
Compiling a new French frequency dictionary

Deryle Lonsdale, Yvon Le Bras,
Amy Berglund, and Fritz Abélard

Brigham Young University
lonz@byu.edu



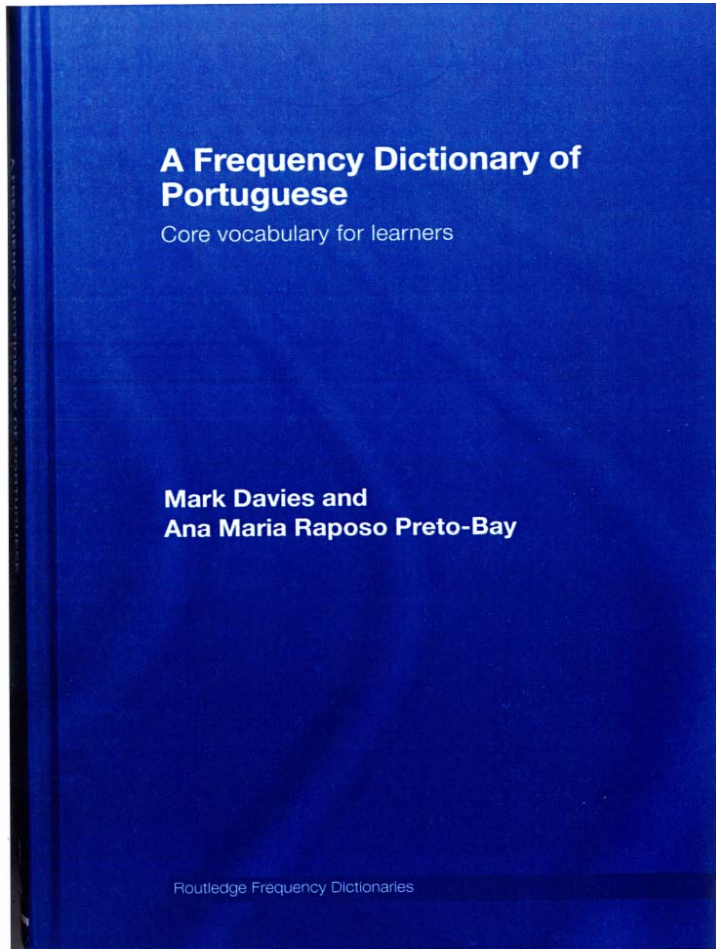
Our work

- To develop yet another French dictionary
- Why?
 - Second most taught foreign language worldwide (after English)
 - Second most globally used language (after English)
 - Lack of a modern, corpus-derived, frequency-based dictionary for learners

Prior literature

- Some pioneering but outdated lists
 - Henmon 1924, Juilland et al. 1970
- Some narrow-scope dialectal work
 - Beauchemin et al. 1992
- Some intended for scholarly reference
 - Brunet 1981, Imbs 1971-1994
- Some exclusively online or via subscription
 - ARTFL FRANTEXT, TLFi
- Some beginner dictionaries, without frequency information
 - Oxford, Larousse, etc.

Analogous efforts



- Previous frequency-listed learner dictionaries
 - Spanish, Portuguese
 - Top 5000 lemmas
- Based on corpus of collected materials
- Extensive text processing and annotation
- Arranged for ease of use

Corpus-based rank value

36 A Frequency Dictionary of Portuguese

597 **comer** *v* to eat

- não comeu nada, só uma garfada de arroz com ovo – *He didn't eat anything, just a forkful of rice with eggs.*
94 | 3639 +s -n

598 **fugir** *v* to flee, run away

- muita gente já começou a deixar a região para fugir da onda de violência – *A lot of people have already begun to leave the region to flee from the wave of violence.*
95 | 3204

599 **motivo** *nm* reason, motive

- não havia motivo para alarme – *There was no reason for alarm.*
95 | 2540

600 **som** *nm* sound

- os instrumentos emitem um som semelhante ao de uma flauta – *The instruments emit a sound similar to that of a flute.*
96 | 2279

Lemma and part of speech

597 **comer** *v* to eat

- não comeu nada, só uma garfada de arroz com ovo – *He didn't eat anything, just a forkful of rice with eggs.*
94 | 3639 +s -n

598 **fugir** *v* to flee, run away

- muita gente já começou a deixar a região para fugir da onda de violência – *A lot of people have already begun to leave the region to flee from the wave of violence.*
95 | 3204

599 **motivo** *nm* reason, motive

- não havia motivo para alarme – *There was no reason for alarm.*
95 | 2540

600 **som** *nm* sound

- os instrumentos emitem um som semelhante ao de uma flauta – *The instruments emit a sound similar to that of a flute.*
96 | 2279

Indicative English gloss(es)

36 A Frequency Dictionary of Portuguese

597 comer *v* to eat

- não comeu nada, só uma garfada de arroz com ovo – *He didn't eat anything, just a forkful of rice with eggs.*

94 | 3639 +s -n

598 fugir *v* to flee, run away

- muita gente já começou a deixar a região para fugir da onda de violência – *A lot of people have already begun to leave the region to flee from the wave of violence.*

95 | 3204

599 motivo *nm* reason, motive

- não havia motivo para alarme – *There was no reason for alarm.*

95 | 2540

600 som *nm* sound

- os instrumentos emitem um som semelhante ao de uma flauta – *The instruments emit a sound similar to that of a flute.*

96 | 2279

Corpus-derived usage example

597 comer *v* to eat

- não comeu nada, só uma garfada de arroz com ovo – *He didn't eat anything, just a forkful of rice with eggs.*
94 | 3639 +s -n

598 fugir *v* to flee, run away

- muita gente já começou a deixar a região para fugir da onda de violência – *A lot of people have already begun to leave the region to flee from the wave of violence.*
95 | 3204

599 motivo *nm* reason, motive

- não havia motivo para alarme – *There was no reason for alarm.*
95 | 2540

600 som *nm* sound

- os instrumentos emitem um som semelhante ao de uma flauta – *The instruments emit a sound similar to that of a flute.*
96 | 2279

English translation of usage eg.

36 A Frequency Dictionary of Portuguese

597 comer *v* to eat

- não comeu nada, só uma garfada de arroz com ovo – *He didn't eat anything, just a forkful of rice with eggs.*
94 | 3639 +s -n

598 fugir *v* to flee, run away

- muita gente já começou a deixar a região para fugir da onda de violência – *A lot of people have already begun to leave the region to flee from the wave of violence.*
95 | 3204

599 motivo *nm* reason, motive

- não havia motivo para alarme – *There was no reason for alarm.*
95 | 2540

600 som *nm* sound

- os instrumentos emitem um som semelhante ao de uma flauta – *The instruments emit a sound similar to that of a flute.*
96 | 2279

Frequency/range values

36 A Frequency Dictionary of Portuguese

597 comer *v* to eat

- não comeu nada, só uma garfada de arroz com ovo – *He didn't eat anything, just a forkful of rice with eggs.*
94 | 3639 +s -n

598 fugir *v* to flee, run away

- muita gente já começou a deixar a região para fugir da onda de violência – *A lot of people have already begun to leave the region to flee from the wave of violence.*
95 | 3204

599 motivo *nm* reason, motive

- não havia motivo para alarme – *There was no reason for alarm.*
95 | 2540

600 som *nm* sound

- os instrumentos emitem um som semelhante ao de uma flauta – *The instruments emit a sound similar to that of a flute.*
96 | 2279

Corpus collection

Corpus design

- 23 million words
 - Half oral/spoken language
 - Half textual/written language
- No materials earlier than 1950
- No attempt to proportion data demographically
- Some balance across genres, not perfect
- Some exhaustive content, some sampled

Spoken texts

- 11.5 million words
 - Transcripts of governmental debates, hearings (Canada, Europe)
 - Transcripts of media interviews with writers, entertainment figures, business leaders, athletes, academicians, etc.
 - Transcripts of telephone calls, face-to-face dialogues
 - Movie scripts/subtitles, theatrical plays
 - Transcripts of broadcast news, media chat

Written texts

- 11.5 million words
 - Newswire stories
 - Daily, weekly newspapers
 - Literature: fiction, nonfiction essays, memoirs, novels, etc.
 - Magazines: popular science, technical
 - Newsletters, bulletins, business correspondence
 - Technical manuals

Corpus annotation

Tokenization and cleaning

- Wide range of character representations
 - EBCDIC, MACROMAN, ISO, UTF-8, HTML...
- Much extraneous information to remove
- Perl scripting, Unix tools (iconv, tr, make), SGML/XML parsers
- Linguistic issues:
 - Accentuation of capital letters
 - Hyphenation/apostrophe breaks (dis-moi vs. week-end, l'homme vs. aujourd'hui)

POS tagging

- A dozen or so extant POS taggers for French
 - Various approaches, theoretical frameworks
 - Each has strengths, weaknesses
 - Issue: which tagset(s) to use
 - Installed and tested several
 - Best results obtained via combination of approaches, own tagset since fine-grained distinctions not required
 - Editing/hand-correcting: questionable value

Lemmatization

- Required for frequency computations
- Some cases are tricky
 - Non-finite forms (p.part.: verb? gerund? adj?)
 - Morphological ambiguity
 - Abbreviations, case folding, symbols (&, %, X^e)
- Lexical resources are very helpful (e.g. BDLEX)
- Some taggers do this too
- Again, best strategy is hybrid approach

The target list

- Top 5000 lemmas
 - What to reject: some proper nouns (personal names, geographic locations)
 - How to merge: inflectional variants are straightforward (others aren't!)
 - Maintain POS distinction, but abstract away from polysemy
- Figuring out weights for linear combination of raw frequency, dispersion for final ranking value

Spoken vs. written language

- Text only: aéronautique, bouleversement, coïncider, crépuscule, guérilla, itinéraire, jadis, laïque, logistique, météorologique, microphone, solennel
- Spoken only: abusif, allô, bah, cafard, cingler, clown, copine, dingue, flic, fric, hockey, lucratif, machin, météo, micro, ouais, porno, sexy, sympa

Glossing the terms

- Mostly manual effort, to avoid copyright infringement
- Largely human process, little technology needed
- Difficulty: limiting the range of possibilities, expressing the core meaning(s) succinctly

Finding usage contexts (1)

- Important part of entry:
 - Intuitive: meaning should be clear in short (!) context
 - Self-contained: should form a reasonable syntactic/semantic unit
 - Consistent: reflect content of English gloss(es)
- Must come from the underlying corpus
- Indexing is indispensable

Finding usage contexts (2)

- Concordancing will be required for human inspection
- Large number of instances to examine
- Solution: generate as many as possible automatically, score them for applicability
 - Syntactic status (chunking, shallow parsing)
 - Use of target vocabulary within context
 - Present thresholded list of possibilities

Sample contexts

- ✓ les Bulls, qui comptent désormais quatre victoires
- ✗ il se le fût désormais tenu pour dit
- ✓ ils m'ont posé des questions. des tas de questions.
- ✗ la question se pose de manière inverse!
- ✓ la mise en place progressive d'un nouveau système
- ✗ aucune mise aux enchères prochaine n'est prévue
- ✓ j'ai soulevé des questions assez simples
- ✗ dieux anciens, rois médiévaux ou simples esclaves, tous sont bien vivants

Generating usage translations

- Again, mostly human, manual effort
- Some terms, usage contexts come from aligned bitexts
 - When this is the case, retrieve English translation automatically
- Possibly gisting, drafting via (HA)MT

Thematic lists

- Terms for food, body parts, weather, professions, opposites, subjunctive triggers, etc.
- Leveraging lexical resources (e.g. FrnWordNet)
- Using WSD techniques
 - Probabilistic, exemplar-based approaches

Current status

- Corpus collected, cleaned, annotated
- Glossing is ongoing
- Currently generating usage contexts
- Migrating to relational database architecture
- Packaging infrastructure for continued updating over time
- Web deployment?

Conclusions

- Dictionary calibrated to learners' needs
- Corpus linguistics at core of effort
- Wide array of human skills, computational linguistic techniques
 - Exploring tradeoffs is an interesting enterprise
- Compelling work at the intersection of humanities and sciences

Questions?
