Japan Association for English Corpus Studies (JAECS)
46th Annual Conference (Oct 3-4, 2020)

Extending vocabulary profiling to languages other than English

Laurence ANTHONY (Waseda University), Natalie FINLAYSON, Emma MARSDEN, Rachel HAWKES, and Nick AVERY (National Centre of Excellence for Language Pedagogy, University of York)

Contact: anthony@waseda.jp; https://www.laurenceanthony.net/; @antlabjp
natalie.finlayson@york.ac.uk; https://ncelp.org/; @natalie_eloise

Overview

- Background
  - Resources for vocabulary-related research
  - Current desktop and online profiling tools
- Vocabulary profiling in a multilingual context
  - Limitations of existing profiling tools
  - Challenges when processing non-English languages
  - Mission of the National Centre of Excellence for Language Pedagogy (NCELP)
  - Vocabulary teaching strand at NCELP
  - Adapting AntWordProfiler to NCELP goals
- MultiLingProfiler (by NCELP)
  - Open-access, online, multilingual profiling

Background

Background

Resources for vocabulary-related research

- Four types of resources available for vocabulary-related research (Anthony, 2019)
  - corpora (as primary data sources)
  - software tools for collecting/building primary corpus sources
  - pre-built word lists (as secondary data sources)
  - software tools for probing/profiling/analyzing primary and secondary data sources

Anthony, L. (2019). Resources for researching vocabulary, in S. Webb (Ed.)
The Routledge Handbook of Vocabulary Studies. Abingdon: UK. Routledge Press.

Background

Resources for vocabulary-related research

"It is hard to imagine any area of vocabulary research into acquisition, processing, pedagogy, or assessment where the insights available from corpus analysis would not be valuable. In fact, it is probably not too extreme to say that most sound vocabulary research will have some corpus element." (Schmitt, 2010, p. 307)

Schmitt, N. (2010). Researching vocabulary: A vocabulary research manual. New York, NY: Palgrave Macmillan.

Background

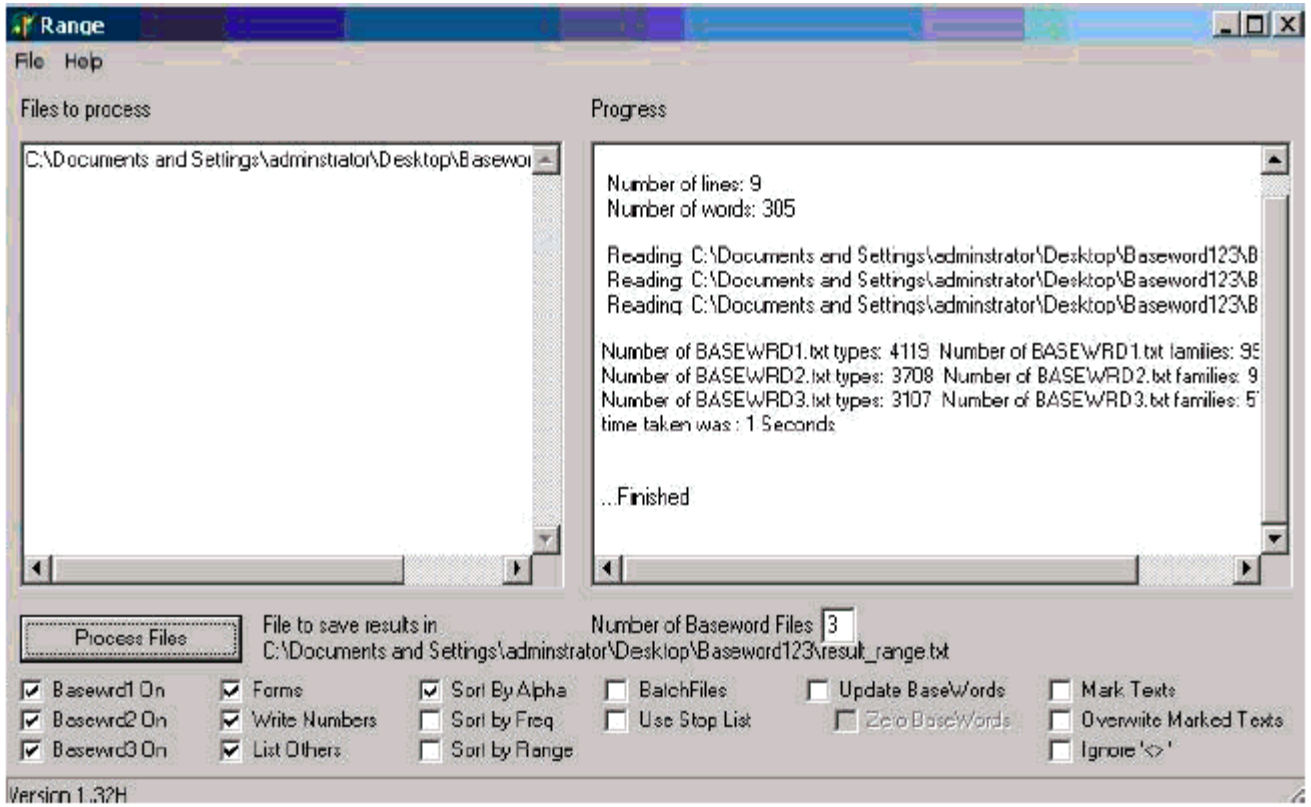
Resources for vocabulary-related research

- Resources available for vocabulary-related research (Anthony, 2019)
  - corpora (as primary data sources)
  - software tools for collecting/building primary corpus sources
  - pre-built word lists (as secondary data sources)
  - software tools for probing/profiling/analyzing primary and secondary data sources

Anthony, L. (2019). Resources for researching vocabulary, in S. Webb (Ed.)
The Routledge Handbook of Vocabulary Studies. Abingdon: UK. Routledge Press.

Background

Current desktop profiling tools



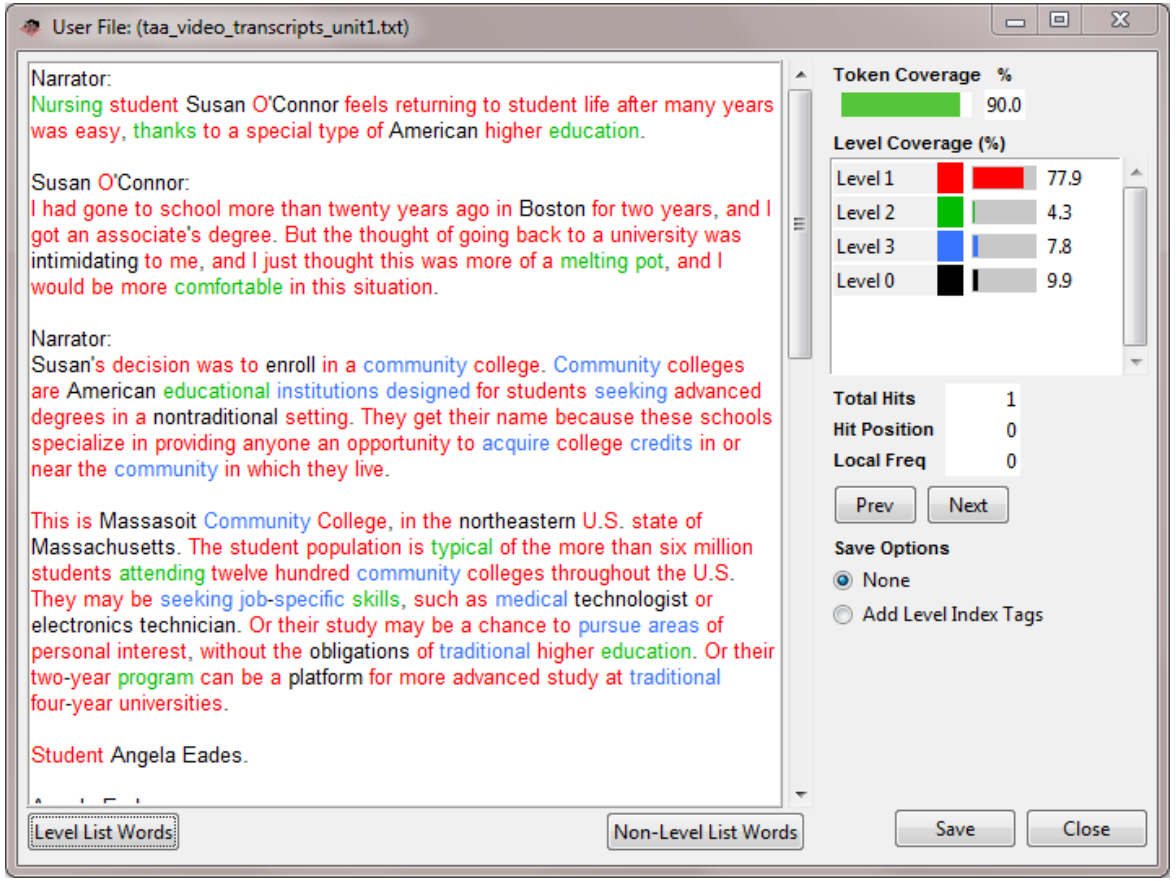
Range

(2002)

Heatley, A., Nation, I.S.P. and Coxhead, A. (2002). *RANGE* and *FREQUENCY* programs. http://www.vuw.ac.nz/lals/staff/Paul_Nation

Background

Current desktop profiling tools



AntWordProfiler

(2014)

Anthony, L. (2014). *AntWordProfiler* (Version 1.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Background

Current online profiling tools

Compleat Lexical Tutor. Accessed at: www.lextutor.ca/

Compleat Lexical Tutor (Web VP)

(2020)

Japan Association for English Corpus Studies (JAECS)
46th Annual Conference (Oct 3-4, 2020)

Vocabulary profiling in a multilingual context

Vocabulary profiling in a multilingual context

Limitations of existing profiling tools for teachers of languages other than English

- Questions that teachers of KS3 (pre-GCSE) languages in the UK need to ask to assess the suitability of L2 texts for their classes:
  - Q1. What percentage of the words in this text are high-frequency words?
  - Q2. Does this text contain any low-frequency words which I will need to gloss or replace?
  - Q3. How many of the words in this text have students learned so far?
- Existing lexical profiling tools limited in terms of:
  - languages available (Q1 and Q2)
  - compatibility with a syllabus (Q3). Existing profiling tools assume knowledge of all inflections of a word, which is not the case at lower levels.

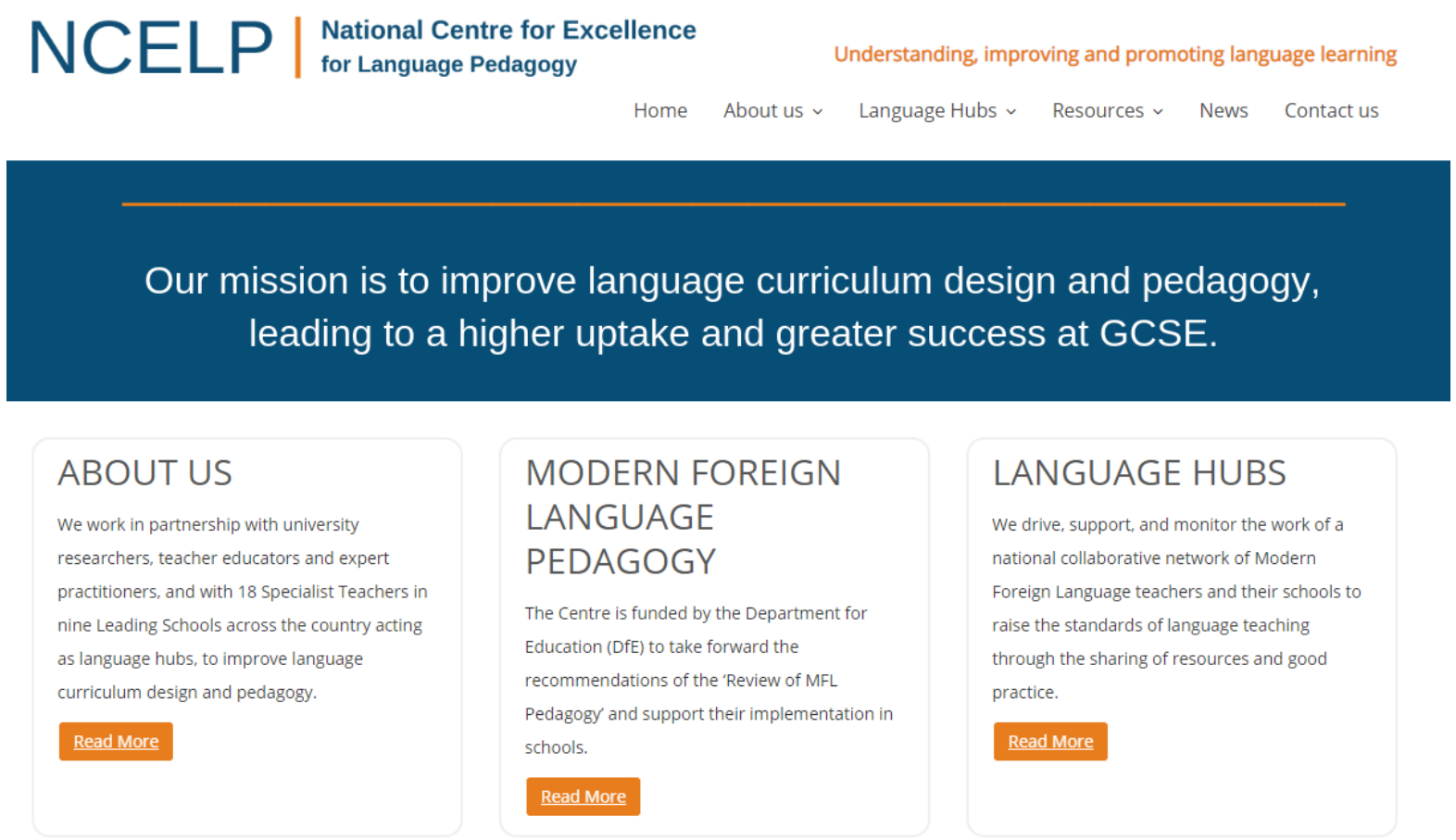
Vocabulary profiling in a multilingual context

Challenges when processing non-English languages

- Rich inflectional morphology
  (German: singen, singe, singt, singst, gesungen)
  - level lists take longer to create
  - fewer existing sources to adapt
- Morphological concerns
  - punctuation as part of the morphology (French: *aujourd'hui, peut-être*)
  - case sensitivity (German: ***E****ssen* (noun), ***e****ssen* (verb))
- Definition of a word
  - e.g. *aujourd'hui* – one word or two?

Vocabulary profiling in a multilingual context

Mission of the National Centre for Excellence for Language Pedagogy (NCELP)



www.ncelp.org

Vocabulary profiling in a multilingual context

NCELP vocabulary teaching strand – principles

- Principles of vocabulary selection and sequencing:
  - average of **10 new words** introduced per **week** (Schmitt 2004)
  - most words chosen from the **2,000 most frequent words** in the language (**adapted** to **needs** of target learner group)
    [Davies, M., & Davies, K. (2018); Jones, R., & Tschirner, L. (2006); Lonsdale, D. & Le Bras. Y. (2009)]
  - words are not linked to a particular topic when introduced
  - focus on high-frequency verbs with complementary items from **different word classes**
  - 5 – 20 encounters to remember a word; spaced repetition
- NCELP SOW approach includes:
  - self-access pre-learning (Quizlet or audio learning HW)
  - multiple opportunities to use and revisit within a week (listening, reading, writing, speaking)
  - further systematic recycling within a month, within a term, within a year

Vocabulary profiling in a multilingual context

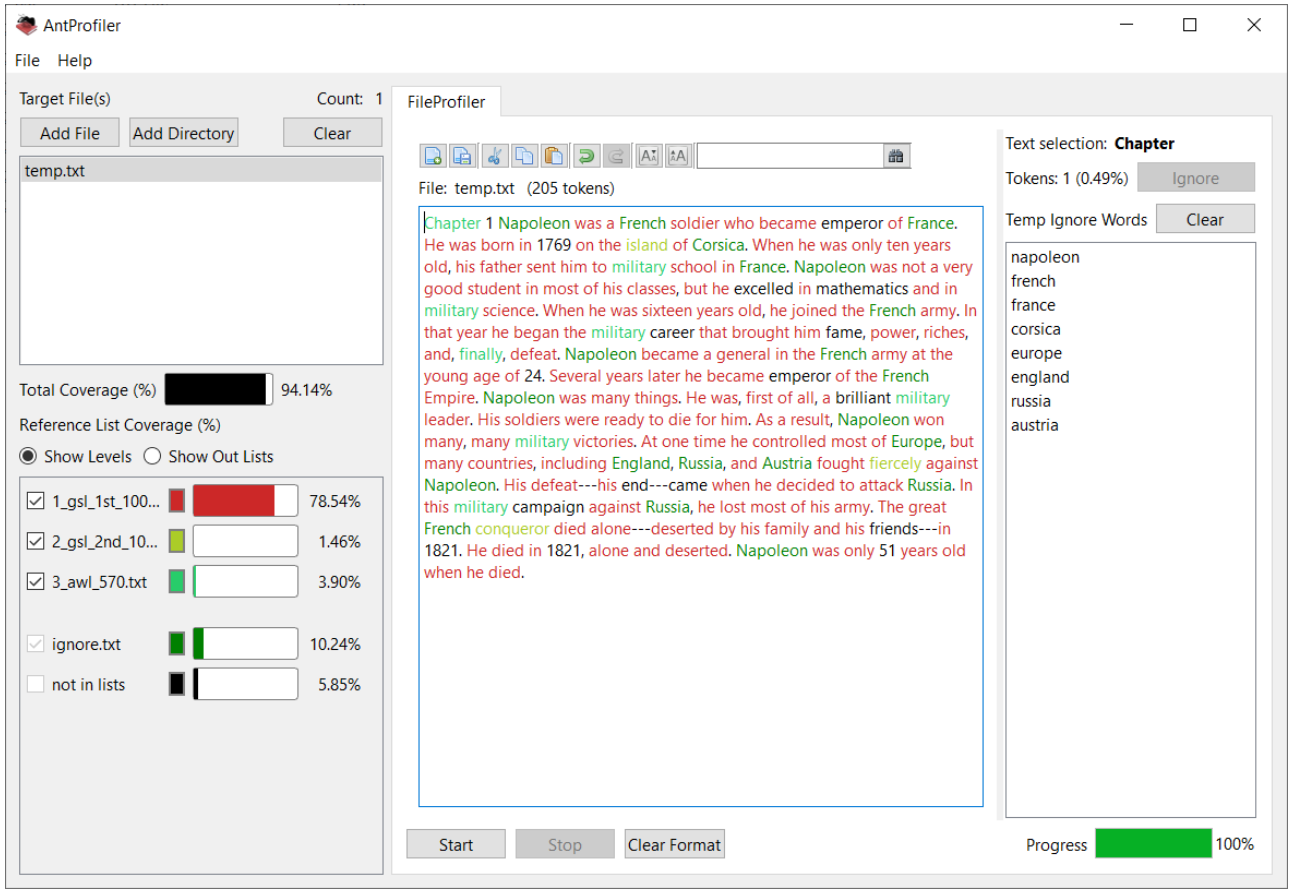
NCELP vocabulary teaching strand – systematic vocabulary revisiting

wk 10	19 20	**-ER verbs (je, tu, il/elle)** **present simple used with its continuous meaning** **two-verb structures: *aimer* + infinitive**	aimer [242], **cocher** [>5000], passer[1] [90], porter [105], rester [100], trouver [83], école [477], moment [148], semaine [245], **solution** [608], uniforme [1801], chaque [151], à[1] [4], avec [23]	faire [25], fais [25], fait [25], ça [54], activité [452], courses [1289], cuisine [2618], devoirs [39], lit [1837], ménage [2326], modèle [958], quoi ? [297]	est [5], il1 [13], elle1 [38], amusant [4695], calme [1731], content [1841], intelligent [2509], malade [1066], méchant [3184], triste [1843], mais [30], ou [33], merci [1070]
wk 5	9 10	être & avoir (je, tu, il/elle) **feminisation of job titles (-e)** **subject pronouns il/elle meaning 'it'** *indefinite articles*	il[2] [13], elle[2] [38], **ami** [467], **amie** [467], chanteur [3251], chanteuse [3251], femme [154], homme [136], **professeur** [1150], **professeure** [1150], drôle [2166], intéressant [1244], **faux** [555], sympa(thique) [4164], **vrai** [292]	est [5], il1 [13], elle1 [38], amusant [4695], calme [1731], content [1841], intelligent [2509], malade [1066], méchant [3184], triste [1843], mais [30], ou [33], merci [1070]	
wk 2	3 4	être (je, tu, **il/elle**) regular adjective gender agreement (as complement to verb only); **intonation questions**	est [5], il[1] [13], elle[1] [38], amusant [4695], calme [1731], content [1841], intelligent [2509], malade [1066], méchant [3184], triste [1843], mais [30], ou [33], **merci** [1070]	N/A	

- spring test (wk 20)
- summer test (wk 33)
- follow-up revisiting and testing every year until GCSE

Vocabulary profiling in a multilingual context

Adapting AntWordProfiler to NCELP goals



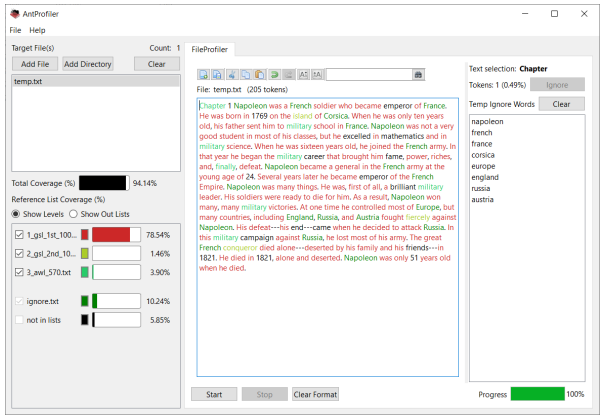
AntProfiler (2020)

Anthony, L. (2020). AntProfiler (Version 1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Vocabulary profiling in a multilingual context

Adapting AntWordProfiler to NCELP goals

- **AntProfiler** (Anthony, 2020)
  https://www.laurenceanthony.net/software/antprofiler/
  - freeware
  - multiplatform (Windows, Macintosh, Linux)
  - portable (can run directly from a USB stick)
  - standard editor features
    (load, save, cut, copy, paste, undo, redo, font increase/decrease, search)

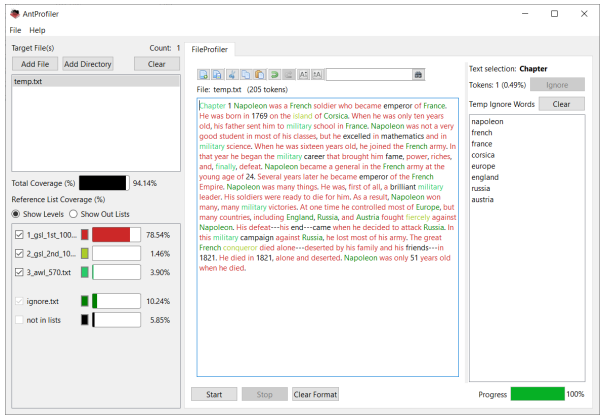


Anthony, L. (2020). AntProfiler (Version 1.0) [Computer Software].
Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Vocabulary profiling in a multilingual context

Adapting AntWordProfiler to NCELP goals

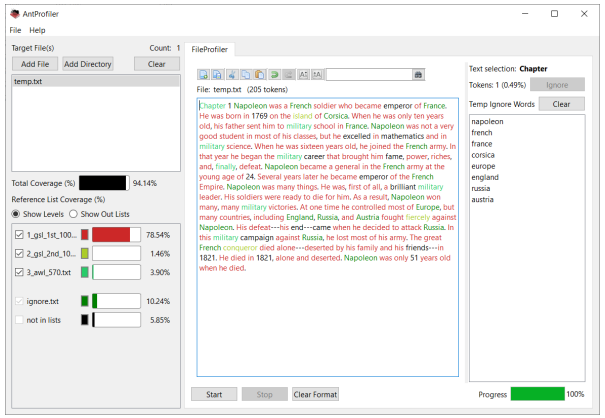
- **AntProfiler** (Anthony, 2020)
  https://www.laurenceanthony.net/software/antprofiler/
  - profiling of 'in-lists', 'out-lists', and 'ignore lists'
  - use of "row-format" and "range-format" level-lists
  - language independent: works with all Unicode-supported languages
    (English, French, German, Spanish, Chinese, Japanese, Korean, …)
  - export of 'in-lists' and color-coded profiles



Anthony, L. (2020). AntProfiler (Version 1.0) [Computer Software].
Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Vocabulary profiling in a multilingual context:
Adapting AntWordProfiler to NCELP goals

- **AntProfiler-NCELP** (Anthony, 2020)
  https://www.laurenceanthony.net/software/antprofiler/
  - embedded in-lists
    - most frequent 2000 words of French, German, Spanish
      [Davies, M., & Davies, K. (2018); Jones, R., & Tschirner, L. (2006); Lonsdale, D. & Le Bras. Y. (2009)]
  - import/export of custom user-defined in-lists



Anthony, L. (2020). AntProfiler (Version 1.0) [Computer Software].
Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Japan Association for English Corpus Studies (JAECS)
46th Annual Conference (Oct 3-4, 2020)

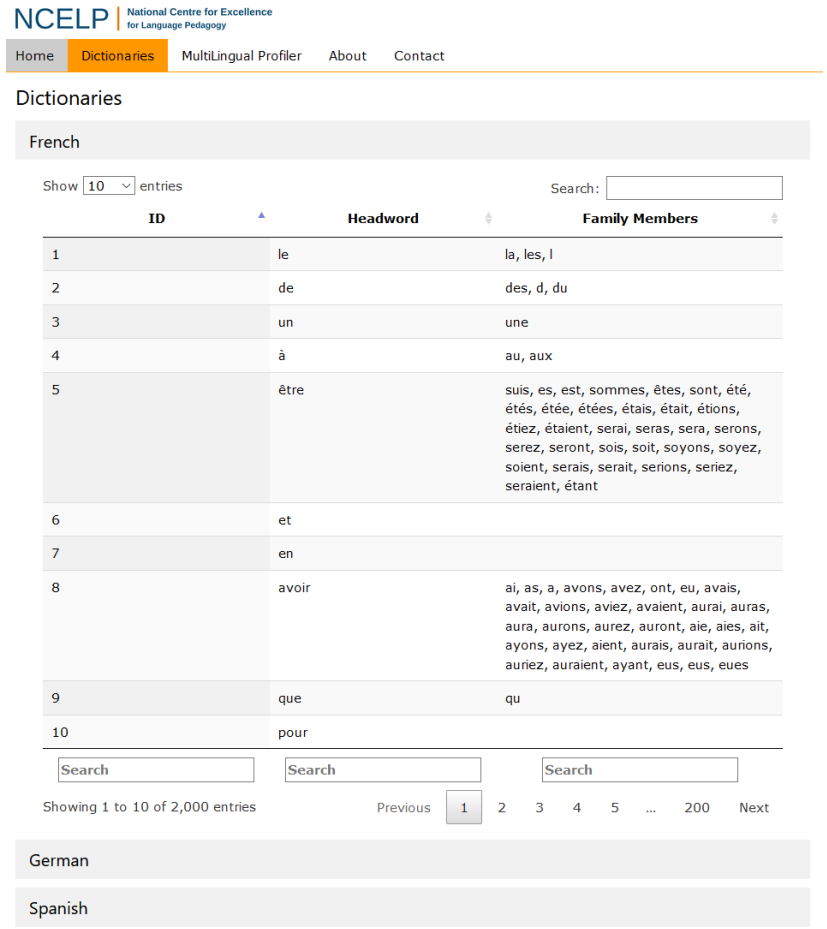
MultiLingProfiler (by NCELP)

open-access, online, multilingual profiling

MultiLingProfiler

Overview and features

- **MultiLingProfiler** (NCELP, 2020)
  https://www.multilingprofiler.net/ (coming soon)



Davies, M., & Davies, K. (2018). *A frequency dictionary of Spanish: Core vocabulary for learners* (2nd ed.). London: Routledge

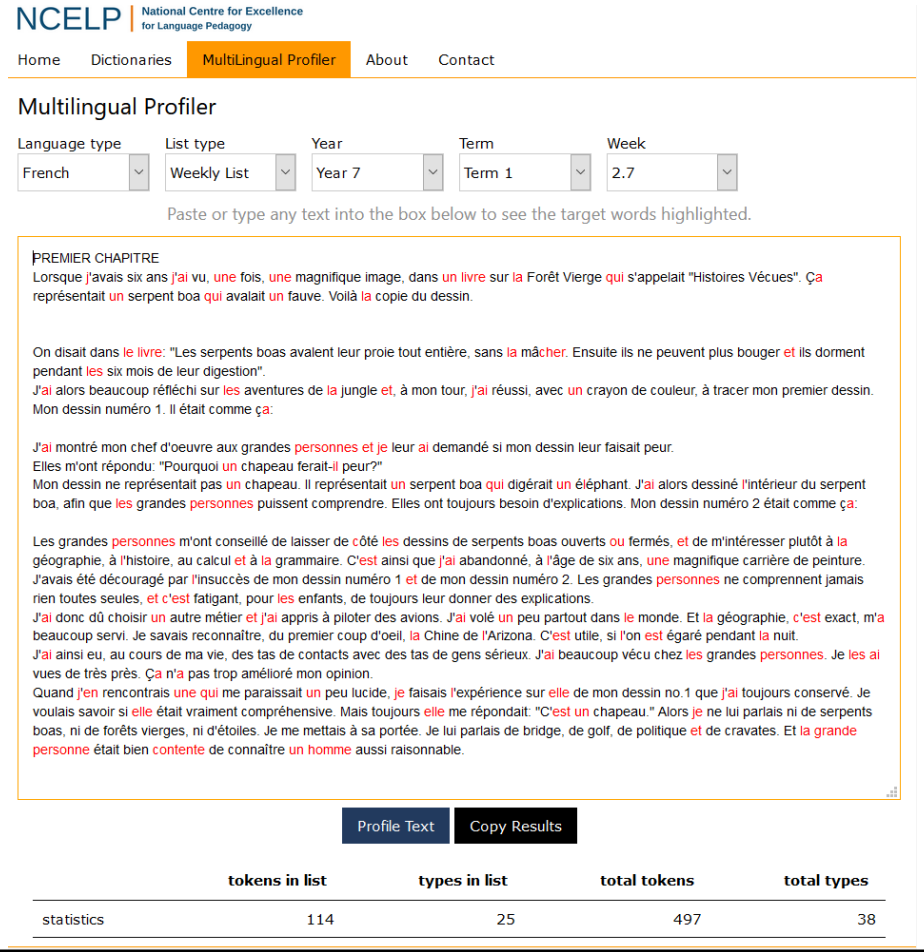
Jones, R., & Tschirner, L. (2006). *A frequency dictionary of German: Core vocabulary for learners.* Oxford: Routledge

Lonsdale, D. & Le Bras. Y. (2009). *A frequency dictionary of French: Core vocabulary for learners.* Oxford: Routledge

MultiLingProfiler

Overview and features

- **MultiLingProfiler** (NCELP, 2020)
  https://www.multilingprofiler.net/ (coming soon)



MultiLingProfiler

Overview and features

- **MultiLingProfiler** (NCELP, 2020)
  https://www.multilingprofiler.net/ (coming soon)
  - frequency list viewer (with dynamic search)
    - most frequent 2000 words of French, German, Spanish
  - multilingual profiler
    - most frequent 2000 words of French, German, Spanish
    - NCELP course weekly lists
  - type/token statistics (in-list vs complete file)

Summary

Summary

- Many tools offer vocabulary profiling features
- Profiling multilingual texts can be a challenge
  - definition of a word (at the tokenization stage)
  - general morphological concerns (at the list creation stage)
  - inflectional morphology choices (at the list member grouping stage)
- AntProfiler (desktop) and MultiLingProfiler (online)
  - viewing of general and curriculum-based word-lists
  - profiling of multilingual texts
  - exporting of lists and profiles
    for use classroom materials development and
    research