



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



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



Background

Background

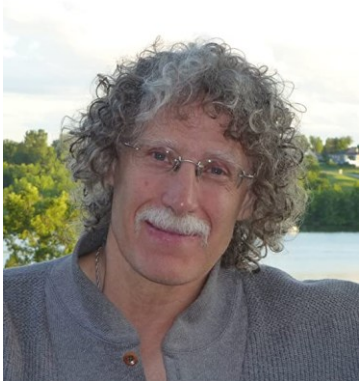
Resources for vocabulary-related research

- 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

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

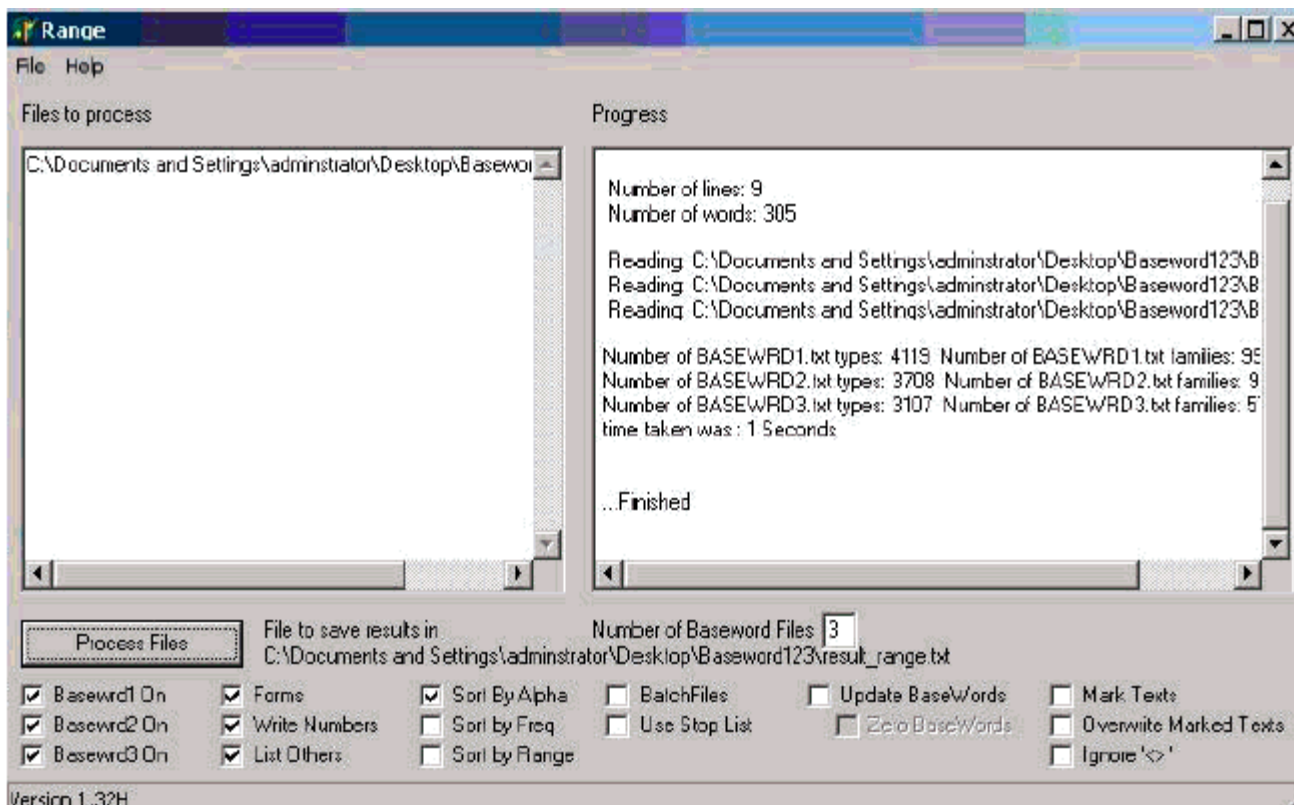
Resources for vocabulary-related research

- 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



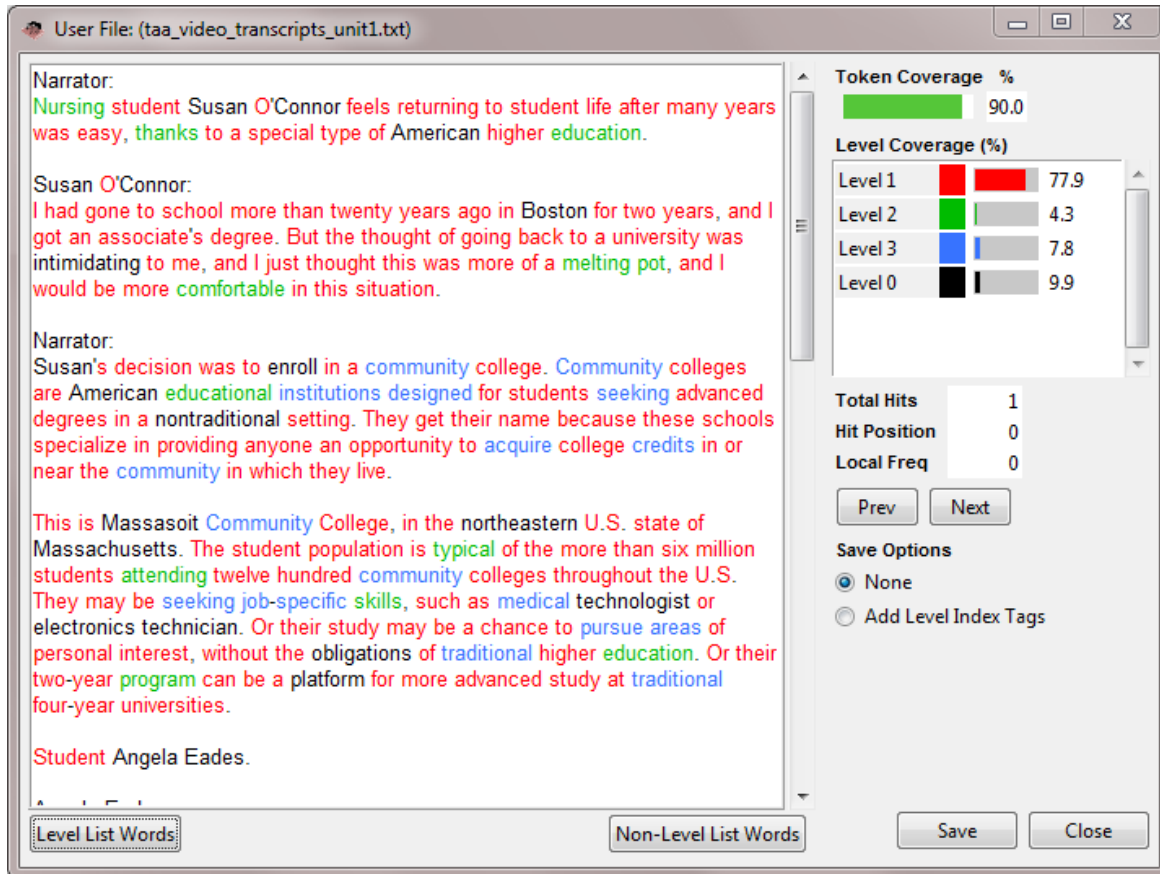
Range (2002)



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 Web VP v.2!
Nine list frameworks at one interface for clear comparisons

Note that BNL, Coca-Core, CEFR, and Classic AWL are not full 1000-family lists (see [?] details) and that NGSL and French are Lemmas not Families

How to make list framework comparisons? Demo 8 [here](#)
Lex Frequency predicts Text Complexity? Check [these](#)

FRAMEWORKS

- CEFR - English NEW [?] Lists
- NGSL + NAWL OR + TOEIC OR + BIZ [?] Lists
- CLASSIC (GSL/AWL) Lists
- BNL [?] Lists
- BNC 1-20k Lists
- BNC-COCA Core-4 [?] Lists
- BNC-COCA 1-25k [?] Lists
- >> BNC-COCA 1-25 "c" (100-family lists) [?] Lists
- FRENCH v.5, 1-25k [?] Lists

Input mode A Type or paste small to medium size text (max 350,000 chars - about 60,000 words) and click yellow Submit button for Frequency Profile.

Title: Untitled | Eng+Fr| Cognates (See Lists) | Edit-to-a-Profile | Sentence Count | Bar Chart | Count Index

INSTRUCTIONS: Type or paste your text here and click the yellow SUBMIT button. VocabProfile will tell you how many words the text contains from frequency bands as determined by analysing research corpora. For a demonstration, enter this text, or one of the sample texts below.

TEXT SET-UP
General: Include an empty space after every comma or full stop.
Research: Deal with spelling errors and proper nouns.

SIZE LIMITS: Web form input is currently max about 400 000 characters - use UPLOAD method below for larger files (must be ~ txt send in straight from your own

Demos : Isogram | Lit (1) (2) | Science (1) (2) | News (1) (2) | Speech Adults Kids | Rex M. | LEGAL | GSL+AWL 1k 2k AWL | French | Highlight | Count | No returns

Words to recategorize => 1k (type or dbl-click) (E.g. Cognates, specific props)
PROPER NOUN HANDLING ?
Mid-sentence capped words...
Class as offlist
Class as 1k/1c
Eliminate from anal.
Plus specific props at sentence boundary => 1k (E.g., "Paul Martin")

SUBMIT

Input mode B Upload larger text files. => LARGER TEXT = LESS RICH INFO

DEMO SPACE to see how upload works, time it takes, etc
Lady Chatterly 1-5 (20,000) | Caps=>1k | Submit

UPLOAD SPACE (for ~ txt files, encoded UTF8 if French NO CAPS/PROPS HANDLING)
1. Browse... No file selected. 2. Choose Freq Scheme at top; 3. Submit_File

About V. Big Texts

Compleat Lexical Tutor (Web VP) (2020)

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



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: *Essen* (noun), *essen* (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)

Our mission is to improve language curriculum design and pedagogy,
leading to a higher uptake and greater success at GCSE.

ABOUT US

We work in partnership with university researchers, teacher educators and expert practitioners, and with 18 Specialist Teachers in nine Leading Schools across the country acting as language hubs, to improve language curriculum design and pedagogy.

[Read More](#)

MODERN FOREIGN LANGUAGE PEDAGOGY

The Centre is funded by the Department for Education (DfE) to take forward the recommendations of the 'Review of MFL Pedagogy' and support their implementation in schools.

[Read More](#)

LANGUAGE HUBS

We drive, support, and monitor the work of a national collaborative network of Modern Foreign Language teachers and their schools to raise the standards of language teaching through the sharing of resources and good practice.

[Read More](#)

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	<p>-ER verbs (je, tu, il/elle)</p> <p>present simple used with its continuous meaning</p> <p>two-verb structures: aimer + infinitive</p>	<p>aimer [242], cocher [>5000], passer¹ [90], porter [105], rester [100], trouver [83], école [477], moment [148], semaine [245], solution [608], uniforme [1801], chaque [151], à¹ [4], avec [23]</p>	<p>faire [25], fais [25], fait [25], ça [54], activité [452], courses [1289], cuisine [2618], devoirs [39], lit [1837], ménage [2326], modèle [958], quoi ? [297]</p>	<p>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]</p>
----------	---	---	---	--

wk 5

9 10	<p>être & avoir (je, tu, il/elle)</p> <p>feminisation of job titles (-e)</p> <p>subject pronouns il/elle meaning 'it'</p> <p><i>indefinite articles</i></p>	<p>il² [13], elle² [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]</p>	<p>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]</p>
---------	---	--	--

wk 2

3 4	<p>être (je, tu, il/elle)</p> <p>regular adjective gender agreement (as complement to verb only);</p> <p>intonation questions</p>	<p>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]</p>	<p>N/A</p>
--------	--	---	------------

- 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>

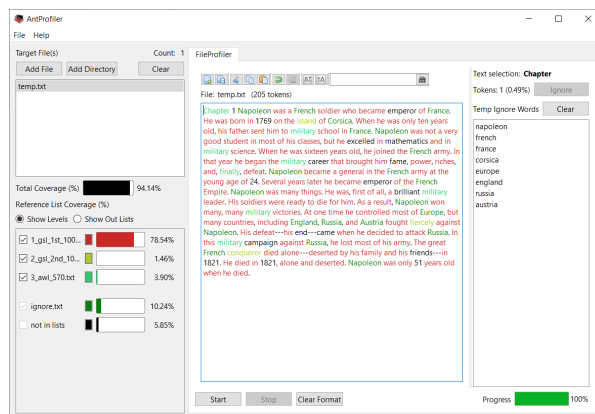
Vocabulary profiling in a multilingual context

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>

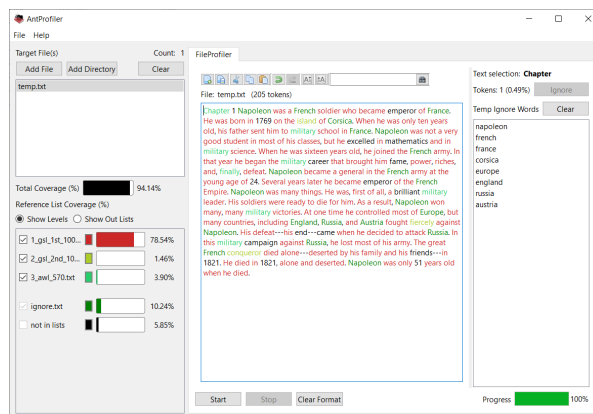
Vocabulary profiling in a multilingual context

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)

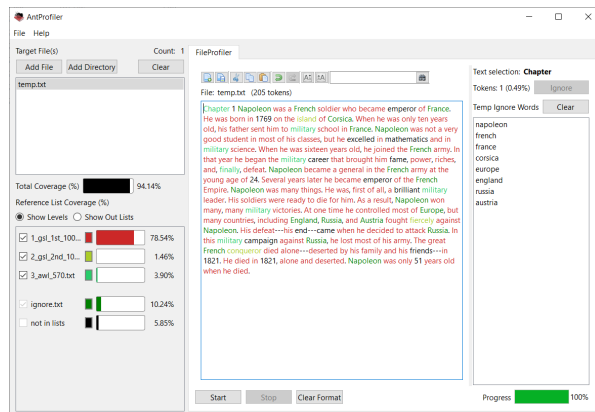
<https://www.laurenceanthony.net/software/antprofiler/>

■ 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)

NCELP National Centre for Excellence
for Language Pedagogy

Home **Dictionaries** MultiLing Profiler About Contact

Dictionaries

French

Show entries Search:

ID	Headword	Family Members
1	le	la, les, l
2	de	dés, d, du
3	un	une
4	à	au, aux
5	être	suis, es, est, sommes, êtes, sont, été, êtes, étée, étées, étais, était, étions, étiez, étaient, serai, seras, sera, serons, serez, seront, sois, soit, soyons, soyez, soient, serais, serait, serions, seriez, seraient, étant
6	et	
7	en	
8	avoir	ai, as, a, avons, avez, ont, eu, avais, avait, avions, aviez, avaient, aurai, auras, aura, aurons, aurez, auront, aie, aies, ait, ayons, ayez, aient, aurais, aurait, aurions, auriez, auraient, ayant, eus, eus, eues
9	que	qu
10	pour	

Search Search Search

Showing 1 to 10 of 2,000 entries Previous **1** 2 3 4 5 ... 200 Next

German

Spanish

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)

NCELP | National Centre for Excellence
for Language Pedagogy

Home Dictionaries **MultiLing Profiler** About Contact

Multilingual Profiler

Language type: French | List type: Weekly List | Year: Year 7 | Term: Term 1 | Week: 2.7

Paste or type any text into the box below to see the target words highlighted.

PREMIER CHAPITRE
Lorsque j'avais six ans j'ai vu, une fois, une magnifique image, dans un livre sur la Forêt Vierge qui s'appelait "Histoires Vécues". Ça représentait un serpent boa qui avalait un fauve. Voilà la copie du dessin.

On disait dans le livre: "Les serpents boas avalent leur proie tout entière, sans la mâcher. Ensuite ils ne peuvent plus bouger et ils dorment pendant les six mois de leur digestion".
J'ai alors beaucoup réfléchi sur les aventures de la jungle et, à mon tour, j'ai réussi, avec un crayon de couleur, à tracer mon premier dessin. Mon dessin numéro 1. Il était comme ça:

J'ai montré mon chef d'oeuvre aux grandes personnes et je leur ai demandé si mon dessin leur faisait peur.
Elles m'ont répondu: "Pourquoi un chapeau ferait-il peur?"
Mon dessin ne représentait pas un chapeau. Il représentait un serpent boa qui digérait un éléphant. J'ai alors dessiné l'intérieur du serpent boa, afin que les grandes personnes puissent comprendre. Elles ont toujours besoin d'explications. Mon dessin numéro 2 était comme ça:

Les grandes personnes m'ont conseillé de laisser de côté les dessins de serpents boas ouverts ou fermés, et de m'intéresser plutôt à la géographie, à l'histoire, au calcul et à la grammaire. C'est ainsi que j'ai abandonné, à l'âge de six ans, une magnifique carrière de peintre. J'avais été découragé par l'insuccès de mon dessin numéro 1 et de mon dessin numéro 2. Les grandes personnes ne comprennent jamais rien toutes seules, et c'est fatigant, pour les enfants, de toujours leur donner des explications.
J'ai donc dû choisir un autre métier et j'ai appris à piloter des avions. J'ai volé un peu partout dans le monde. Et la géographie, c'est exact, m'a beaucoup servi. Je savais reconnaître, du premier coup d'oeil, la Chine de l'Arizona. C'est utile, si l'on est égaré pendant la nuit.
J'ai ainsi eu, au cours de ma vie, des tas de contacts avec des tas de gens sérieux. J'ai beaucoup vécu chez les grandes personnes. Je les ai vues de très près. Ça n'a pas trop amélioré mon opinion.
Quand j'en rencontrais une qui me paraissait un peu lucide, je faisais l'expérience sur elle de mon dessin no.1 que j'ai toujours conservé. Je voulais savoir si elle était vraiment compréhensive. Mais toujours elle me répondait: "C'est un chapeau." Alors je ne lui parlais ni de serpents boas, ni de forêts vierges, ni d'étoiles. Je me mettais à sa portée. Je lui parlais de bridge, de golf, de politique et de cravates. Et la grande personne était bien contente de connaître un homme aussi raisonnable.

Profile Text Copy Results

	tokens in list	types in list	total tokens	total types
statistics	114	25	497	38

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)



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



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