Association Internationale pour la Recherche en Didactique du Français (AIRDF)

Quick start, slow finish:
Learning the lexis of French is like learning to play the guitar

DS-1540

15h10 – 15h50 Bloc J-6

Tom Cobb

http://lextutor.ca/AiRDF_2016.pdf

Association Internationale pour la Recherche en Didactique du Français (AIRDF)

Vite à demarrer, lente à finir :

Acquérir le lexique du français est comme apprendre à jouer au guitare

DS-1540

15h10 – 15h50 Bloc J-6

Tom Cobb

http://lextutor.ca/AiRDF_2016.pptx

Association Internationale pour la Recherche en Didactique du Français (AIRDF)

Earlier title

Profiling French Vocabulary:

The shape of lexicons by frequency & coverage

DS-1540

15h10 – 15h50 Bloc J-6

Tom Cobb

http://lextutor.ca/AiRDF_2016.pptx

Resumé

- Le **profilage de la fréquence lexicale** (PFL, Laufer et Nation, 1995), très influent dans la recherche et l'instruction du vocabulaire en anglais langue seconde (English as a Second Language, ou ESL), a eu un début plutôt lent en français. Ceci est dû notamment au manque d'accès à des grands corpus français à partir desquels des informations pédagogiquement pertinentes sur la fréquence des mots pourraient être dérivées. Des efforts pionniers dans les années 1990 (Goodfellow et Lamy, 2002) ont facilité des comparaisons prometteuses de la **couverture lexicale** des textes en français et en anglais (Cobb & Horst, 2004), ce qui a eu des implications pédagogiques qui étaient à la fois intéressantes et pratiques (Ovtcharov, Cobb & Halter, 2006), mais non concluantes, en raison de l'incomplétude de l'information sur les**fréquences (des mots)**. En revanche, présentement le travail le travail qui sous-tend le **Dictionnaire des fréquences du français** de Lonsdale et Lebras (Routledge 2009) a produit et mis à disposition des informations sur la fréquence des mots autant complète que lemmatisée, tirée de corpus français. Cela signifie que les chercheurs et les enseignants peuvent désormais, en principe, utiliser la méthode de**PFL** pour explorer en profondeur la composition lexicale, la sophistication, et la «richesse» des textes français.

  À être discuté sera la méthode d'intégration des informations sur la fréquence au sein d'une méthodologie **PFL**, des exemples des types de recherche qui rendent possible ce profilage, et les moyens par lesquels les chercheurs peuvent accéder aux outils de cette analyse afin de les utiliser pour leurs propres fins. Les premiers résultats représentatifs de l'application de cette méthodologie en français seront offerts, y compris une suggestion que le français déploie ses ressources lexicales différemment de l'anglais et peut présenter des défis lexicaux nouveaux et précédemment indéfinis à ses apprenants.

Key assumptions

(1) Reading competence is largely lexical competence

(2) Lexical competence includes but is not limited to knowing words

(3) The big problem is WHICH words are most important to know

(4) That word **frequency** is the best available guide to the utility of knowing a word

– And essential to any discussion of "lexical competence"

(5) That learning starts with **recognition** of form and main meaning

--- which is largely sufficient for reading

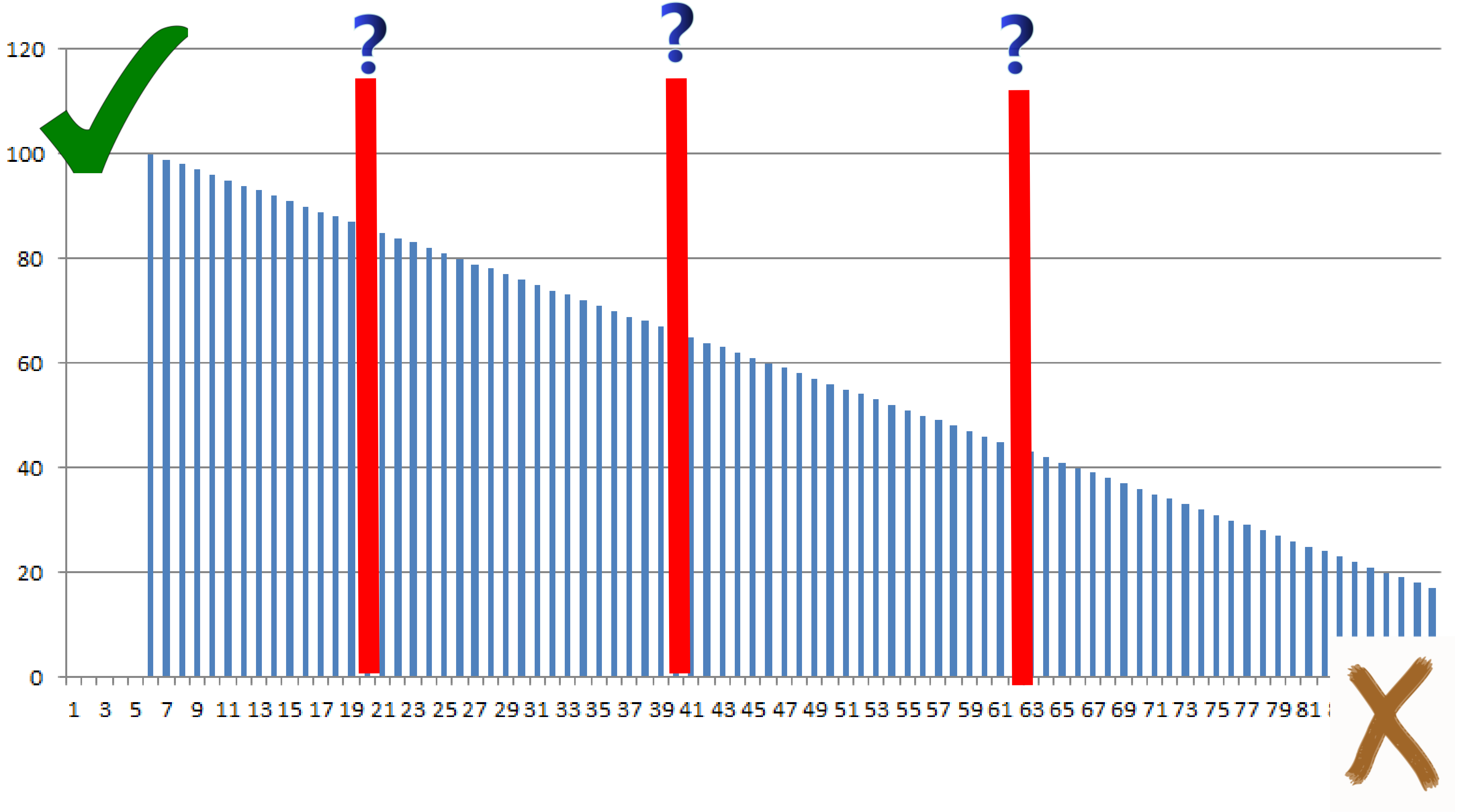
--- with **production** coming later

Frequency - the main new idea of the "vocab revolution" 1990- in ESL/FL...

Is Zipf's old idea that some words get ***way*** more use in any language

But now made useable by corpus technology

Computer + empirical research = **where to draw the line** on frequency



Key Concepts

- **Frequency**
  - Word: The number of occurrences of a word in a corpus
  - Family: The occurrences of a whole word family in a corpus
    - Family = Word + Inflections + derivations
- **Frequency Band**
  - Groups of (usually 1,000) word families (or *k-lists*)
- **Frequency profile**
  - The % of word tokens in a particular text that are from each band
    - E.g, 70% from first 1,000, 10% from 2nd 1,000, etc.

Example

- "The cat sat on the mat"

The	1k
Cat	1k
Sat	1k
On	1k
The	1k
Mat	4k

- Six words = 100% of text
  - 1k items = 5/6 of text = 83%
- So 1k gives 83% ***coverage*** in this text
  - Or "accounts for" 83% of the tokens"

So the profile is:

- 1k=83%
- 4k=17%

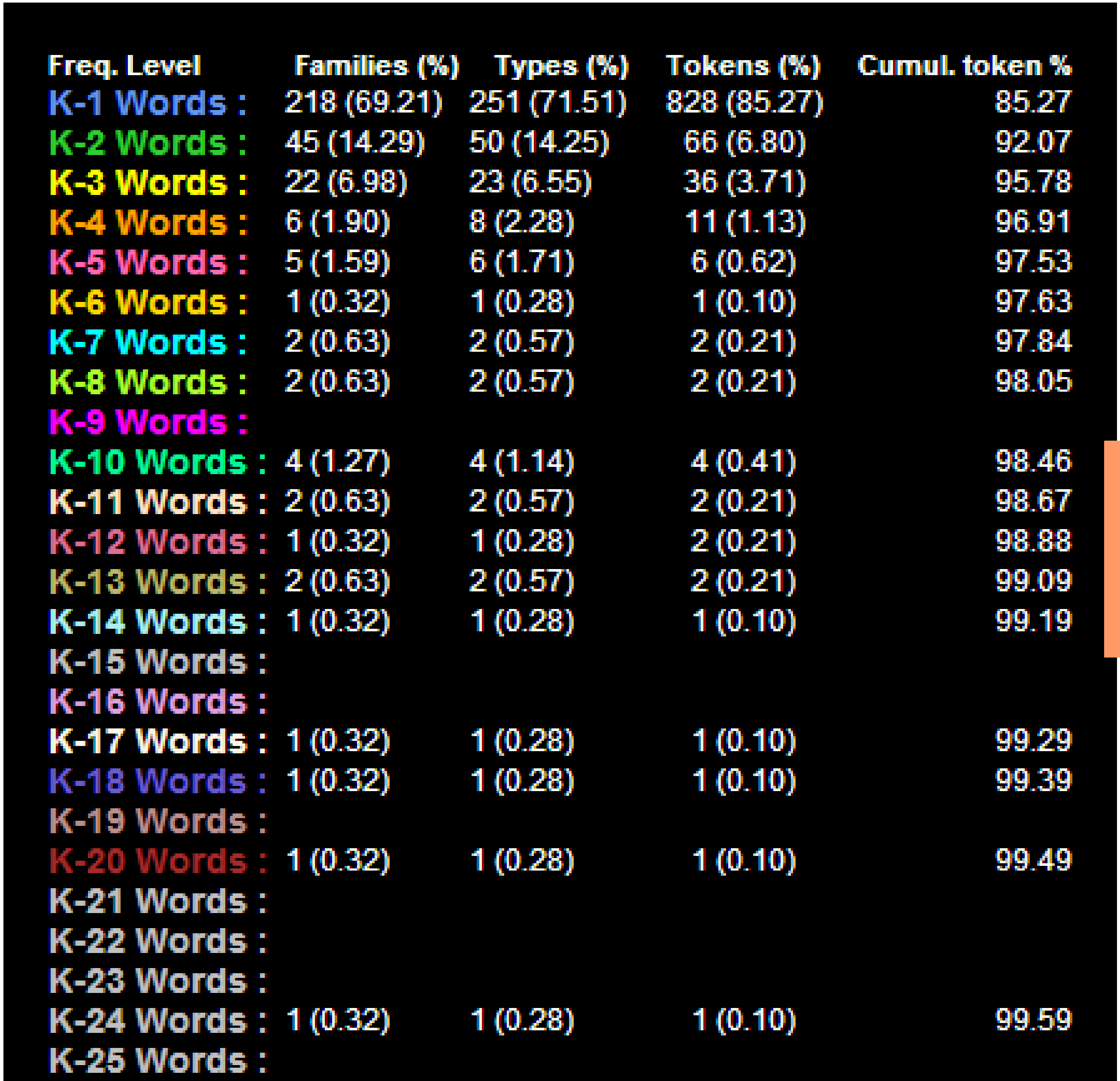
The pedagogical question is:

- Can a learner with 1,000 words 'read' this text?
  - I.e. infer the meaning of 'mat' to build a semantic model of the entire proposition

The empirical research is:

- 95% coverage is needed for reliable inference
  - So 'mat' here would be Maybe

Visual of a VP for Text “x” (v. 2016)



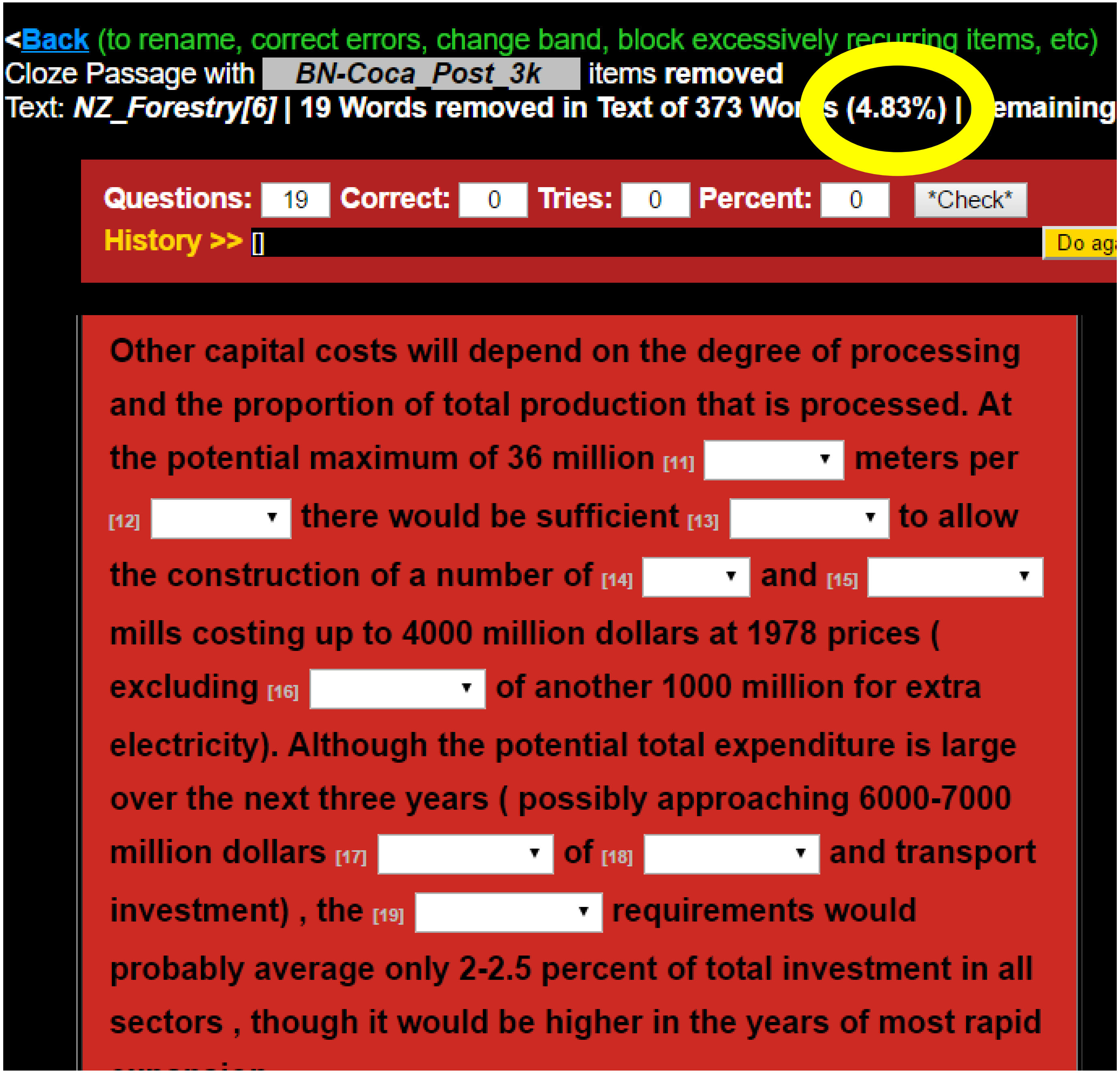
Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	218 (69.21)	251 (71.51)	828 (85.27)	85.27
K-2 Words :	45 (14.29)	50 (14.25)	66 (6.80)	92.07
K-3 Words :	22 (6.98)	23 (6.55)	36 (3.71)	95.78
K-4 Words :	6 (1.90)	8 (2.28)	11 (1.13)	96.91
K-5 Words :	5 (1.59)	6 (1.71)	6 (0.62)	97.53
K-6 Words :	1 (0.32)	1 (0.28)	1 (0.10)	97.63
K-7 Words :	2 (0.63)	2 (0.57)	2 (0.21)	97.84
K-8 Words :	2 (0.63)	2 (0.57)	2 (0.21)	98.05
K-9 Words :				
K-10 Words :	4 (1.27)	4 (1.14)	4 (0.41)	98.46
K-11 Words :	2 (0.63)	2 (0.57)	2 (0.21)	98.67
K-12 Words :	1 (0.32)	1 (0.28)	2 (0.21)	98.88
K-13 Words :	2 (0.63)	2 (0.57)	2 (0.21)	99.09
K-14 Words :	1 (0.32)	1 (0.28)	1 (0.10)	99.19
K-15 Words :				
K-16 Words :				
K-17 Words :	1 (0.32)	1 (0.28)	1 (0.10)	99.29
K-18 Words :	1 (0.32)	1 (0.28)	1 (0.10)	99.39
K-19 Words :				
K-20 Words :	1 (0.32)	1 (0.28)	1 (0.10)	99.49
K-21 Words :				
K-22 Words :				
K-23 Words :				
K-24 Words :	1 (0.32)	1 (0.28)	1 (0.10)	99.59
K-25 Words :				

http://lextutor.ca/vp/

Key concept:
Minimal Lexical Competence for reading

- Defined in English as knowing 95% of the words in a text
  - Or, when your lexical knowledge cover 95% of the words in a text
    - Or, your knowledge gives you 95% coverage of a text

–As determined how?



<Back (to rename, correct errors, change band, block excessively recurring items, etc)

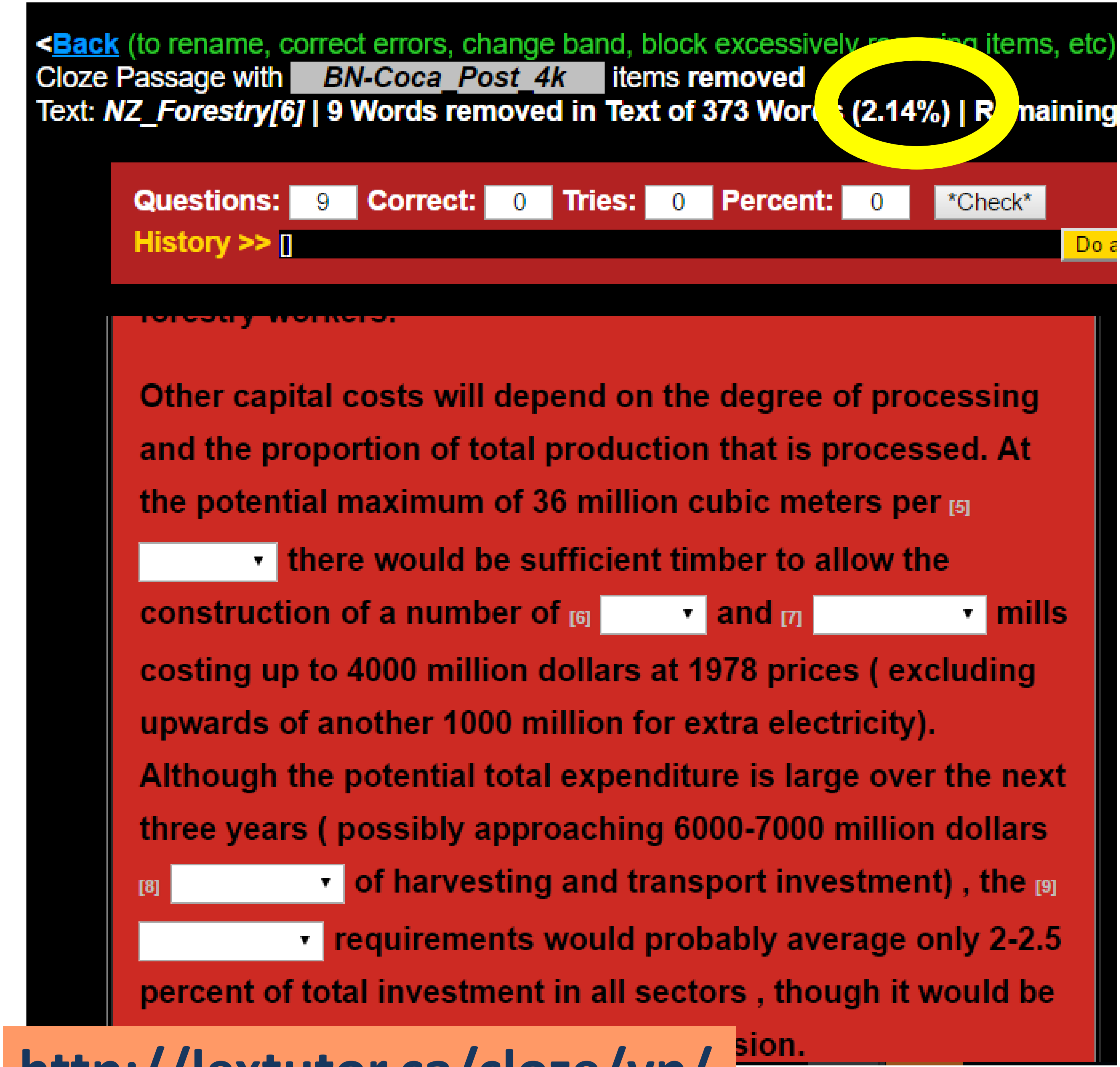
Cloze Passage with BN-Coca_Post_3k items removed

Text: *NZ_Forestry[6]* | 19 Words removed in Text of 373 Wor s (4.83%) | emaining

Questions: 19 Correct: 0 Tries: 0 Percent: 0 *Check*

History >> []

Other capital costs will depend on the degree of processing and the proportion of total production that is processed. At the potential maximum of 36 million [11] ▾ meters per [12] ▾ there would be sufficient [13] ▾ to allow the construction of a number of [14] ▾ and [15] ▾ mills costing up to 4000 million dollars at 1978 prices (excluding [16] ▾ of another 1000 million for extra electricity). Although the potential total expenditure is large over the next three years (possibly approaching 6000-7000 million dollars [17] ▾ of [18] ▾ and transport investment) , the [19] ▾ requirements would probably average only 2-2.5 percent of total investment in all sectors , though it would be higher in the years of most rapid



<Back (to rename, correct errors, change band, block excessively r ing items, etc)

Cloze Passage with BN-Coca_Post_4k items removed

Text: *NZ_Forestry[6]* | 9 Words removed in Text of 373 Wor (2.14%) | R maining

forestry workers.

Other capital costs will depend on the degree of processing and the proportion of total production that is processed. At the potential maximum of 36 million cubic meters per [5] ▾ there would be sufficient timber to allow the construction of a number of [6] ▾ and [7] ▾ mills costing up to 4000 million dollars at 1978 prices (excluding upwards of another 1000 million for extra electricity). Although the potential total expenditure is large over the next three years (possibly approaching 6000-7000 million dollars [8] ▾ of harvesting and transport investment) , the [9] ▾ requirements would probably average only 2-2.5 percent of total investment in all sectors , though it would be

sion.

http://lextutor.ca/cloze/vp/

Classic coverage figures for English

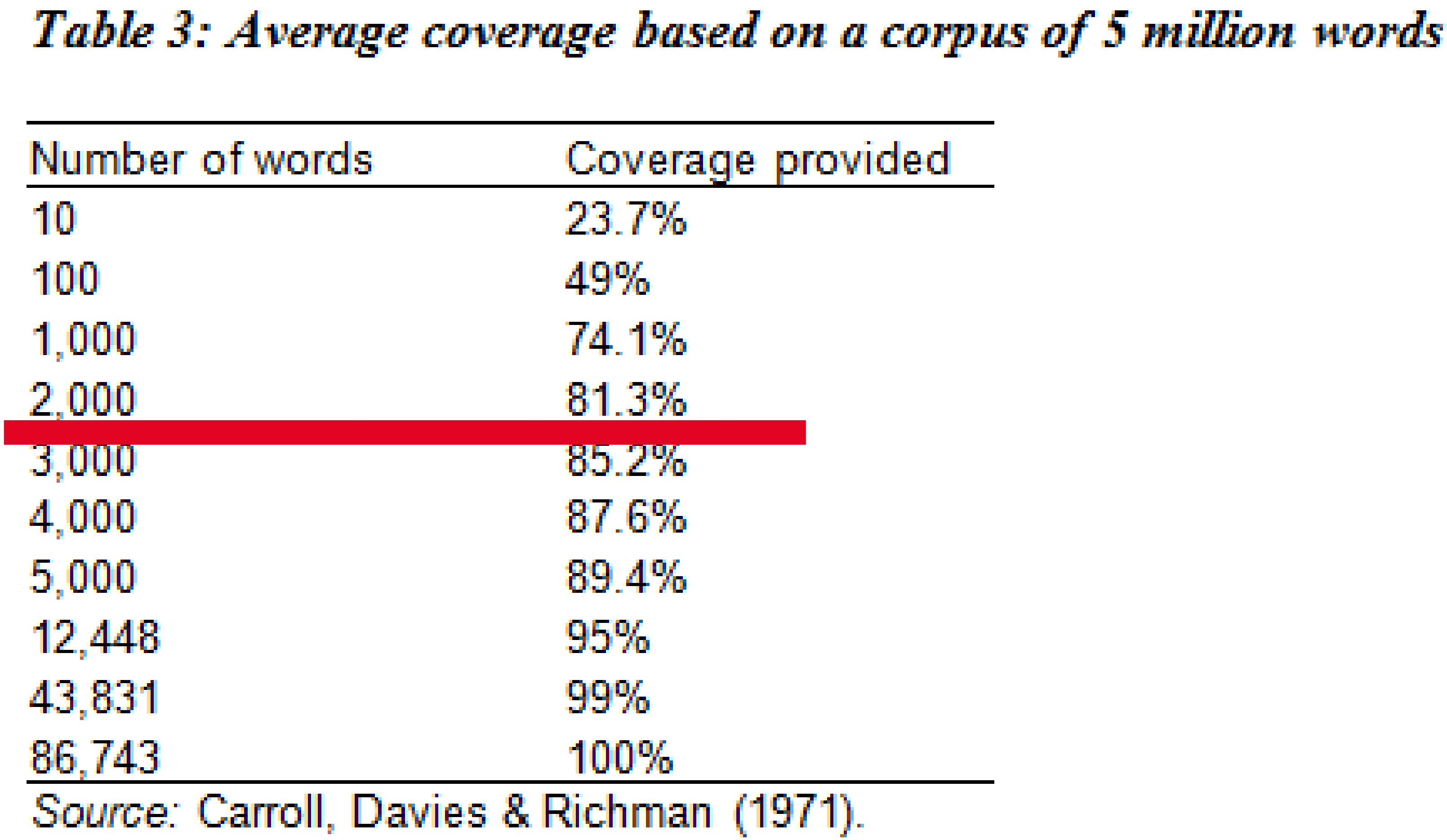
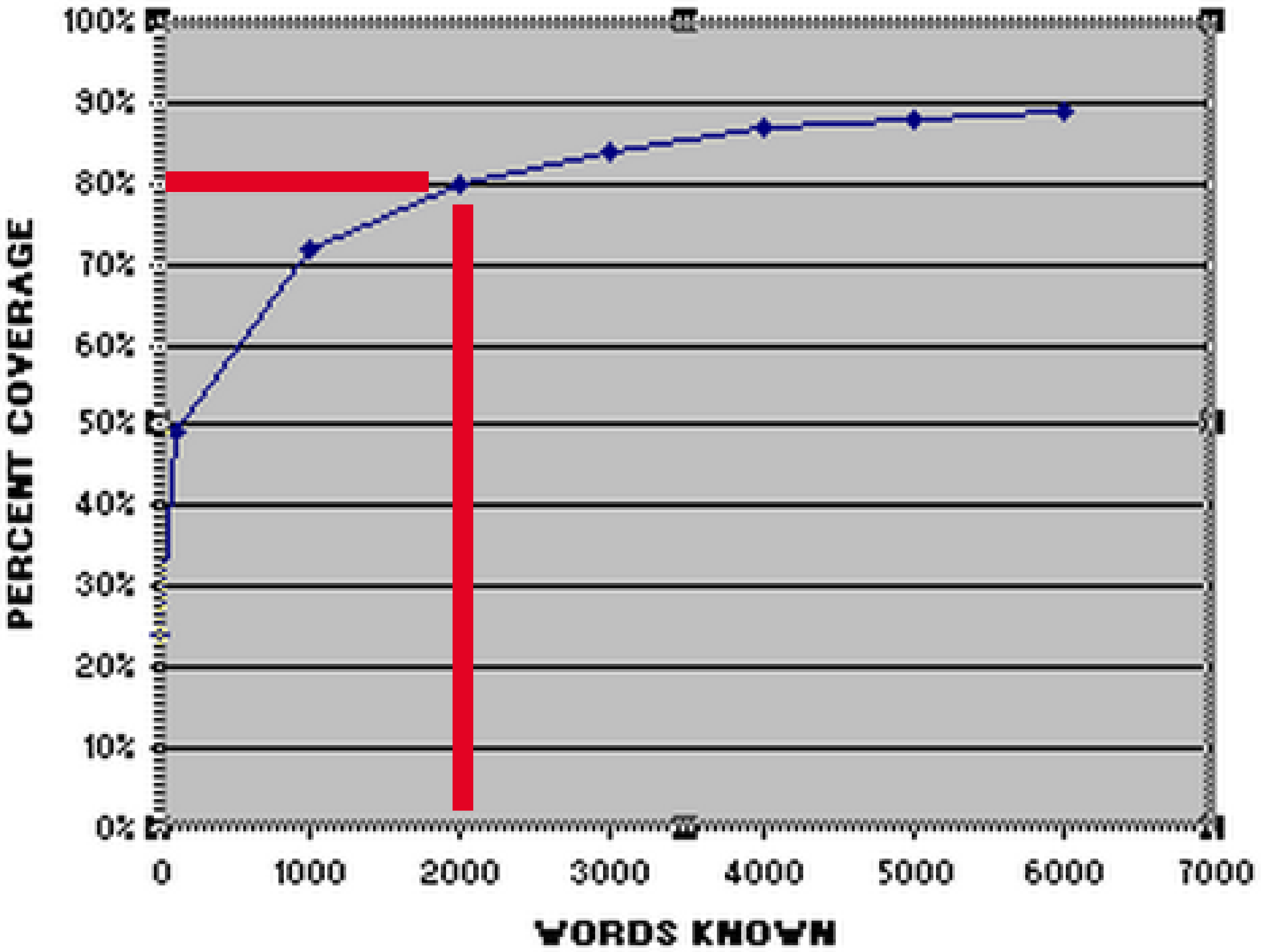
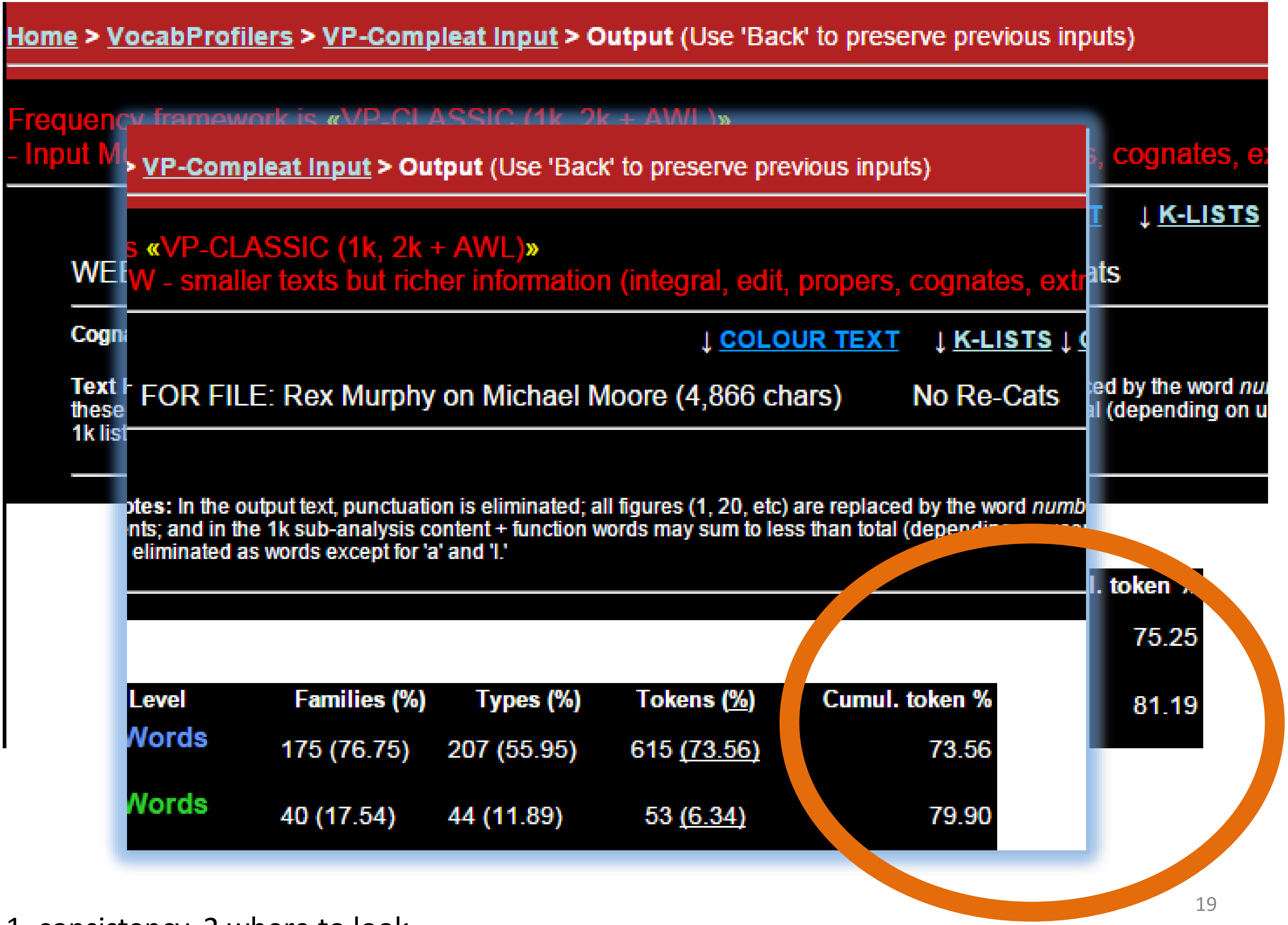


Table 3: Average coverage based on a corpus of 5 million words

Number of words	Coverage provided
10	23.7%
100	49%
1,000	74.1%
2,000	81.3%
3,000	85.2%
4,000	87.6%
5,000	89.4%
12,448	95%
43,831	99%
86,743	100%

Source: Carroll, Davies & Richman (1971).


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
PERCENT COVERAGE
0
1000
2000
3000
4000
5000
6000
7000
WORDS KNOWN



Home > VocabProfilers > VP-Compleat Input > Output (Use 'Back' to preserve previous inputs)

Frequency framework is «VP-CLASSIC (1k, 2k + AWL)»

- Input M

s, cognates, ex

T ↓ K-LISTS

WEB

ats

Cogn

Text

these

1k list

ed by the word *nu*

al (depending on u

. token %

75.25

81.19

> VP-Compleat Input > Output (Use 'Back' to preserve previous inputs)

s «VP-CLASSIC (1k, 2k + AWL)»

W - smaller texts but richer information (integral, edit, propers, cognates, extr

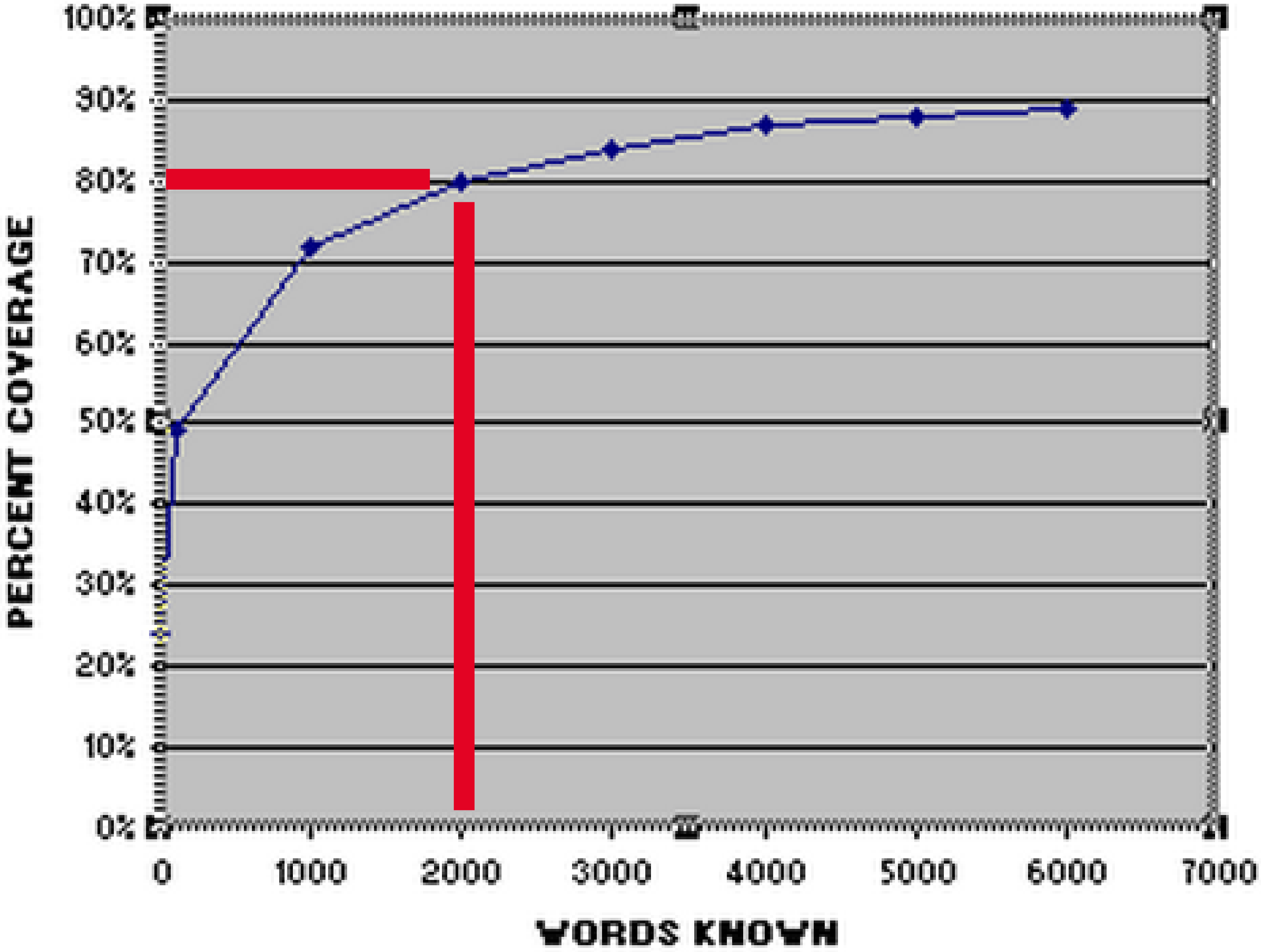
↓ COLOUR TEXT ↓ K-LISTS ↓ C

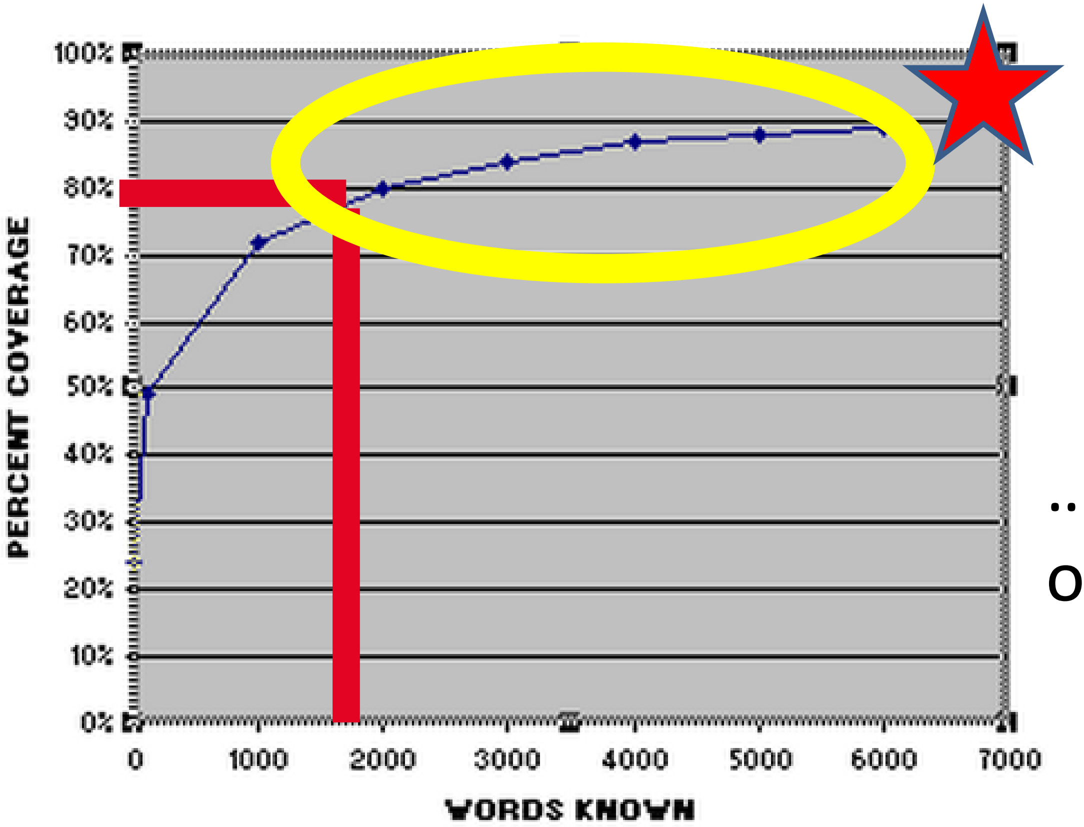
FOR FILE: Rex Murphy on Michael Moore (4,866 chars) No Re-Cats

otes: In the output text, punctuation is eliminated; all figures (1, 20, etc) are replaced by the word *numb*
nts; and in the 1k sub-analysis content + function words may sum to less than total (dependi
eliminated as words except for 'a' and 'I.'

Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
Words	175 (76.75)	207 (55.95)	615 (73.56)	73.56
Words	40 (17.54)	44 (11.89)	53 (6.34)	79.90

1, consistency, 2 where to look


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
PERCENT COVERAGE
0
1000
2000
3000
4000
5000
6000
7000
WORDS KNOWN

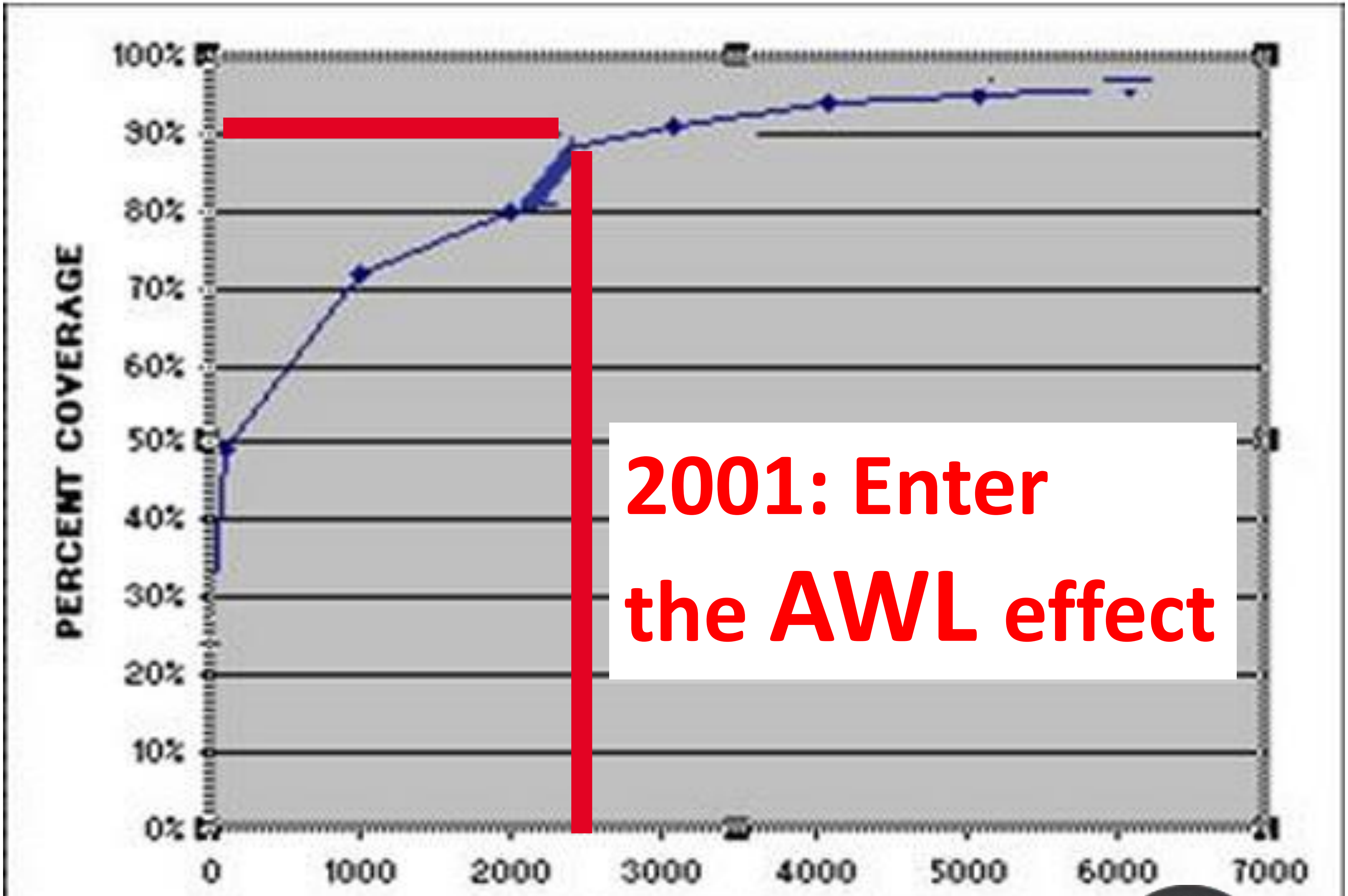


So while 2,000 families
= 80% coverage
is good news …

…Attention soon focused on the flat curve beyond

Especially as empirical research showed basic comprehension depends on **95%** words known

-e.g. Laufer 1989



2001: Enter the **AWL** effect

Averil Coxhead (2001), New Academic word List, TESOL Quarterly

Fairly uniform across disciplines

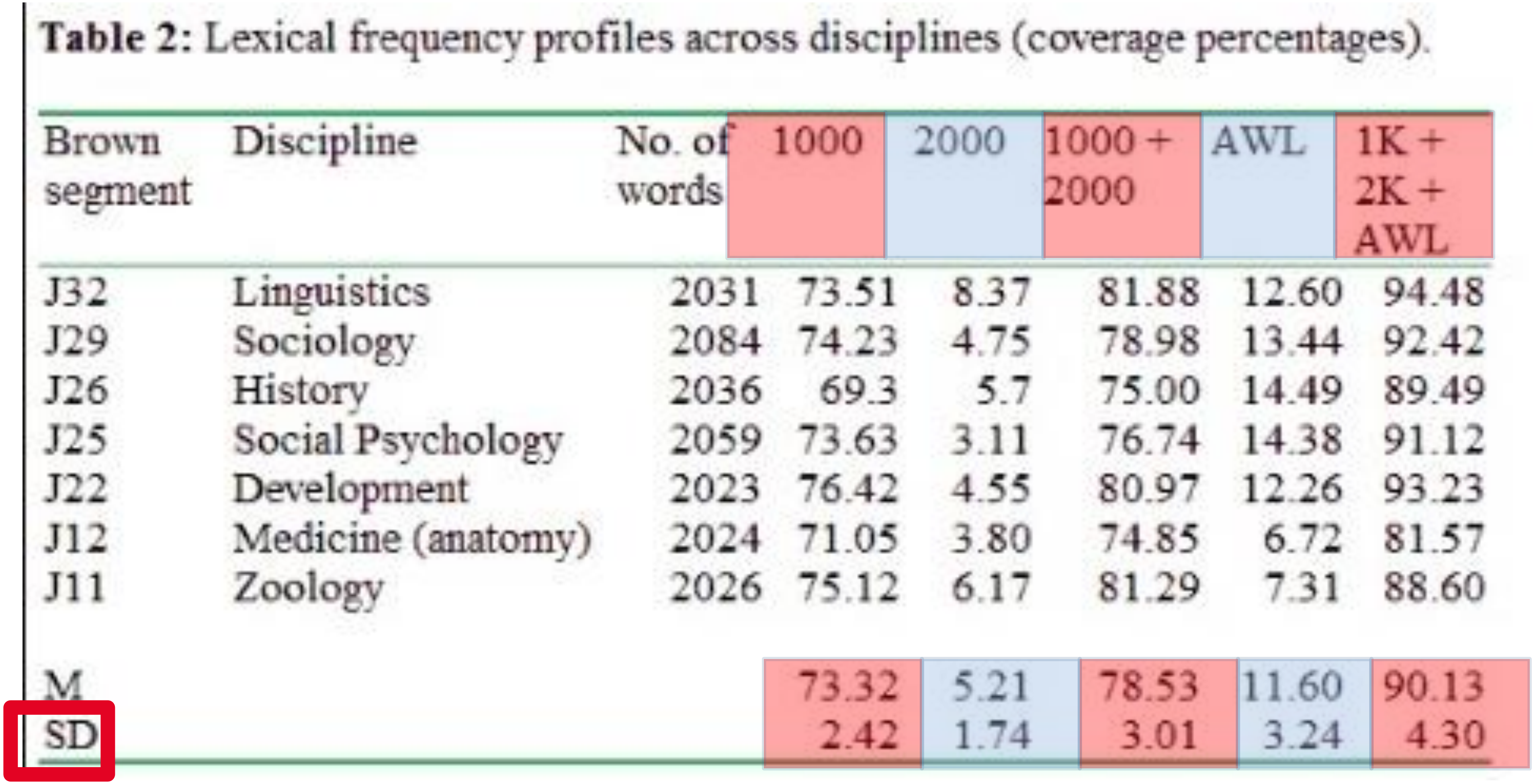


Table 2: Lexical frequency profiles across disciplines (coverage percentages).

Brown segment	Discipline	No. of words	1000	2000	1000 + 2000	AWL	1K + 2K + AWL
J32	Linguistics	2031	73.51	8.37	81.88	12.60	94.48
J29	Sociology	2084	74.23	4.75	78.98	13.44	92.42
J26	History	2036	69.3	5.7	75.00	14.49	89.49
J25	Social Psychology	2059	73.63	3.11	76.74	14.38	91.12
J22	Development	2023	76.42	4.55	80.97	12.26	93.23
J12	Medicine (anatomy)	2024	71.05	3.80	74.85	6.72	81.57
J11	Zoology	2026	75.12	6.17	81.29	7.31	88.60
M			73.32	5.21	78.53	11.60	90.13
SD			2.42	1.74	3.01	3.24	4.30

So it was a reasonable question to ask, "Is there an AWL in French?"

An interesting question for several reasons…

<u>1 PRACTICE</u>:

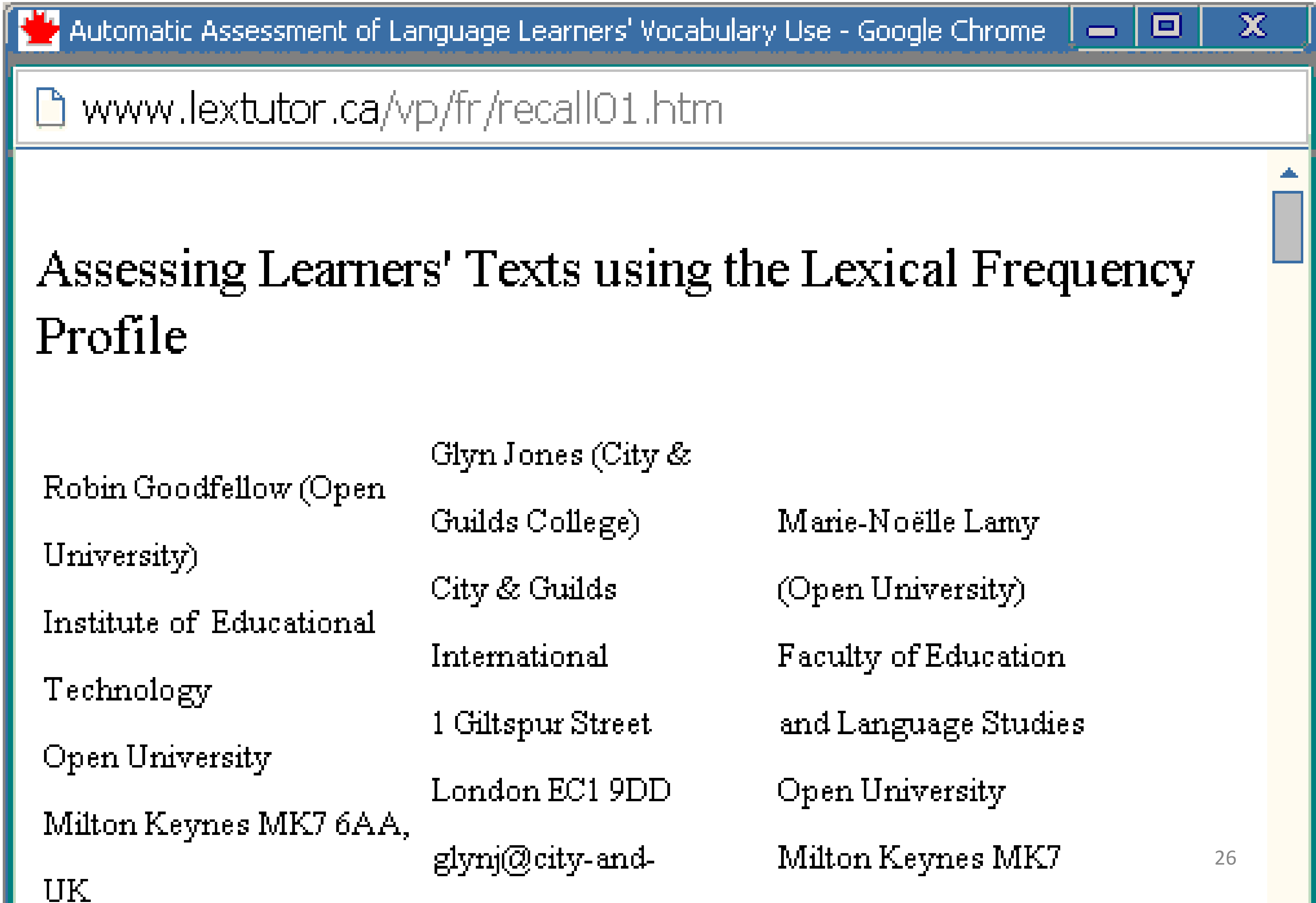
Investigate lexical competence in French
on behalf of FL2 learners

<u>2 THEORY</u>:

Investigate a curious puzzle
Since English AWL basically = French cognates…
So in French are these terms "academic words" or common words?
Within or beyond 2k?

An interesting question …

Which it gradually became possible to answer

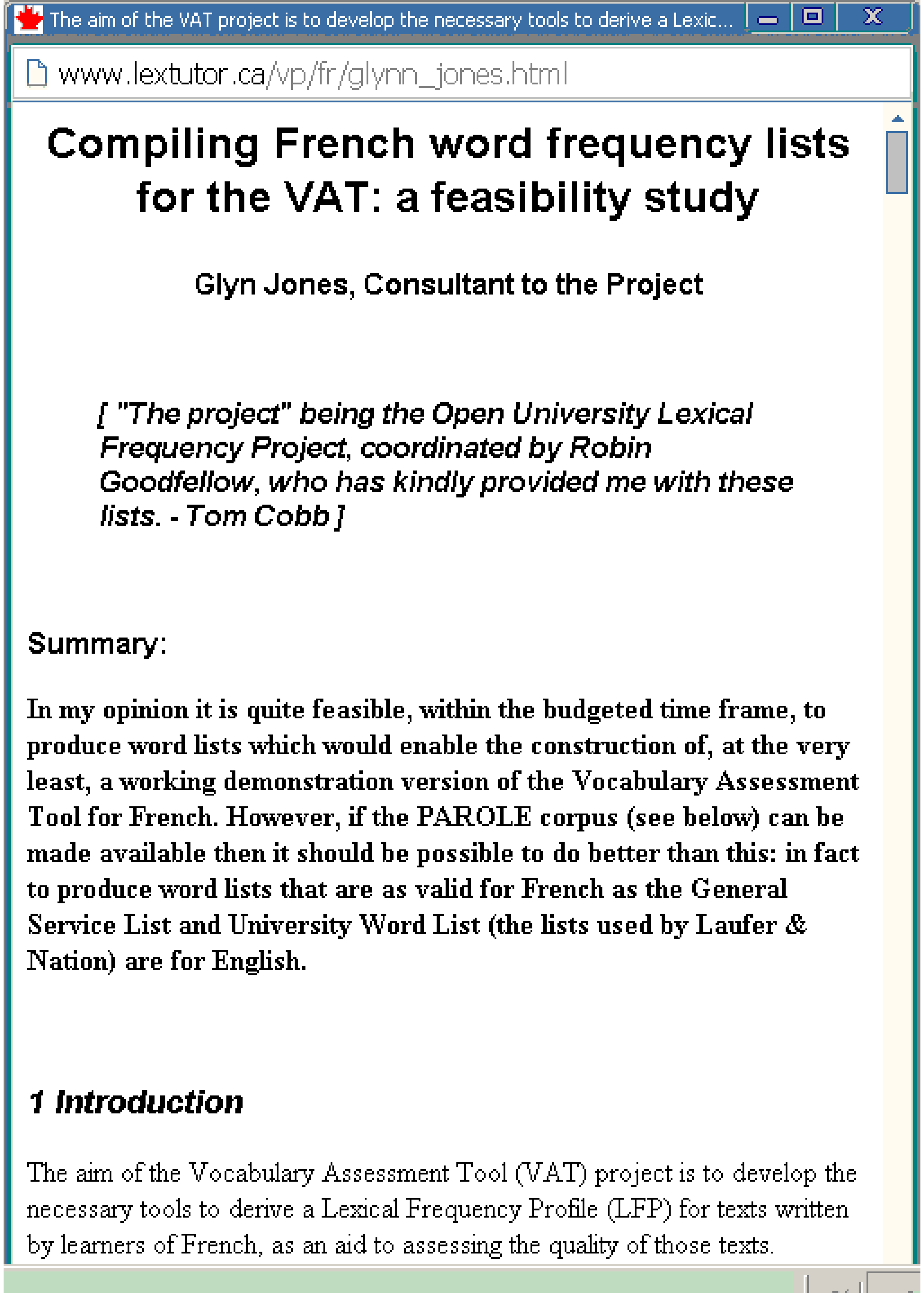

Automatic Assessment of Language Learners' Vocabulary Use - Google Chrome
www.lextutor.ca/vp/fr/recall01.htm

Assessing Learners' Texts using the Lexical Frequency Profile

Robin Goodfellow (Open University)
Institute of Educational Technology
Open University
Milton Keynes MK7 6AA, UK

Glyn Jones (City & Guilds College)
City & Guilds International
1 Giltspur Street
London EC1 9DD
glynj@city-and-

Marie-Noëlle Lamy
(Open University)
Faculty of Education
and Language Studies
Open University
Milton Keynes MK7


The aim of the VAT project is to develop the necessary tools to derive a Lexic...
www.lextutor.ca/vp/fr/glynn_jones.html

Compiling French word frequency lists for the VAT: a feasibility study

Glyn Jones, Consultant to the Project

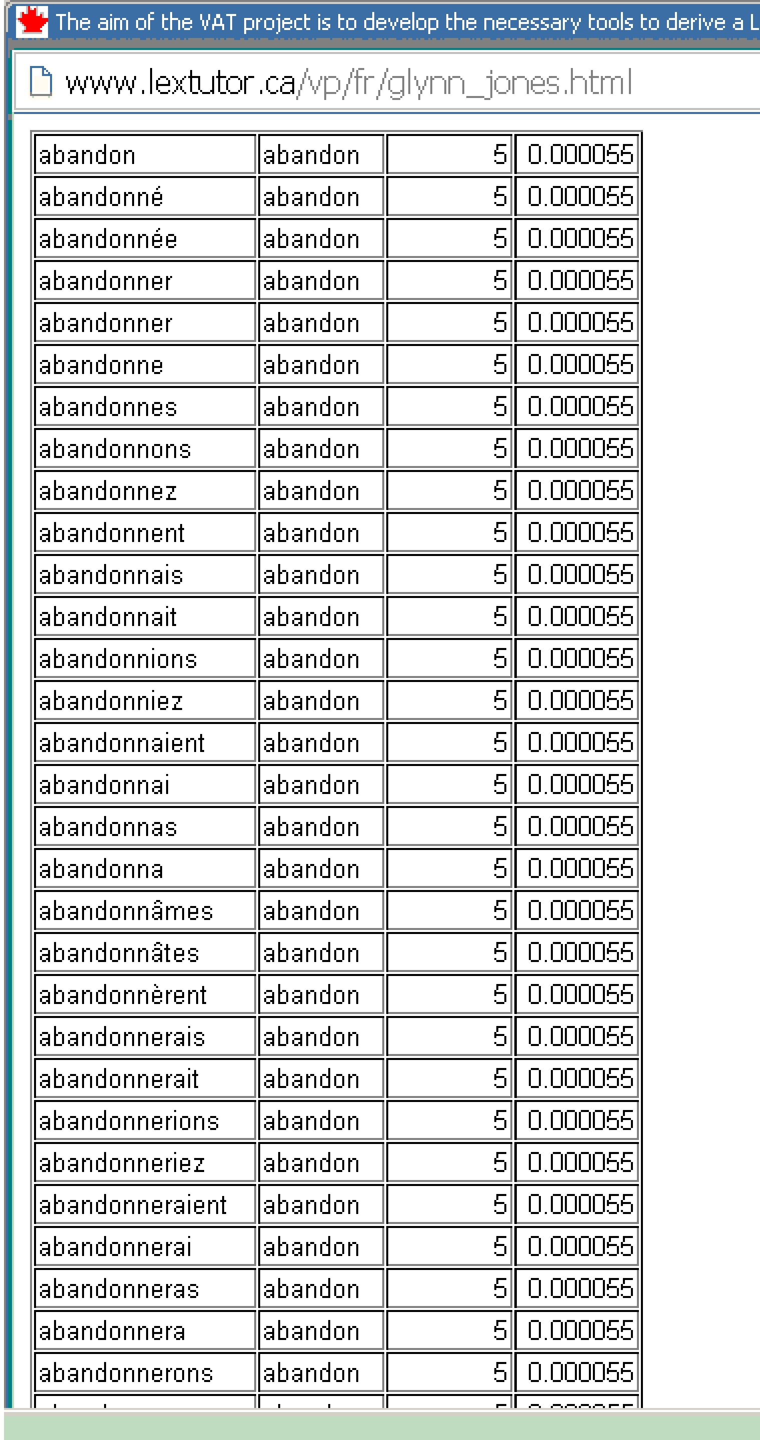
["The project" being the Open University Lexical Frequency Project, coordinated by Robin Goodfellow, who has kindly provided me with these lists. - Tom Cobb]

Summary:

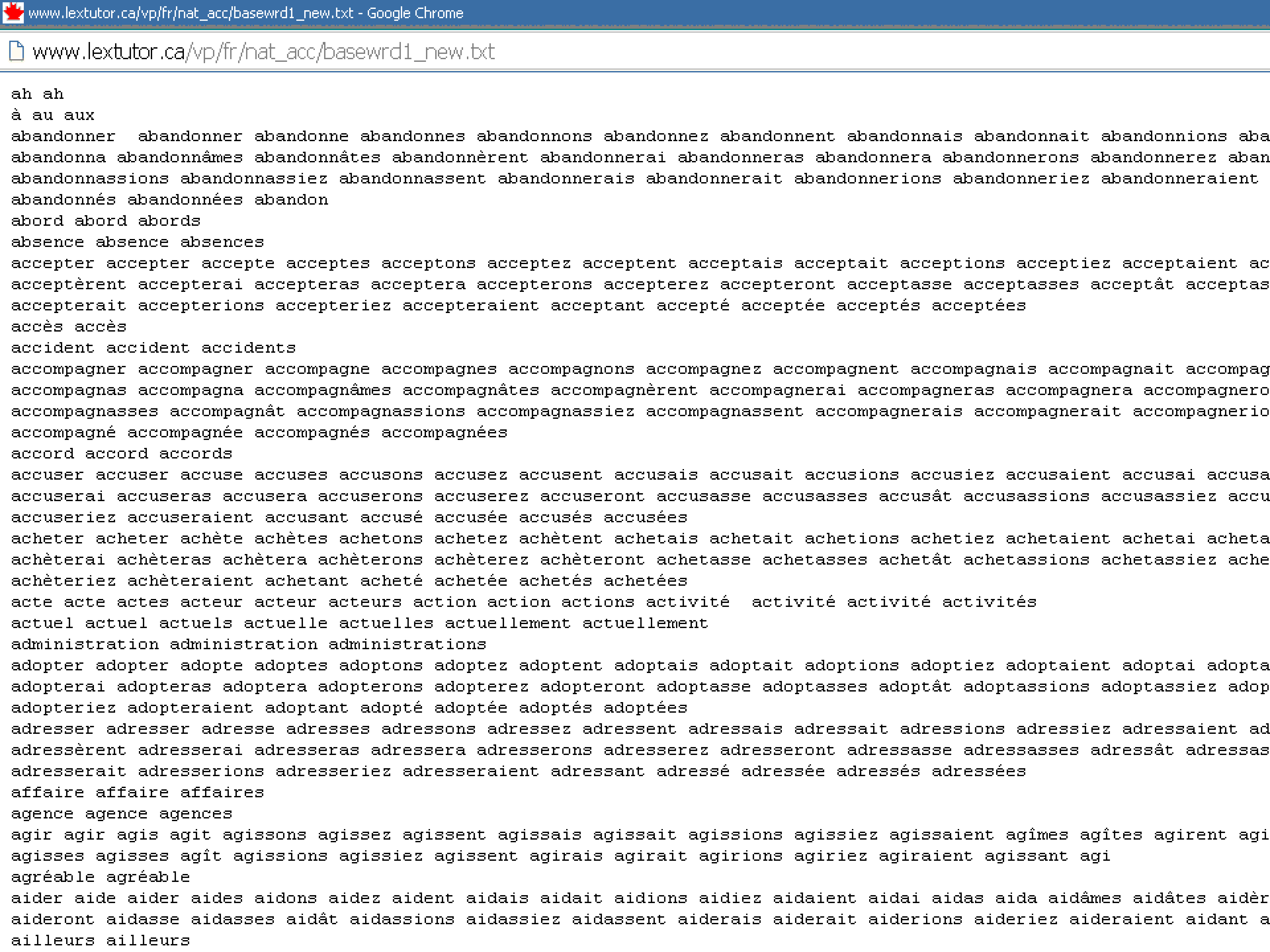
In my opinion it is quite feasible, within the budgeted time frame, to produce word lists which would enable the construction of, at the very least, a working demonstration version of the Vocabulary Assessment Tool for French. However, if the PAROLE corpus (see below) can be made available then it should be possible to do better than this: in fact to produce word lists that are as valid for French as the General Service List and University Word List (the lists used by Laufer & Nation) are for English.

1 Introduction

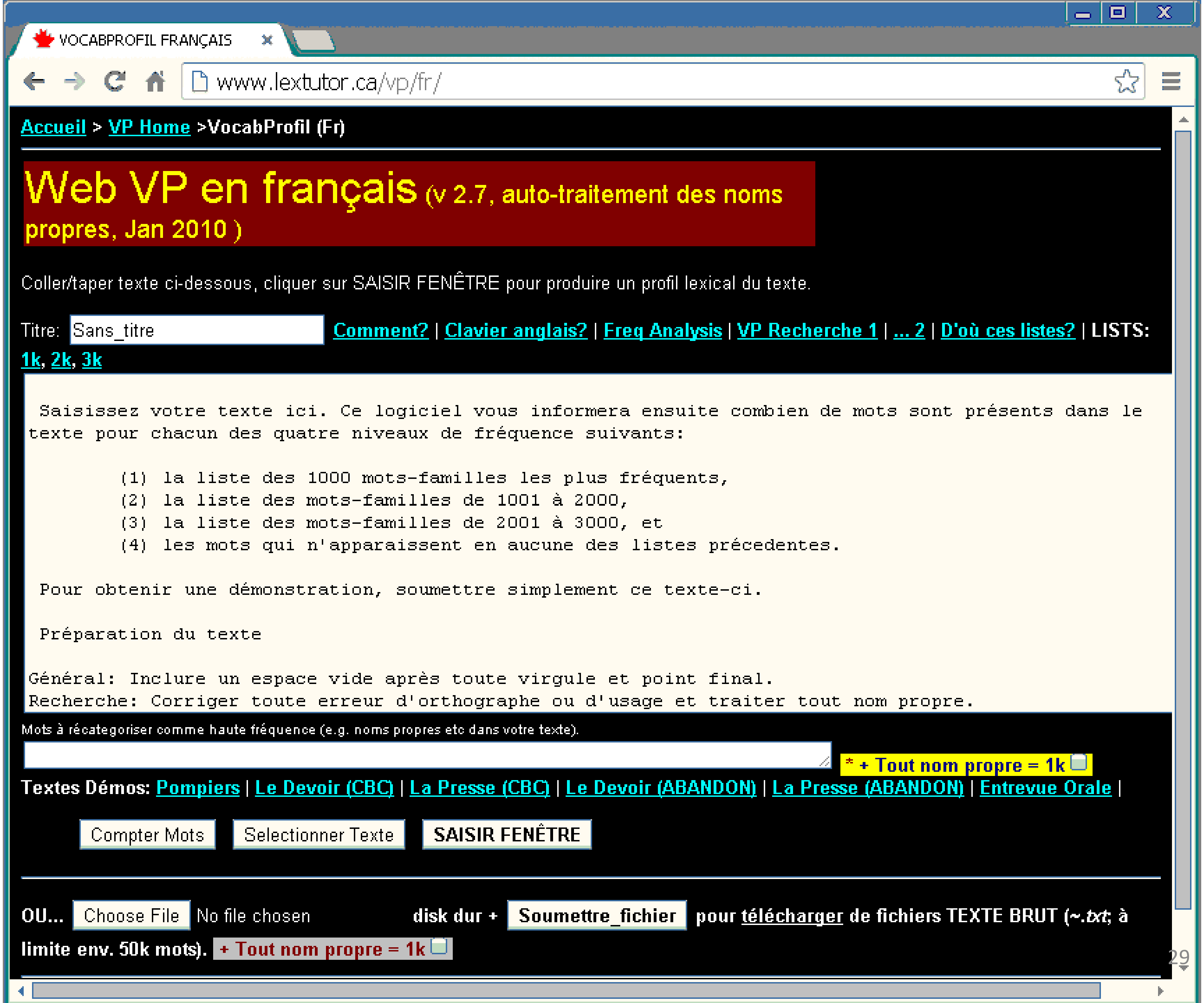
The aim of the Vocabulary Assessment Tool (VAT) project is to develop the necessary tools to derive a Lexical Frequency Profile (LFP) for texts written by learners of French, as an aid to assessing the quality of those texts.


The aim of the VAT project is to develop the necessary tools to derive a L
www.lextutor.ca/vp/fr/glynn_jones.html

abandon	abandon	5	0.000055
abandonné	abandon	5	0.000055
abandonnée	abandon	5	0.000055
abandonner	abandon	5	0.000055
abandonner	abandon	5	0.000055
abandonne	abandon	5	0.000055
abandonnes	abandon	5	0.000055
abandonnons	abandon	5	0.000055
abandonnez	abandon	5	0.000055
abandonnent	abandon	5	0.000055
abandonnais	abandon	5	0.000055
abandonnait	abandon	5	0.000055
abandonnions	abandon	5	0.000055
abandonniez	abandon	5	0.000055
abandonnaient	abandon	5	0.000055
abandonnai	abandon	5	0.000055
abandonnas	abandon	5	0.000055
abandonna	abandon	5	0.000055
abandonnâmes	abandon	5	0.000055
abandonnâtes	abandon	5	0.000055
abandonnèrent	abandon	5	0.000055
abandonnerais	abandon	5	0.000055
abandonnerait	abandon	5	0.000055
abandonnerions	abandon	5	0.000055
abandonneriez	abandon	5	0.000055
abandonneraient	abandon	5	0.000055
abandonnerai	abandon	5	0.000055
abandonneras	abandon	5	0.000055
abandonnera	abandon	5	0.000055
abandonnerons	abandon	5	0.000055


www.lextutor.ca/vp/fr/nat_acc/basewrd1_new.txt - Google Chrome
www.lextutor.ca/vp/fr/nat_acc/basewrd1_new.txt

```
ah ah
à au aux
abandonner  abandonner abandonne abandonnes abandonnons abandonnez abandonnent abandonnais abandonnait abandonnions aba
abandonna abandonnâmes abandonnâtes abandonnèrent abandonnerai abandonneras abandonnera abandonnerons abandonnerez aban
abandonnassions abandonnassiez abandonnassent abandonnerais abandonnerait abandonnerions abandonneriez abandonneraient
abandonnés abandonnées abandon
abord abord abords
absence absence absences
accepter accepter accepte acceptes acceptons acceptez acceptent acceptais acceptait acceptions acceptiez acceptaient ac
acceptèrent accepterai accepteras acceptera accepterons accepterez accepteront acceptasse acceptasses acceptât acceptas
accepterait accepterions accepteriez accepteraient acceptant accepté acceptée acceptés acceptées
accès accès
accident accident accidents
accompagner accompagner accompagne accompagnes accompagnons accompagnez accompagnent accompagnais accompagnait accompag
accompagnas accompagna accompagnâmes accompagnâtes accompagnèrent accompagnerai accompagneras accompagnera accompagnero
accompagnasses accompagnât accompagnassions accompagnassiez accompagnassent accompagnerais accompagnerait accompagnerio
accompagné accompagnée accompagnés accompagnées
accord accord accords
accuser accuser accuse accuses accusons accusez accusent accusais accusait accusions accusiez accusaient accusai accusa
accuserai accuseras accusera accuserons accuserez accuseront accusasse accusasses accusât accusassions accusassiez accu
accuseriez accuseraient accusant accusé accusée accusés accusées
acheter acheter achète achètes achetons achetez achètent achetais achetait achetions achetiez achetaient achetai acheta
achèterai achèteras achètera achèterons achèterez achèteront achetasse achetasses achetât achetassions achetassiez ache
achèteriez achèteraient achetant acheté achetée achetés achetées
acte acte actes acteur acteur acteurs action action actions activité  activité activité activités
actuel actuel actuels actuelle actuelles actuellement actuellement
administration administration administrations
adopter adopter adopte adoptes adoptons adoptez adoptent adoptais adoptait adoptions adoptiez adoptaient adoptai adopta
adopterai adopteras adoptera adopterons adopterez adopteront adoptasse adoptasses adoptât adoptassions adoptassiez adop
adopteriez adopteraient adoptant adopté adoptée adoptés adoptées
adresser adresser adresse adresses adressons adressez adressent adressais adressait adressions adressiez adressaient ad
adressèrent adresserai adresseras adressera adresserons adresserez adresseront adressasse adressasses adressât adressas
adresserait adresserions adresseriez adresseraient adressant adressé adressée adressés adressées
affaire affaire affaires
agence agence agences
agir agir agis agit agissons agissez agissent agissais agissait agissions agissiez agissaient agîmes agîtes agirent agi
agisses agisses agît agissions agissiez agissent agirais agirait agirions agiriez agiraient agissant agi
agréable agréable
aider aide aider aides aidons aidez aident aidais aidait aidions aidiez aidaient aidai aidas aida aidâmes aidâtes aidèr
aideront aidasse aidasses aidât aidassions aidassiez aidassent aiderais aiderait aiderions aideriez aideraient aidant a
ailleurs ailleurs
```



Accueil > VP Home >VocabProfil (Fr)

Web VP en français (v 2.7, auto-traitement des noms propres, Jan 2010)

Coller/taper texte ci-dessous, cliquer sur SAISIR FENÊTRE pour produire un profil lexical du texte.

Titre: Sans_titre Comment? | Clavier anglais? | Freq Analysis | VP Recherche 1 | ... 2 | D'où ces listes? | LISTS: 1k, 2k, 3k

```
 Saisissez votre texte ici. Ce logiciel vous informera ensuite combien de mots sont présents dans le
texte pour chacun des quatre niveaux de fréquence suivants:

        (1) la liste des 1000 mots-familles les plus fréquents,
        (2) la liste des mots-familles de 1001 à 2000,
        (3) la liste des mots-familles de 2001 à 3000, et
        (4) les mots qui n'apparaissent en aucune des listes précedentes.

 Pour obtenir une démonstration, soumettre simplement ce texte-ci.

 Préparation du texte

Général: Inclure un espace vide après toute virgule et point final.
Recherche: Corriger toute erreur d'orthographe ou d'usage et traiter tout nom propre.
```

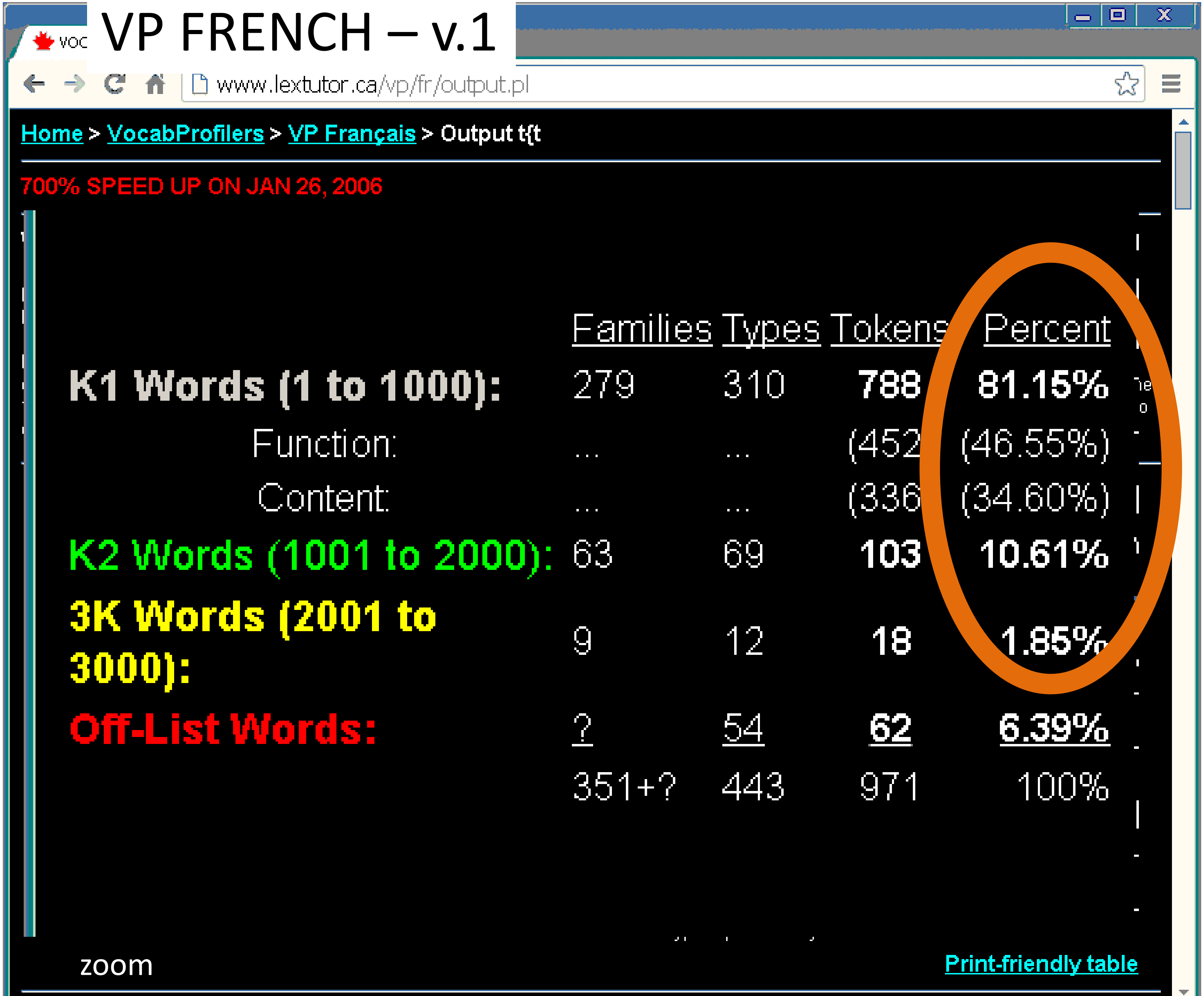
Mots à récategoriser comme haute fréquence (e.g. noms propres etc dans votre texte).

* + Tout nom propre = 1k □

Textes Démos: Pompiers | Le Devoir (CBC) | La Presse (CBC) | Le Devoir (ABANDON) | La Presse (ABANDON) | Entrevue Orale |

OU... Choose File No file chosen disk dur + Soumettre_fichier pour télécharger de fichiers TEXTE BRUT (~.txt; à limite env. 50k mots). + Tout nom propre = 1k □

VP FRENCH – v.1



Home > VocabProfilers > VP Français > Output t{t

700% SPEED UP ON JAN 26, 2006

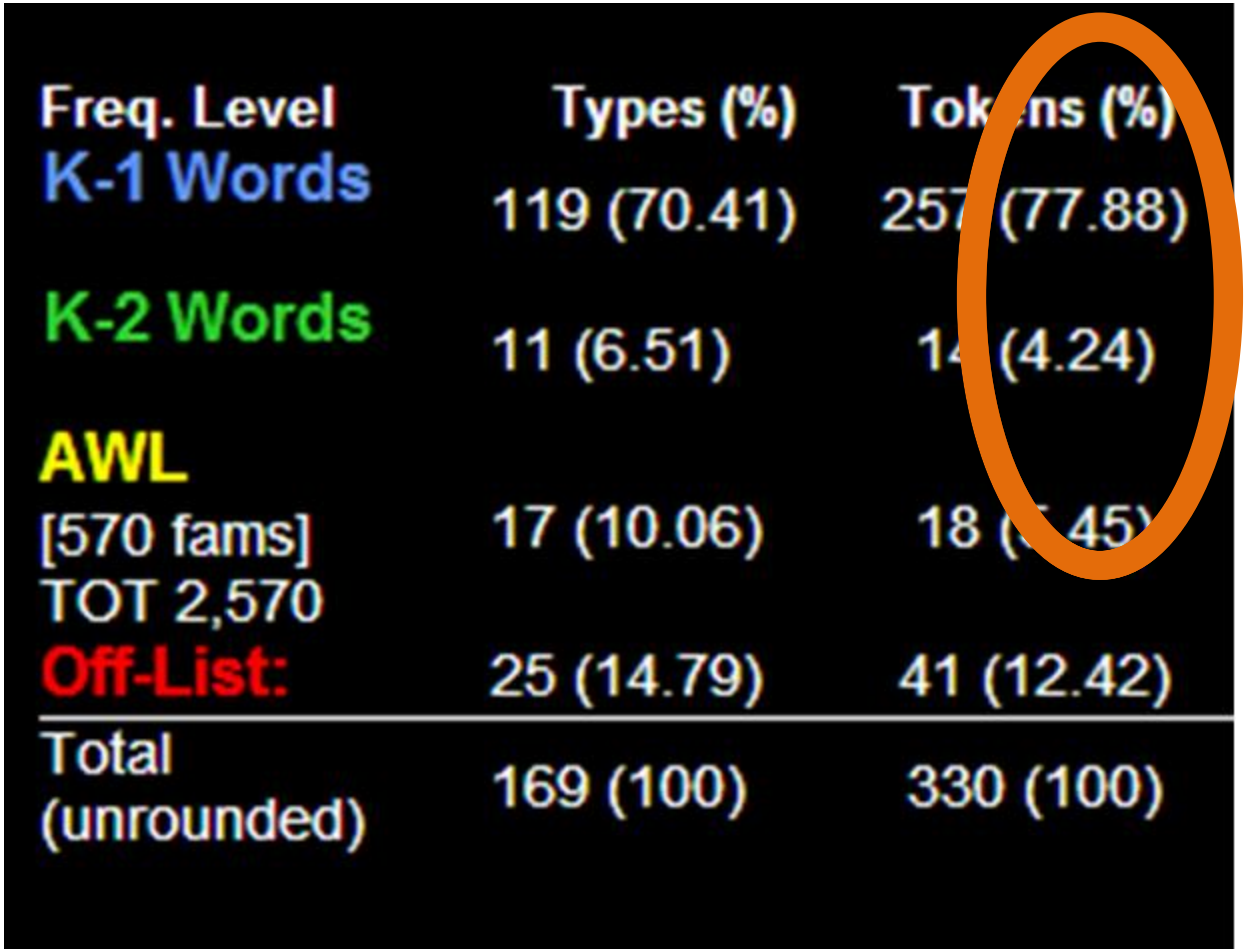
	Families	Types	Tokens	Percent
K1 Words (1 to 1000):	279	310	788	81.15%
Function:	...	...	(452	(46.55%)
Content:	...	...	(336	(34.60%)
K2 Words (1001 to 2000):	63	69	103	10.61%
3K Words (2001 to 3000):	9	12	18	1.85%
Off-List Words:	?	54	62	6.39%
	351+?	443	971	100%

zoom

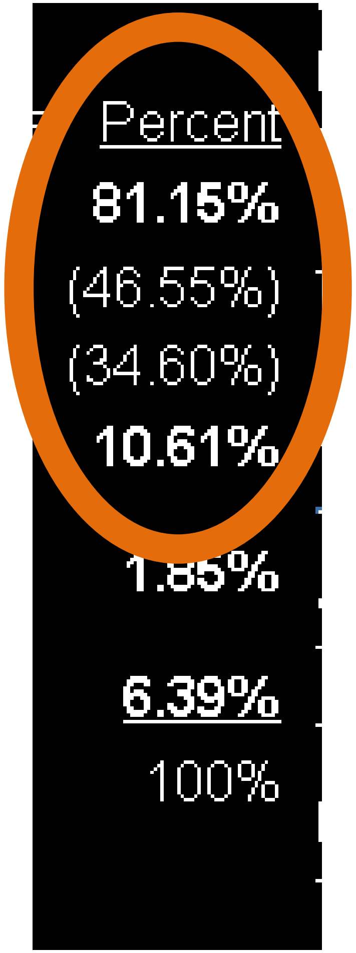
Print-friendly table

English

French



Freq. Level	Types (%)	Tokens (%)
K-1 Words	119 (70.41)	257 (77.88)
K-2 Words	11 (6.51)	14 (4.24)
AWL [570 fams] TOT 2,570	17 (10.06)	18 (5.45)
Off-List:	25 (14.79)	41 (12.42)
Total (unrounded)	169 (100)	330 (100)

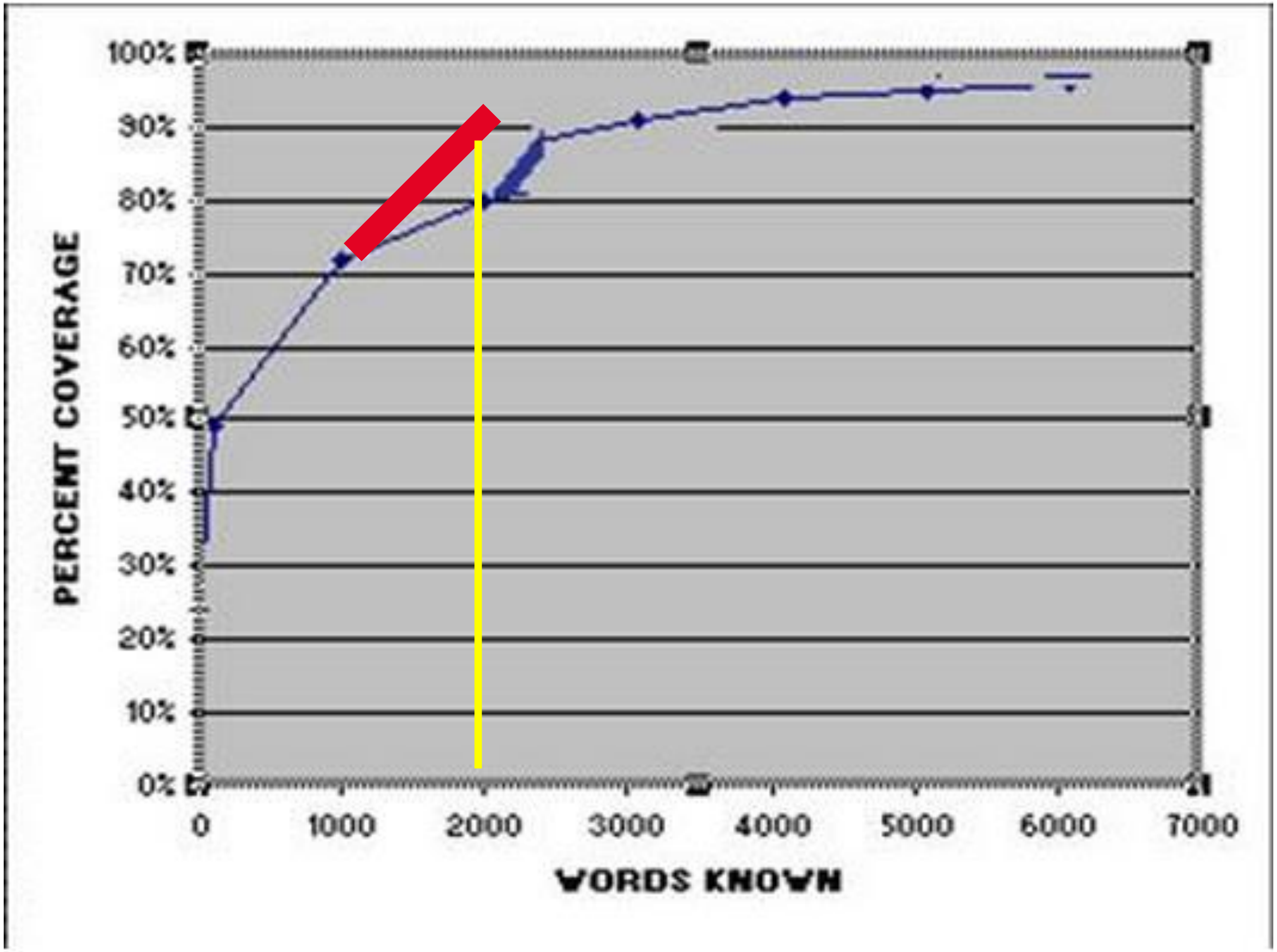


Percent
81.15%
(46.55%)
(34.60%)
10.61%
1.85%
6.39%
100%

ENG 1+2=80, FR 1+2=90

80% **90%**

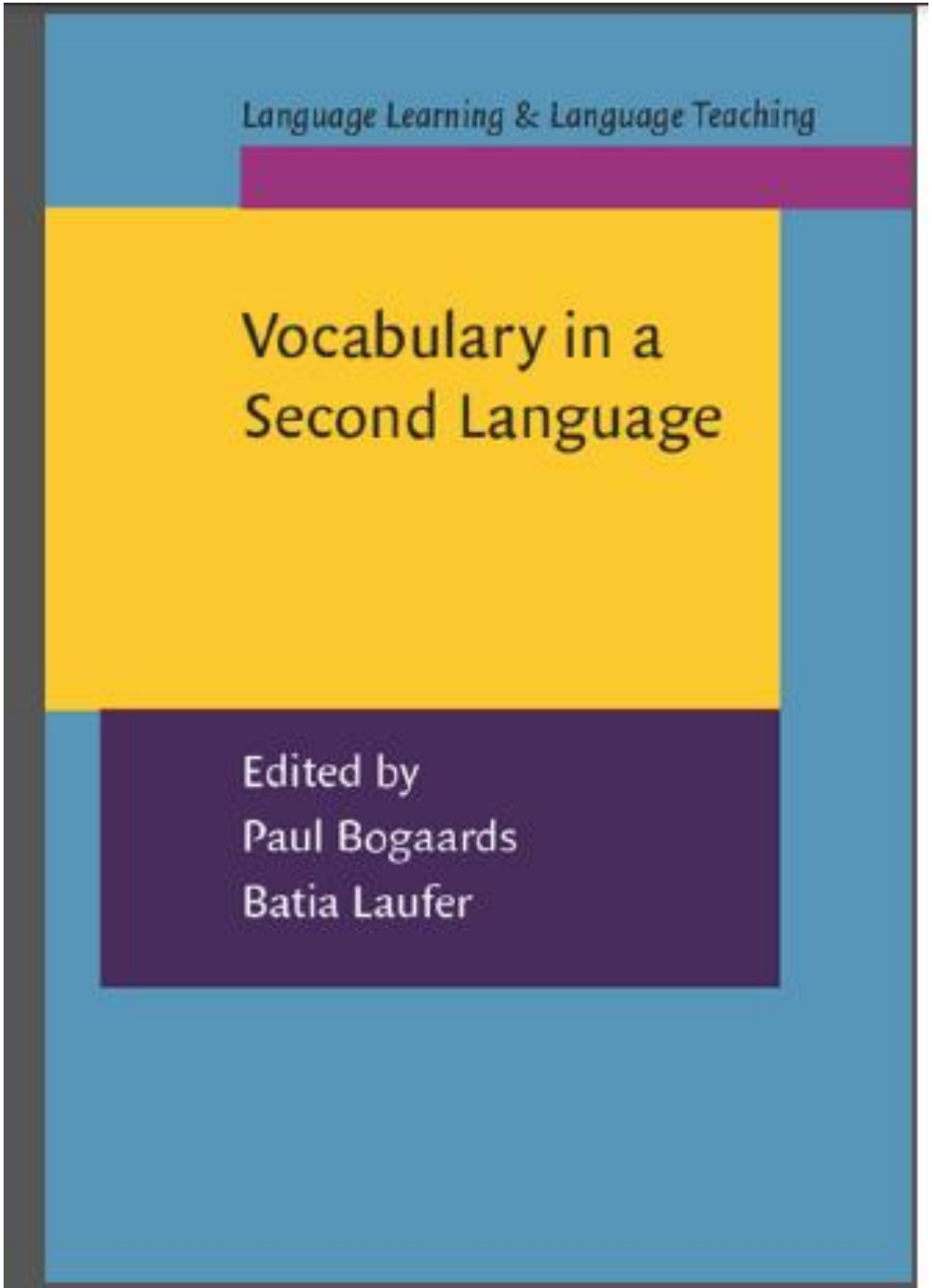
So is French getting the AWL effect for free?



The question was gradually reformulated:

~~Is there an AWL in French?~~

“Is there ***room*** for an AWL In French?”

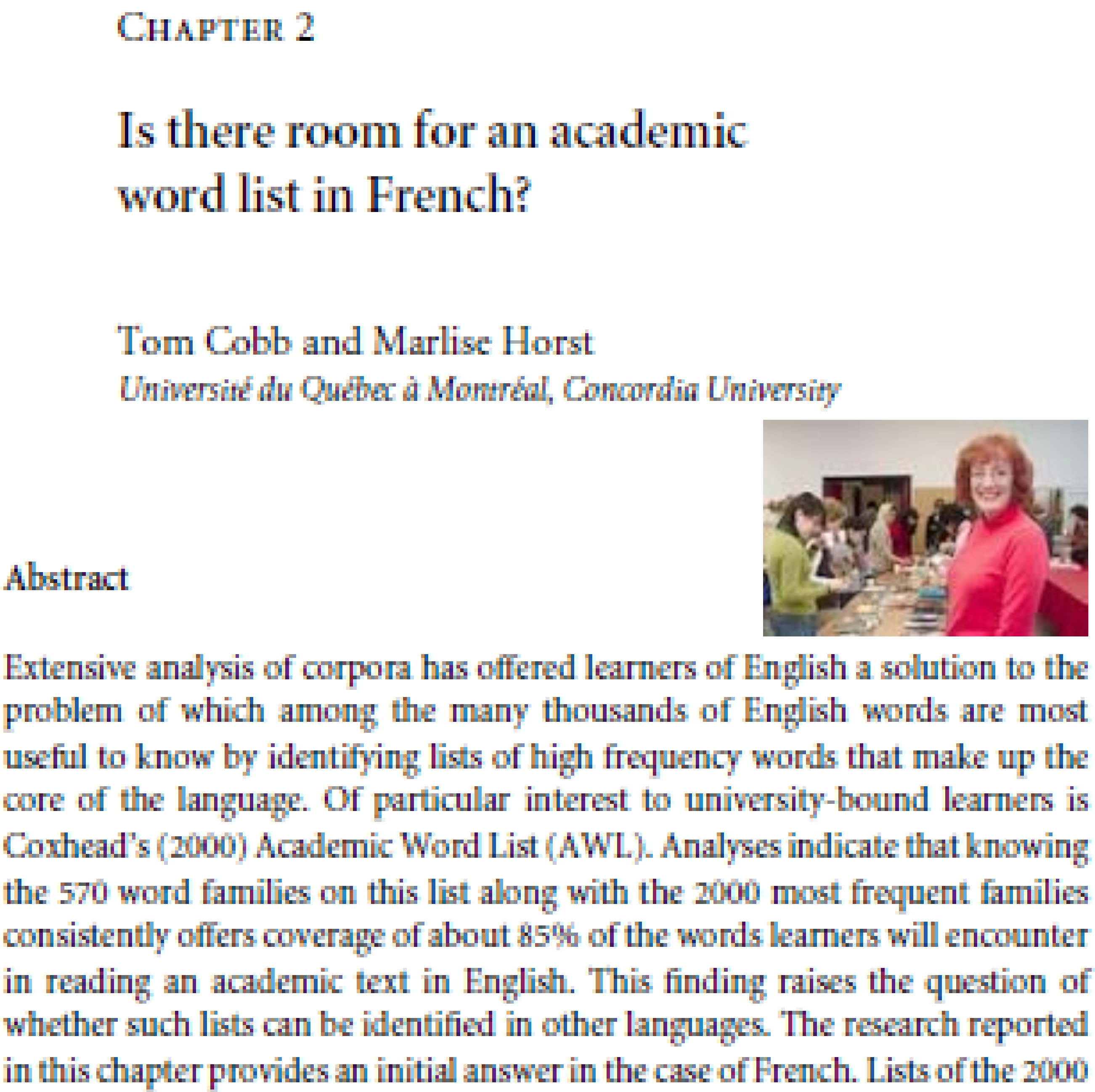


Language Learning & Language Teaching

Vocabulary in a Second Language

Edited by
Paul Bogaards
Batia Laufer

2004



Chapter 2

Is there room for an academic word list in French?

Tom Cobb and Marlise Horst
Université du Québec à Montréal, Concordia University

Abstract

Extensive analysis of corpora has offered learners of English a solution to the problem of which among the many thousands of English words are most useful to know by identifying lists of high frequency words that make up the core of the language. Of particular interest to university-bound learners is Coxhead's (2000) Academic Word List (AWL). Analyses indicate that knowing the 570 word families on this list along with the 2000 most frequent families consistently offers coverage of about 85% of the words learners will encounter in reading an academic text in English. This finding raises the question of whether such lists can be identified in other languages. The research reported in this chapter provides an initial answer in the case of French. Lists of the 2000

The answered seemed, **“No”**

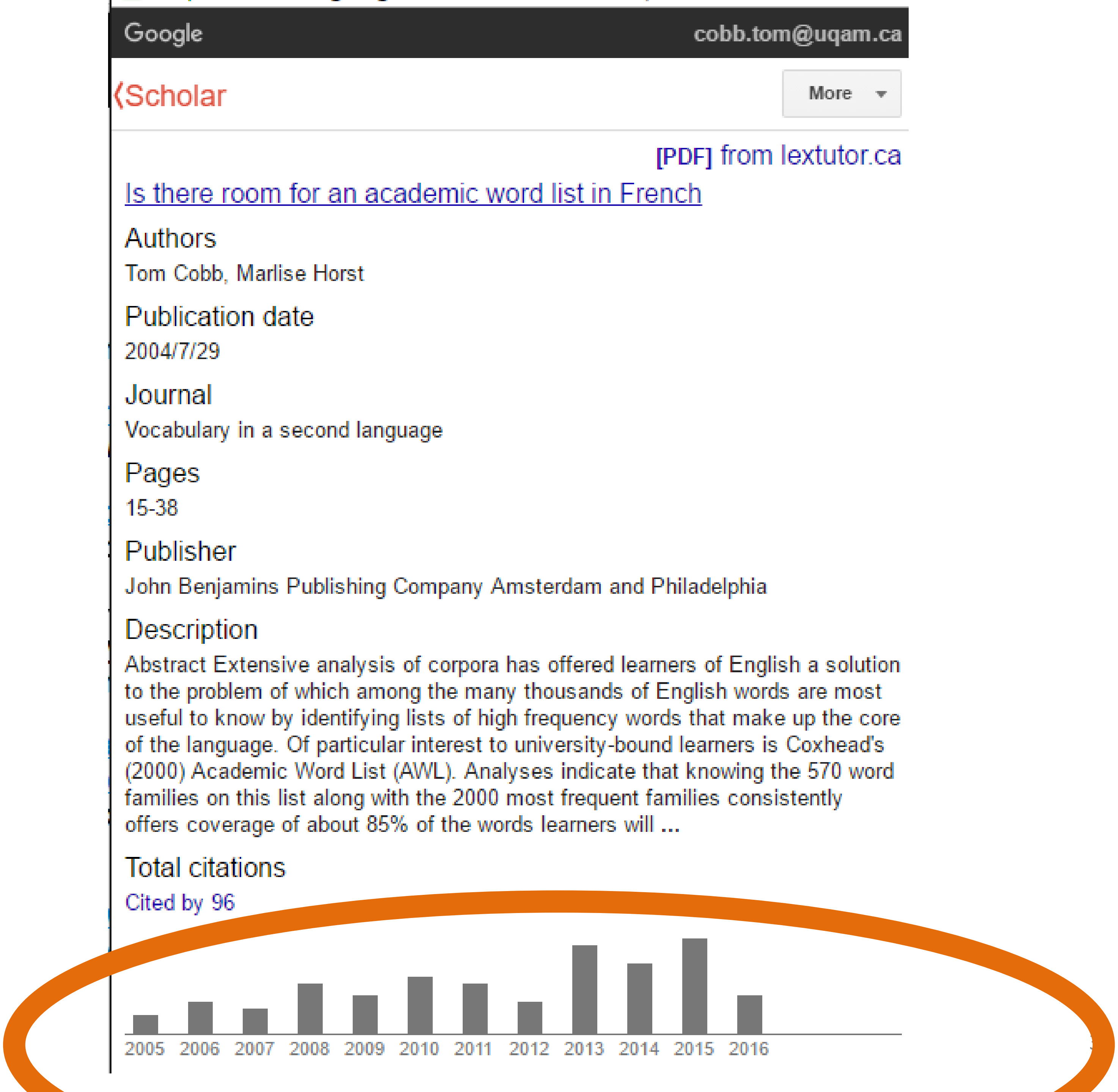
1k+2k is already giving 90% coverage in French

(Because French contains its AWL within its common lexis?)

And the remaining 10% is presumably needed for technical, archaic, oddball, & misspelled items

With the implication that acquiring a functional lexical competence was ***easier*** in French

Less to learn for = coverage



[PDF] from lextutor.ca

Is there room for an academic word list in French

Authors
Tom Cobb, Marlise Horst

Publication date
2004/7/29

Journal
Vocabulary in a second language

Pages
15-38

Publisher
John Benjamins Publishing Company Amsterdam and Philadelphia

Description
Abstract Extensive analysis of corpora has offered learners of English a solution to the problem of which among the many thousands of English words are most useful to know by identifying lists of high frequency words that make up the core of the language. Of particular interest to university-bound learners is Coxhead's (2000) Academic Word List (AWL). Analyses indicate that knowing the 570 word families on this list along with the 2000 most frequent families consistently offers coverage of about 85% of the words learners will ...

Total citations
Cited by 96

Meanwhile, back in English

1995-2005, a happy picture in ESL vocab

2k+AWL=90% (+ ‘technical’= 95%*!*)

BUT SHORT LIVED

1. Definition of basic competence recalculated :

The Comprehension-Bar is raised
95% coverage → 98% coverage (Nation, 2006)

2. Definition of technical lexis became less clear

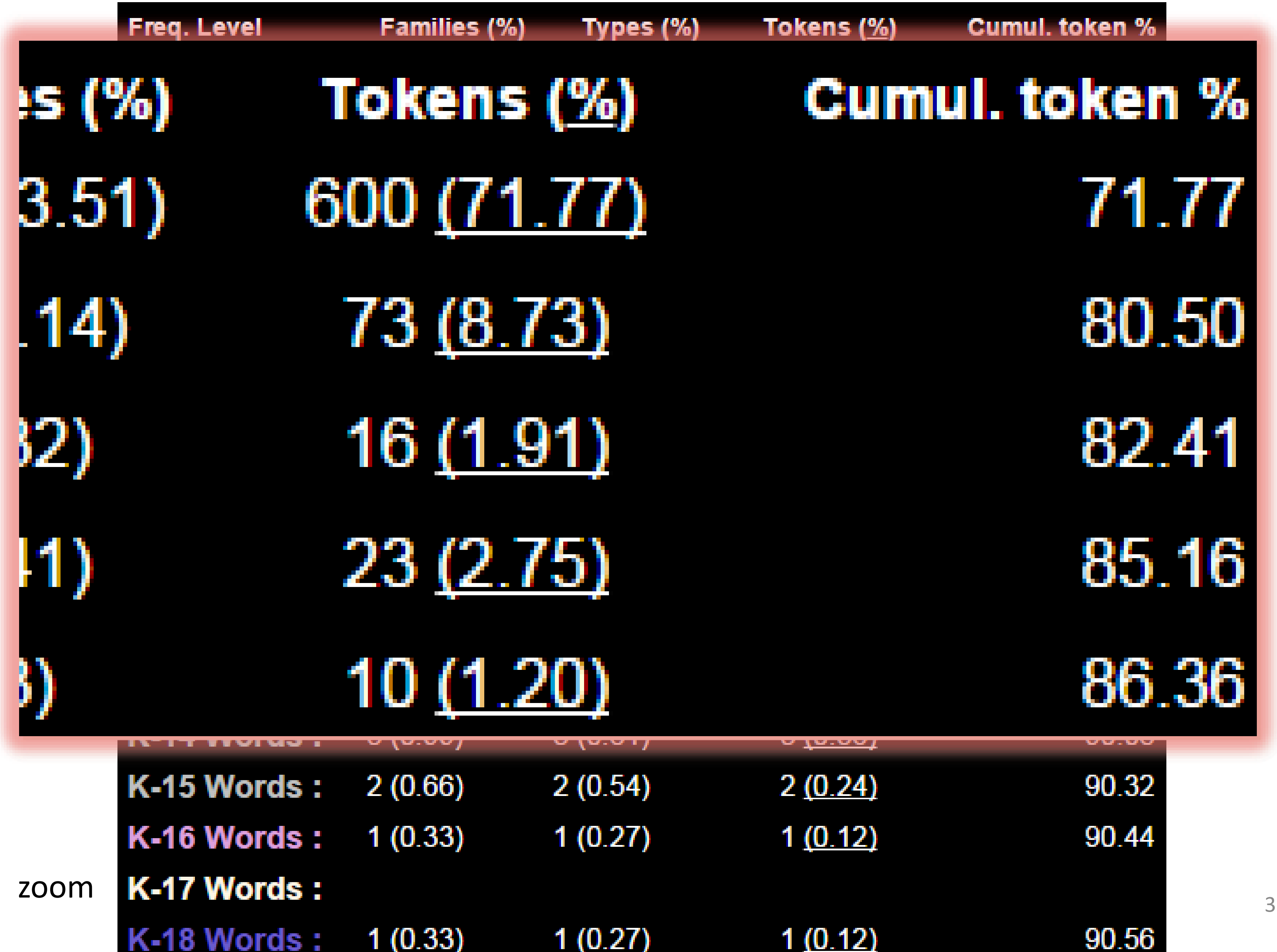
Some domains just use common words (‘needle’ in nursing)

3. New corpora put the existence of AWL in question

- BNC lists (2005)
- BNC-COCA lists (2012)

- AWL just an artefact of the old pre-corpus 1k-2k frequency lists?

VP-BNC-Coca – new type of profile



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-15 Words :	2 (0.66)	2 (0.54)	2 (0.24)	90.32
K-16 Words :	1 (0.33)	1 (0.27)	1 (0.12)	90.44
K-17 Words :				
K-18 Words :	1 (0.33)	1 (0.27)	1 (0.12)	90.56

es (%)	Tokens (%)	Cumul. token %
3.51)	600 (71.77)	71.77
.14)	73 (8.73)	80.50
2)	16 (1.91)	82.41
1)	23 (2.75)	85.16
)	10 (1.20)	86.36

zoom

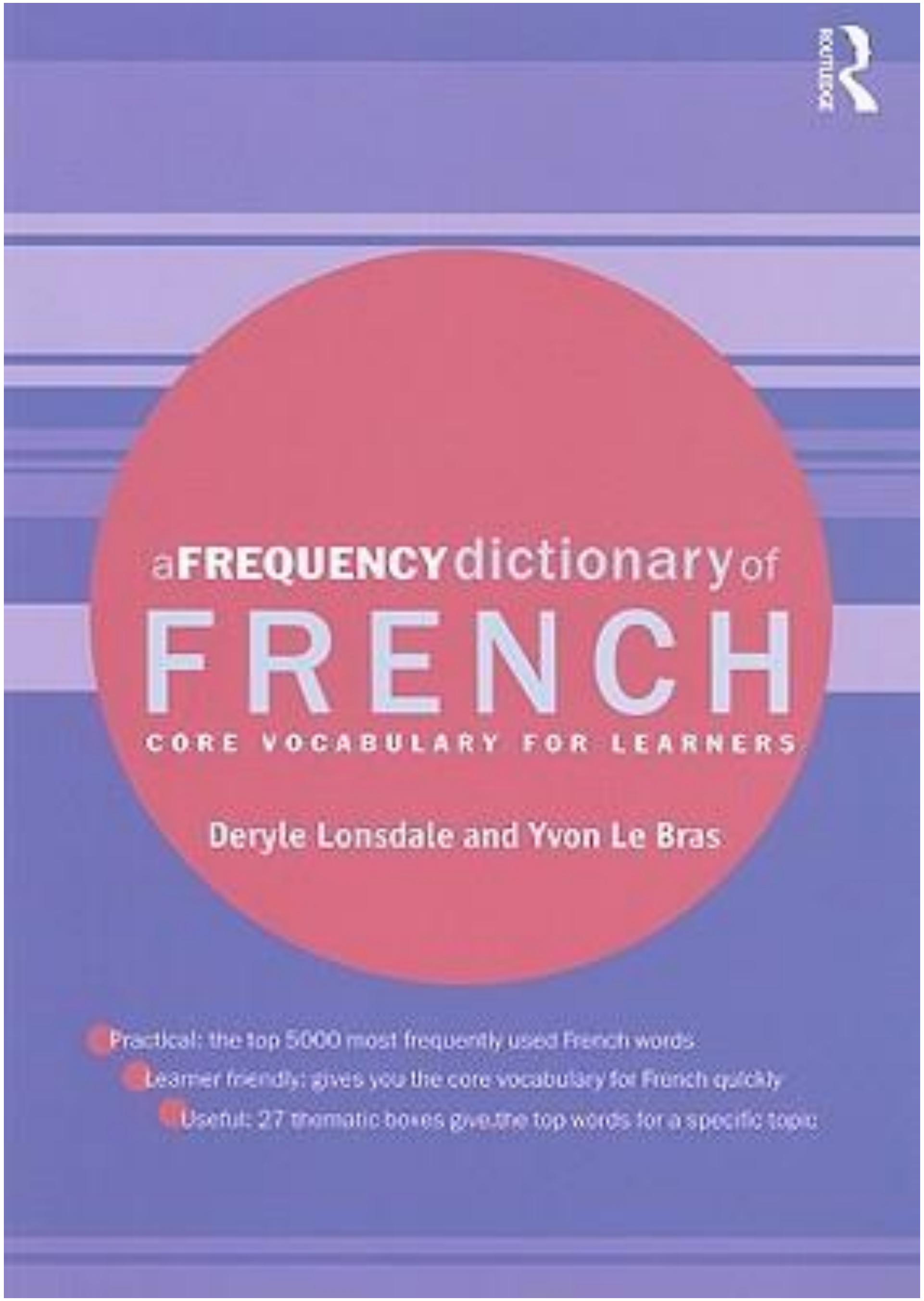
So the new question about French is ~

~~Is there room for an AWL In French?~~

“How are the medium and low frequency lexical resources of French deployed in the remaining 10% space available?”

What does this imply for learning French?

Again, the question gradually became answerable →



a **FREQUENCY** dictionary of

FRENCH

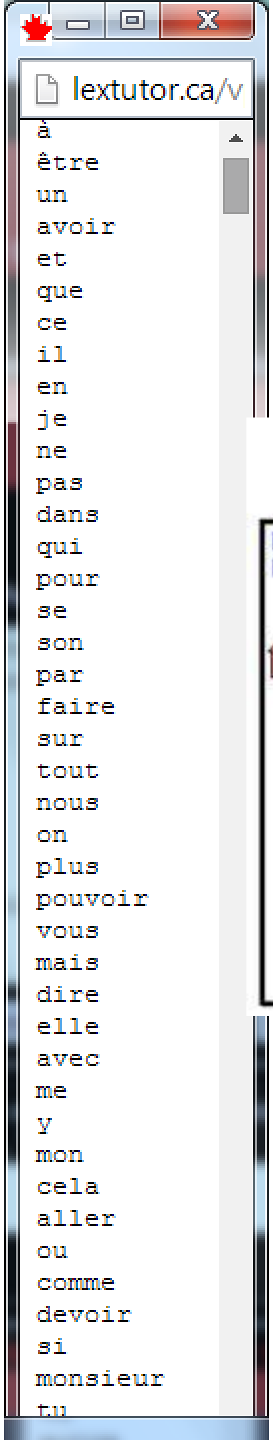
CORE VOCABULARY FOR LEARNERS

Deryle Lonsdale and Yvon Le Bras

- Practical: the top 5000 most frequently used French words
- Learner friendly: gives you the core vocabulary for French quickly
- Useful: 27 thematic boxes give the top words for a specific topic

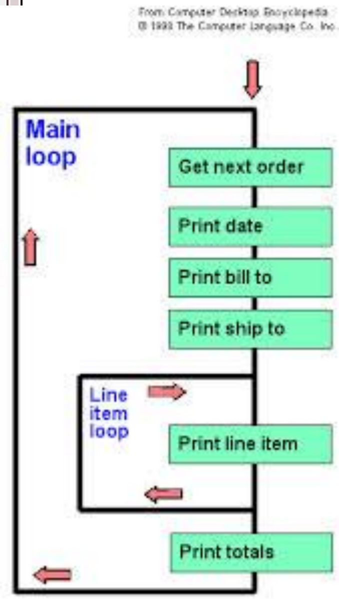
25 lemmatized French k-lists

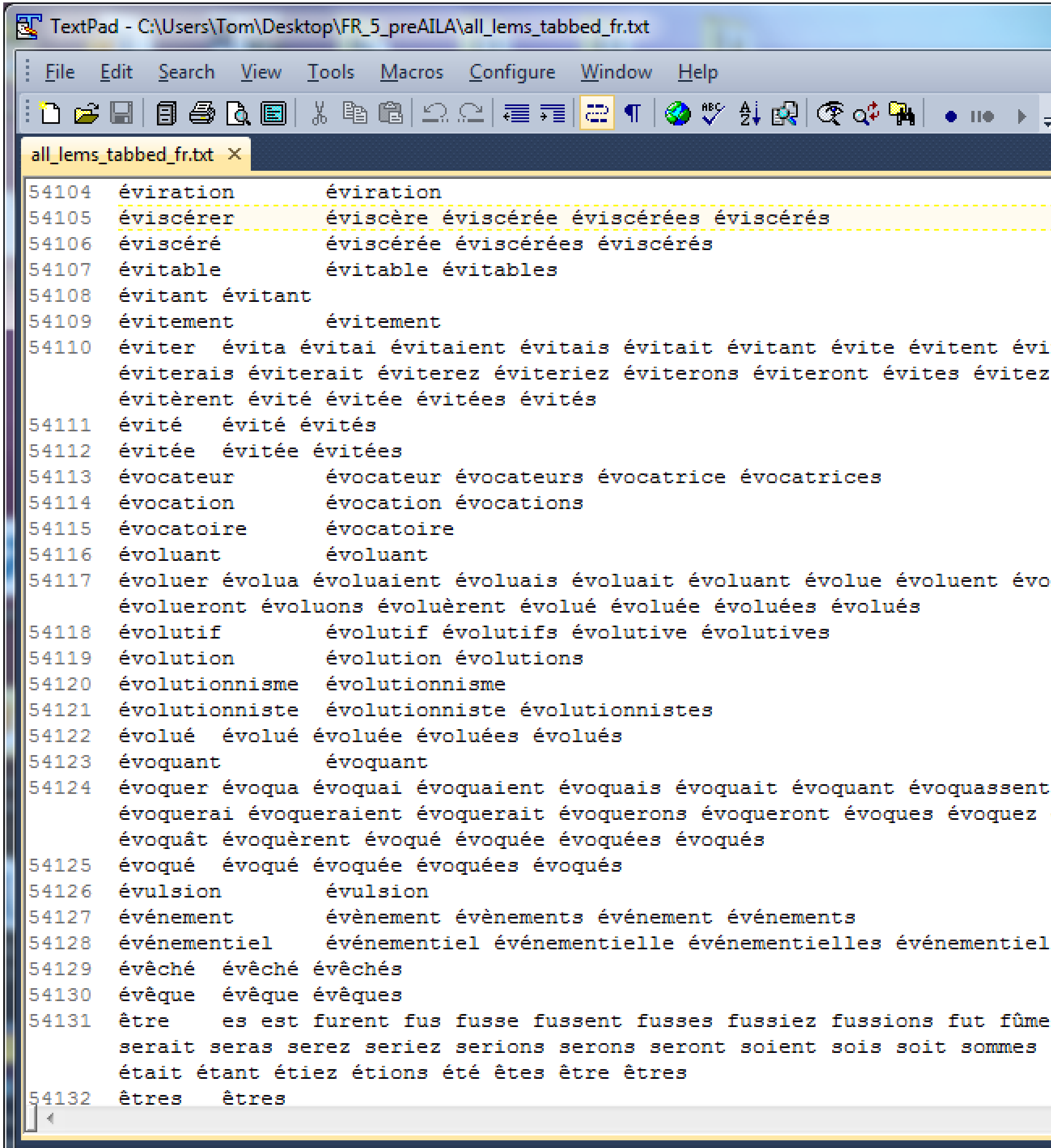
- From Lonsdale & Le Bras dictionary project at BYU
- Based on 23-million word corpus
- Continental + International French 50/50
- Spoken and written 50/50
- Literary 40%, expository 60%
- List-crunched for RANGE + FREQ



lextutor.ca/v

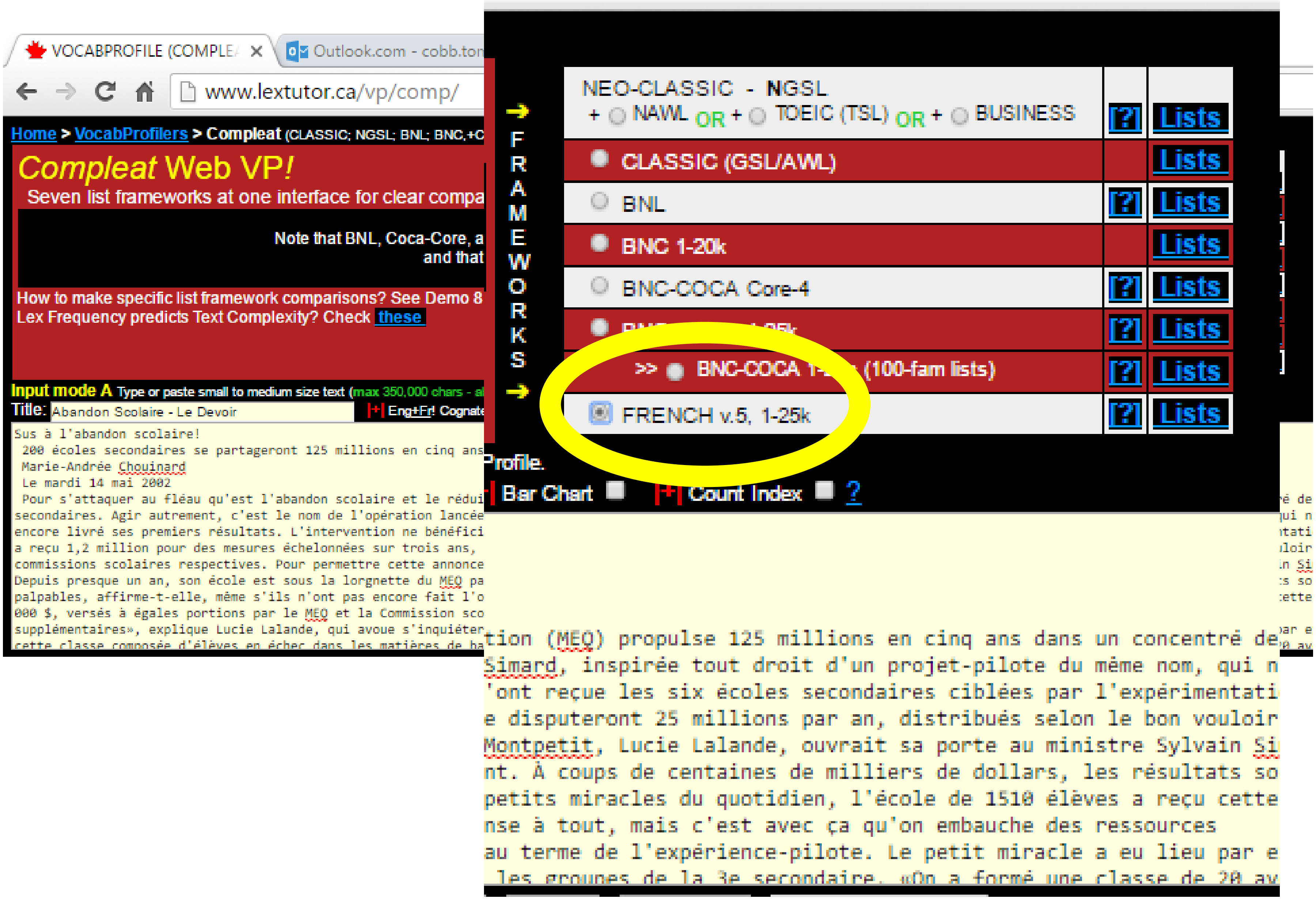
```
à
être
un
avoir
et
que
ce
il
en
je
ne
pas
dans
qui
pour
se
son
par
faire
sur
tout
nous
on
plus
pouvoir
vous
mais
dire
elle
avec
me
y
mon
cela
aller
ou
comme
devoir
si
monsieur
tu
```





TextPad - C:\Users\Tom\Desktop\FR_5_preAILA\all_lems_tabbed_fr.txt
File Edit Search View Tools Macros Configure Window Help
all_lems_tabbed_fr.txt

```
54104 éviration        éviration
54105 éviscérer        éviscère éviscérée éviscérées éviscérés
54106 éviscéré         éviscérée éviscérées éviscérés
54107 évitable         évitable évitables
54108 évitant évitant
54109 évitement        évitement
54110 éviter  évita évitai évitaient évitais évitait évitant évite évitent évi
      éviterais éviterait éviterez éviteriez éviterons éviteront évites évitez
      évitèrent évité évitée évitées évités
54111 évité   évité évités
54112 évitée  évitée évitées
54113 évocateur        évocateur évocateurs évocatrice évocatrices
54114 évocation        évocation évocations
54115 évocatoire       évocatoire
54116 évoluant         évoluant
54117 évoluer évolua évoluaient évoluais évoluait évoluant évolue évoluent évo
      évolueront évoluons évoluèrent évolué évoluée évoluées évolués
54118 évolutif         évolutif évolutifs évolutive évolutives
54119 évolution        évolution évolutions
54120 évolutionnisme   évolutionnisme
54121 évolutionniste   évolutionniste évolutionnistes
54122 évolué  évolué évoluée évoluées évolués
54123 évoquant         évoquant
54124 évoquer évoqua évoquai évoquaient évoquais évoquait évoquant évoquassent
      évoquerai évoqueraient évoquerait évoquerons évoqueront évoques évoquez
      évoquât évoquèrent évoqué évoquée évoquées évoqués
54125 évoqué  évoqué évoquée évoquées évoqués
54126 évulsion         évulsion
54127 événement        évènement évènements événement événements
54128 événementiel     événementiel événementielle événementielles événementiel
54129 évêché  évêché évêchés
54130 évêque  évêque évêques
54131 être    es est furent fus fusse fussent fusses fussiez fussions fut fûme
      serait seras serez seriez serions serons seront soient sois soit sommes
      était étant étiez étions été êtes être êtres
54132 êtres   êtres
```



VOCABPROFILE (COMPLE
Outlook.com - cobb.tor
www.lextutor.ca/vp/comp/

Home > VocabProfilers > Compleat (CLASSIC; NGSL; BNL; BNC,+C

Compleat Web VP!

Seven list frameworks at one interface for clear compa

Note that BNL, Coca-Core, a
and that

How to make specific list framework comparisons? See Demo 8
Lex Frequency predicts Text Complexity? Check these

FRAMEWORKS

NEO-CLASSIC - NGSL + ○ NAWL OR + ○ TOEIC (TSL) OR + ○ BUSINESS	[?]	Lists
● CLASSIC (GSL/AWL)		Lists
○ BNL	[?]	Lists
● BNC 1-20k		Lists
○ BNC-COCA Core-4	[?]	Lists
● BNC-COCA 1-25k	[?]	Lists
>> ● BNC-COCA 1-25k (100-fam lists)	[?]	Lists
◉ FRENCH v.5, 1-25k	[?]	Lists

Profile.

Bar Chart □ Count Index □ ?

Input mode A Type or paste small to medium size text (max 350,000 chars - al

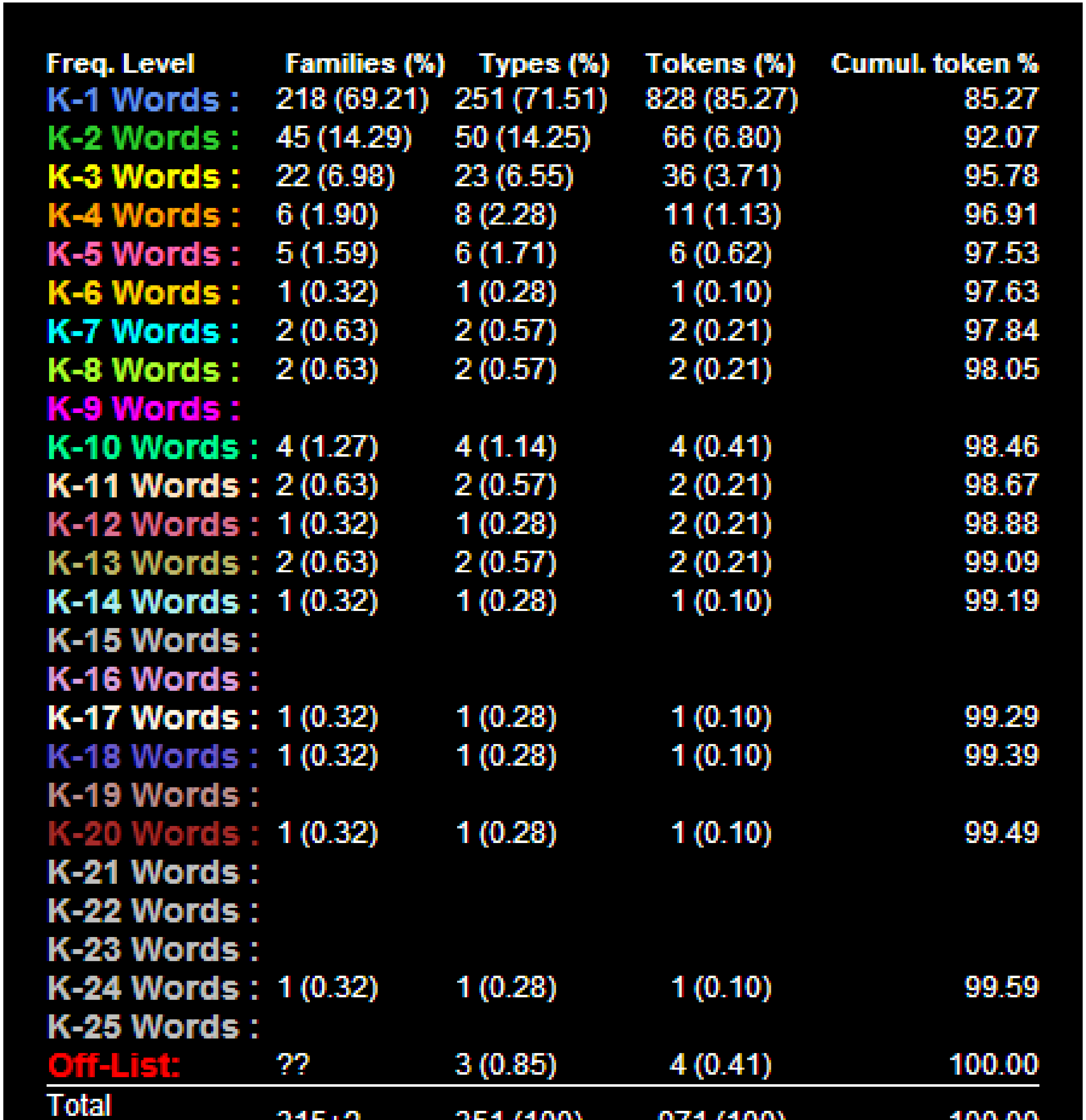
Title: Abandon Scolaire - Le Devoir

Eng+Fr! Cognate

```
Sus à l'abandon scolaire!
 200 écoles secondaires se partageront 125 millions en cinq ans
 Marie-Andrée Chouinard
 Le mardi 14 mai 2002
 Pour s'attaquer au fléau qu'est l'abandon scolaire et le rédui
secondaires. Agir autrement, c'est le nom de l'opération lancée
encore livré ses premiers résultats. L'intervention ne bénéfici
a reçu 1,2 million pour des mesures échelonnées sur trois ans,
commissions scolaires respectives. Pour permettre cette annonce
Depuis presque un an, son école est sous la lorgnette du MEQ pa
palpables, affirme-t-elle, même s'ils n'ont pas encore fait l'o
000 $, versés à égales portions par le MEQ et la Commission sco
supplémentaires», explique Lucie Lalande, qui avoue s'inquiéter
cette classe composée d'élèves en échec dans les matières de ba
```

```
tion (MEQ) propulse 125 millions en cinq ans dans un concentré de
Simard, inspirée tout droit d'un projet-pilote du même nom, qui n
'ont reçue les six écoles secondaires ciblées par l'expérimentati
e disputeront 25 millions par an, distribués selon le bon vouloir
Montpetit, Lucie Lalande, ouvrait sa porte au ministre Sylvain Si
nt. À coups de centaines de milliers de dollars, les résultats so
petits miracles du quotidien, l'école de 1510 élèves a reçu cette
nse à tout, mais c'est avec ça qu'on embauche des ressources
au terme de l'expérience-pilote. Le petit miracle a eu lieu par e
 les groupes de la 3e secondaire. «On a formé une classe de 20 av
```

FRENCH – v.5



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	218 (69.21)	251 (71.51)	828 (85.27)	85.27
K-2 Words :	45 (14.29)	50 (14.25)	66 (6.80)	92.07
K-3 Words :	22 (6.98)	23 (6.55)	36 (3.71)	95.78
K-4 Words :	6 (1.90)	8 (2.28)	11 (1.13)	96.91
K-5 Words :	5 (1.59)	6 (1.71)	6 (0.62)	97.53
K-6 Words :	1 (0.32)	1 (0.28)	1 (0.10)	97.63
K-7 Words :	2 (0.63)	2 (0.57)	2 (0.21)	97.84
K-8 Words :	2 (0.63)	2 (0.57)	2 (0.21)	98.05
K-9 Words :				
K-10 Words :	4 (1.27)	4 (1.14)	4 (0.41)	98.46
K-11 Words :	2 (0.63)	2 (0.57)	2 (0.21)	98.67
K-12 Words :	1 (0.32)	1 (0.28)	2 (0.21)	98.88
K-13 Words :	2 (0.63)	2 (0.57)	2 (0.21)	99.09
K-14 Words :	1 (0.32)	1 (0.28)	1 (0.10)	99.19
K-15 Words :				
K-16 Words :				
K-17 Words :	1 (0.32)	1 (0.28)	1 (0.10)	99.29
K-18 Words :	1 (0.32)	1 (0.28)	1 (0.10)	99.39
K-19 Words :				
K-20 Words :	1 (0.32)	1 (0.28)	1 (0.10)	99.49
K-21 Words :				
K-22 Words :				
K-23 Words :				
K-24 Words :	1 (0.32)	1 (0.28)	1 (0.10)	99.59
K-25 Words :				
Off-List:	??	3 (0.85)	4 (0.41)	100.00
Total	315+?	351 (100)	971 (100)	100.00

So with this we can investigate the shape of the mid-frequency French lexicon

<u>And make plausible comparisons with English</u>

- What lies between 90% and 95% coverage in French texts?
  - Or between 90% and 98%?
- Is there “less to learn” in French than in English ?
  - (Remembering that lemmas ≠ families)

3 tests

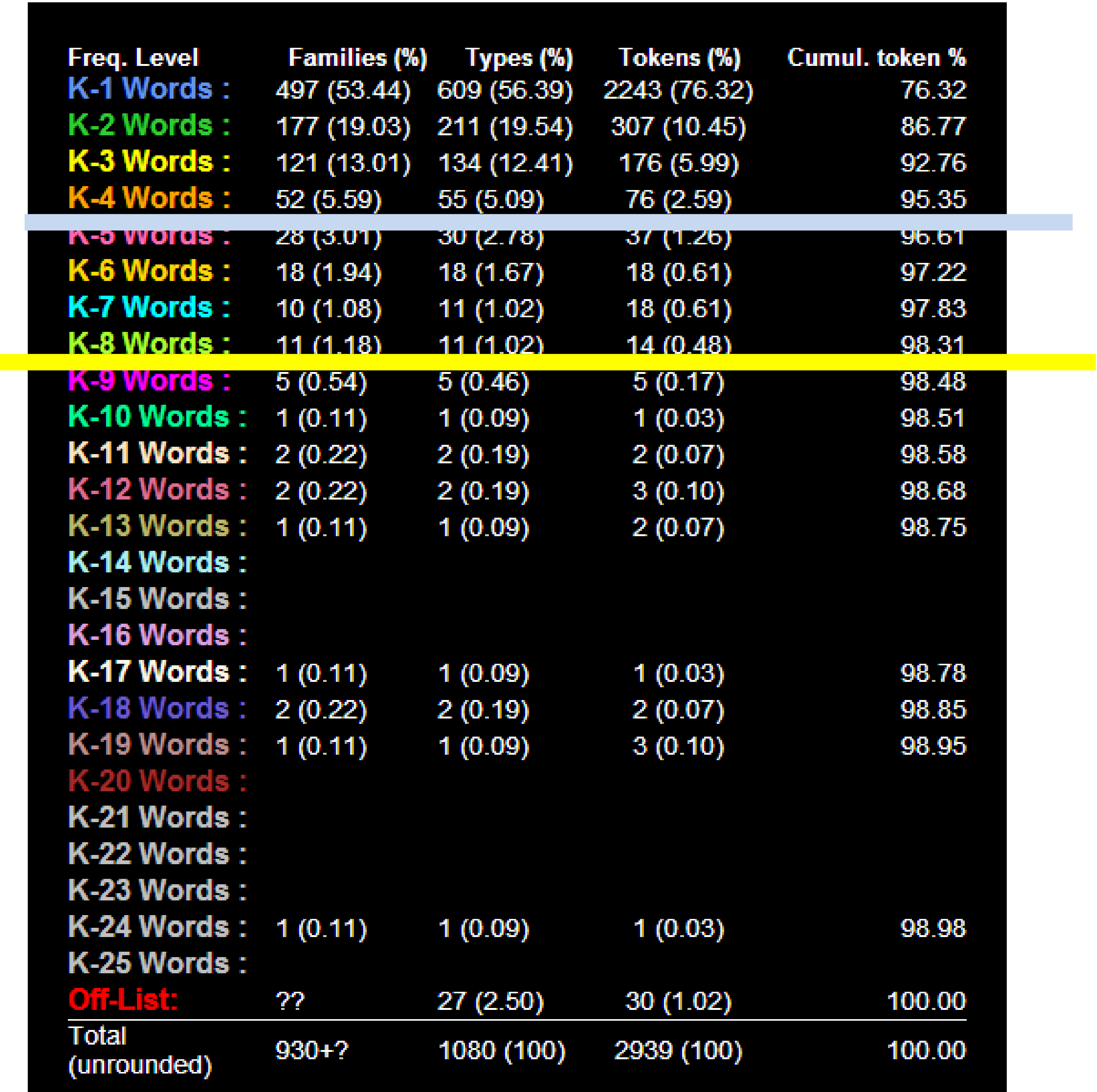
Test 1

Translated popular texts

20 translated Readers' Digest texts
→ 20 Fr, 20 Eng
Half translated **Eng->Fr**, half **Fr-> Eng**
Total 2939 words Eng, 3650 words Fr

Run through VP-Fr as a mini-corpus (as a single file)

ENGLISH

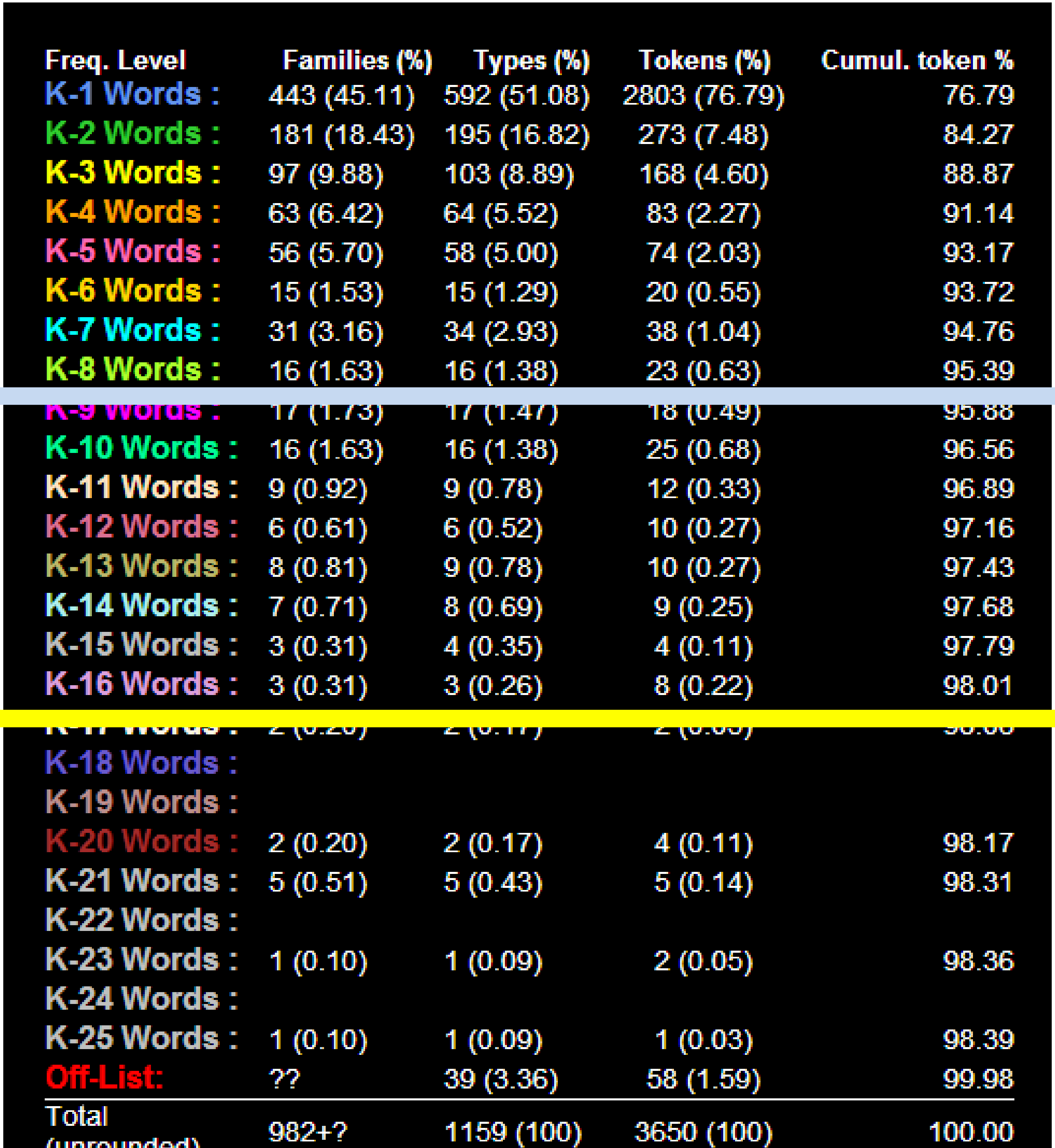


Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	497 (53.44)	609 (56.39)	2243 (76.32)	76.32
K-2 Words :	177 (19.03)	211 (19.54)	307 (10.45)	86.77
K-3 Words :	121 (13.01)	134 (12.41)	176 (5.99)	92.76
K-4 Words :	52 (5.59)	55 (5.09)	76 (2.59)	95.35
K-5 Words :	28 (3.01)	30 (2.78)	37 (1.26)	96.61
K-6 Words :	18 (1.94)	18 (1.67)	18 (0.61)	97.22
K-7 Words :	10 (1.08)	11 (1.02)	18 (0.61)	97.83
K-8 Words :	11 (1.18)	11 (1.02)	14 (0.48)	98.31
K-9 Words :	5 (0.54)	5 (0.46)	5 (0.17)	98.48
K-10 Words :	1 (0.11)	1 (0.09)	1 (0.03)	98.51
K-11 Words :	2 (0.22)	2 (0.19)	2 (0.07)	98.58
K-12 Words :	2 (0.22)	2 (0.19)	3 (0.10)	98.68
K-13 Words :	1 (0.11)	1 (0.09)	2 (0.07)	98.75
K-14 Words :				
K-15 Words :				
K-16 Words :				
K-17 Words :	1 (0.11)	1 (0.09)	1 (0.03)	98.78
K-18 Words :	2 (0.22)	2 (0.19)	2 (0.07)	98.85
K-19 Words :	1 (0.11)	1 (0.09)	3 (0.10)	98.95
K-20 Words :				
K-21 Words :				
K-22 Words :				
K-23 Words :				
K-24 Words :	1 (0.11)	1 (0.09)	1 (0.03)	98.98
K-25 Words :				
Off-List:	??	27 (2.50)	30 (1.02)	100.00
Total (unrounded)	930+?	1080 (100)	2939 (100)	100.00

95%

98%

FRENCH



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	443 (45.11)	592 (51.08)	2803 (76.79)	76.79
K-2 Words :	181 (18.43)	195 (16.82)	273 (7.48)	84.27
K-3 Words :	97 (9.88)	103 (8.89)	168 (4.60)	88.87
K-4 Words :	63 (6.42)	64 (5.52)	83 (2.27)	91.14
K-5 Words :	56 (5.70)	58 (5.00)	74 (2.03)	93.17
K-6 Words :	15 (1.53)	15 (1.29)	20 (0.55)	93.72
K-7 Words :	31 (3.16)	34 (2.93)	38 (1.04)	94.76
K-8 Words :	16 (1.63)	16 (1.38)	23 (0.63)	95.39
K-9 Words :	17 (1.73)	17 (1.47)	18 (0.49)	95.88
K-10 Words :	16 (1.63)	16 (1.38)	25 (0.68)	96.56
K-11 Words :	9 (0.92)	9 (0.78)	12 (0.33)	96.89
K-12 Words :	6 (0.61)	6 (0.52)	10 (0.27)	97.16
K-13 Words :	8 (0.81)	9 (0.78)	10 (0.27)	97.43
K-14 Words :	7 (0.71)	8 (0.69)	9 (0.25)	97.68
K-15 Words :	3 (0.31)	4 (0.35)	4 (0.11)	97.79
K-16 Words :	3 (0.31)	3 (0.26)	8 (0.22)	98.01
K-17 Words :	2 (0.20)	2 (0.17)	2 (0.05)	98.06
K-18 Words :				
K-19 Words :				
K-20 Words :	2 (0.20)	2 (0.17)	4 (0.11)	98.17
K-21 Words :	5 (0.51)	5 (0.43)	5 (0.14)	98.31
K-22 Words :				
K-23 Words :	1 (0.10)	1 (0.09)	2 (0.05)	98.36
K-24 Words :				
K-25 Words :	1 (0.10)	1 (0.09)	1 (0.03)	98.39
Off-List:	??	39 (3.36)	58 (1.59)	99.98
Total (unrounded)	982+?	1159 (100)	3650 (100)	100.00

95%

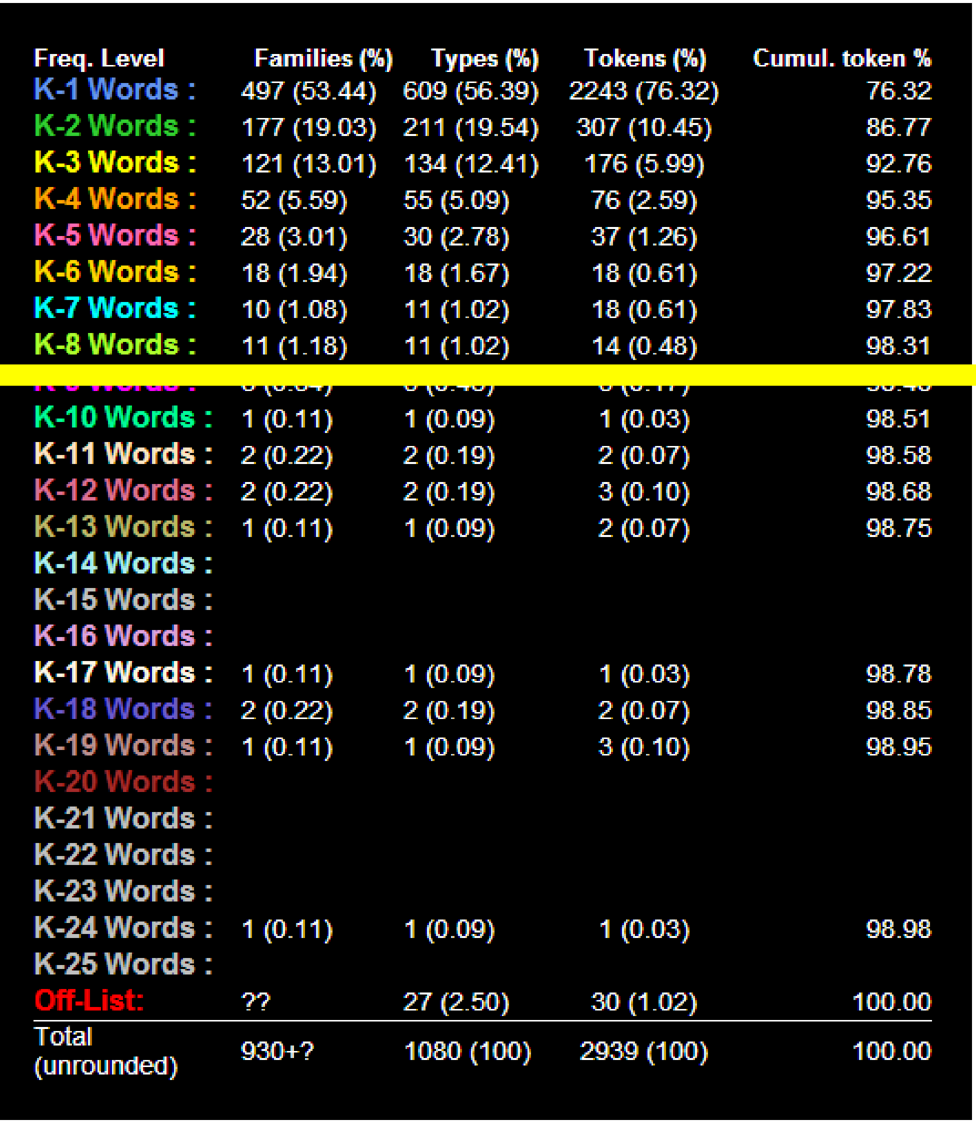
98%

Eng
(fams)

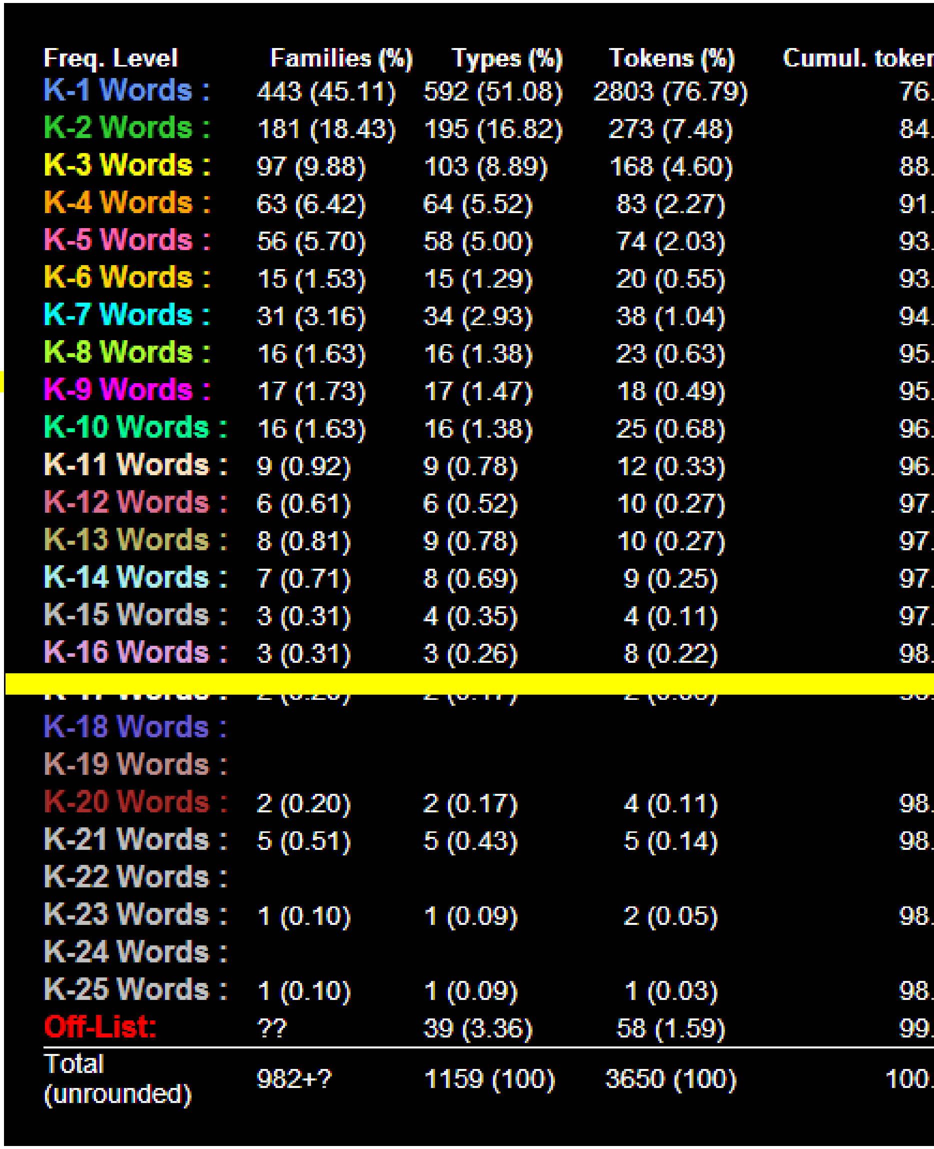
Side by side

Using 98% criterion

Fr
(lemmas)



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	497 (53.44)	609 (56.39)	2243 (76.32)	76.32
K-2 Words :	177 (19.03)	211 (19.54)	307 (10.45)	86.77
K-3 Words :	121 (13.01)	134 (12.41)	176 (5.99)	92.76
K-4 Words :	52 (5.59)	55 (5.09)	76 (2.59)	95.35
K-5 Words :	28 (3.01)	30 (2.78)	37 (1.26)	96.61
K-6 Words :	18 (1.94)	18 (1.67)	18 (0.61)	97.22
K-7 Words :	10 (1.08)	11 (1.02)	18 (0.61)	97.83
K-8 Words :	11 (1.18)	11 (1.02)	14 (0.48)	98.31
K-9 Words :	5 (0.54)	5 (0.46)	5 (0.17)	98.48
K-10 Words :	1 (0.11)	1 (0.09)	1 (0.03)	98.51
K-11 Words :	2 (0.22)	2 (0.19)	2 (0.07)	98.58
K-12 Words :	2 (0.22)	2 (0.19)	3 (0.10)	98.68
K-13 Words :	1 (0.11)	1 (0.09)	2 (0.07)	98.75
K-14 Words :				
K-15 Words :				
K-16 Words :				
K-17 Words :	1 (0.11)	1 (0.09)	1 (0.03)	98.78
K-18 Words :	2 (0.22)	2 (0.19)	2 (0.07)	98.85
K-19 Words :	1 (0.11)	1 (0.09)	3 (0.10)	98.95
K-20 Words :				
K-21 Words :				
K-22 Words :				
K-23 Words :				
K-24 Words :	1 (0.11)	1 (0.09)	1 (0.03)	98.98
K-25 Words :				
Off-List:	??	27 (2.50)	30 (1.02)	100.00
Total (unrounded)	930+?	1080 (100)	2939 (100)	100.00



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. toker
K-1 Words :	443 (45.11)	592 (51.08)	2803 (76.79)	76.
K-2 Words :	181 (18.43)	195 (16.82)	273 (7.48)	84.
K-3 Words :	97 (9.88)	103 (8.89)	168 (4.60)	88.
K-4 Words :	63 (6.42)	64 (5.52)	83 (2.27)	91.
K-5 Words :	56 (5.70)	58 (5.00)	74 (2.03)	93.
K-6 Words :	15 (1.53)	15 (1.29)	20 (0.55)	93.
K-7 Words :	31 (3.16)	34 (2.93)	38 (1.04)	94.
K-8 Words :	16 (1.63)	16 (1.38)	23 (0.63)	95.
K-9 Words :	17 (1.73)	17 (1.47)	18 (0.49)	95.
K-10 Words :	16 (1.63)	16 (1.38)	25 (0.68)	96.
K-11 Words :	9 (0.92)	9 (0.78)	12 (0.33)	96.
K-12 Words :	6 (0.61)	6 (0.52)	10 (0.27)	97.
K-13 Words :	8 (0.81)	9 (0.78)	10 (0.27)	97.
K-14 Words :	7 (0.71)	8 (0.69)	9 (0.25)	97.
K-15 Words :	3 (0.31)	4 (0.35)	4 (0.11)	97.
K-16 Words :	3 (0.31)	3 (0.26)	8 (0.22)	98.
K-17 Words :	2 (0.20)	2 (0.17)	2 (0.05)	98.
K-18 Words :				
K-19 Words :				
K-20 Words :	2 (0.20)	2 (0.17)	4 (0.11)	98.
K-21 Words :	5 (0.51)	5 (0.43)	5 (0.14)	98.
K-22 Words :				
K-23 Words :	1 (0.10)	1 (0.09)	2 (0.05)	98.
K-24 Words :				
K-25 Words :	1 (0.10)	1 (0.09)	1 (0.03)	98.
Off-List:	??	39 (3.36)	58 (1.59)	99.
Total (unrounded)	982+?	1159 (100)	3650 (100)	100.

So ~

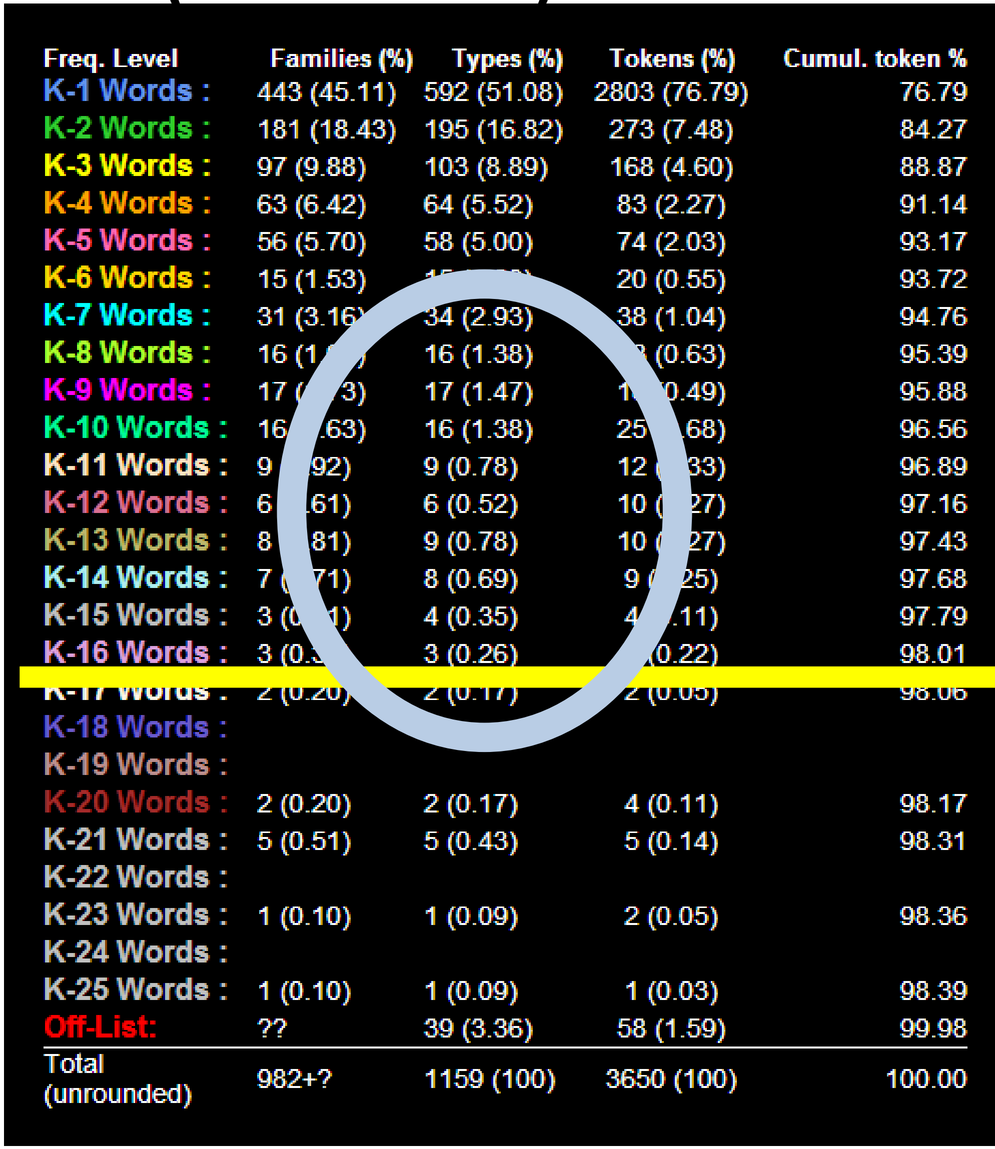
With the new lists and definitions

  - (Note that 98% figure has never actually been established for French)

- While English and French both get to 90% at about 3,000 families/lemmas
  - – English gets to 98% at 8,000 known words
  - – French gets to 98% at 16,000 known words!

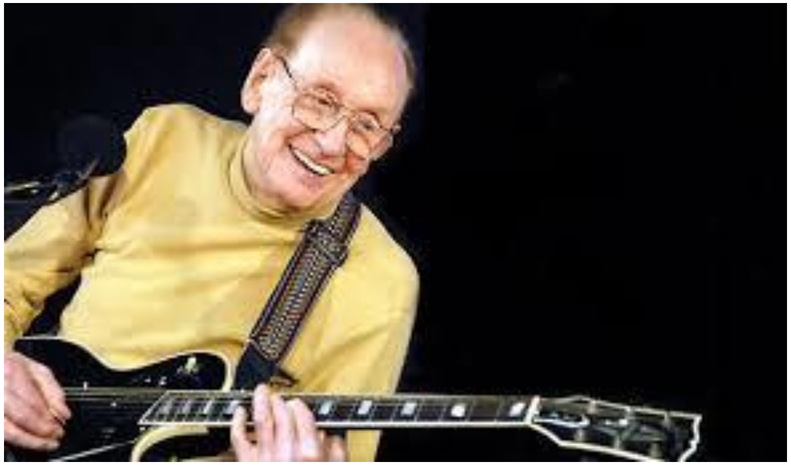
Fr (lemmas)

- A lot of words lie behind that circle!
- The difference between k8 to k16 is only **100** word types in ***this*** mini-corpus
- … these 100 words are drawn from a pool of 8,000 lemmas
  - So for generalizability…



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	443 (45.11)	592 (51.08)	2803 (76.79)	76.79
K-2 Words :	181 (18.43)	195 (16.82)	273 (7.48)	84.27
K-3 Words :	97 (9.88)	103 (8.89)	168 (4.60)	88.87
K-4 Words :	63 (6.42)	64 (5.52)	83 (2.27)	91.14
K-5 Words :	56 (5.70)	58 (5.00)	74 (2.03)	93.17
K-6 Words :	15 (1.53)	[illegible]	20 (0.55)	93.72
K-7 Words :	31 (3.16)	34 (2.93)	38 (1.04)	94.76
K-8 Words :	16 (1[illegible])	16 (1.38)	[illegible] (0.63)	95.39
K-9 Words :	17 ([illegible]3)	17 (1.47)	1[illegible]0.49)	95.88
K-10 Words :	16[illegible].63)	16 (1.38)	25[illegible].68)	96.56
K-11 Words :	9 [illegible]92)	9 (0.78)	12 ([illegible]33)	96.89
K-12 Words :	6 [illegible].61)	6 (0.52)	10 ([illegible]27)	97.16
K-13 Words :	8 [illegible]81)	9 (0.78)	10 ([illegible]27)	97.43
K-14 Words :	7 ([illegible]71)	8 (0.69)	9 ([illegible]25)	97.68
K-15 Words :	3 ([illegible]1)	4 (0.35)	4 [illegible].11)	97.79
K-16 Words :	3 (0.[illegible]	3 (0.26)	[illegible](0.22)	98.01
K-17 Words :	2 (0.20)	2 (0.17)	2 (0.05)	98.06
K-18 Words :				
K-19 Words :				
K-20 Words :	2 (0.20)	2 (0.17)	4 (0.11)	98.17
K-21 Words :	5 (0.51)	5 (0.43)	5 (0.14)	98.31
K-22 Words :				
K-23 Words :	1 (0.10)	1 (0.09)	2 (0.05)	98.36
K-24 Words :				
K-25 Words :	1 (0.10)	1 (0.09)	1 (0.03)	98.39
Off-List:	??	39 (3.36)	58 (1.59)	99.98
Total (unrounded)	982+?	1159 (100)	3650 (100)	100.00

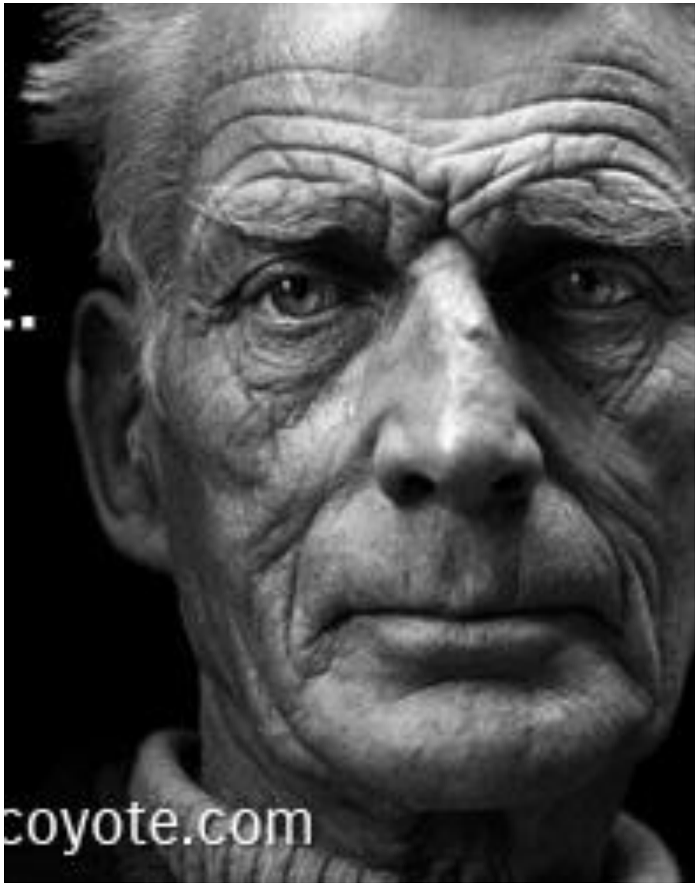
Vite à demarrer ~ lente à finir

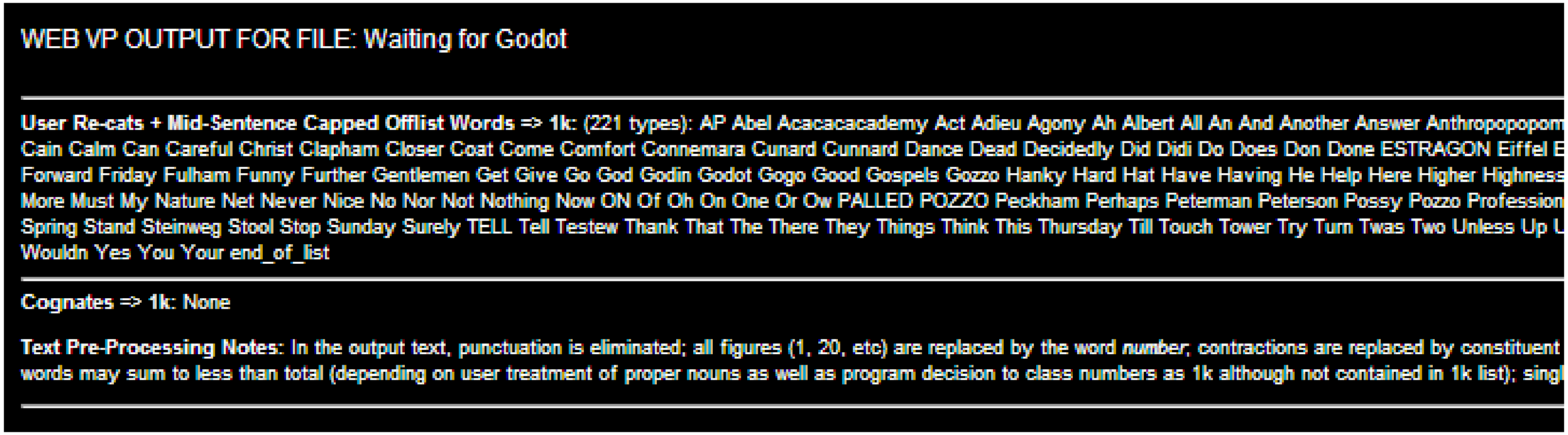


Test 2

Translation of extended literary work

- Samuel Beckett's idea - French as "an impoverished lexicon"?
  - Actually he never said this
- But he did write in French, and "use stark language to con-vey a stark world"
  - How stark is Beckett's French?



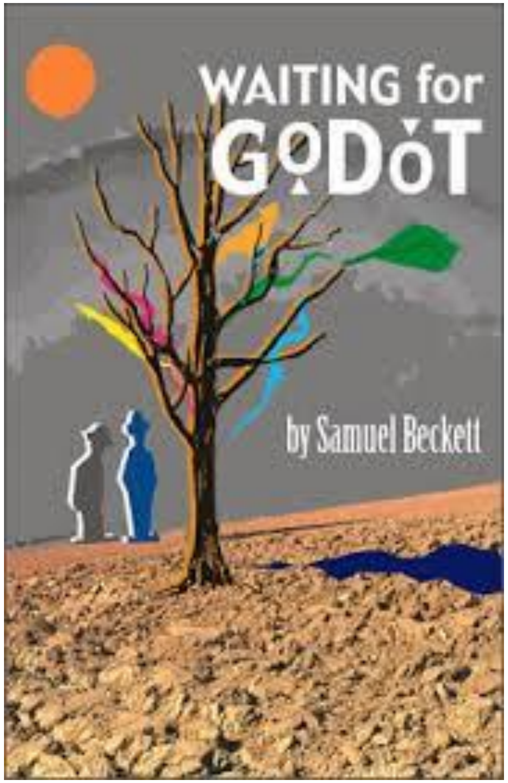


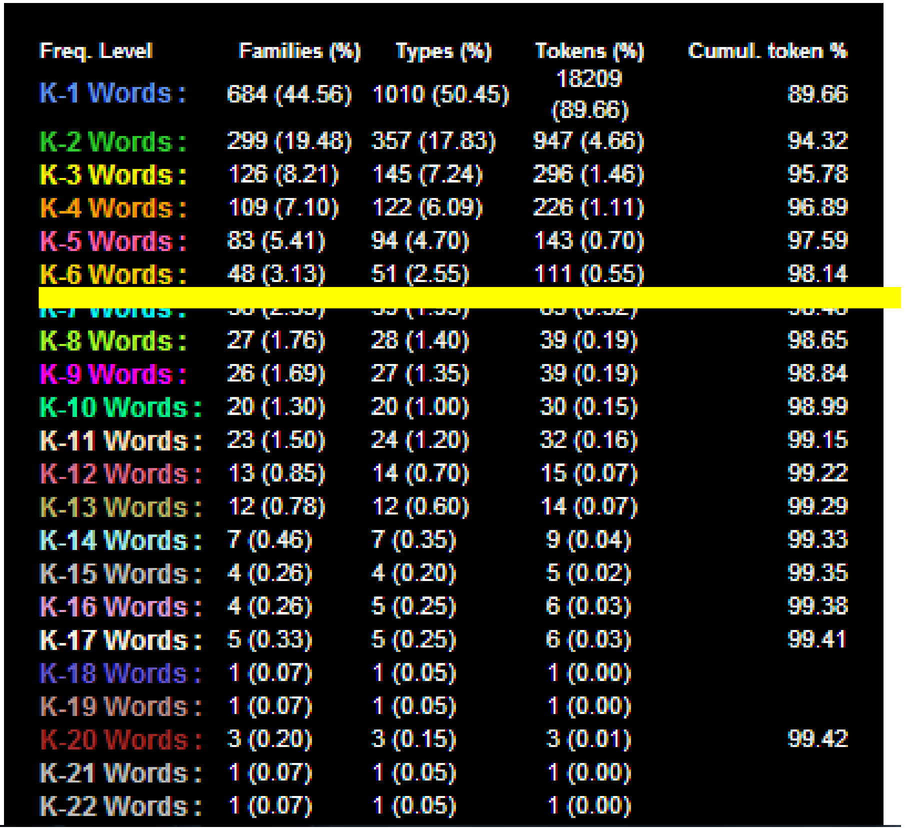
WEB VP OUTPUT FOR FILE: Waiting for Godot

User Re-cats + Mid-Sentence Capped Offlist Words => 1k: (221 types): AP Abel Acacacacademy Act Adieu Agony Ah Albert All An And Another Answer Anthropopopom
Cain Calm Can Careful Christ Clapham Closer Coat Come Comfort Connemara Cunard Cunnard Dance Dead Decidedly Did Didi Do Does Don Done ESTRAGON Eiffel E
Forward Friday Fulham Funny Further Gentlemen Get Give Go God Godin Godot Gogo Good Gospels Gozzo Hanky Hard Hat Have Having He Help Here Higher Highness
More Must My Nature Net Never Nice No Nor Not Nothing Now ON Of Oh On One Or Ow PALLED POZZO Peckham Perhaps Peterman Peterson Possy Pozzo Profession
Spring Stand Steinweg Stool Stop Sunday Surely TELL Tell Testew Thank That The There They Things Think This Thursday Till Touch Tower Try Turn Twas Two Unless Up U
Wouldn Yes You Your end_of_list

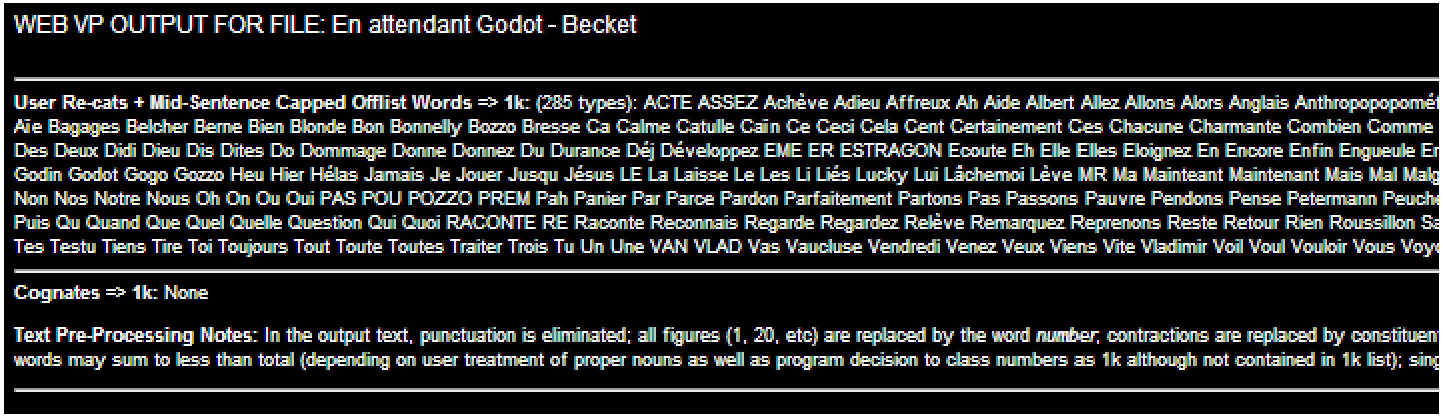
Cognates => 1k: None

Text Pre-Processing Notes: In the output text, punctuation is eliminated; all figures (1, 20, etc) are replaced by the word *number*; contractions are replaced by constituent
words may sum to less than total (depending on user treatment of proper nouns as well as program decision to class numbers as 1k although not contained in 1k list); singl





Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	684 (44.56)	1010 (50.45)	18209 (89.66)	89.66
K-2 Words :	299 (19.48)	357 (17.83)	947 (4.66)	94.32
K-3 Words :	126 (8.21)	145 (7.24)	296 (1.46)	95.78
K-4 Words :	109 (7.10)	122 (6.09)	226 (1.11)	96.89
K-5 Words :	83 (5.41)	94 (4.70)	143 (0.70)	97.59
K-6 Words :	48 (3.13)	51 (2.55)	111 (0.55)	98.14
K-7 Words :	[illegible]	[illegible]	[illegible]	[illegible]
K-8 Words :	27 (1.76)	28 (1.40)	39 (0.19)	98.65
K-9 Words :	26 (1.69)	27 (1.35)	39 (0.19)	98.84
K-10 Words :	20 (1.30)	20 (1.00)	30 (0.15)	98.99
K-11 Words :	23 (1.50)	24 (1.20)	32 (0.16)	99.15
K-12 Words :	13 (0.85)	14 (0.70)	15 (0.07)	99.22
K-13 Words :	12 (0.78)	12 (0.60)	14 (0.07)	99.29
K-14 Words :	7 (0.46)	7 (0.35)	9 (0.04)	99.33
K-15 Words :	4 (0.26)	4 (0.20)	5 (0.02)	99.35
K-16 Words :	4 (0.26)	5 (0.25)	6 (0.03)	99.38
K-17 Words :	5 (0.33)	5 (0.25)	6 (0.03)	99.41
K-18 Words :	1 (0.07)	1 (0.05)	1 (0.00)	
K-19 Words :	1 (0.07)	1 (0.05)	1 (0.00)	
K-20 Words :	3 (0.20)	3 (0.15)	3 (0.01)	99.42
K-21 Words :	1 (0.07)	1 (0.05)	1 (0.00)	
K-22 Words :	1 (0.07)	1 (0.05)	1 (0.00)	

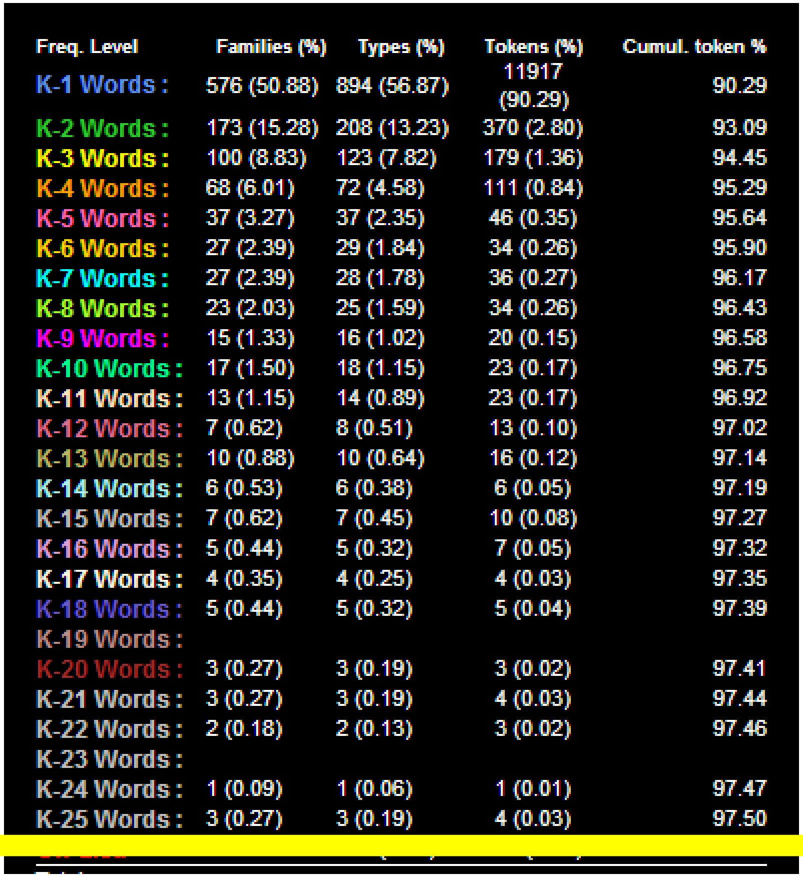


WEB VP OUTPUT FOR FILE: En attendant Godot - Becket

User Re-cats + Mid-Sentence Capped Offlist Words => 1k: (285 types): ACTE ASSEZ Achève Adieu Affreux Ah Aide Albert Allez Allons Alors Anglais Anthropopopomét
Aïe Bagages Belcher Berne Bien Blonde Bon Bonnelly Bozzo Bresse Ca Calme Catulle Caïn Ce Ceci Cela Cent Certainement Ces Chacune Charmante Combien Comme
Des Deux Didi Dieu Dis Dites Do Dommage Donne Donnez Du Durance Déj Développez EME ER ESTRAGON Ecoute Eh Elle Elles Eloignez En Encore Enfin Engueule Er
Godin Godot Gogo Gozzo Heu Hier Hélas Jamais Je Jouer Jusqu Jésus LE La Laisse Le Les Li Liés Lucky Lui Lâchemoi Lève MR Ma Mainteant Maintenant Mais Mal Malg
Non Nos Notre Nous Oh On Ou Oui PAS POU POZZO PREM Pah Panier Par Parce Pardon Parfaitement Partons Pas Passons Pauvre Pendons Pense Petermann Peuche
Puis Qu Quand Que Quel Quelle Question Qui Quoi RACONTE RE Raconte Reconnais Regarde Regardez Relève Remarquez Reprenons Reste Retour Rien Roussillon Sa
Tes Testu Tiens Tire Toi Toujours Tout Toute Toutes Traiter Trois Tu Un Une VAN VLAD Vas Vaucluse Vendredi Venez Veux Viens Vite Vladimir Voil Voul Vouloir Vous Voye

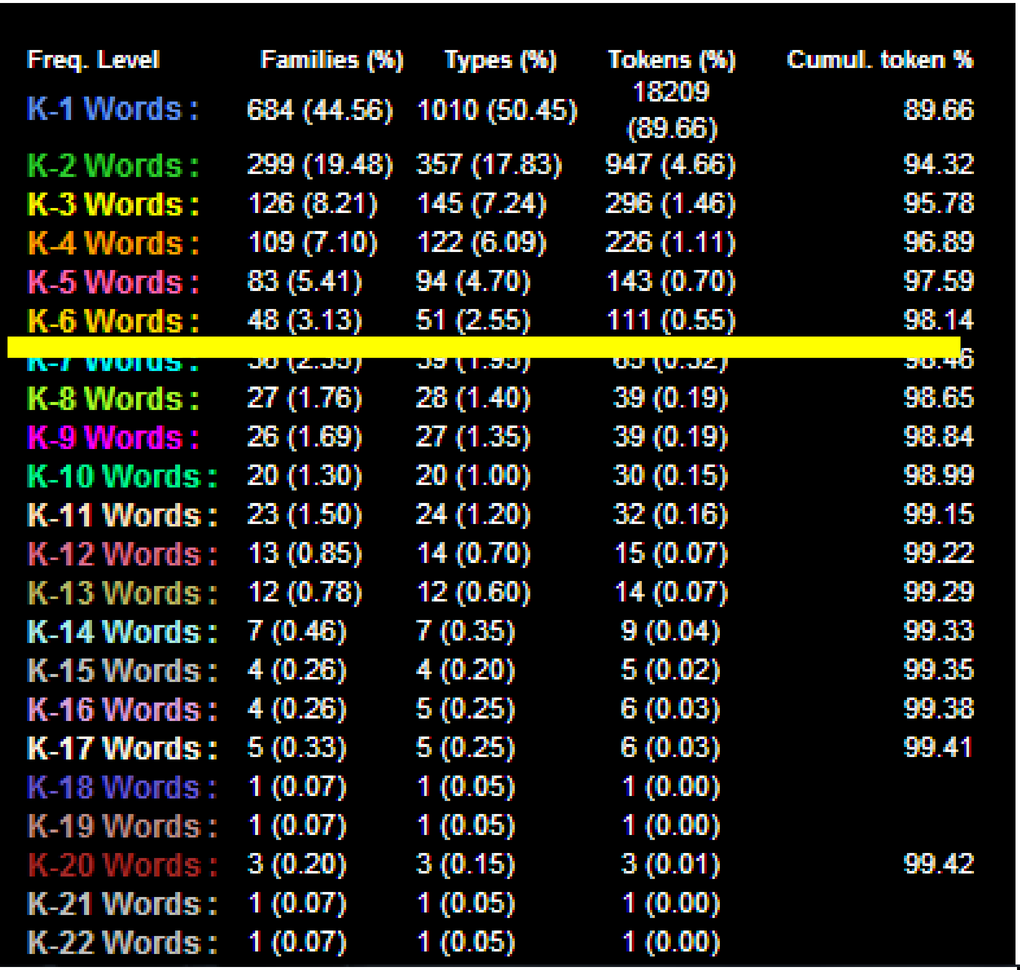
Cognates => 1k: None

Text Pre-Processing Notes: In the output text, punctuation is eliminated; all figures (1, 20, etc) are replaced by the word *number*; contractions are replaced by constituen
words may sum to less than total (depending on user treatment of proper nouns as well as program decision to class numbers as 1k although not contained in 1k list); sing

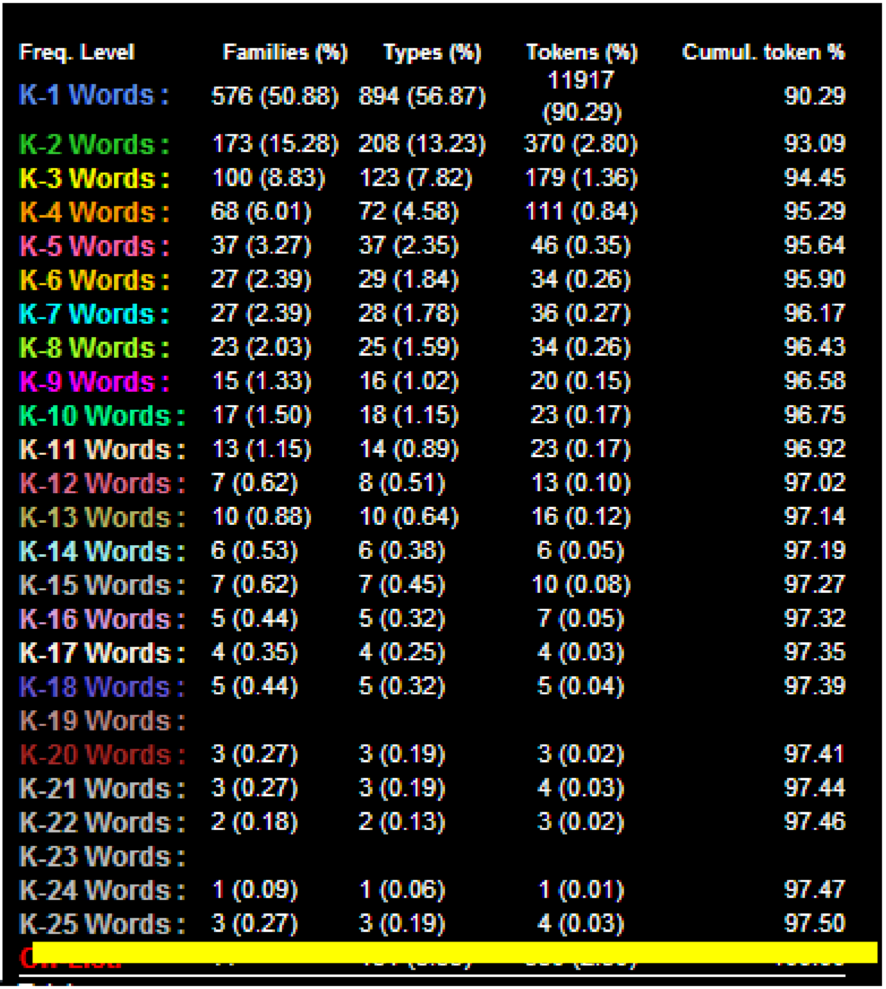


Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	576 (50.88)	894 (56.87)	11917 (90.29)	90.29
K-2 Words :	173 (15.28)	208 (13.23)	370 (2.80)	93.09
K-3 Words :	100 (8.83)	123 (7.82)	179 (1.36)	94.45
K-4 Words :	68 (6.01)	72 (4.58)	111 (0.84)	95.29
K-5 Words :	37 (3.27)	37 (2.35)	46 (0.35)	95.64
K-6 Words :	27 (2.39)	29 (1.84)	34 (0.26)	95.90
K-7 Words :	27 (2.39)	28 (1.78)	36 (0.27)	96.17
K-8 Words :	23 (2.03)	25 (1.59)	34 (0.26)	96.43
K-9 Words :	15 (1.33)	16 (1.02)	20 (0.15)	96.58
K-10 Words :	17 (1.50)	18 (1.15)	23 (0.17)	96.75
K-11 Words :	13 (1.15)	14 (0.89)	23 (0.17)	96.92
K-12 Words :	7 (0.62)	8 (0.51)	13 (0.10)	97.02
K-13 Words :	10 (0.88)	10 (0.64)	16 (0.12)	97.14
K-14 Words :	6 (0.53)	6 (0.38)	6 (0.05)	97.19
K-15 Words :	7 (0.62)	7 (0.45)	10 (0.08)	97.27
K-16 Words :	5 (0.44)	5 (0.32)	7 (0.05)	97.32
K-17 Words :	4 (0.35)	4 (0.25)	4 (0.03)	97.35
K-18 Words :	5 (0.44)	5 (0.32)	5 (0.04)	97.39
K-19 Words :				
K-20 Words :	3 (0.27)	3 (0.19)	3 (0.02)	97.41
K-21 Words :	3 (0.27)	3 (0.19)	4 (0.03)	97.44
K-22 Words :	2 (0.18)	2 (0.13)	3 (0.02)	97.46
K-23 Words :				
K-24 Words :	1 (0.09)	1 (0.06)	1 (0.01)	97.47
K-25 Words :	3 (0.27)	3 (0.19)	4 (0.03)	97.50

“Waiting for Godot” «En attendant Godot»



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	684 (44.56)	1010 (50.45)	18209 (89.66)	89.66
K-2 Words :	299 (19.48)	357 (17.83)	947 (4.66)	94.32
K-3 Words :	126 (8.21)	145 (7.24)	296 (1.46)	95.78
K-4 Words :	109 (7.10)	122 (6.09)	226 (1.11)	96.89
K-5 Words :	83 (5.41)	94 (4.70)	143 (0.70)	97.59
K-6 Words :	48 (3.13)	51 (2.55)	111 (0.55)	98.14
K-7 Words :	36 (2.35)	39 (1.95)	65 (0.32)	98.46
K-8 Words :	27 (1.76)	28 (1.40)	39 (0.19)	98.65
K-9 Words :	26 (1.69)	27 (1.35)	39 (0.19)	98.84
K-10 Words :	20 (1.30)	20 (1.00)	30 (0.15)	98.99
K-11 Words :	23 (1.50)	24 (1.20)	32 (0.16)	99.15
K-12 Words :	13 (0.85)	14 (0.70)	15 (0.07)	99.22
K-13 Words :	12 (0.78)	12 (0.60)	14 (0.07)	99.29
K-14 Words :	7 (0.46)	7 (0.35)	9 (0.04)	99.33
K-15 Words :	4 (0.26)	4 (0.20)	5 (0.02)	99.35
K-16 Words :	4 (0.26)	5 (0.25)	6 (0.03)	99.38
K-17 Words :	5 (0.33)	5 (0.25)	6 (0.03)	99.41
K-18 Words :	1 (0.07)	1 (0.05)	1 (0.00)	
K-19 Words :	1 (0.07)	1 (0.05)	1 (0.00)	
K-20 Words :	3 (0.20)	3 (0.15)	3 (0.01)	99.42
K-21 Words :	1 (0.07)	1 (0.05)	1 (0.00)	
K-22 Words :	1 (0.07)	1 (0.05)	1 (0.00)	



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	576 (50.88)	894 (56.87)	11917 (90.29)	90.29
K-2 Words :	173 (15.28)	208 (13.23)	370 (2.80)	93.09
K-3 Words :	100 (8.83)	123 (7.82)	179 (1.36)	94.45
K-4 Words :	68 (6.01)	72 (4.58)	111 (0.84)	95.29
K-5 Words :	37 (3.27)	37 (2.35)	46 (0.35)	95.64
K-6 Words :	27 (2.39)	29 (1.84)	34 (0.26)	95.90
K-7 Words :	27 (2.39)	28 (1.78)	36 (0.27)	96.17
K-8 Words :	23 (2.03)	25 (1.59)	34 (0.26)	96.43
K-9 Words :	15 (1.33)	16 (1.02)	20 (0.15)	96.58
K-10 Words :	17 (1.50)	18 (1.15)	23 (0.17)	96.75
K-11 Words :	13 (1.15)	14 (0.89)	23 (0.17)	96.92
K-12 Words :	7 (0.62)	8 (0.51)	13 (0.10)	97.02
K-13 Words :	10 (0.88)	10 (0.64)	16 (0.12)	97.14
K-14 Words :	6 (0.53)	6 (0.38)	6 (0.05)	97.19
K-15 Words :	7 (0.62)	7 (0.45)	10 (0.08)	97.27
K-16 Words :	5 (0.44)	5 (0.32)	7 (0.05)	97.32
K-17 Words :	4 (0.35)	4 (0.25)	4 (0.03)	97.35
K-18 Words :	5 (0.44)	5 (0.32)	5 (0.04)	97.39
K-19 Words :				
K-20 Words :	3 (0.27)	3 (0.19)	3 (0.02)	97.41
K-21 Words :	3 (0.27)	3 (0.19)	4 (0.03)	97.44
K-22 Words :	2 (0.18)	2 (0.13)	3 (0.02)	97.46
K-23 Words :				
K-24 Words :	1 (0.09)	1 (0.06)	1 (0.01)	97.47
K-25 Words :	3 (0.27)	3 (0.19)	4 (0.03)	97.50
Off-List:	[illegible]	[illegible]	[illegible]	[illegible]

Proper nouns-<1k has changed the 1k-2k thing

Test 3

Maybe Tests 1+2 were something about translated texts?

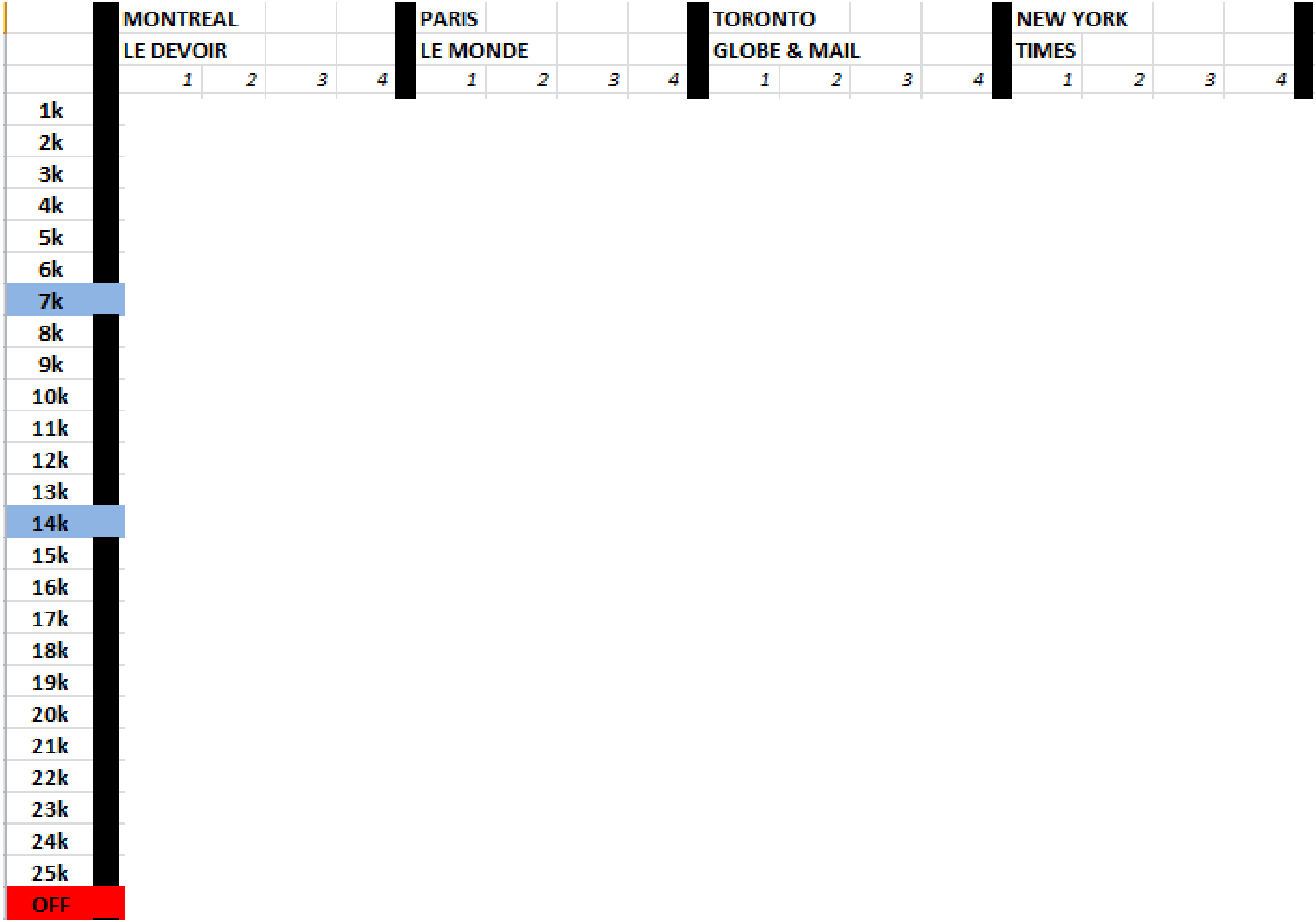
Ok, then let’s compare

4 random original editorial texts

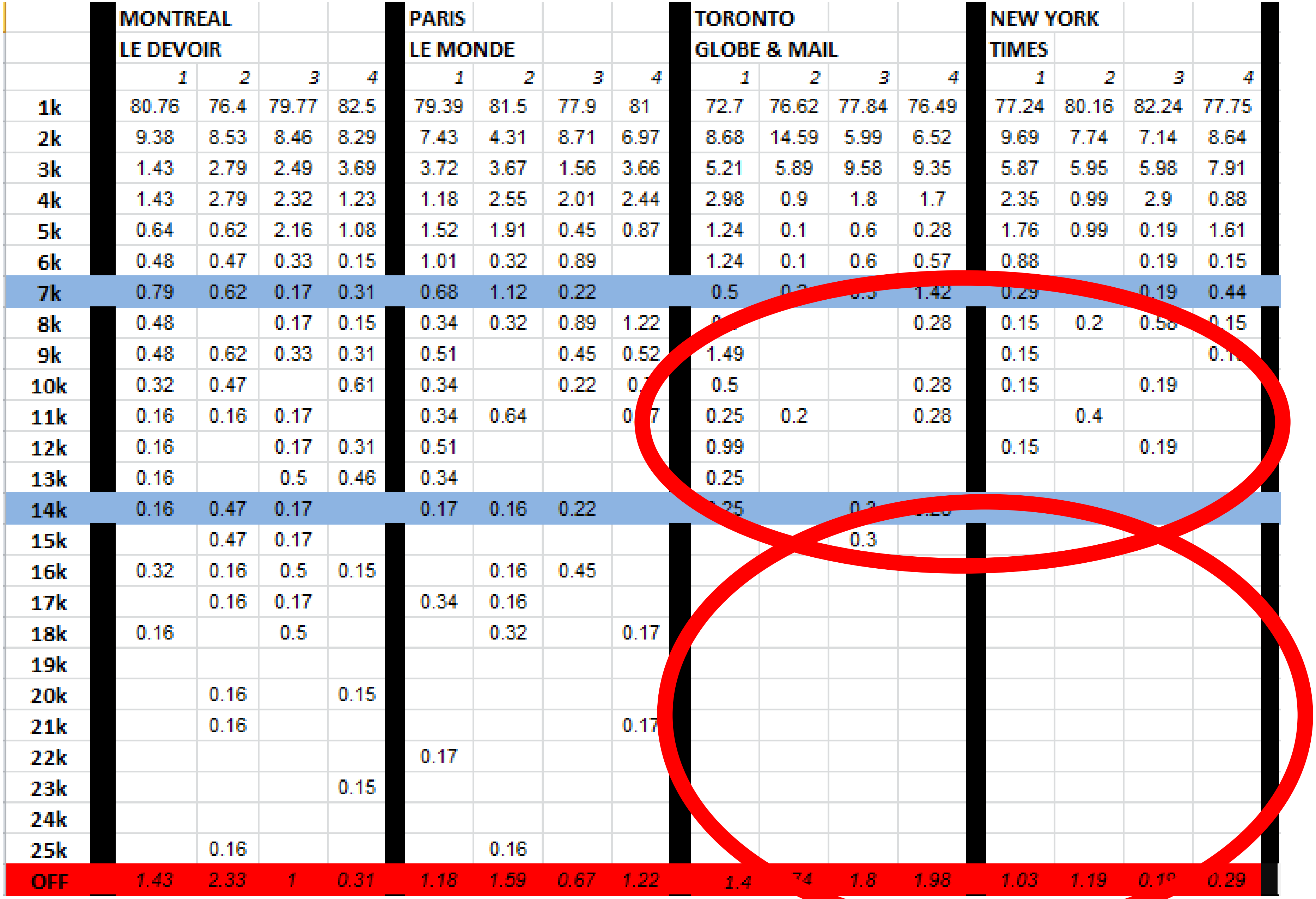
From each of ~

(1) Le Devoir – Montreal
(2) Le Monde - Paris
(3) The Globe & Mail – Toronto
(4) The NY Times – New York

Chosen 14-15 August, 2016



	MONTREAL				PARIS				TORONTO				NEW YORK			
	LE DEVOIR				LE MONDE				GLOBE & MAIL				TIMES			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1k																
2k																
3k																
4k																
5k																
6k																
7k																
8k																
9k																
10k																
11k																
12k																
13k																
14k																
15k																
16k																
17k																
18k																
19k																
20k																
21k																
22k																
23k																
24k																
25k																
OFF																



	MONTREAL				PARIS				TORONTO				NEW YORK			
	LE DEVOIR				LE MONDE				GLOBE & MAIL				TIMES			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1k	80.76	76.4	79.77	82.5	79.39	81.5	77.9	81	72.7	76.62	77.84	76.49	77.24	80.16	82.24	77.75
2k	9.38	8.53	8.46	8.29	7.43	4.31	8.71	6.97	8.68	14.59	5.99	6.52	9.69	7.74	7.14	8.64
3k	1.43	2.79	2.49	3.69	3.72	3.67	1.56	3.66	5.21	5.89	9.58	9.35	5.87	5.95	5.98	7.91
4k	1.43	2.79	2.32	1.23	1.18	2.55	2.01	2.44	2.98	0.9	1.8	1.7	2.35	0.99	2.9	0.88
5k	0.64	0.62	2.16	1.08	1.52	1.91	0.45	0.87	1.24	0.1	0.6	0.28	1.76	0.99	0.19	1.61
6k	0.48	0.47	0.33	0.15	1.01	0.32	0.89		1.24	0.1	0.6	0.57	0.88		0.19	0.15
7k	0.79	0.62	0.17	0.31	0.68	1.12	0.22		0.5	0.2	0.3	1.42	0.29		0.19	0.44
8k	0.48		0.17	0.15	0.34	0.32	0.89	1.22	[illegible]			0.28	0.15	0.2	0.58	0.15
9k	0.48	0.62	0.33	0.31	0.51		0.45	0.52	1.49				0.15			[illegible]
10k	0.32	0.47		0.61	0.34		0.22	[illegible]	0.5			0.28	0.15		0.19	
11k	0.16	0.16	0.17		0.34	0.64		0.17	0.25	0.2		0.28		0.4		
12k	0.16		0.17	0.31	0.51				0.99				0.15		0.19	
13k	0.16		0.5	0.46	0.34				0.25							
14k	0.16	0.47	0.17		0.17	0.16	0.22		0.25		0.3	[illegible]				
15k		0.47	0.17								0.3					
16k	0.32	0.16	0.5	0.15		0.16	0.45									
17k		0.16	0.17		0.34	0.16										
18k	0.16		0.5			0.32		0.17								
19k																
20k		0.16		0.15												
21k		0.16						0.17								
22k					0.17											
23k				0.15												
24k																
25k		0.16				0.16										
OFF	*1.43*	*2.33*	*1*	*0.31*	*1.18*	*1.59*	*0.67*	*1.22*	*1.4*	*[illegible]*	*1.8*	*1.98*	*1.03*	*1.19*	*0.19*	*0.29*

Conclusion

(1) Comparing languages:

- French makes ***slightly*** more use of its common words than English does
- But it makes ***far*** more use of its mid- and low-frequency lexical resources (3k to 20k+)
- So, Yes, **languages are distinct** in the way they deploy their lexical resources
  - So Cobb & Horst (2004) was right as far as it went, but incomplete
    - Old technology, fledgling paradigm,…

Conclusion

(2) Comparing learning tasks:

Learning enough vocab for 90% coverage looks ***slightly*** easier in French than English

But learning enough words for 98% or even 95% coverage looks ***far*** more difficult

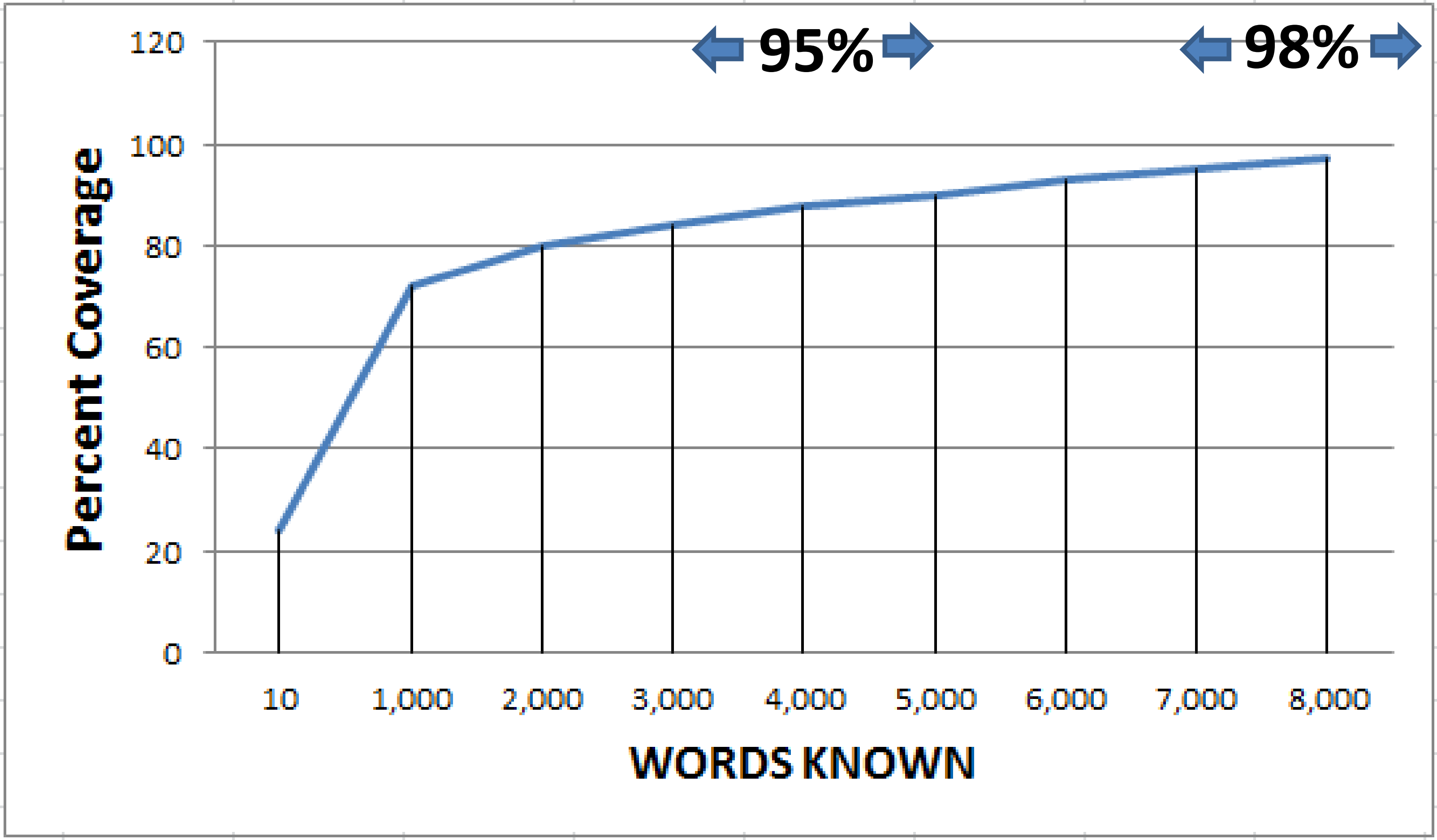
95% is best guess at basic lexical competence for reading

98% for full competence

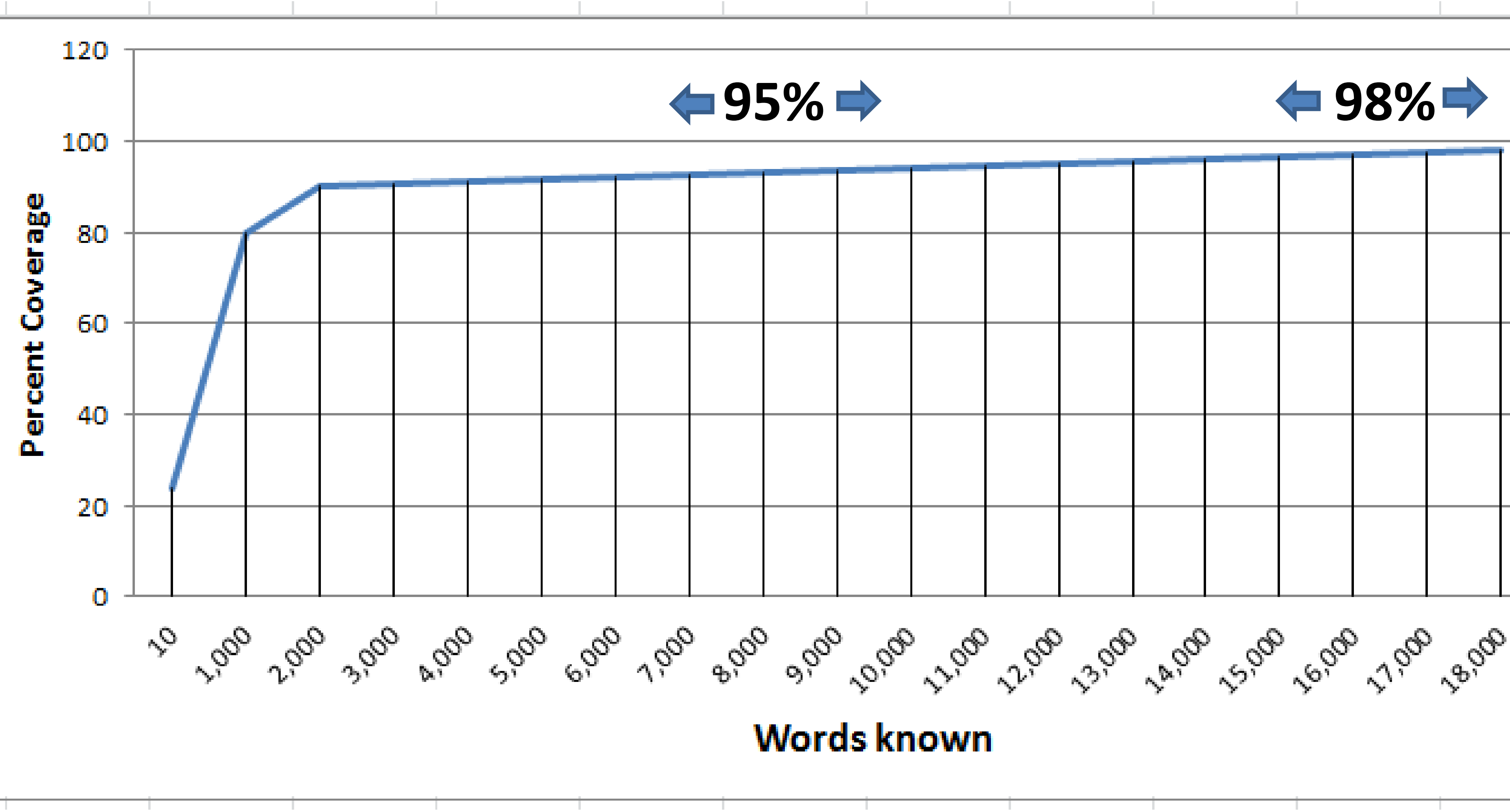
How many FL2-S's ever get to basic ?

(3) The **shape**s of the two lexicons seem to be like this:

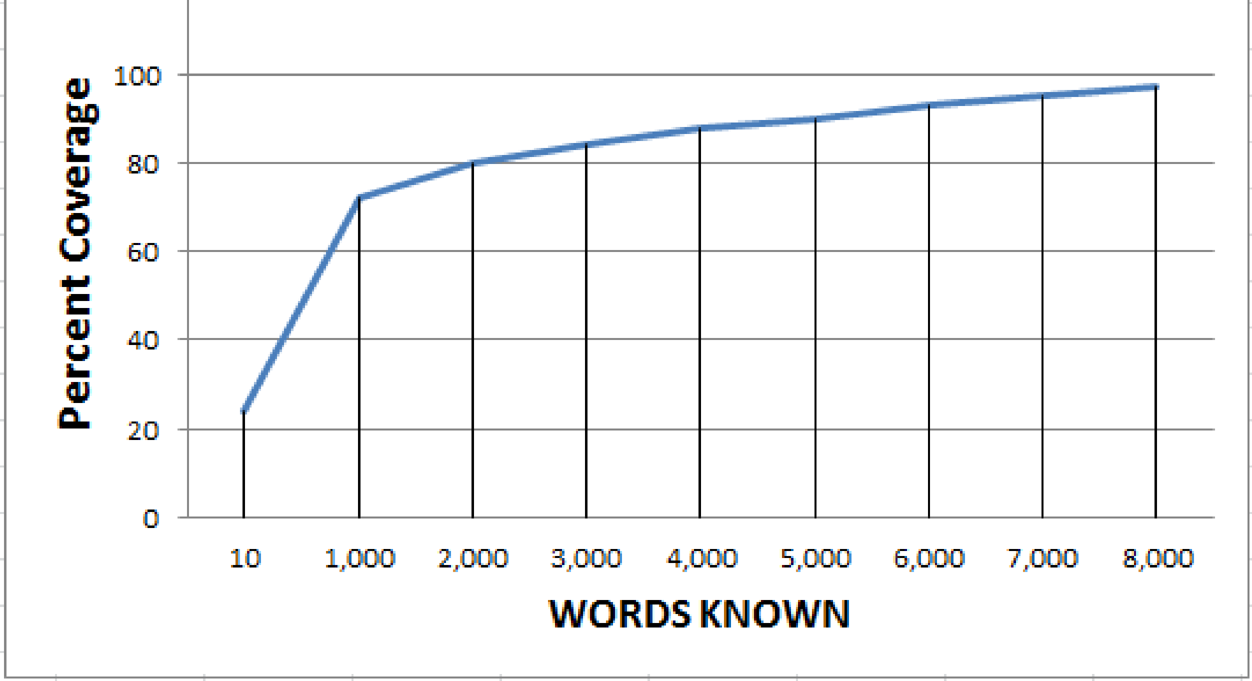
English



French



95%
98%
120
100
80
60
40
20
0
Percent Coverage
10
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,000
11,000
12,000
13,000
14,000
15,000
16,000
17,000
18,000
Words known

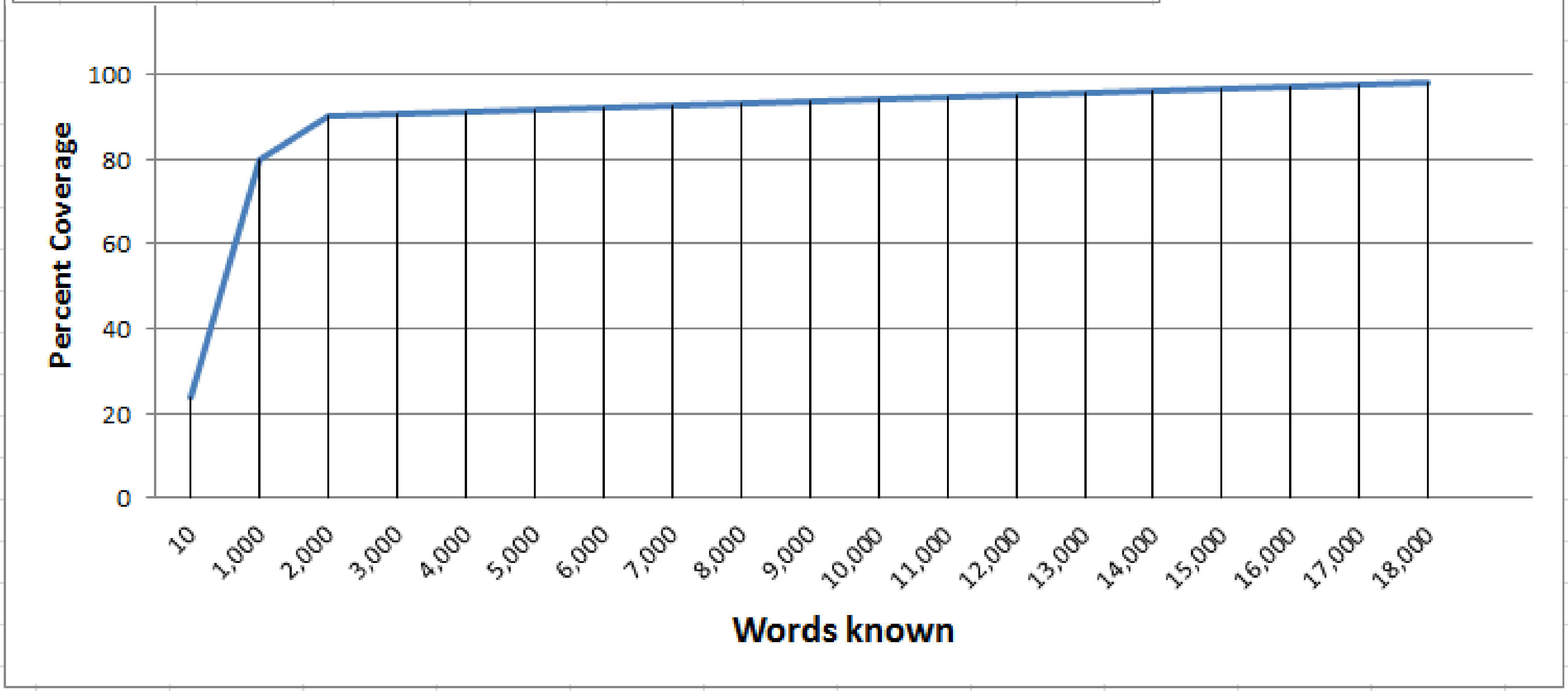


TOGETHER:

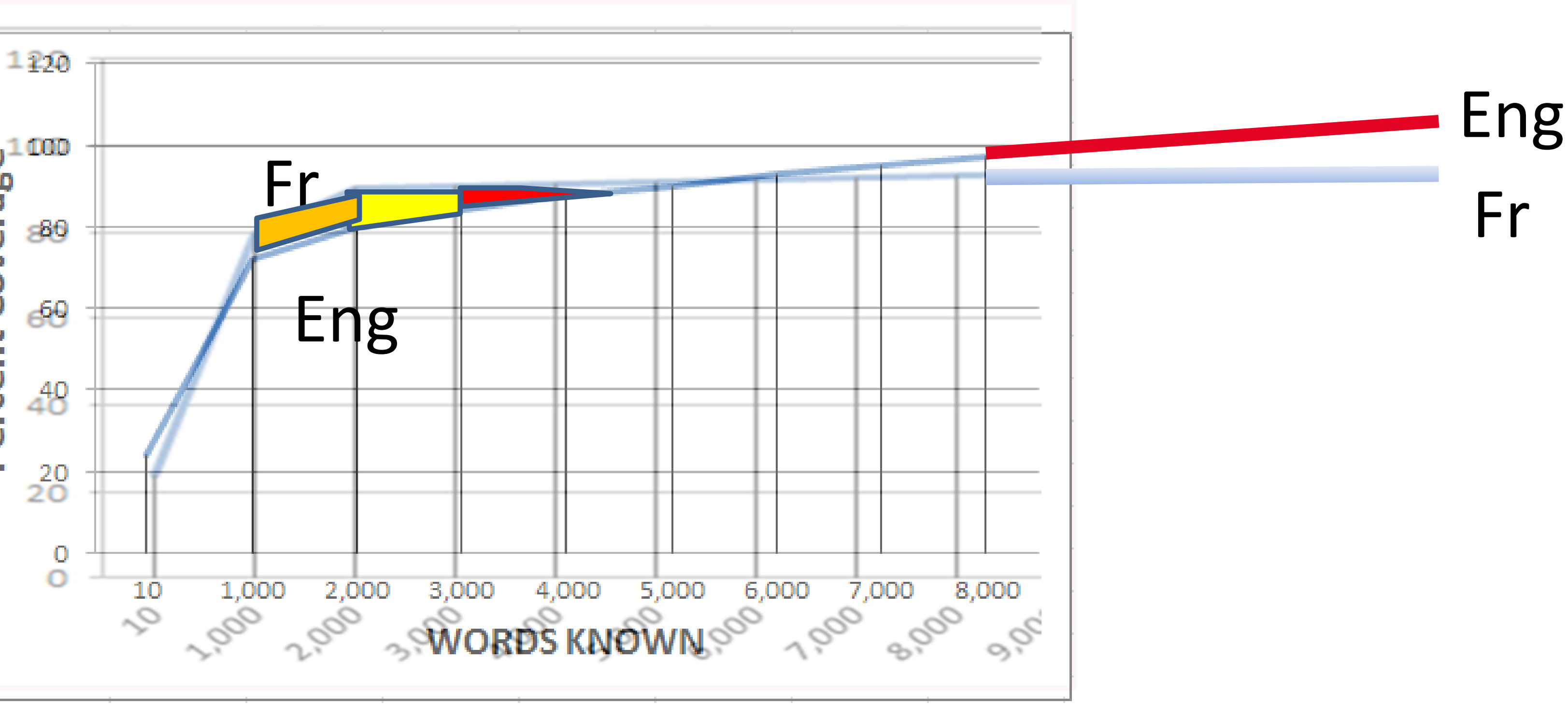
← Eng

Fr

↓



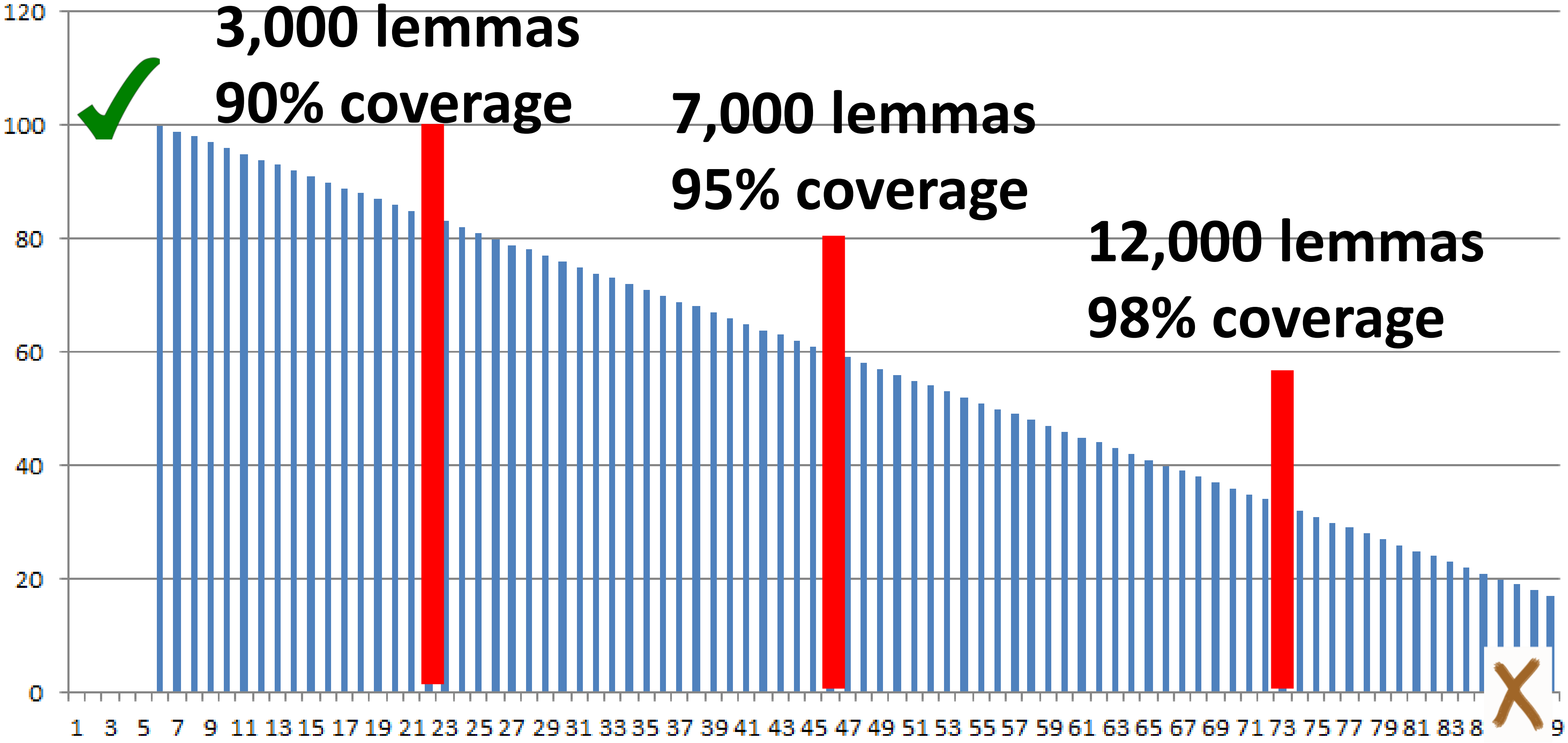
Superimposed



But notice that the French early advantage (higher coverage) persists to about 4k

(So 3k words in French gives better coverage than in English)

So our best guess (v.2016) at *basic lexical competence* for reading in FL_2 ?



3,000 lemmas
90% coverage

7,000 lemmas
95% coverage

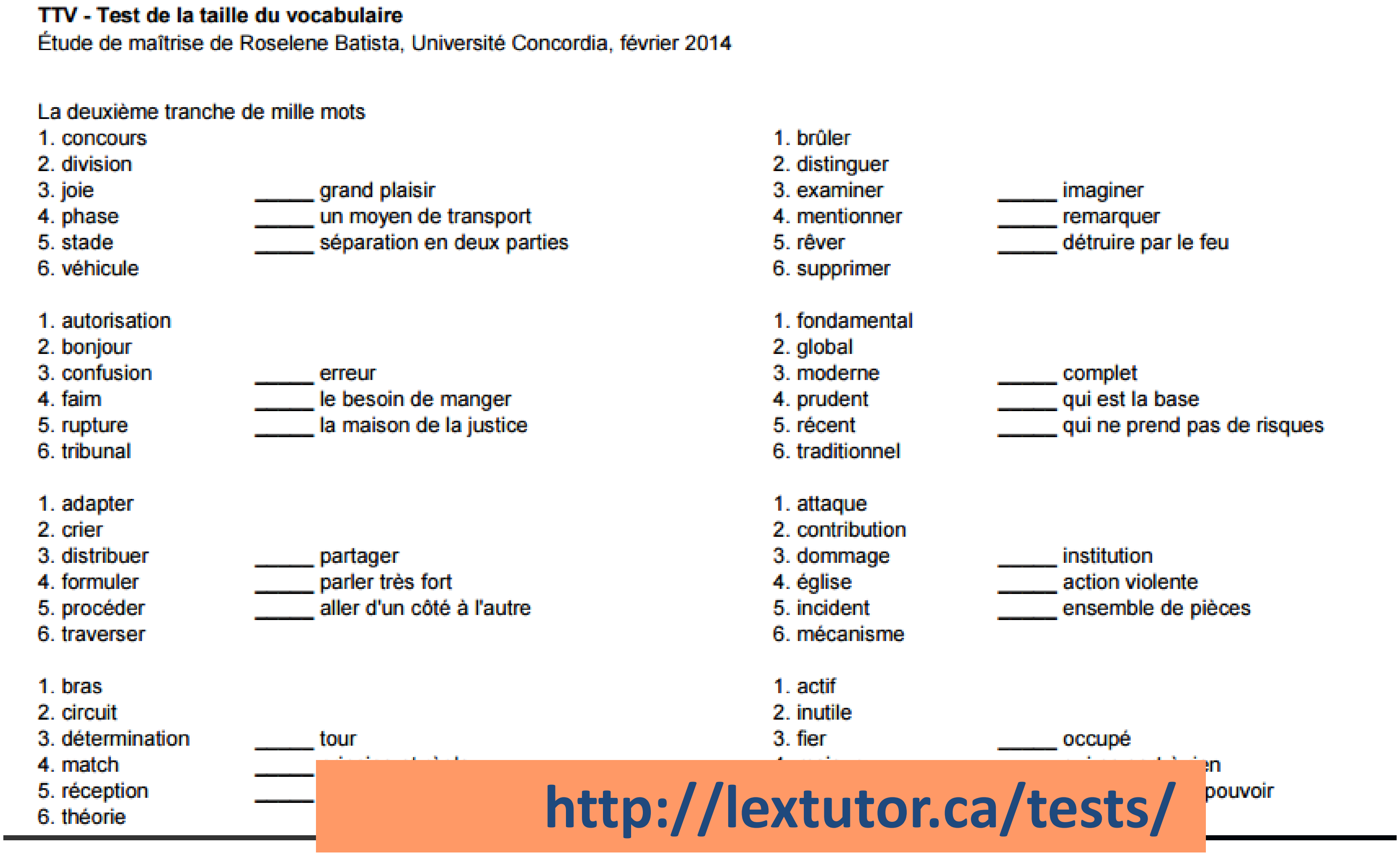
12,000 lemmas
98% coverage

1 word in 10 unknown

1 word in 20 unknown

1 word in 50 unknown

Where to start? How many words do our students know already?



TTV - Test de la taille du vocabulaire
Étude de maîtrise de Roselene Batista, Université Concordia, février 2014

La deuxième tranche de mille mots

1. concours
2. division
3. joie
4. phase
5. stade
6. véhicule

_____ grand plaisir
_____ un moyen de transport
_____ séparation en deux parties

1. autorisation
2. bonjour
3. confusion
4. faim
5. rupture
6. tribunal

_____ erreur
_____ le besoin de manger
_____ la maison de la justice

1. adapter
2. crier
3. distribuer
4. formuler
5. procéder
6. traverser

_____ partager
_____ parler très fort
_____ aller d'un côté à l'autre

1. bras
2. circuit
3. détermination
4. match
5. réception
6. théorie

_____ tour
_____
_____

1. brûler
2. distinguer
3. examiner
4. mentionner
5. rêver
6. supprimer

_____ imaginer
_____ remarquer
_____ détruire par le feu

1. fondamental
2. global
3. moderne
4. prudent
5. récent
6. traditionnel

_____ complet
_____ qui est la base
_____ qui ne prend pas de risques

1. attaque
2. contribution
3. dommage
4. église
5. incident
6. mécanisme

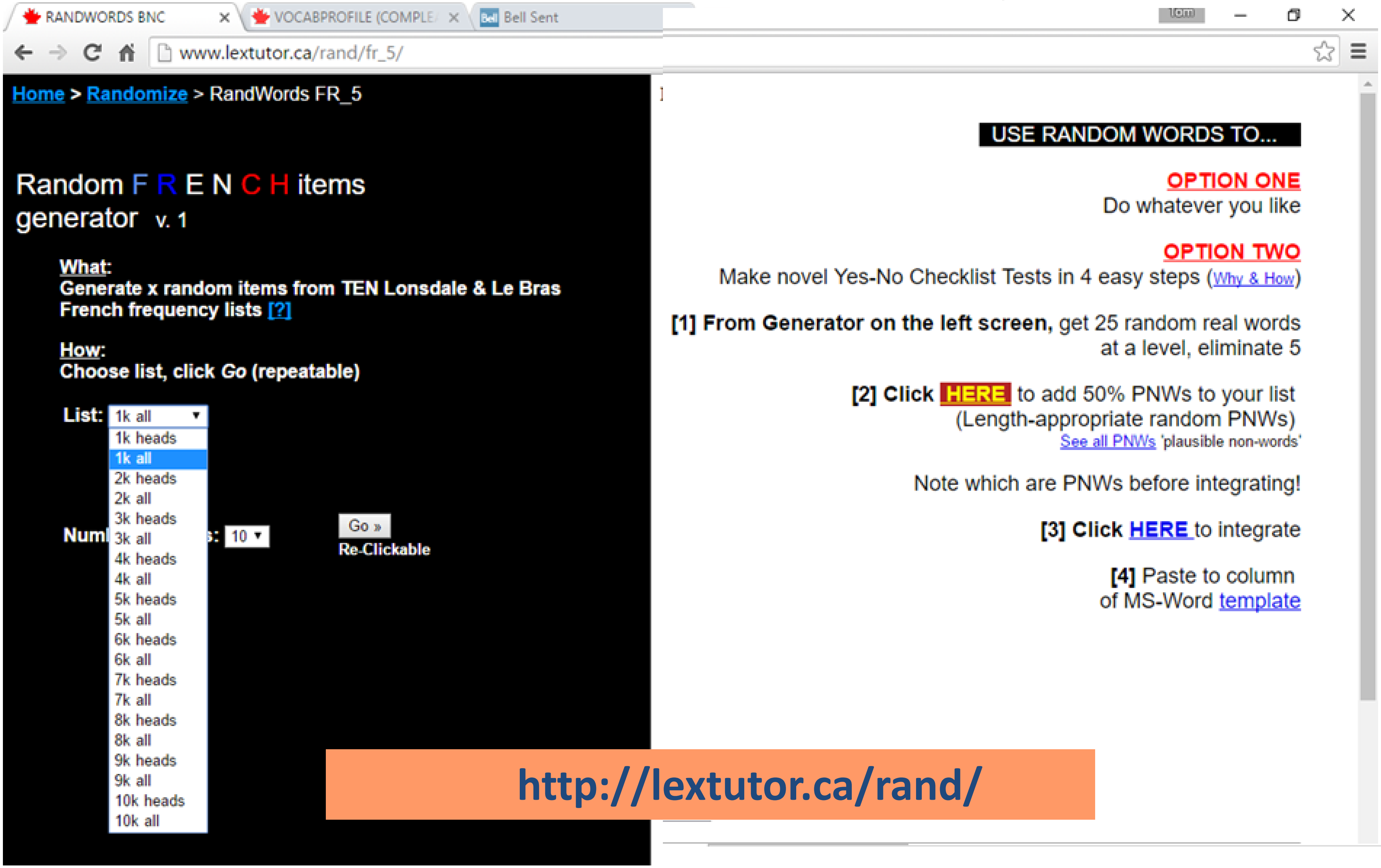
_____ institution
_____ action violente
_____ ensemble de pièces

1. actif
2. inutile
3. fier

_____ occupé
en
pouvoir

http://lextutor.ca/tests/

Where to start? How many words do our students know already? (2)



Home > Randomize > RandWords FR_5

Random F R E N C H items generator v. 1

What:
Generate x random items from TEN Lonsdale & Le Bras French frequency lists [?]

How:
Choose list, click *Go* (repeatable)

List: 1k all

Num[...]s: 10

Re-Clickable

USE RANDOM WORDS TO...

OPTION ONE
Do whatever you like

OPTION TWO
Make novel Yes-No Checklist Tests in 4 easy steps (Why & How)

[1] From Generator on the left screen, get 25 random real words at a level, eliminate 5

[2] Click HERE to add 50% PNWs to your list
(Length-appropriate random PNWs)
See all PNWs 'plausible non-words'

Note which are PNWs before integrating!

[3] Click HERE to integrate

[4] Paste to column of MS-Word template

http://lextutor.ca/rand/

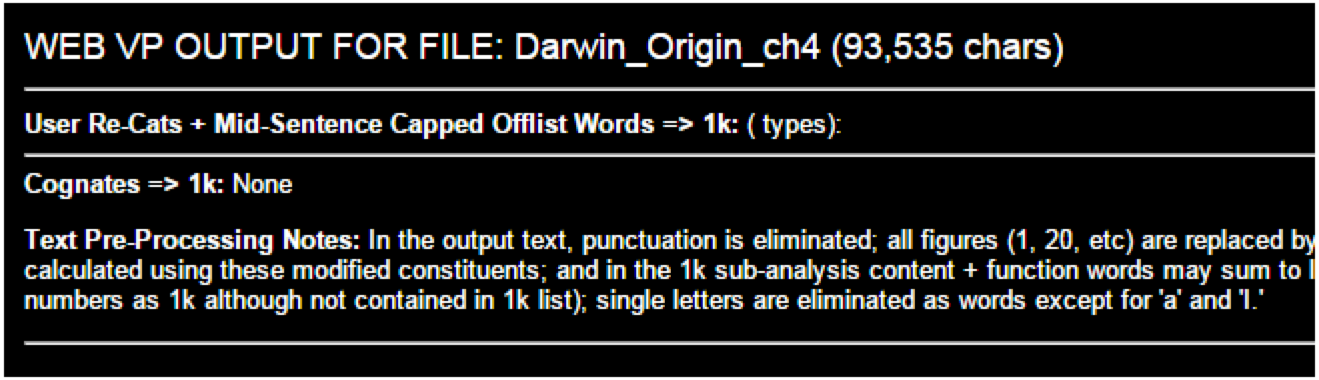
Afterthought

- Which language is out of step here – English or French?
  - Few languages have a separate academic lexicon

- Hazenburg & Hulstijn (c. 2005) calculated basic lexical competence in Dutch at 10,000 lemmas

- Maybe the shape of English reflects the *lingua franca* role the language has come to play
  - Such that its writers use *circumlocution* for complex ideas, rather than seeking « le mot juste »?

ENGLISH AS A LINGUA FRANCA? BUT SURELY NOT IN 19th CENT.

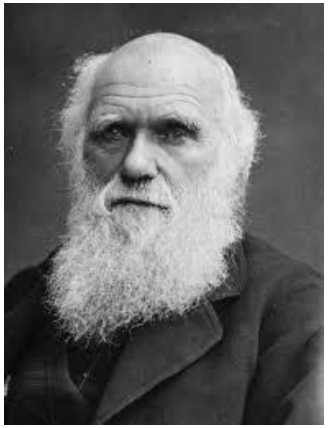


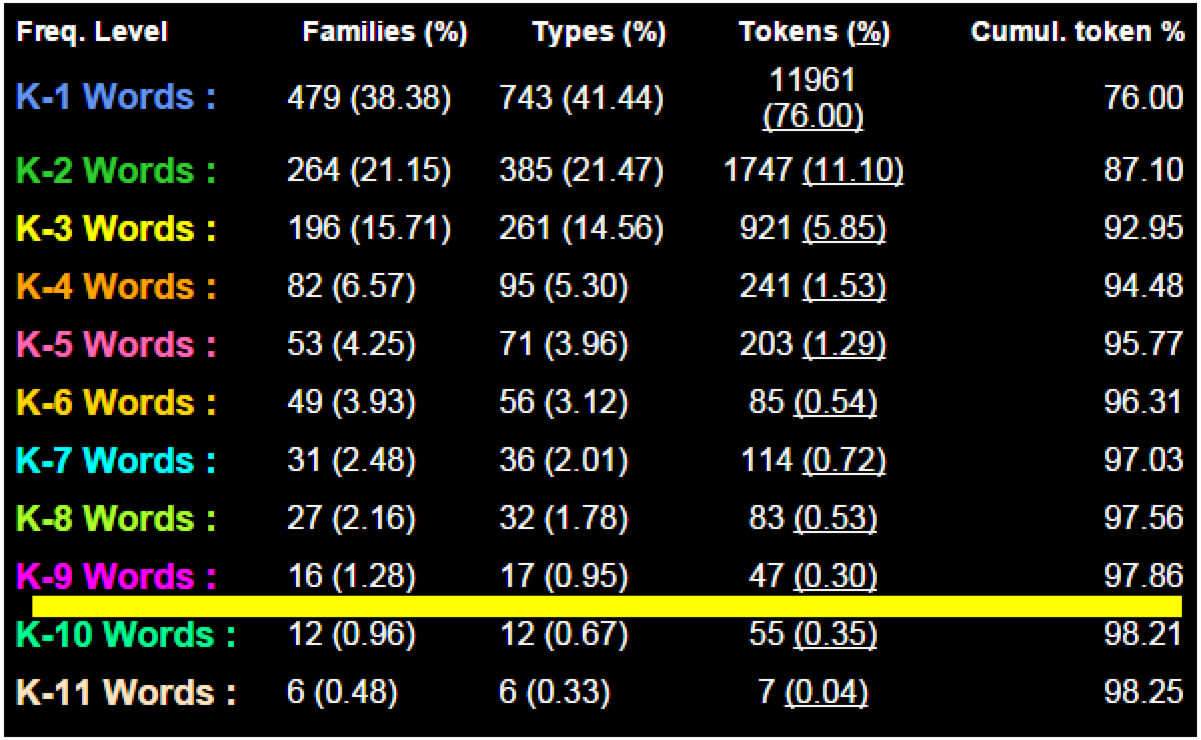
WEB VP OUTPUT FOR FILE: Darwin_Origin_ch4 (93,535 chars)

User Re-Cats + Mid-Sentence Capped Offlist Words => 1k: (types):

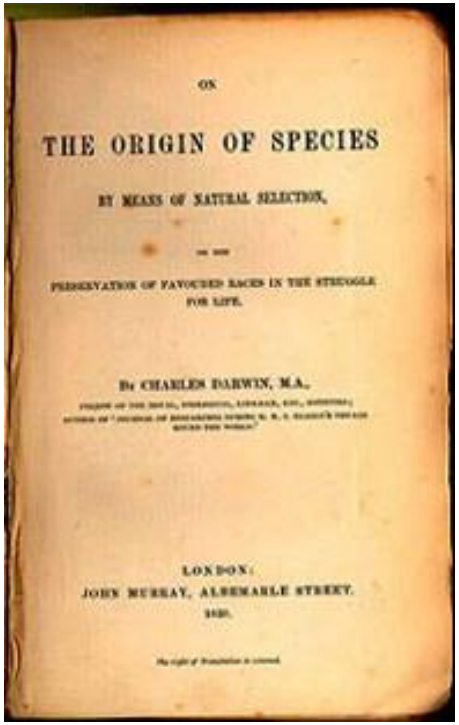
Cognates => 1k: None

Text Pre-Processing Notes: In the output text, punctuation is eliminated; all figures (1, 20, etc) are replaced by calculated using these modified constituents; and in the 1k sub-analysis content + function words may sum to l numbers as 1k although not contained in 1k list); single letters are eliminated as words except for 'a' and 'I.'





Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	479 (38.38)	743 (41.44)	11961 (76.00)	76.00
K-2 Words :	264 (21.15)	385 (21.47)	1747 (11.10)	87.10
K-3 Words :	196 (15.71)	261 (14.56)	921 (5.85)	92.95
K-4 Words :	82 (6.57)	95 (5.30)	241 (1.53)	94.48
K-5 Words :	53 (4.25)	71 (3.96)	203 (1.29)	95.77
K-6 Words :	49 (3.93)	56 (3.12)	85 (0.54)	96.31
K-7 Words :	31 (2.48)	36 (2.01)	114 (0.72)	97.03
K-8 Words :	27 (2.16)	32 (1.78)	83 (0.53)	97.56
K-9 Words :	16 (1.28)	17 (0.95)	47 (0.30)	97.86
K-10 Words :	12 (0.96)	12 (0.67)	55 (0.35)	98.21
K-11 Words :	6 (0.48)	6 (0.33)	7 (0.04)	98.25



ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

[illegible]

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE
FOR LIFE.

BY CHARLES DARWIN, M.A.,
[illegible]

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
[illegible]

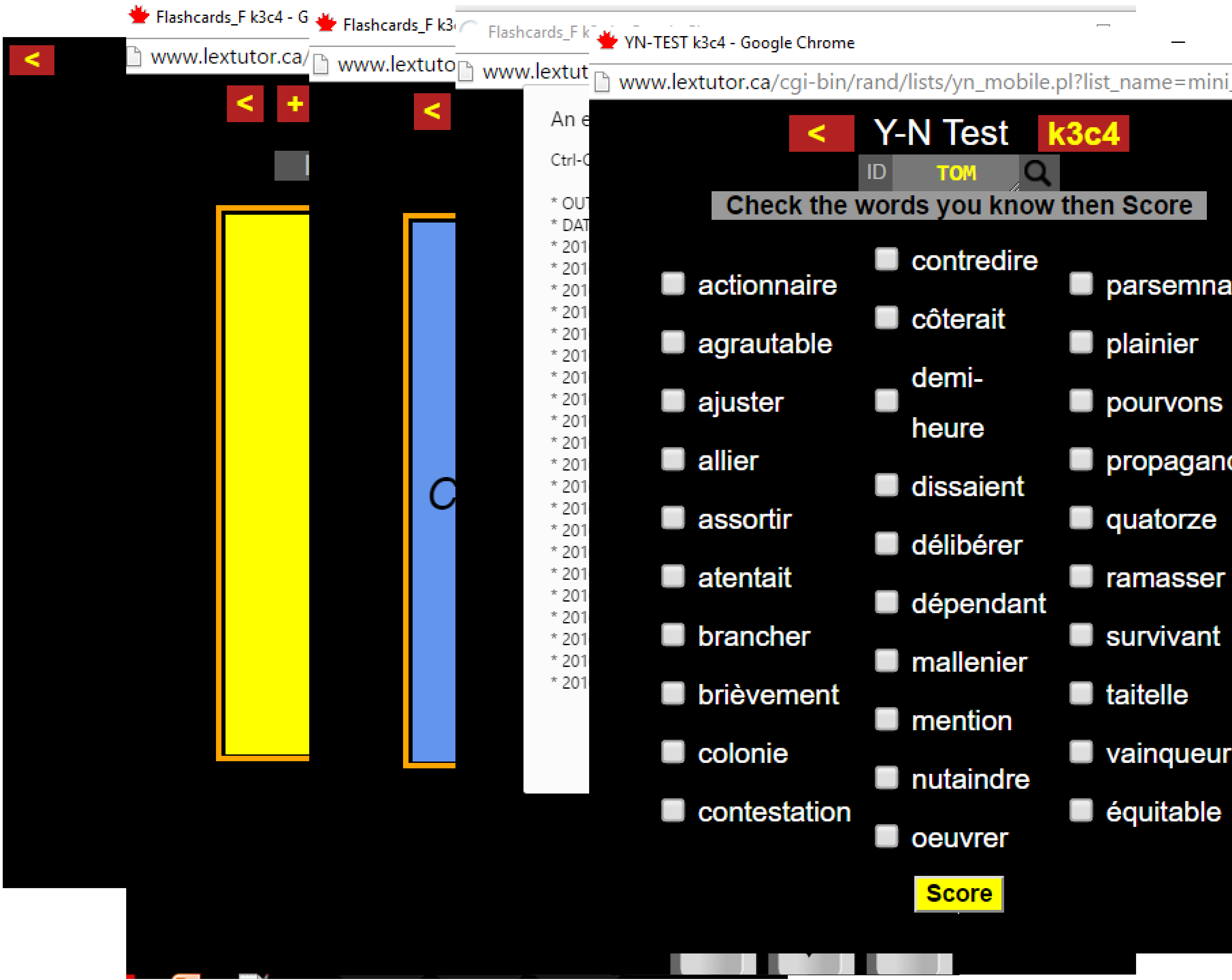
[illegible]

Further work

- As ever in corpus work, this needs empirical validation
  - Do FL2 readers with 5k=95% lexicons **actually** experience a comprehension deficit?
    - Or just have to look up a few more words?
  - Is it worth teaching vocab up to 98% general coverage?

- As ever in corpus work, newer better bigger lists are probably just around the next corner
  - Any picture is strictly provisional (yet we must do ***something*** Monday morning…)

- **Perspective needed:**
  My presentation deals with **advanced** learner issues, while 90% of vocab work is getting over the 5k hump
  - Establishing a basic form-meaning link ASAP so the true learning can begin (from reading, etc.)



Check the words you know then Score

□ actionnaire
□ agrautable
□ ajuster
□ allier
□ assortir
□ atentait
□ brancher
□ brièvement
□ colonie
□ contestation

□ contredire
□ côterait
□ demi-heure
□ dissaient
□ délibérer
□ dépendant
□ mallenier
□ mention
□ nutaindre
□ oeuvrer

□ parsemna
□ plainier
□ pourvons
□ propagan
□ quatorze
□ ramasser
□ survivant
□ taitelle
□ vainqueur
□ équitable

All references & software available

@

www.lextutor.ca

facebook.com/groups/lextutor

twitter.com/lextutor

Merci !

cobb.tom@sympatico.ca

A method note

- But wait!
- We are comparing lemmas v. families

Cat cats **v.** **cat cats *catty***

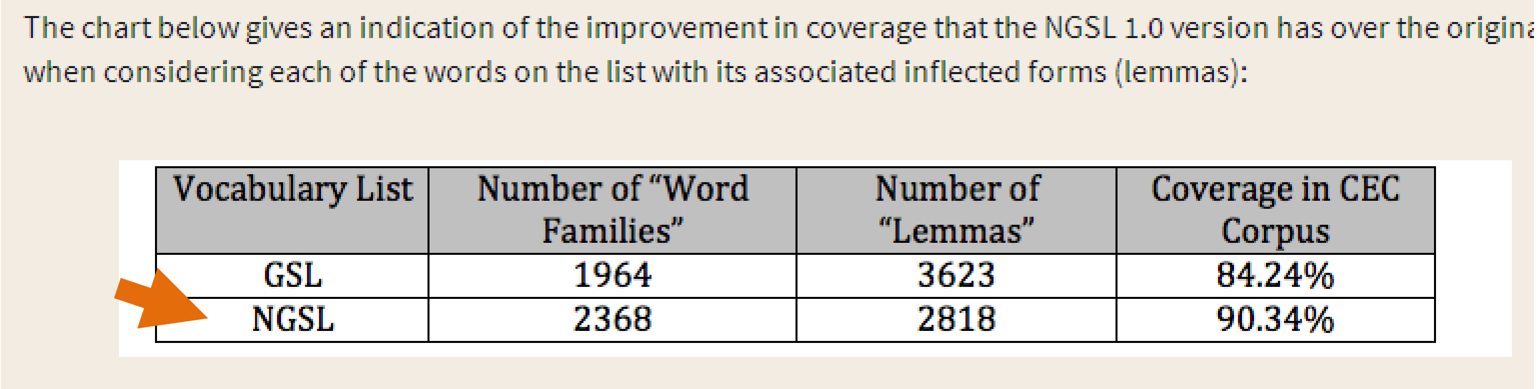
- **1000 families give more coverage than 1000 lemmas**
  - **How much more?**
    - Some recent work by Charles Browne suggests an answer

A NEW GENERAL SERVICE LIST (1.01)

the most important words for second language learners of English

CONTACT: BROWNE@LTR.MEIJIGAKUIN.AC.JP

http://www.newgeneralservicelist.org/



The chart below gives an indication of the improvement in coverage that the NGSL 1.0 version has over the origina when considering each of the words on the list with its associated inflected forms (lemmas):

Vocabulary List	Number of "Word Families"	Number of "Lemmas"	Coverage in CEC Corpus
GSL	1964	3623	84.24%
NGSL	2368	2818	90.34%

2368 / 2818 *100 = 84%

1000 lems have ~**16%** less coverage than 1000 fams in Eng

At High-Frequency NGSL zone (1k+2k)

(probably less at lower frequency zones)

But even assuming (1) a 16% difference that (2) was maintained at lower-frequency zones

- About every six lemma lists (6 x 16% = 96%) we would lose a k-level to maintain lemma-family equivalence
  - So in 18 levels we would lose 3

- The picture would not change greatly
  - Even in exaggerated worst-case scenario

Eng

(fams)

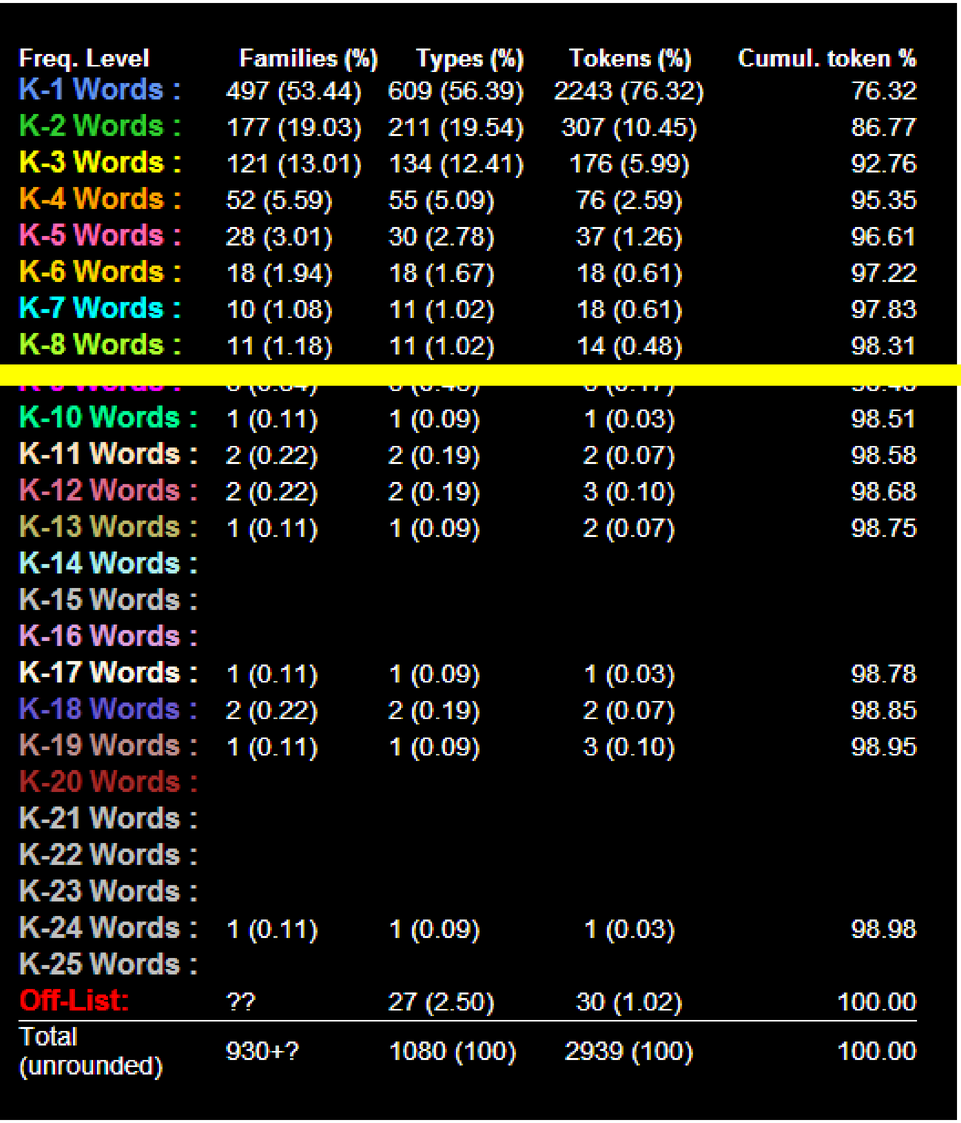
K8 E-fams = ~~k16~~ F-lems for 98% ?

→K8 E-fams = k13 F-lems for 98%

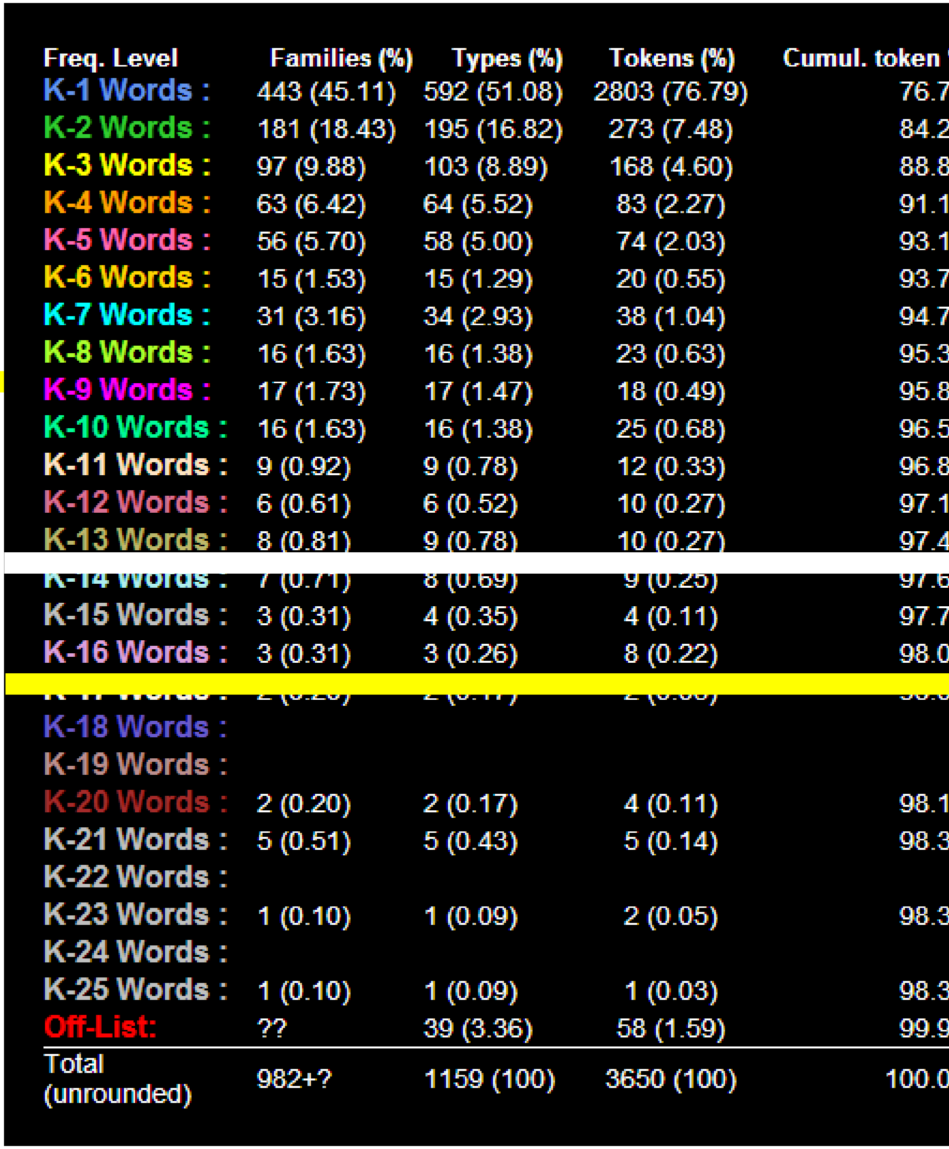
Pattern is the same

Fr

(lemmas)



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	497 (53.44)	609 (56.39)	2243 (76.32)	76.32
K-2 Words :	177 (19.03)	211 (19.54)	307 (10.45)	86.77
K-3 Words :	121 (13.01)	134 (12.41)	176 (5.99)	92.76
K-4 Words :	52 (5.59)	55 (5.09)	76 (2.59)	95.35
K-5 Words :	28 (3.01)	30 (2.78)	37 (1.26)	96.61
K-6 Words :	18 (1.94)	18 (1.67)	18 (0.61)	97.22
K-7 Words :	10 (1.08)	11 (1.02)	18 (0.61)	97.83
K-8 Words :	11 (1.18)	11 (1.02)	14 (0.48)	98.31
K-9 Words :	5 (0.54)	5 (0.46)	5 (0.17)	98.48
K-10 Words :	1 (0.11)	1 (0.09)	1 (0.03)	98.51
K-11 Words :	2 (0.22)	2 (0.19)	2 (0.07)	98.58
K-12 Words :	2 (0.22)	2 (0.19)	3 (0.10)	98.68
K-13 Words :	1 (0.11)	1 (0.09)	2 (0.07)	98.75
K-14 Words :				
K-15 Words :				
K-16 Words :				
K-17 Words :	1 (0.11)	1 (0.09)	1 (0.03)	98.78
K-18 Words :	2 (0.22)	2 (0.19)	2 (0.07)	98.85
K-19 Words :	1 (0.11)	1 (0.09)	3 (0.10)	98.95
K-20 Words :				
K-21 Words :				
K-22 Words :				
K-23 Words :				
K-24 Words :	1 (0.11)	1 (0.09)	1 (0.03)	98.98
K-25 Words :				
Off-List:	??	27 (2.50)	30 (1.02)	100.00
Total (unrounded)	930+?	1080 (100)	2939 (100)	100.00



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token
K-1 Words :	443 (45.11)	592 (51.08)	2803 (76.79)	76.7
K-2 Words :	181 (18.43)	195 (16.82)	273 (7.48)	84.2
K-3 Words :	97 (9.88)	103 (8.89)	168 (4.60)	88.8
K-4 Words :	63 (6.42)	64 (5.52)	83 (2.27)	91.1
K-5 Words :	56 (5.70)	58 (5.00)	74 (2.03)	93.1
K-6 Words :	15 (1.53)	15 (1.29)	20 (0.55)	93.7
K-7 Words :	31 (3.16)	34 (2.93)	38 (1.04)	94.7
K-8 Words :	16 (1.63)	16 (1.38)	23 (0.63)	95.3
K-9 Words :	17 (1.73)	17 (1.47)	18 (0.49)	95.8
K-10 Words :	16 (1.63)	16 (1.38)	25 (0.68)	96.5
K-11 Words :	9 (0.92)	9 (0.78)	12 (0.33)	96.8
K-12 Words :	6 (0.61)	6 (0.52)	10 (0.27)	97.1
K-13 Words :	8 (0.81)	9 (0.78)	10 (0.27)	97.4
K-14 Words :	7 (0.71)	8 (0.69)	9 (0.25)	97.6
K-15 Words :	3 (0.31)	4 (0.35)	4 (0.11)	97.7
K-16 Words :	3 (0.31)	3 (0.26)	8 (0.22)	98.0
K-17 Words :	2 (0.20)	2 (0.17)	2 (0.05)	98.0
K-18 Words :				
K-19 Words :				
K-20 Words :	2 (0.20)	2 (0.17)	4 (0.11)	98.1
K-21 Words :	5 (0.51)	5 (0.43)	5 (0.14)	98.3
K-22 Words :				
K-23 Words :	1 (0.10)	1 (0.09)	2 (0.05)	98.3
K-24 Words :				
K-25 Words :	1 (0.10)	1 (0.09)	1 (0.03)	98.3
Off-List:	??	39 (3.36)	58 (1.59)	99.9
Total (unrounded)	982+?	1159 (100)	3650 (100)	100.0