 Description of the resource

2.1 File format

Flexique is distributed as a family of csv files encoded in Unicode (UTF-8). Except for the first row which lists labels, each row of the file provides:

1. A unique lexeme identifier derived from the orthography of the citation form.
2. A list of orthographic variants of the citation form.
3. In the case of nouns, gender information: 'm' for masculines, 'f' for feminines. The symbol 'b' ('both') is used for nouns that exist both in the masculine and the feminine with identical forms (e.g. *secrétaire* /səkʁetɛʁ/ 'secretary').
4. A list of inflected forms in quasi-IPA transcription; see below section 2.3.

Gaps in the paradigm of defective lexemes are noted by the sequence '#DEF#'.

2.2 Content

In its current state Flexique has no provision for overabundance (i.e. cases where more than one form may be used in a paradigm cell): in cases of overabundance a decision was made as to what the more likely form is. Thus each lexeme is described by a single row and each row contains only one form per column.

One important design feature of Lexique is the association of pairs of masculine and feminine nouns to a single lemma; e.g. *directeur* 'male director' and *directrice* 'female director' are treated as two forms of a single lexeme. Since this is a vividly debated issue in morphology, the decision in Flexique has been to:

- Code pairs such as *directeur/directrice* as two separate entries;
- Use a single entry for pairs of homophonous masculines and feminines that are semantically equivalent modulo sex (e.g. *secrétaire/secrétaire*), noting gender as 'b' for 'both'.
- Use two entries for cases of homophonous but semantically nonequivalent masculine-feminine pairs (e.g. *pendule*)

The advantage of that convention is that the lexicon can easily be automatically adjusted to treat cases such as *secrétaire* either as one or two lexemes.

The issue of lexeme identity should be handled with care. In its current state, `Flexique` uses one line per (phonological) inflectional paradigm; thus it makes no difference between the case of a single lexeme with an orthographic variant (e.g. *shaman* vs. *chamane*) and the case of true homonyms (e.g. *verre* vs. *vers*). The only exception to this is the case of homophonous nouns of different genders: as stated above, there are two separate entries for *pendule*.

This is clearly a poor compromise that needs to be improved upon in future versions.

2.3 Transcription conventions

Phonetic transcriptions use the following IPA symbols:

p t k b d g f s ʃ v z ʒ m n ɲ ŋ l ʁ w ɥ j i y u e ø o ɛ œ ɔ a ə ɛ̃ ɑ̃ ɔ̃

Three symbols that are not part of the IPA are used to transcribe neutralized mid vowels:

- E for the vowel neutral between e and ɛ
- O for the vowel neutral between o and ɔ
- Ø for the vowel neutral between ø and œ

IPA symbols have their standard interpretation. As is standard in studies of French, ə notes a vowel alternating between ø, œ and no realization.

For forms with phonologically regular multiple realizations, arbitrations were made so that a single transcription be proposed that is as close as possible to a likely surface form while allowing inference of the other possible forms. Thus:

- Except in word final position, schwas have been included wherever they are possible, even in cases where the actual realization of a schwa is very unlikely. For instance for the future 3rd singular of AIMER 'like', the transcription is ɛməʁa, which is a lot less likely than ɛmʁa in normal speech. This decision is motivated by the fact that it is possible to predict the range of possible realizations of a form from the form containing the maximal number of word-internal schwas, but it is not possible to predict where schwas will be possible from a form without schwas—e.g. the form kɔ̃tʁa is ambiguous between the future 2nd or 3rd singular of COMPTER 'count' and the simple past 2nd or 3rd singular of CONTRER 'counter', whereas kɔ̃təʁa is unambiguously the former. For this reason including all schwas is the only solution if one is to give a single phonological representation for a word form, but one should be aware of the fact that it artificially diminishes the prevalence of homophony; thus in many applications it is advisable to process the `Lexique` data so as to suppress some schwas.

  Word-final schwas are not included because their distribution is entirely dependent on the phonological context.[6]

- Transcriptions do not include liaison consonants. This is mainly due to a lack of data: at least in masculine singular adjectives and in verbs in the present, there is quite a bit of uncertainty as to whether, where and when a liaison consonant is available. Inclusion of relevant information will have to await a large scale empirical study of the lexical distribution of liaison.

[6] See e.g. François Dell (1995), 'Consonant clusters and phonological syllables in French', *Lingua* 95:5–26.

- For non-first conjugation verbs, when the stem ends in ʁ, the future and conditional forms have a geminate ʁ in conservative varieties (e.g. MOURIR 'die', future 3rd singular: muʁʁa. This is definitely not the only possibility: degemination is very common (muʁa), and regularizations by vowel epenthesis, while frowned upon, are quite frequent (muʁəʁa,muʁiʁa). Flexique records the conservative form.

- French morphophonology leads to frequent alternations between high vowels i, y and u, the corresponding glides j, ɥ and w, and the vowel-glide sequence ij, yɥ, uw. The transcription convention is to use:

  - A vowel wherever it is the only possible form, e.g. *elle relie* 'she links': ɛlʁəli
  - A glide wherever it is the only possible form, e.g. *elle paye* 'she pays': ɛlpɛj
  - A vowel-glide sequence ij wherever it is the only possible form, e.g. *elle priait* 'she prayed': ɛlpʁijE
  - A single glide where there is alternation between glide and vowel-glide sequence, e.g. *elle reliait* 'she linked', which can be realized both ɛlʁəljɛ and ɛlʁəlijɛ, is transcribed as ɛlʁəljE.

  Where morphology would warrant a geminate glide (e.g. *nous payions*: pɛj+j+ɔ̃) this is realized as a single glide in standard French. Hypercorrect forms such as pɛjjɔ̃ are sometimes heard, but are ignored in *Flexique*: *payions* is transcribed pEjɔ̃.

3 Resource construction

For each part of speech we followed the same procedure:

1. Determination of a set of principal parts from which the full paradigm can be deduced for (almost) all lexemes in that category
2. Filtering of items that are incorrectly categorized in Lexique
3. Hand-correction of the transcription given by Lexique for all principal parts of all lexemes
4. Automatic generation of full paradigms for all lexemes
5. Semi-automatic coherence check of the full paradigms, by examination of implicative relations between paradigm cells.[7]
6. Examination by hand of the full paradigms of a selected subset of lexemes
7. Iterative correction of both the principal parts of lexemes and generation scripts
8. Reduction of homonyms to a single description

Given this procedure, one can be quite confident that (almost) all lexemes are inflected coherently in Flexique: that is, if there are errors, they are likely to be systematic across the whole paradigm. Such errors have little impact on the study of the inflection system, which is the primary purpose of Flexique.

However there certainly remain some systematic errors. We expect to correct them incrementally in future releases and are thus eager to receive suggested corrections from users.

[7] The tools used for that semiautomatic check are partially documented in Olivier Bonami (2012), *Discovering implicative morphology*, invited talk at the Huitièmes Décembrettes International Conference (Bordeaux, December 2012).

3.1 Adjectives

The principal parts for adjectives are the masculine singular and the feminine singular. Generation of plurals from singulars is easy, as the two forms are identical except in the masculine for a small and well-understood set of cases (most adjectives in *-al*).

The masculine plural of all adjectives in *-al* was thus corrected by hand.

For all other adjectives, wherever `Lexique` contains matching singular and plural forms, we checked whether the generated plurals match the one found in `Lexique`.

There are two known limitations to the adjective database. First, we did not generate masculine singular liaison forms, despite the fact that those arguably fill a distinct cell in adjectival paradigms.[8] Second, there are some adjectives that arguably are defective in one of the two genders, e.g. *enceinte* 'pregnant'. This is not taken into account, and all adjectives are treated as having full paradigms.

3.2 Verbs

The verbal part of the resource was built in three steps.

1. With a handful of exceptions, first and second conjugation verbs have a paradigm that is fully predictable from the phonetic transcription of the infinitive and its orthographic form.[9] Thus for most verbs only the infinitive's transcription was corrected by hand, and the rest of the paradigm was generated directly.

2. Verbs in the third conjugation have highly unpredictable paradigms: according to Bonami & Boyé however,[10] all forms can be predicted from 12 principal parts for all but a handful of verbs. Thus the transcription of all 12 principal parts was corrected by hand, and the remaining 39 paradigm cells were generated automatically. For the handful of deeply irregular verbs, irregular forms are directly listed in the generation script.

3. Defective verbs were added as a last step: since the list of missing cells can not practically be predicted from form alone, and the list of relevant verbs is quite small, full paradigms were generated using the procedures defined above, and a list of defective cells was established by hand for each relevant verb.[11]

3.3 Nouns

Nouns do not have reliable principal parts: a noun has only two paradigm cells (singular and plural), and neither is fully predictable from the other. However the very high number of nouns (nearly 35,000

[8] See e.g. Yves-Charles Morin (1992), 'Un cas méconnu de la déclinaison de l'adjectif en français: les formes de liaison de l'adjectif antéposé', in *Le mot, les mots, les bons mots. Word, words, witty words. Hommage à Igor A. Mel'čuk*, Presses de l'Université de Montréal.

[9] The orthographic form is needed to disambiguate

- Verbs like GRILLER 'grill' `gʁije`, with a stable stem-final glide (present 1st singular *grille*: `gʁij`), from verbs like CRIER 'shout' `kʁije`, whose stem-final glide alternates with a vowel (present 1st singular *crie*: `kʁi`).
- Verbs like DÉJEUNER 'have lunch': `deʒəne`, with no vowel alternation in the stem-final syllable (present 1st singular *déjeune*: `deʒn`), from verbs like DÉMENER 'struggle': `deməne`, whose schwa alternates with `ɛ` (present 1st singular *démène*: `demɛn`).

[10] Olivier Bonami & Gilles Boyé (2002), 'Suppletion and dependency in inflectional morphology'. In F. Van Eynde, L. Hellan et D. Beerman (eds), *The Proceedings of the HPSG'01 Conference*. Stanford : CSLI Publications.

[11] There is quite a bit of uncertainty as to the exact list of defective lexemes, and for each defective lexeme, the set of defective cells. Arbitrations were made taking into account the discussions in various publications (most notably *Le Bon unsage* 14th edition, *Le Trésor de la Langue Française*, *Bescherelle: la conjugaison pour tous*, and Boyé (2000), *Problèmes de morphophonologie verbale*, PhD thesis, Université Paris 7), and usage observed in the corpora underlying `Lexique` and on the web.

lexemes in Lexique) makes the hand correction of full paradigms quite costly.

Thus only the transcriptions of the singulars in Lexique was checked by hand. The plurals were uniformly generated as identical to the singulars, but a number of checks were then performed:

- All nouns with an orthographic singular in *-al* or *-ail* were checked by hand.
- All nouns that only occur in the plural in Lexique are likely to be *pluralia tantum*, nouns that are defective in the singular. They were all checked by hand.
- For all nouns both of whose forms are documented in Lexique, it was checked whether Lexique lists identical singulars and plurals, both in orthography and in phonetic transcription. All lexemes for which this is not the case were checked by hand.

It remains possible that a few irregular nouns were missed, but that number is likely to be very low, as there are few irregulars that are infrequent enough to not be listed both in the singular and in the plural in Lexique.

It is in the transcription of nouns that Flexique is most likely to be inaccurate: in the case of verbs and adjectives, more than one form was transcribed by hand, so that any errors are likely to have been caught by the study of implicative relations (they give rise to unexpected unreliability). In the case of nouns however, we have no easy way to check that there is no error in the single used principal part. Users should thus proceed with caution if they are heavily dependent on transcriptions of nouns.

4 Limitations and future work

Known limitations:

- Move from the simple csv format to an xml representation, e.g. using LMF.
- Include overabundant forms.
- Distinguish orthographic variants from true homonyms
- Include frequency information for each form filling each paradigm cell of each lexeme.[12]
- Where possible, port back relevant corrections into Lexique.

[12] Lexique only provides frequency information for unique forms, but not for paradigm cells: if a lexeme has two syncretic forms, a single frequency value for that form is provided.