英語コーパス学会第48回大会

2022年10月1日（土）9時30分～17時40分 On Zoom

Workshop : Data-Driven Course Design

Tom Cobb

Université du Québec à Montréal

cobb.tom.3@gmail.com

Sat Oct 1 2022 9:40-10:40

UQÀM

Running a Vocabulary Course With Lextutor

Dr. Tom Cobb (Université du Québec à Montréal)

The importance of vocabulary knowledge in any type of language course or curriculum is now acknowledged but still not easy to incorporate in a systematic manner. There are few dedicated vocabulary courses, and the ones there are do not match the increasing specialisation of many learner programs. Lextutor has been designed to basically take on the whole job of a dedicated vocabulary supplement from a practical corpus perspective. My workshop will show the main steps in this process, from building a corpus of learning materials, to placement testing, to text selection and adaptation, to assuring a manageable supply of new and appropriate items, to test writing that reflects what learners have actually been exposed to sufficiently to be tested on. The theme is 'corpora for courses' and I will share results from locales where this approach and technology is being deployed.

Premise and RQ

Much learning research has been done

- And replicated

Many good corpus tools based on it are available

- And online or free to download

How often/systematically are these be brought together for the benefit of learners?

- IE, as courses? Curricula?

What would doing so look like? SOME CONCRETE IDEAS→

Definitions, context, rationale

Corpus

- Representative collection of texts
  - Ex Brown Corpus (1960s)
    - 1 million words
    - 500 textes on 15 topics of 2000 words each
  - BNC (1980s)
    - 100 million words
    - >4000 texts on 100 topics of 25,000 words (approx.)
  - COCA, Subtlex, TenTen …

Corpus – read and interpreted how ?

001. 0 décembre, 134 personnes 53 hommes, 55 FEMMES **26** enfants de un à treize ans ét
002. de ces deux affaires, une homme et une FEMME : Alfred Sirven, ancien directeur
003. ns cent quatre-vingt-dix hommes et deux FEMMES ? Un peu plus tôt, a-t-il partici
004. chorégraphie. La pièce met en scène une FEMME **agressée** par deux hommes et appela
005. pide peur de l'homme, des hommes et des FEMMES **aussi**... Un besoin maladif d'igno
006. août 1942. Déjà, de nombreux hommes et FEMMES **avaient** été astreints à des pério
007. amenant vers l'emploi des hommes et des FEMMES **dans** la force de l'âge, souvent c
008. 'hématocrite chez 46 athlètes hommes et FEMMES **de** haut niveau (participant à des
009. . Les hommes d'aujourd'hui traitent les FEMMES **de** manière aussi primitive, aussi
010. de vingt-cinq ans et aux hommes et aux FEMMES **de** plus de vingt-cinq ans. En rev
011. long desquelles se succèdent hommes et FEMMES **de** toutes conditions et origines
012. ne se reproduisent, tous les hommes et FEMMES **du** groupe qui travaillent à la re
013. hommes d'affaires tout-puissants... "La FEMME **du** président" a encore six heures
014. ravaillait comme un homme, ou comme une FEMME **Elle** croit ce qu'elle dit et dit
015. onnage principal, l'homme trompé par sa FEMME **employant** maints stratagèmes pour
016. ports ambigus d'un vieil homme et d'une FEMME **en** fleur. Une tragédie intimiste s
017. janvier, traite ainsi des hommes et des FEMMES **en** politique dans le monde, du "m
018. nt ce partage de la vie entre hommes et FEMMES **en** termes politiques et culturels
019. emportent maintenant des hommes et des FEMMES **encore** jeunes qui laissent des en
020. e malade et doit subir la cruauté de sa FEMME **Enlevé** par ses hommes, il finit p
021. rieusement atteinte ? "J'espère que les FEMMES **et** aussi les hommes de ce pays e
022. vant ces ouvrages historiques signés de FEMMES **et** d'hommes publics, dont on imag
023. du Soudan ROUX OLIVIER DES hommes, des FEMMES **et** des enfants meurent de faim au
024. le monde a toujours fermé les yeux. Des FEMMES **et** des hommes n'ont cessé de tire
025. es", "filières criminelles") contre des FEMMES **et** des hommes prêts à se laisser

Definitions…

"Data driven learning" or "corpus based learning"

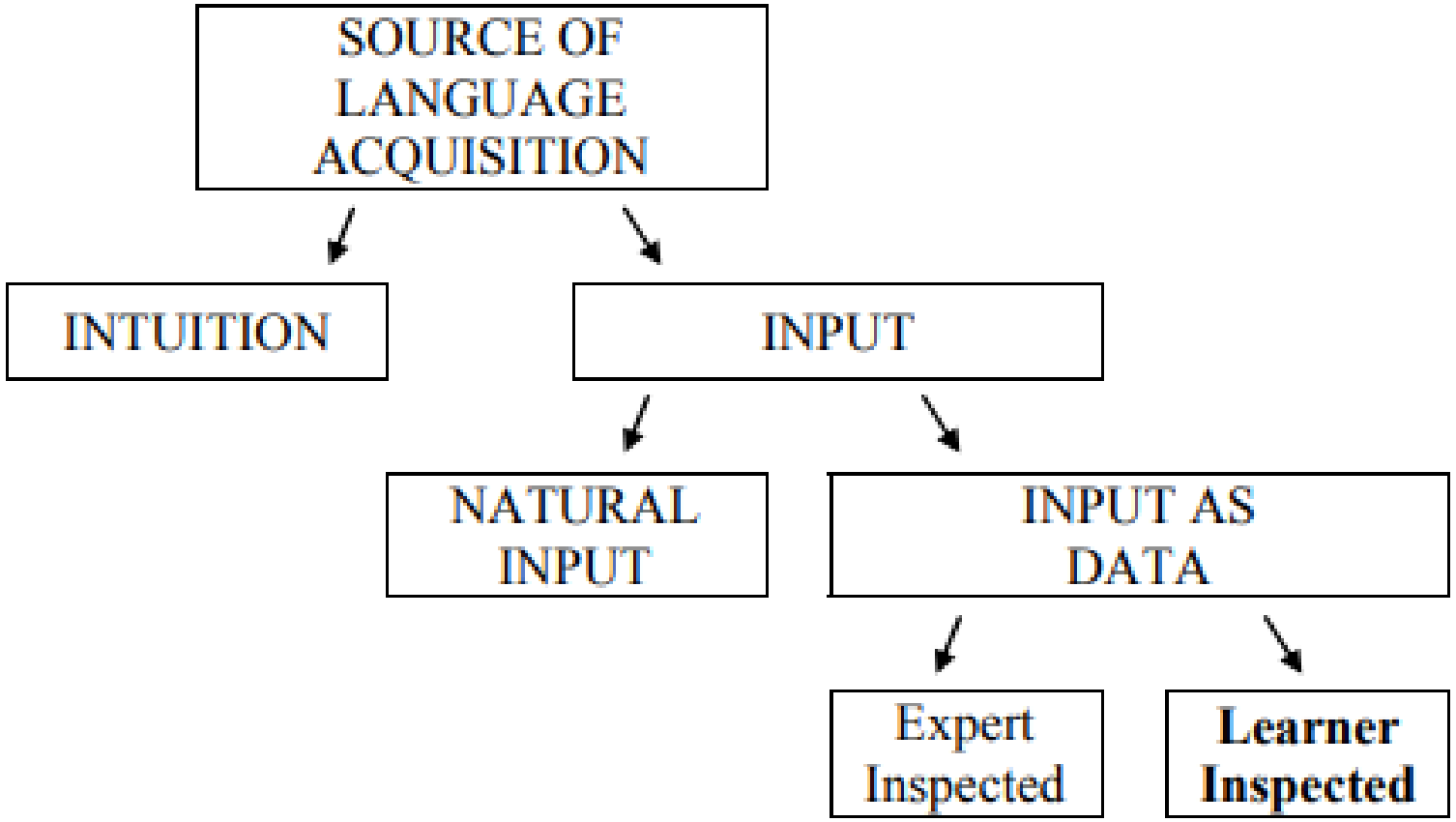


Figure 1. The place of data-driven language learning in the broader scheme

Main idea of DDL

Language learning happens through input,

BUT

- Language input reveals its patterns slowly
  - Too slowly to be effective in most L2 learning
- Input, however, can be treated by computer programs to expose patterns more quickly
  - more clearly
- To create more transparent, ‘learnable’ input
  - Patterns exposed through data ‘agglomeration’
    - As compared to rules or explanation

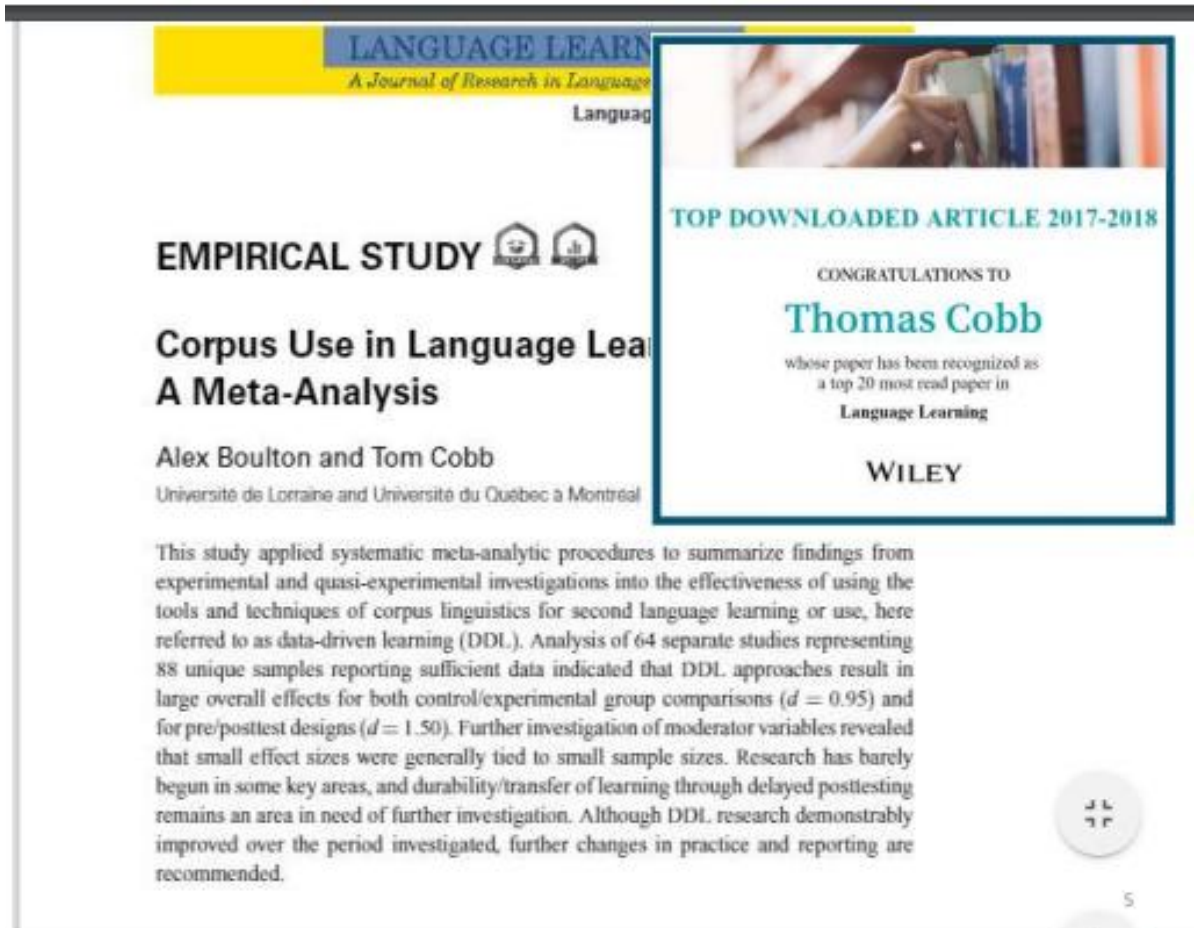
DDL (Data driven learning) is well established as a set of learning tools

- Concordances, frequency lists, learner-as-linguist, discovery learning, pattern extraction, exemplar-based not rule-based learning, etc.
  - But does it work?
- Boulton & Cobb (2017) meta-analysis of DDL
  - Effect size of **1.5** cf 'traditional' learning method
    - Meaning scores 1.5 standard deviations higher
- Many ESL courses now employ these tools piecemeal
  - But can they be integrated systematically?
    - At the point of course conception?

Learning from corpora (=DDL)

Apprendre à travers l'utilisation des corpus / Apprentissage faisant appel aux corpus

- Récemment montré efficace de façon concluante (2017)


LANGUAGE LEARN

A Journal of Research in Language

Languag

EMPIRICAL STUDY

Corpus Use in Language Lea
A Meta-Analysis

Alex Boulton and Tom Cobb

Université de Lorraine and Université du Québec à Montréal

This study applied systematic meta-analytic procedures to summarize findings from experimental and quasi-experimental investigations into the effectiveness of using the tools and techniques of corpus linguistics for second language learning or use, here referred to as data-driven learning (DDL). Analysis of 64 separate studies representing 88 unique samples reporting sufficient data indicated that DDL approaches result in large overall effects for both control/experimental group comparisons ($d = 0.95$) and for pre/posttest designs ($d = 1.50$). Further investigation of moderator variables revealed that small effect sizes were generally tied to small sample sizes. Research has barely begun in some key areas, and durability/transfer of learning through delayed posttesting remains an area in need of further investigation. Although DDL research demonstrably improved over the period investigated, further changes in practice and reporting are recommended.

Comparé à une gamme d'autres moyens d'apprendre…

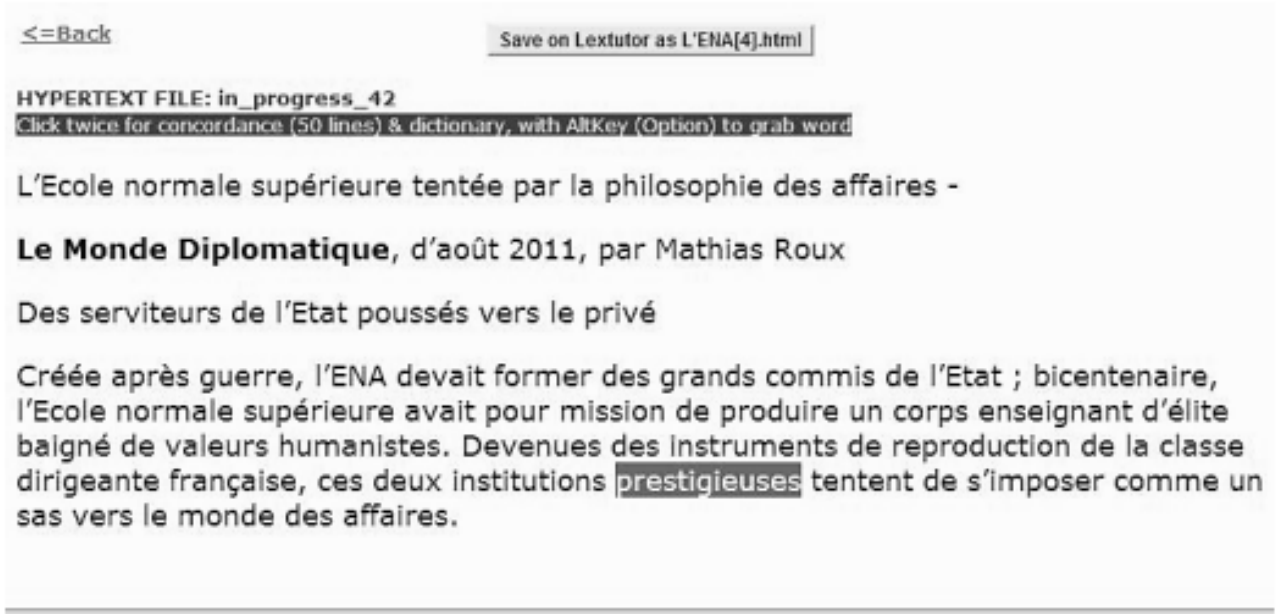
Le vocabulaire
Les locutions
Le syntaxe
La grammaire
La culture

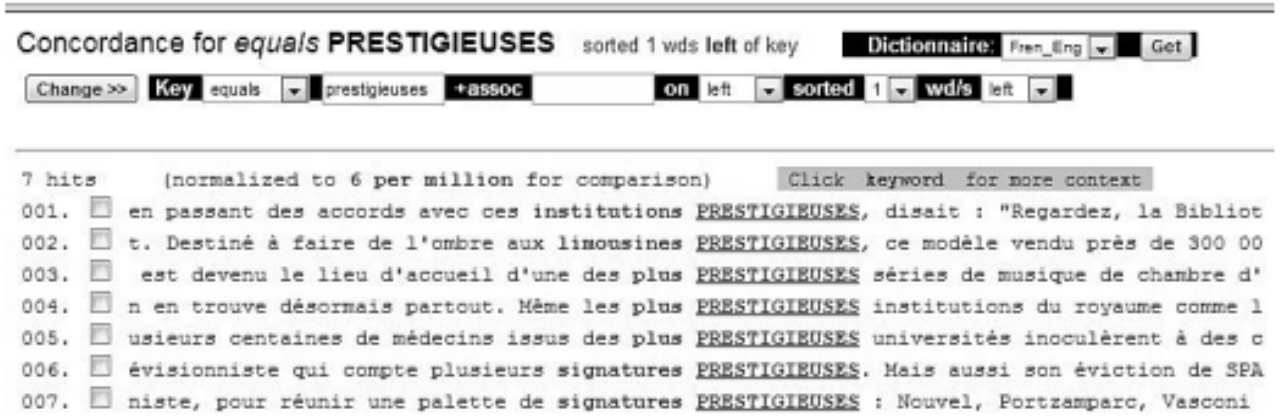
… DDL 'GAGNE' par la taille de l'effet de 1,5 écarts-types (**effect size of 1.5)** pour les études pré-post

Ex. Control Group Mean **65**, SD 15
DDL Group Mean **80**, SD 15

Example of a DDL learning activity?

Constructing word meanings from raw data vs pre-constructed knowledge





←vs.→

Qu'est-ce qu'une activité d'apprentissage basée sur un corpus?

Exemple 2 – sensibilisation à la culture

Tâche

- Do concordances for ‘woman’ with collocation ‘man’ on either side in English, French, and Spanish corpora
  - Woman - man
  - Femme – homme
  - Mujer – hombre

Then in a one-page report:

- Is the sequence always man first then woman second across these three languages?
- In which language is there most variation in the sequence?
- Is the common sequence formulaic or indicate a belief in the superiority of one gender?
  - Or with one gender acting and the other being acted upon?

001. 0 décembre, 134 personnes 53 hommes, 55 FEMMES 26 enfants de un à treize ans ét
002. de ces deux affaires, une homme et une FEMME : Alfred Sirven, ancien directeur
003.
004.
005.
006.
007.
008.
009.
010.
011.
012.
013.
014.
015.
016.
017.
018.
019.
020.
021.
022.
023.
024.
025.

001. ult life of the average American man or WOMAN. Following a vigorous campaign of
002. not standing up to them. An aggressive WOMAN wants a man to demand, not knuckle
003. all the way home. The ordinary man and WOMAN, however, saw little of the great
004. to the life of the middle-class man or WOMAN, dictating the methods of child re
005. t exasperating to men, who expect every WOMAN to verify their preconceived notio
006. er, he muttered, to meditate on man (or WOMAN) than on God. David finished shavi
007. dresser". "I guess it's children make a WOMAN old. A man gets old anyhow". After
008. e. There was a very old man and a young WOMAN and a brood of children ranging fr
009. r tilted back. "So you're looking for a WOMAN who married a man who might have l
010. ecial reason to believe that the man or WOMAN they sought had stayed only overni
011. at? The door opened and three men and a WOMAN in a sari swept past him and down
012. a scout ship, partnered with a man or a WOMAN, whichever she chose, as the mobil
013. house. Now he saw that both the man and WOMAN were moving slowly and irregularly
014. many affairs of the heart, a man and a WOMAN meet and something clicks. Somethi
015. in stormy weather I marry this man and WOMAN together. Let none but Him who rul
016. who rules the thunder Put this man and WOMAN asunder". Absolution for his lie?
017. fteen miles east of here. Eleven men, a WOMAN and a teen-age boy tramped over co
018. ." is a problem piece about a man and a WOMAN and the three "figures" that bothe
019. hem somehow. Unfortunately, the man and WOMAN were not made to appear very inter
020. the rain came more heavily, and men and WOMEN in light summer clothes began to d
021. en my experience to find as many men as WOMEN in church, and to hear almost ever
022. he church school, missions, men's work, WOMEN's work, youth program, social acti

Why ***should*** this learning method be effective ?

The case in principle (1)

Data-driven learning (DDL): using the tools and techniq
for teaching/learning/using a foreign/second language

- Greater effort
  better retention/deeper learning
- Strategy training
  learn to observe patterns in environment (exposure)
- Incidental learning
  "pick up" collocations while assembling word meanings
- Appeal to research-oriented clientele
  as opposed to ESL textbooks; PhD students, etc.

The case in principle (2)

Data-driven learning (DDL): using the tools and techniques of CL
for teaching/learning/using a foreign/second language

- A basis for CALL
  that goes beyond read/listen + MCQ with fixed answer
- Show not tell (constructivism)
  complex, dynamic, probabilistic, usage-based
- Autonomy
  own questions, life-long learning, multiple affordances
- Shapes up what students do anyway
  use Google for language research

But ***learning*** with corpora is not the topic of today's lesson

Rather ***teaching*** with corpora

Or, the design of courses conceived as corpora

And for this we need some DDL research tools

- Frequency lists based on both standard and course corpora
  - Frequency-list based tests
  - Frequency-list based text profiler
    - Lexical frequency profiler (LFP, Laufer & Nation)

- Some empirical findings to guide tool development
  - E.g. that reading comprehension depends on 95% know-word coverage
    - Laufer (1989 … 2020)

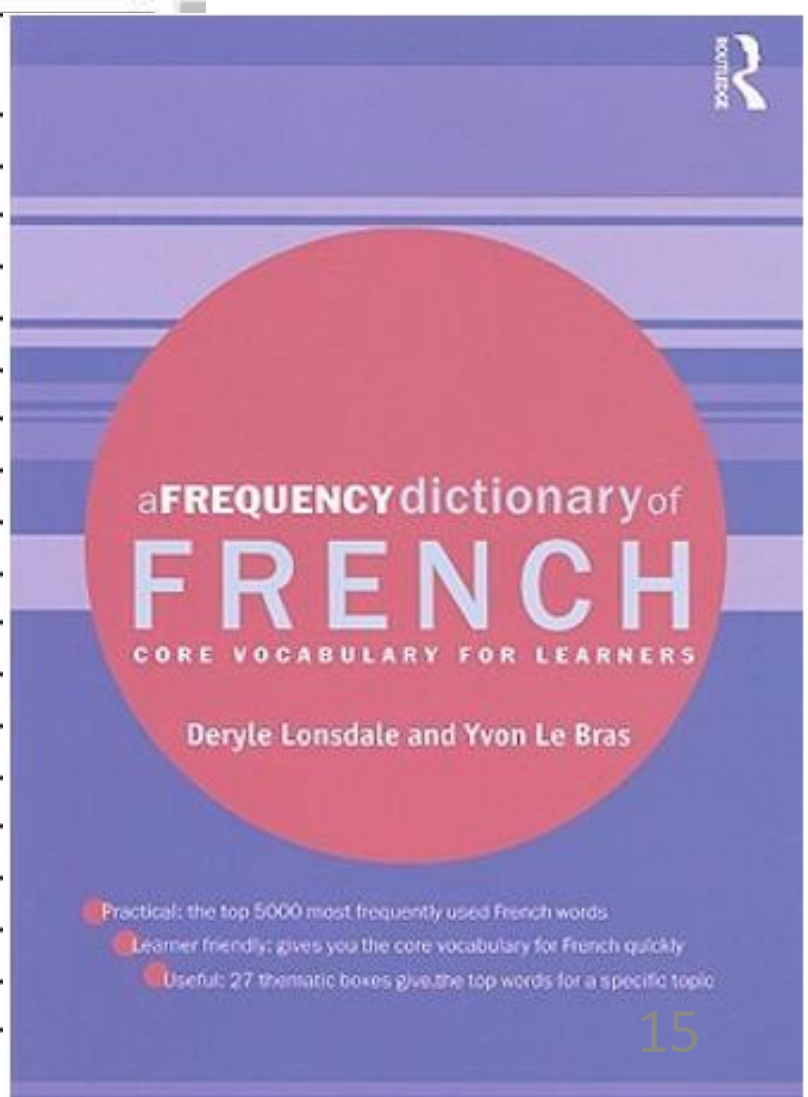
The standard French corpus used on Lextutor

Corpus

	# de mots (approx.)	Genre
Oral		
	175,000	Conversations
	3,750,000	Hansard canadien
	3,020,000	Transcrits d'interviews
	1,000,000	Débats parlementaires de l'UE
	855,000	Appels téléphoniques
	470,000	Dialogues de théâtre
	2,230,000	Sous-titres de films
TOTAL	11,500,000	
Écrit		
	3,000,000	Agences de presse
	2,015,000	Articles de quotidiens
	4,734,000	Œuvres littéraires(fiction, non-fiction)
	434,000	Magazines de vulgarisation technique
	1,317,000	Bulletins de presse, manuels d'emploi
TOTAL	11,500,000	
AU TOTAL	23,000,000	

Tableau 1 : Composition du corpus

List



FREQUENCY LIST

1. de 64006
2. la 31617
3. le 23365
4. et 20555
5. les 19532
6. des 19202
7. en 15185
8. du 14915
9. un 12781
10. a 12091
11. une 11168
12. est 10026
13. qui 9128
14. que 8796
15. dans 8124
16. pour 7600
17. par 7469
18. au 6991
19. pr 6428
20. sur 6224
21. pas 5382
22. ne 5303
23. plus 4987

Lemmatized and sorted by LEMMA frequency (not individual word freq.)

Aller	10,000
vais	3,000
vas	750
allé	500
etc.	
TOTAL	**14,250**

… and divided into groups of 1,000 lemmas

1k

https://lextutor.ca/vp/c

lextutor.ca/vp/com

à
âge
école
économie
économique
écouter
écrire
éducation
également
égard
élément
élection
élever
émission
énergie
époque
équipe
équipement
établir
établissement
état
étranger
étude
étudiant
étudier
événement
éviter
être
accès
accepter

2k

lextutor.ca/vp/c

âgé
âme
écarter
échange
échapper
échec
échelle
écho
éclater
écrivain
édition
égal
égalité
église
élève
élaborer
électeur
électoral
éliminer
élire
éloigner
émettre
énorme
épaule
époux
épreuve
éprouver
équilibre
équivalent
été
étape
étendre
étonner

3k

lextutor.ca/vp/comp/fr

écart
échanger
échouer
éclairer
écouler
écran
écraser
écrit
écriture
écu
éditeur
élargir
électricité
électrique
électronique
élite
émotion
énormément
épargne
épouser
équilibrer
équité
équitable
étage
éteindre
éternel
éthique
étoile
étonnant
étroit
évacuer
éventuellement
aîné

4k

lextutor.ca/vp/comp/fr

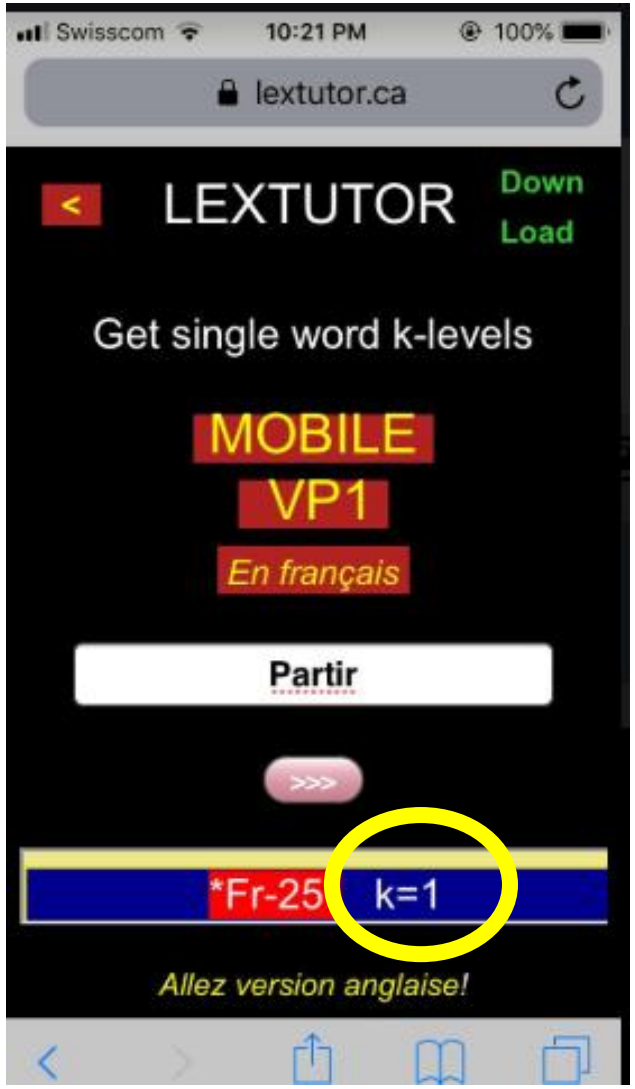
ère
échéance
échantillon
éclat
écologique
économiste
écoute
écrier
édifice
éditorial
égyptien
élémentaire
élaboration
élan
élargissement
élevé
élimination
élu
émaner
émerger
émetteur
émouvoir
énergétique
énoncer
épais
épargner
épidémie
épisode
épuiser
équipage
équiper
étaler
état major

5k etc … →25k

lextutor.ca/vp/comp/fr_5

échéant
échelon
éclaircir
économiquement
économiser
écrasant
écrouler
élégant
éloge
émeute
éminent
énergique
énumérer
épanouir
épouvantable
équation
équivaloir
ériger
étroitement
évasion
éventail
éventualité
abord
aboutissement
abstenir
accélération
acceptation
accessoire
accompli
accumulation
acharner
acheteur
acide

Such that every word has its ‘serial number’



Reflecting the order they are typically found in the environment

+ the importance of their being learned

(depending on the corpus used, obviously)

So, with corpus and numbered word-lists in hand… What can we do with this?

- Mainly, we have the means of providing learners with systematic vocabulary development
- Starting with systematic tests of vocabulary level in any language

With these lists and some tools anyone can make reliable level-based tests

- ONE
  Format Oui-Non

List framework IS **FR_5**

fr_5_heads(02).txt

vide
fabriquer
pétrole
revoir
impression
ignorer
acteur
entretien
avocat
côte

List framework IS **FR_5**

fr_5_heads(02).txt

poulard
sconber
radeurs
cageait
atérait

vide
fabriquer
pétrole
revoir
impression
ignorer
acteur
entretien
avocat
côte

USE RANDOM WORDS TO...

OPTION ONE
Do whatever you like

OPTION TWO
Make novel Yes-No Checklist Tests in four 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] Make an MS-Word document with three or four columns or use template

List framework IS **FR_5**

fr_5_heads(02).txt

acteur
atérait
avocat
cageait
côte
entretien
fabriquer
ignorer
impression
poulard
pétrole
radeurs
revoir
sconber
vide

USE RANDOM WORDS TO...

OPTION ONE
Do whatever you like

OPTION TWO
Make novel Yes-No Checklist Tests in four 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

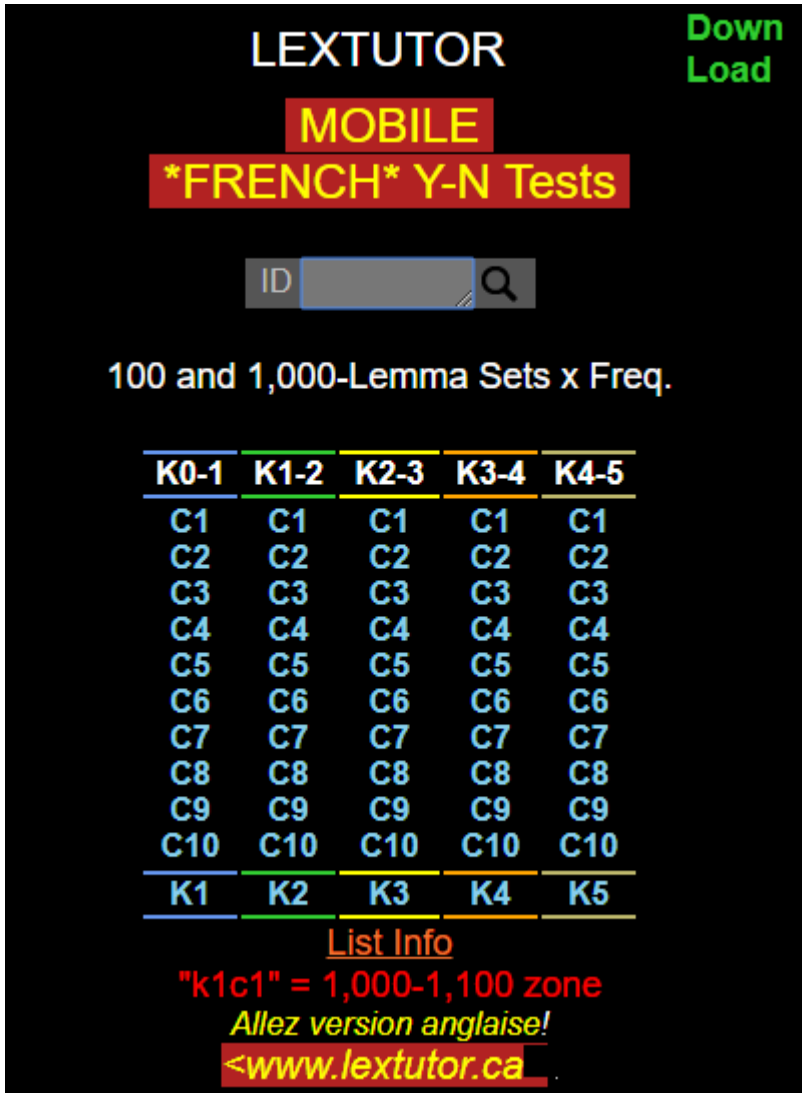
[4] Make an MS-Word document with three or four columns or use template

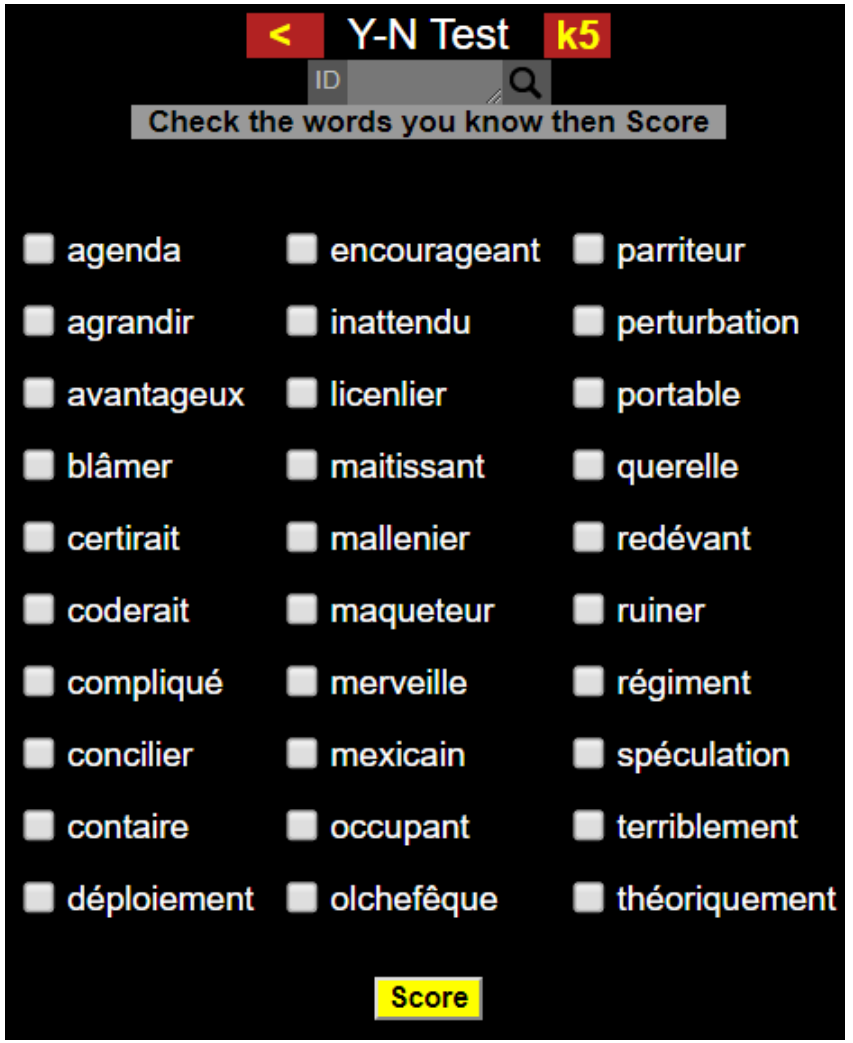
[5] Select, Copy, & Paste your words to one column

FRANCAIS 101 - 2k receptive vocab. Knowledge/diagnostic test
PROF BATISTA / student _____ 21 MAR 2019

Put a check mark beside each word you know, however slightly. Some of the words are not real words, do not check those.

2k	3k	4k	5k
acteur	aistrope	abbey	aerial
atérait	axis	abolish	bourgeois
avocat	battalion	acklon	cambule
cageait	berrow	blunt	cramp
côte	beta	bypass	dole
entretien	binoculars	clap	dowrick
fabriquer	charitable	connery	eccentric
ignorer	commemorate	cradle	eldred
impression	dazzle	delete	fatigue
poulard	dispense	dissent	gummer
pétrole	eldred	duffin	haque
radeurs	elusive	elastic	hinge
revoir	forthcoming	immerse	kiley
sconber	galpin	junction	lavish
vide	inhale	lannery	leaflet
	knight	limp	millilitre


LEXTUTOR
Down Load
MOBILE
FRENCH Y-N Tests
ID
100 and 1,000-Lemma Sets x Freq.
K0-1 K1-2 K2-3 K3-4 K4-5
C1 C1 C1 C1 C1
C2 C2 C2 C2 C2
C3 C3 C3 C3 C3
C4 C4 C4 C4 C4
C5 C5 C5 C5 C5
C6 C6 C6 C6 C6
C7 C7 C7 C7 C7
C8 C8 C8 C8 C8
C9 C9 C9 C9 C9
C10 C10 C10 C10 C10
K1 K2 K3 K4 K5
List Info
"k1c1" = 1,000-1,100 zone
Allez version anglaise!
<www.lextutor.ca


< Y-N Test k5
ID
Check the words you know then Score
agenda
agrandir
avantageux
blâmer
certirait
coderait
compliqué
concilier
contaire
déploiement
encourageant
inattendu
licenlier
maitissant
mallenier
maqueteur
merveille
mexicain
occupant
olchefêque
parriteur
perturbation
portable
querelle
redévant
ruiner
régiment
spéculation
terriblement
théoriquement
Score


< Y-N Test k0c1
ID
Check the words you know then Score
aller
bien
bria
colé
comprendre
dernier
dolium
dont
exacté
falloir
fois
guirum
jour
mais
monde
notre
nous
part
passin
polète
pour
pouvoir
rendre
rester
suite
tané
taître
tout
trouver
vignée
Score

[2]
Or something more standard, similar to Nation's VST

- Exemple

  (Le "TTV", RCLV, 2016)

Box 4. A noun cluster from 5K frequency section of the TTV

1. brouillard
2. coïncidence
3. farce _____ une histoire qui fait rire
4. instituteur _____ ce qui empêche de voir loin
5. pneu _____ un professionnel de l'éducation
6. soumission

A New Receptive Vocabulary Size Test for French

Roselene Batista and Marlise Horst

Abstract: Researchers have developed several tests of receptive vocabulary knowledge suitable for use with learners of English, but options are few for learners of French. This situation motivated the authors to create a new vocabulary size measure for French, the *Test de la taille du vocabulaire* (TTV). The measure is closely modelled on Nation's (1983) Vocabulary Levels Test (VLT) and follows the guidelines written by Schmitt, Schmitt, and Clapham (2001). Initially, a pilot version was trialled with 63 participants; then an improved version was administered to 175 participants at four proficiency levels. Results attest to the TTV's validity: mean scores across the four frequency sections decreased as the tested words became less frequent, and more proficient learner groups outperformed less proficient groups. The TTV in its current form is intended to be of practical use to teachers and learners, but it is also expected to evolve; ideas for future improvements are discussed.

Keywords: frequency, French L2 vocabulary, vocabulary size, assessment

Résumé : Des chercheurs ont développé plusieurs tests de vocabulaire réceptif pour les apprenants d'anglais, mais les options pour les apprenants de français ne sont pas nombreuses. Ce scénario a motivé les auteurs à créer un nouvel outil qui mesure la taille du vocabulaire en français, le Test de la taille du vocabulaire (TTV). Cet outil repose sur le modèle du *Vocabulary Levels Test*, conçu par Nation (1983), et suit les directives proposées par Schmitt, Schmitt et Clapham (2001). Premièrement, une version pilote a été testée auprès de 63 participants, ensuite une version améliorée a été complétée par 175 participants de quatre niveaux de compétence distincts. Les résultats confirment la validité du test: les moyennes obtenues par les participants à chacune des qua-

<

LEXTUTOR

Down Load

MOBILE LEVELS TESTS

Score-Recording Tests

Back Button to return to this page

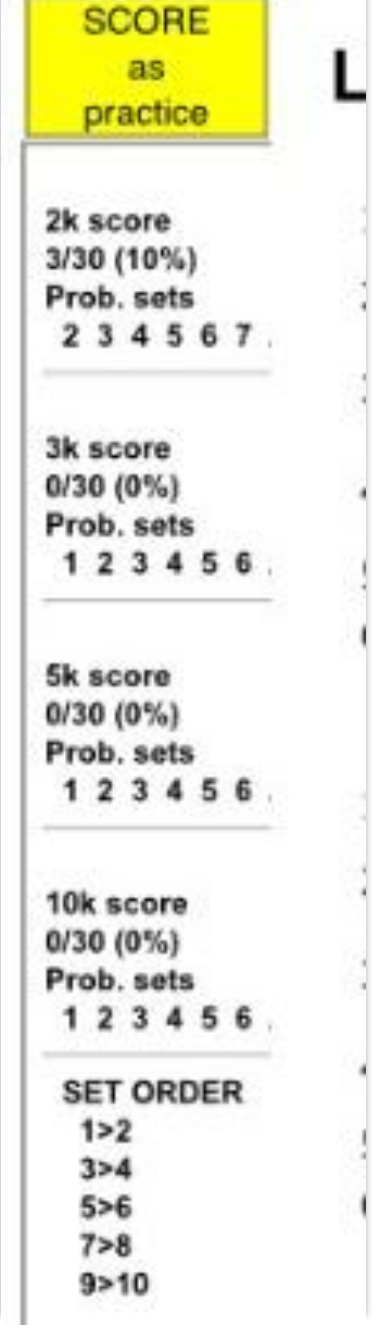
VLT
Vocab Levels Test

VLT v.2

VST
Vocab Size Test

TTV
Test de la taille du vocab

<www.lextutor.ca>



Typical findings from either test are that ~

- Intermediate learners know ≈ 1,000-1,500 lemmas in their L2
- High school graduates know ≈ 2,000-3,000 lemmas

- But is this good or bad?
  - Sufficient or insufficient?
    - To do what?

To answer this, we also need a similar *test* for texts

On the same metric as the tests for learners

- A.k.a. a text profile
  - How many words at each k-level does this text contain?
    - ‘Know’ replaced by ‘contain’

Text profiling by frequency level

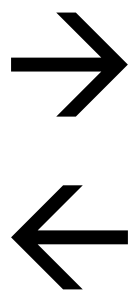
- Also known as
  LFP and Vocabprofil(e) – **“VP”**
- For example, this text (893 words) from a Montreal news story
  - – “L’abandon scolaire”

Sus à l'abandon scolaire! mais je ne pense pas qu'on
200 écoles second le mettre
partageront 125 m e ce que le
cinq ans is.» Est-il trop tôt
Marie-Andrée Cho sque
Le mardi 14 mai 2 pilote n'a pas
Pour s'attaquer au u sa première
qu'est l'abandon s ucun résultat
réduire de 10 % en disponible?
ministère de l'Édu nmes pas encore
(MEQ) propulse 12 ire avec précision
en cinq ans dans u impacts du
de 200 écoles secc mais nous en
Agir autrement, c' pour dire que
de l'opération lanc portante, et le
le ministre de l'Édu ble», notait
Sylvain Simard, ins on, sous-ministre
droit d'un projet-p seignement
même nom, qui n' condaire,
livré ses premiers le suivi de ces
L'intervention ne b serré. Et
toutefois pas de la n de fonds? «Ce
enveloppe qu'ont ement une
écoles secondaires gent», ajoute M.
l'expérimentation: chacune a Bisaillon. «Je souhaiterais que

Profile summary K, #tokens, cumul%		
K-1	816	85.2
K-2	66	92.1
K-3	36	95.9
K-4	11	97.0
K-5	6	97.6
K-6	1	97.7
K-7	2	97.9
K-8	2	98.1
K-10	4	98.5
K-11	2	98.7
K-12	2	98.9
K-13	2	99.1
K-14	1	99.2
K-17	1	99.3
K-18	1	99.4
K-20	1	99.5
K-24	1	99.6
OFF	3	≈100

Colour-coded k-levels

Profile summary K, #tokens, cumul%		
K-1	816	85.2
K-2	66	92.1
K-3	36	95.9
K-4	11	97.0
K-5	6	97.6
K-6	1	97.7
K-7	2	97.9
K-8	2	98.1
K-10	4	98.5
K-11	2	98.7
K-12	2	98.9
K-13	2	99.1
K-14	1	99.2
K-17	1	99.3
K-18	1	99.4
K-20	1	99.5
K-24	1	99.6
OFF	3	≈100

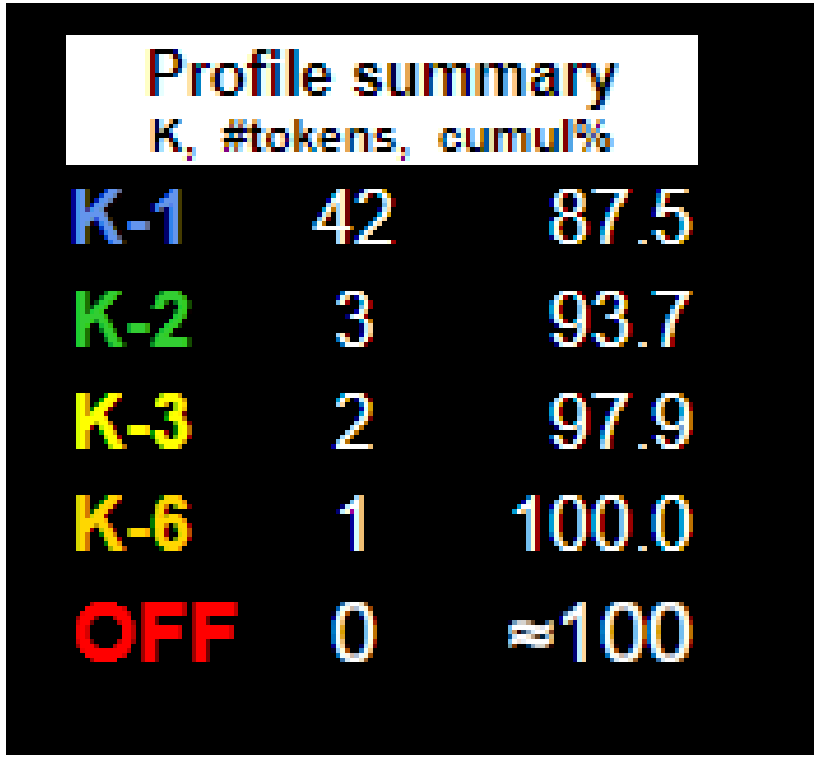


sus à le abandon scolaire chiffre écoles secondaires se partageront chiffre millions en cinq ans marie andrée chouinard
le mardi chiffre mai chiffre
pour se attaquer au fléau que est le abandon scolaire et le réduire de chiffre en dix ans le ministère de le éducation meq propulse chiffre millions en cinq ans dans un concentré de chiffre écoles secondaires agir autrement ce est le nom de le opération lancée hier par le ministre de le éducation sylvain simard inspirée tout droit de un projet pilote du même nom qui ne a pas encore livré ses premiers résultats le intervention ne bénéficiera toutefois pas de la généreuse enveloppe que ont reçue les six écoles secondaires ciblées par le expérimentation chacune a reçu chiffre chiffre million pour des mesures échelonnées sur trois ans tandis que les chiffre écoles ciblées par québec se disputeront chiffre millions par an distribués selon le bon vouloir de leurs commissions scolaires respectives pour permettre cette annonce la directrice de le école secondaire Édouard montpetit lucie lalande ouvrait sa porte au ministre sylvain simard hier depuis presque un an son école est sous la lorgnette du meq parce que elle fait partie du projet agir autrement coups de centaines de milliers de dollars les résultats sont déjà palpables affirme elle

Compare previous profile to this one ~

- Hier Daisy s'était levée tôt ce matin de printemps. Elle travaillait sur une affaire dans la ville voisine. Elle arriva à son bureau à huit heures avec à la main un sac en papier contenant des petits pains. Elle mourrait d'envie d'une tasse de café.

Note – Demo only; short texts = unreliable VP



hier daisy se était levée tôt ce matin de printemps elle travaillait sur une affaire dans la ville voisine elle arriva à son bureau à huit heures avec à la main un sac en papier contenant des petits pains elle mourrait de envie de une tasse de café

Profiles are somewhat different ! ……→

But without these instruments, would we have noticed how much?

To read Daisy with 95% word knowledge, 2000 word families should be know

To read Abandon Scolaire with 95% word knowledge, 3000 word families should be know

Daisy le matin

Profile summary K, #tokens, cumul%		
K-1	42	[illegible]
K-2	3	93.7
K-3	2	[illegible]
K-6	1	100.0
OFF	0	≈100

L'abandon scolaire

Profile summary K, #tokens, cumul%		
K-1	816	85.2
K-2	66	[illegible]
K-3	36	95.9
K-4	11	[illegible]
K-5	6	97.6
K-6	1	97.7
K-7	2	97.9
K-8	2	98.1
K-10	4	98.5
K-11	2	98.7
K-12	2	98.9
K-13	2	99.1
K-14	1	99.2
K-17	1	99.3
K-18	1	99.4
K-20	1	99.5
K-24	1	99.6
OFF	3	≈100

These tests and measures let us compare/contrast profiles of learners & profiles of texts

And determine the quality of the match

- Empirical research has shown
  - Ss need to know 95% of the words on a page to read with basic comprehension
  - **> 95%** of words known = fluent reading for comprehension
  - **85-95%** = intensive reading
    - Many look-ups, comprehension gaps, unresolved passages
  - **< 85%** = unuseful experience
    - (zone where much of school reading lies, in L1 but especially L2)

With these tools we can begin Data-Driven Course/Curriculum design (DDCD)

Start with the raw materials of a course

- E.g., A “reading course” with a set of texts drawn as “interesting” from Internet news

Step 1 – get all course materials into digital form as a corpus

- **In two formats:**
  1. Separate chapters, texts or units (ideally ‘zipped’)
  2. One large single file

- What will this enable us to do?

What would course-as-corpus let us do?

Find out

1. Whether our course level is anywhere near our learners' level
2. If not, what would bring it nearer?
    1. Learners up, texts down?
3. What if anything can legitimately be tested from this course?

* Level in this case refers to lexis but could be other aspects

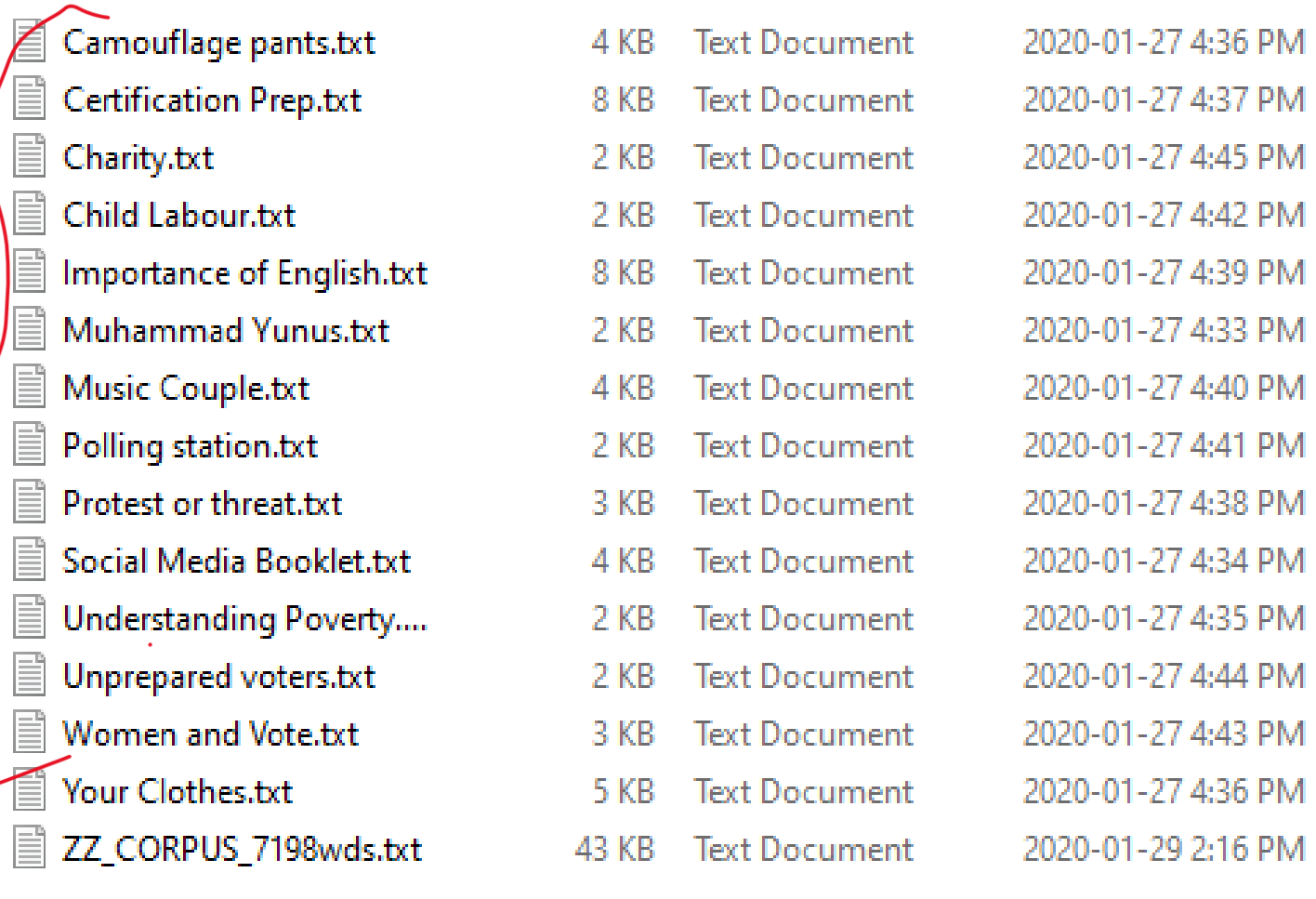
Here is the course-corpus described in one course I re-designed

ESL Adult Reading, 14 units,

- >7000 words,
- in-house course,
- intermediate adults
- high stakes

→ Simple 'flat' corpus

- 2 formats
  - Separate files
  - A combined file



Camouflage pants.txt	4 KB	Text Document	2020-01-27 4:36 PM
Certification Prep.txt	8 KB	Text Document	2020-01-27 4:37 PM
Charity.txt	2 KB	Text Document	2020-01-27 4:45 PM
Child Labour.txt	2 KB	Text Document	2020-01-27 4:42 PM
Importance of English.txt	8 KB	Text Document	2020-01-27 4:39 PM
Muhammad Yunus.txt	2 KB	Text Document	2020-01-27 4:33 PM
Music Couple.txt	4 KB	Text Document	2020-01-27 4:40 PM
Polling station.txt	2 KB	Text Document	2020-01-27 4:41 PM
Protest or threat.txt	3 KB	Text Document	2020-01-27 4:38 PM
Social Media Booklet.txt	4 KB	Text Document	2020-01-27 4:34 PM
Understanding Poverty....	2 KB	Text Document	2020-01-27 4:35 PM
Unprepared voters.txt	2 KB	Text Document	2020-01-27 4:44 PM
Women and Vote.txt	3 KB	Text Document	2020-01-27 4:43 PM
Your Clothes.txt	5 KB	Text Document	2020-01-27 4:36 PM
ZZ_CORPUS_7198wds.txt	43 KB	Text Document	2020-01-29 2:16 PM

Now a new course

For FL2 learners in Sweden

Le Petit Nicolas

(Sempé and Goscinny, c. 1959)

- 18 children's stories
- Similar lengths
- No passé simple/historique
- 28,000 words (≈ 1500 per story)

= a plausible French reading course for 9-year-olds

Name	Date modified	Size	Type
01_souvenir_cherir.txt	2022-09-01 5:29 PM	8 KB	TXT File
02_cow-boys.txt	2022-09-01 5:30 PM	8 KB	TXT File
03_le_bouillon.txt	2022-09-01 5:30 PM	8 KB	TXT File
04_le_football.txt	2022-09-01 5:31 PM	8 KB	TXT File
05_l'inspecteur.txt	2022-09-01 5:31 PM	8 KB	TXT File
06_rex.txt	2022-09-01 5:32 PM	8 KB	TXT File
07_djodjo.txt	2022-09-01 5:32 PM	8 KB	TXT File
08_chouette_bouquet.txt	2022-09-01 5:32 PM	8 KB	TXT File
09_les_carnets.txt	2022-09-01 5:32 PM	8 KB	TXT File
10_louisette.txt	2022-09-01 5:33 PM	14 KB	TXT File
11_je_fume.txt	2022-09-02 11:39 AM	8 KB	TXT File
12_le_petit_poucet.txt	2022-09-01 5:34 PM	8 KB	TXT File
13_le_velo.txt	2022-09-02 11:39 AM	8 KB	TXT File
14_suis_malade.txt	2022-09-01 5:34 PM	8 KB	TXT File
15_bien_rigole.txt	2022-09-01 5:34 PM	7 KB	TXT File
16_frequent_agnan.txt	2022-09-01 5:34 PM	9 KB	TXT File
17_pas_le_soleil.txt	2022-09-02 10:37 AM	8 KB	TXT File
18_quitte_la_maison.txt	2022-09-02 12:01 PM	8 KB	TXT File
Pet_Nic_18.zip	2022-09-02 12:02 PM	61 KB	ZIP File
PN_original_30k.txt	2022-09-05 9:53 AM	136 KB	TXT File

Petit Nicolas (tout) - lexical frequency profile

Home > VocabProfilers > VP-Compleat Input > Output ('Back' preserves inputs) FRAMEWORK IS fr_5

Freq. Level	Flemmas (%)	Types (%)	Tokens (%)	Cumul. token (%)
K-1 :	2451 (72.4)	2651 (65.98)	23531 (86.3)	86.3
K-2 :	248 (7.3)	454 (11.30)	1420 (5.2)	91.5
K-3 :	157 (4.6)	256 (6.37)	660 (2.4)	93.9
K-4 :	110 (3.2)	152 (3.78)	354 (1.3)	95.2
Coverage 95 [?]				
K-5 :	83 (2.5)	111 (2.76)	263 (1.0)	96.2
K-6 :	92 (2.7)	122 (3.04)	274 (1.0)	97.2
K-7 :	36 (1.1)	52 (1.29)	150 (0.5)	97.7
K-8 :	33 (1.0)	39 (0.97)	93 (0.3)	98.0
Coverage 98				
K-9 :	26 (0.8)	29 (0.72)	72 (0.3)	98.3

Result

95% lexical coverage is reached only when 4,000 lemmas are known

Knowing 1,000=86.3% coverage

Knowing 2,000=91.5% coverage

How many French words do Swedish 9-yr-old learners know? Probably < 1,000

So these texts probably correspond to learners' interests but…

- Will be challenging for some/most of them
  - (This could only be established through testing)
  - So here is a prima facie test case for Data-Driven Course Deign
- How can DDL help make this text accessible?

Simplest, DDL can just give us the specific 2k-3k words to teach

- Whether Pre-Teach or emphasize and invite discussion when encountered
- Teacher would choose based on frequency, cognate-ness
- For the whole book or specific story/stories
  - (Next slide is for whole book)

French v.5 -K2 Flemmas: [flems 248 : types 454 : tokens 1420]

VP-negative: fr_5-2

absolument_[3] accompagner_[3] acteur_[1] agent_[5] anglais_[1] animal_[3] apercevoir_[4] appareil_[2] approcher_[7] arbre_[7] arrière_[1] assister_[1] attacher_[3] attaquer_[2] avion_[15] baisser_[1] balle_[15] ba
ca
cie
co
co
de
du
es
fin
ga
inc
joli
ma
mé
pa
pla
pro
rec
rep
rou
se
su
traduire_[
voisin_[4] voler_[2] voyage_[1] zero_[1] ecarter_[1] echapper_[1] eclater_[1] eloigner_[1] etonner_[12]

French v.5 -K3 Flemmas: [flems 157 : types 256 : tokens 660]

VP-negative: fr_5-3

agiter_[7] allumer_[1] amuser_[32] animer_[1] anniversaire_[3] arracher_[1] arranger_[6] bouteille_[8] boîte_[6] brillant_[2] briller_[4] bête_[7] cadeau_[5] calmer_[1] camarade_[13] capitaine_[7] chaise_[3]
chape
corrig
dent_
douce
embr
feuille
gloire
inspe
malhe
morce
oncle
protes
repas
sale_
tranqu
écras

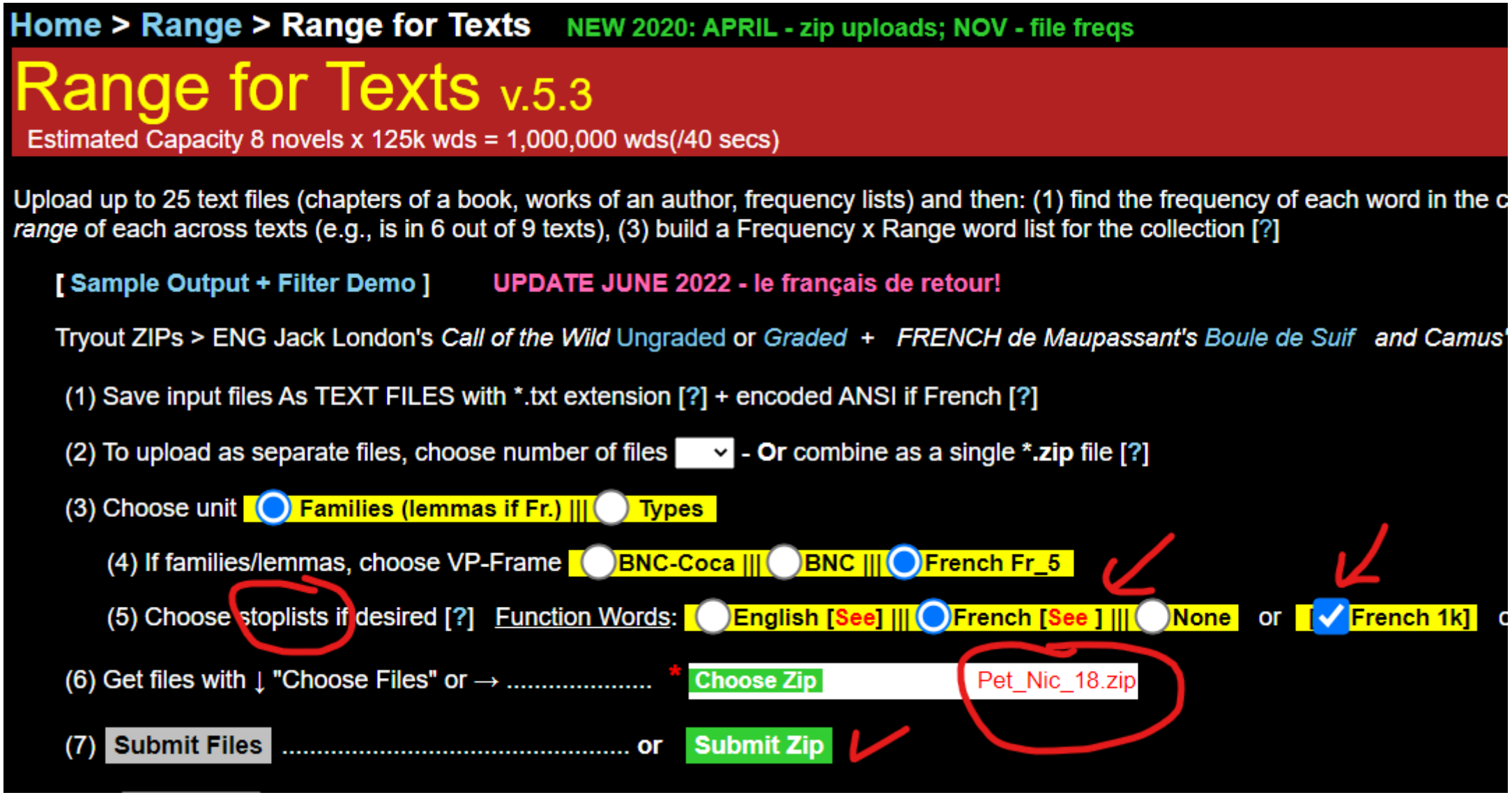
French v.5 -K4 Flemmas: [flems 110 : types 152 : tokens 354]

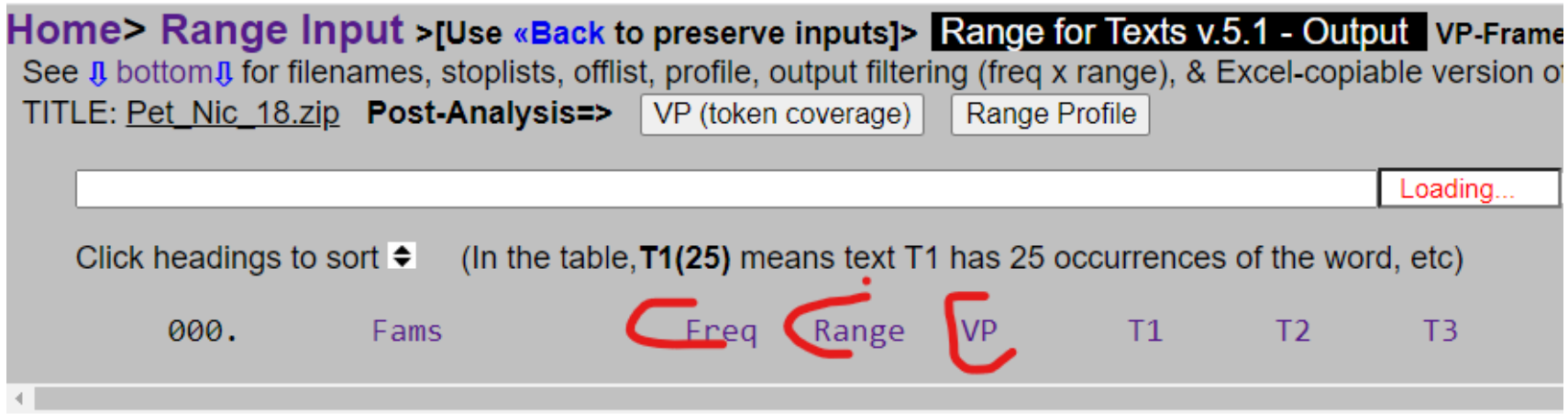
VP-negative: fr_5-4

accueil_[1] affreux_[1] alcool_[2] amateur_[1] applaudir_[1] attraper_[1] auto_[21] avaler_[1] avertissement_[1] bain_[6] ballon_[8] banc_[13] blague_[4] bouton_[2] bravo_[2] brutal_[3] bâton_[1] cahier_[1] canon_[1] caresser_[2] chant_[1] charmant_[1] chasseur_[3] chemise_[10] chimie_[1] château_[1] chèque_[1] coincer_[1] col_[1] combiner_[2] consoler_[2] corde_[2] correctement_[1] costume_[5] cou_[1] crâne_[1] dicter_[1] déboucher_[1] débrouiller_[2] déchirer_[3] dégât_[1] délivrer_[1] dépêcher_[1] désespérer_[2] désordre_[1] détacher_[1] essuyer_[6] fantôme_[2] ficher_[1] fréquent_[2] fréquenter_[2] fusil_[1] fêter_[1] gendarme_[3] gorge_[1] goûter_[9] guérir_[2] habiller_[14] herbe_[3] hurler_[2] jouet_[12] joyeux_[1] lampe_[2] laver_[3] liquide_[2] lunette_[24] maigre_[1] mairie_[1] manche_[2] miroir_[1] mordre_[8] méfier_[4] métal_[1] nerveux_[7] nettoyer_[1] os_[1] pantalon_[4] patte_[2] peindre_[1] pendre_[2] photographe_[15] photographie_[1] piquer_[1] planche_[3] plume_[3] poing_[19] poitrine_[2] pot_[2] proclamer_[1] punir_[6] pâle_[1] queue_[4] rage_[1] rattraper_[4] rembourser_[1] ressort_[1] sourd_[2] stupide_[1] sympathie_[1] tache_[3] tante_[5] taper_[15] thé_[3] transparent_[1] vitre_[1] voleur_[5] élan_[1] élémentaire_[1]

A way to find the most important words Ss must know across a set of texts is with the DDL tool RANGE

- Gives the frequency of each word or lemma in a set of texts
  - Plus the number of texts each features in
- Can boil the lexis down to the frequent PLUS recurring items


Home > Range > Range for Texts
NEW 2020: APRIL - zip uploads; NOV - file freqs
Range for Texts v.5.3
Estimated Capacity 8 novels x 125k wds = 1,000,000 wds(/40 secs)
Upload up to 25 text files (chapters of a book, works of an author, frequency lists) and then: (1) find the frequency of each word in the c
range of each across texts (e.g., is in 6 out of 9 texts), (3) build a Frequency x Range word list for the collection [?]
[Sample Output + Filter Demo]
UPDATE JUNE 2022 - le français de retour!
Tryout ZIPs > ENG Jack London's Call of the Wild Ungraded or Graded + FRENCH de Maupassant's Boule de Suif and Camus'
(1) Save input files As TEXT FILES with *.txt extension [?] + encoded ANSI if French [?]
(2) To upload as separate files, choose number of files - Or combine as a single *.zip file [?]
(3) Choose unit Families (lemmas if Fr.) ||| Types
(4) If families/lemmas, choose VP-Frame BNC-Coca ||| BNC ||| French Fr_5
(5) Choose stoplists if desired [?] Function Words: English [See] ||| French [See] ||| None or French 1k]
(6) Get files with ↓ "Choose Files" or → ..... * Choose Zip Pet_Nic_18.zip
(7) Submit Files ..... or Submit Zip



032.	papa	172	16	2	T1(2)	T2(16)	
036.	maman	154	13	2		T2(5)	
041.	maître	118	12	1	T1(27)		T3(8)
061.	crier	80	16	2	T1(6)	T2(4)	T3(10)
075.	répondre	50	17	1	T1(3)	T2(2)	T3(4)
082.	école	46	15	1	T1(2)		T3(2)
084.	pleurer	44	17	2	T1(2)	T2(1)	T3(6)
088.	fleur	42	2	2			
092.	vélo	40	3	5			
093.	bouillon	39	3	11			T3(32)
095.	rex	39	1	10			
100.	chouette	36	17	7		T2(5)	T3(1)
103.	copain	35	18	3	T1(3)	T2(2)	T3(1)
110.	rigoler	33	13	5	T1(1)	T2(4)	T3(1)
111.	drôlement	32	15	7	T1(3)	T2(1)	T3(2)
115.	inspecteur	30	2	3			
116.	tête	30	12	1	T1(3)	T2(5)	
118.	amuser	29	13	3		T2(4)	
122.	courir	28	13	2	T1(1)	T2(2)	
129.	idée	26	13	1	T1(1)	T2(2)	T3(1)

Offlist

75 OFF-LIST TYPES WITH NO FAMILY HEADER
ac agnan alceste aon arcachon australie blédurt bordenave bouil boxing bril
cyrille dgeorges dinguedingue dissipez djoachim djodjo dubon dég elvire eude

OUTPUT FILTER

Extract data by Frequency >= 7 and Range >= 10
Format: [As above Words only] Go »
Build Frequency- and Range-based word lists; check corpus

These are the key words across these stories –

List can be expanded or shortened by changing the parameters

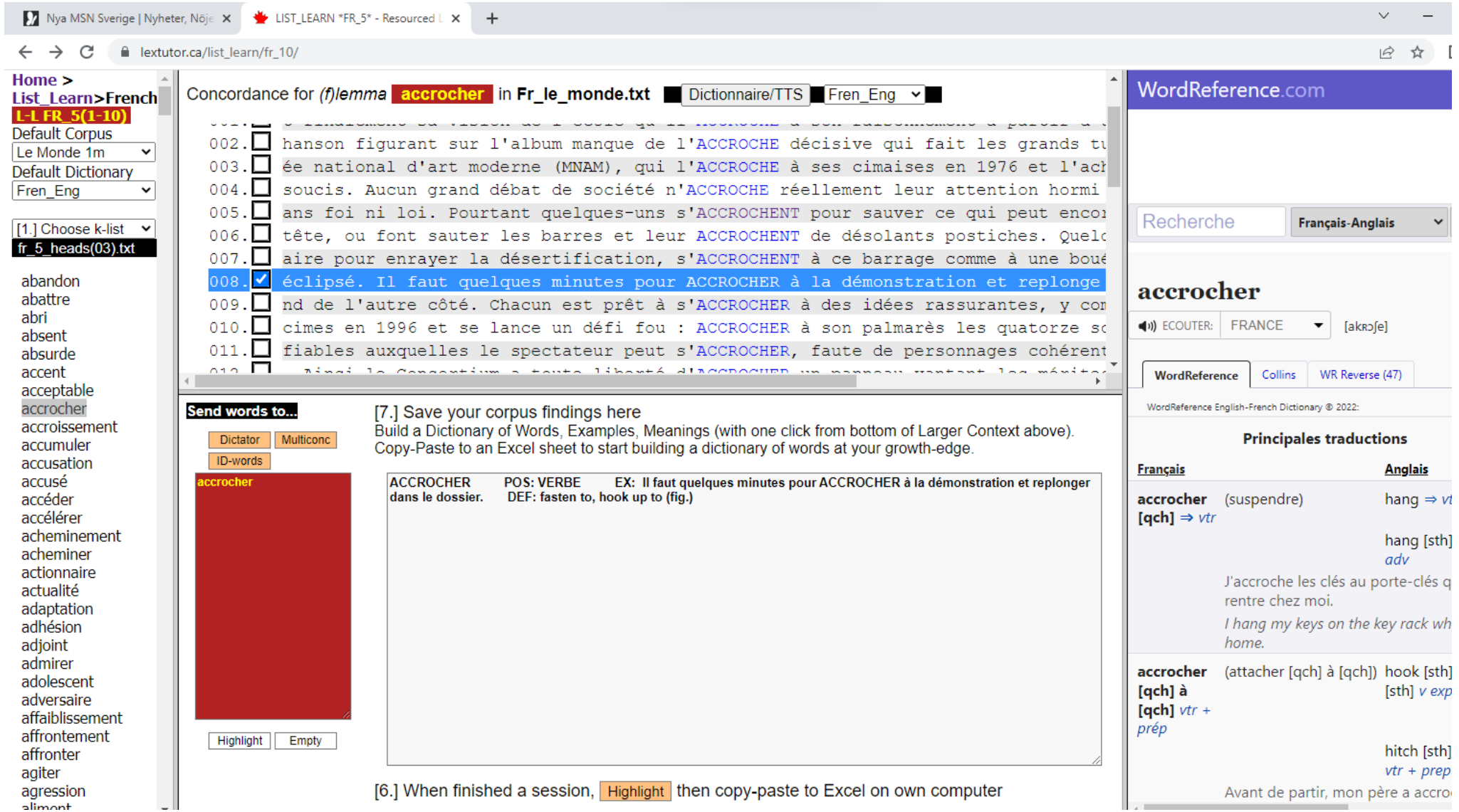
RANGE EXTRACT || FREQUENCY => 7 and RANGE => 10 || **24 i**

	FAM	FREQ	RANGE	V
035.	maman	154	13	2
040.	maître	118	12	1
060.	crier	80	16	2
074.	répondre	50	17	1
081.	école	46	15	1
083.	pleurer	44	17	2
099.	chouette	36	17	7
102.	copain	35	18	3
109.	rigoler	33	13	5
110.	drôlement	32	15	7
115.	tête	30	12	1
117.	amuser	29	13	3
121.	courir	28	13	2
128.	idée	26	13	1
135.	chocolat	24	12	5
151.	arrêter	21	11	1
164.	rouge	19	10	2
165.	terrible	19	11	2
168.	embêter	18	11	6
175.	gentil	17	10	3
187.	sûr	16	13	1
196.	côté	15	10	1
197.	drôle	15	10	2
204.	taper	15	11	4

Once the key words' identities are known…

- Flash Cards are back in style with shiny new credentials

Personally I would have Ss go through lists and check out *all the words* at the next level that they don't already know



More student-centered is to set up a Group Lex database & Ss choose their own words to work on

- And share with others
  - Including by mobile phone
- Leading to a complete lexicon of the whole book with quizzes

#	Qz	NOUVEAU MOT	VP	EXEMPLE	CLASSE DE MOT	DEFINITION	GROUPE	ETUDIA...	DATE/H... SOUMIS...
1	☐	apercevoir	2	« La mi-temps de quoi? a demandé Alceste. Je viens de m'apercevoir que nous n'avons pas de ballon, je l'ai oublié à la maison! »	Verbe	skymta, glimt	Pet_Nic_4	tom_ad...	2022.09.06 8.30
2	☐	musique	1	J'écoute de la musique.	Nom	L'art de la combinaison de sons	Arts	LK1234	2014.12.16 9:31
3	☐	couple	2	Il vit en couple avec Marie.	Nom	deux personnes liées par des sentiments	Economics	LK1234	2014.12.16 11:38
4	☐	couple	2	Seulement un couple sur trois désormais se marie à l'église.	Nom	deux personnes liées par un sentiment	Economics	123French	2014.12.16 11:32
5	☐	pris	1	Elle a pris sa douche antérieurement	Verbe	took (prendre)	Arts	00christi...	2014.09.08 9.35
6	☐	pomme	4	La pomme de pin n'est pas un vrai pomme	Nom	Apple	Arts	tom_ad...	2014.09.08 9.33
7	☐	montre	1	Que dit ta montre?	Nom	(wrist) watch - so 'What time is it?'	Arts	tom_ad...	2014.09.08 9.28
8	☐	nombreux	1	On est demandé de venir 'nombreux' à leur évenement	Nom	in large numbers	Arts	tom_ad...	2014.09.08 9.24
9	☐	salut	2	Dis Salut	Nom	Hello	Arts	tom_ad...	2014.09.08 22:34
10	☐	sourire	1	Donne-moi ta sourire	Nom	A smile	Arts	tom_ad...	2014.09.08 22.23

Accueil
>Group Lex

v.8.3

Beta French interface

OS=Win
BROWSER=Chr

V.Mobile pad/phone

v. English

COMPUTER
Voir les 53 mots

Ajout. mot

Aj. étudiant

Modif. mot

Besoin de Gp Lx?

X-tract-liste

Recherche?

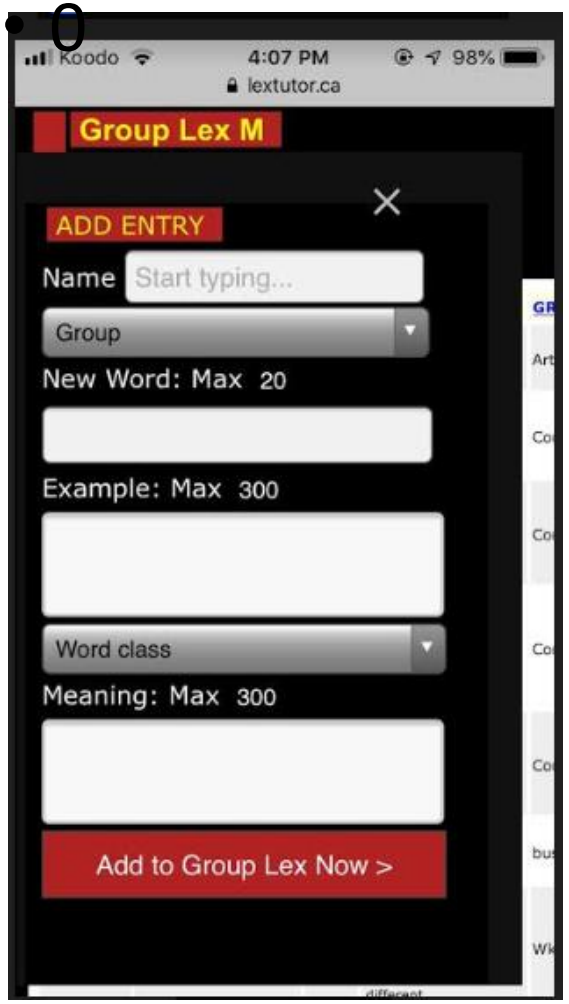
Prononcer la sélection

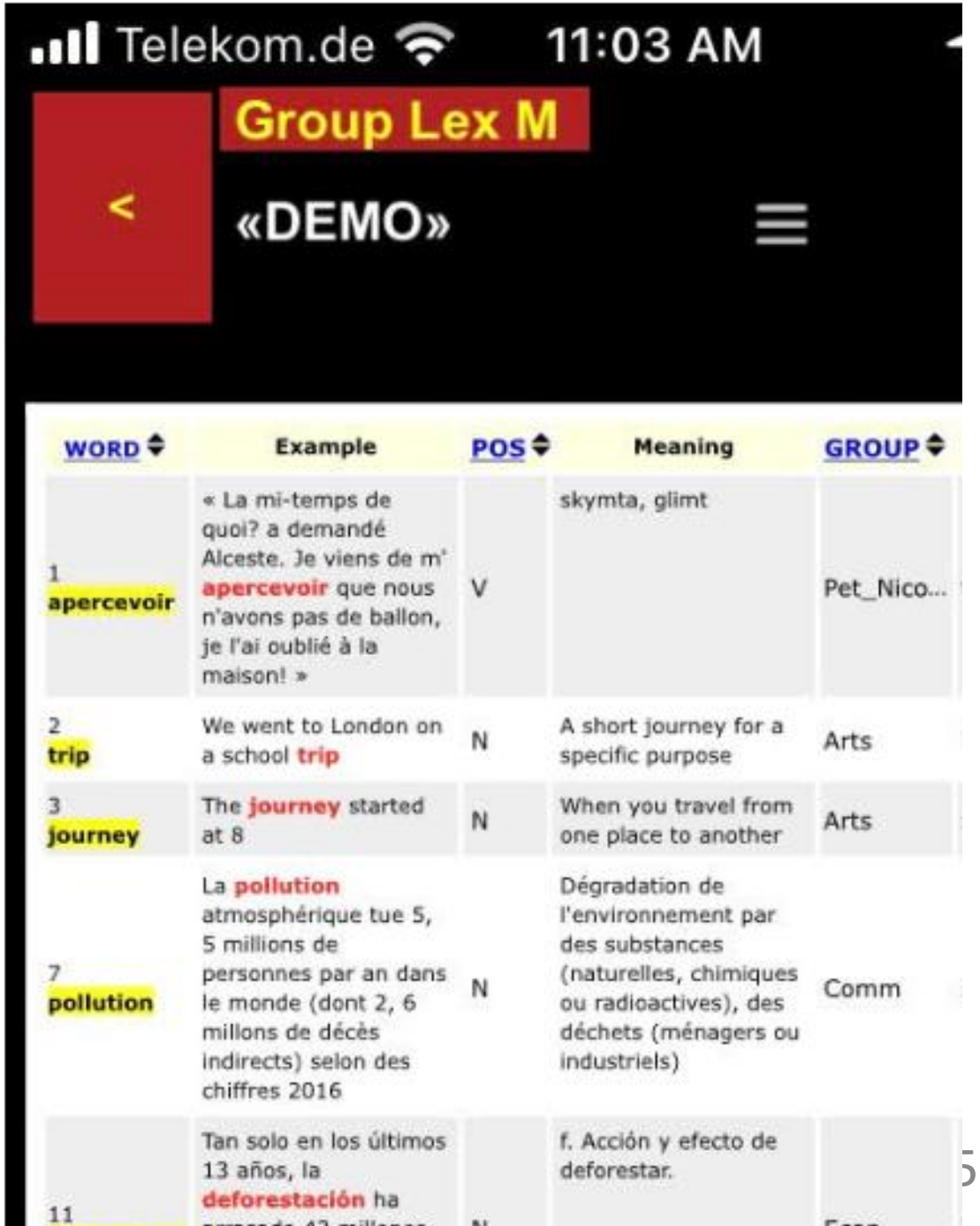
Ceci sera votre entrée :

#	Qz	NOUVEAU MOT	VP	EXEMPLE	CLASSE DE MOT	DÉFINITION	GROUPE	ÉTUDIANT	DATE/HEURE
00	☐	**apercevoir**		« La mi-temps de quoi? a demandé Alceste. Je viens de m' **apercevoir** que nous n'avons pas de ballon, je l'ai oublié à la maison! »	V	skymta, glimt	Pet_Nicolas_4	tom_admin	2022.09.06 04.23

< Retour Sécurité : Taper les chiffres pour trois, deux, un Valider >

xxx





WORD	Example	POS	Meaning	GROUP
1 apercevoir	« La mi-temps de quoi? a demandé Alceste. Je viens de m' apercevoir que nous n'avons pas de ballon, je l'ai oublié à la maison! »	V	skymta, glimt	Pet_Nico...
2 trip	We went to London on a school trip	N	A short journey for a specific purpose	Arts
3 journey	The journey started at 8	N	When you travel from one place to another	Arts
7 pollution	La pollution atmosphérique tue 5, 5 millions de personnes par an dans le monde (dont 2, 6 millons de décès indirects) selon des chiffres 2016	N	Dégradation de l'environnement par des substances (naturelles, chimiques ou radioactives), des déchets (ménagers ou industriels)	Comm
11	Tan solo en los últimos 13 años, la deforestación ha		f. Acción y efecto de deforestar.	

QUIZ 1 - contexts from WHOLE GP LEX 08 Feb 20, 16:49 Quiz 2 -*new contexts - after 100% c*

Check Questions: 10 Correct: 0 Tries: 0 Percent: 0 History >>

Practice

In-Class

	NEW WORD	EXAMPLE	WORD CLASS	DEF
1		Rousseau's fundamental idea was the _____ of nature and culture.	N	A division or contrast l are or are represente
2		The word 'flight' has two different _____s: a plane journey, and the act of running av		
3		He was an _____ person		
4		House prices have been _____ for months.		
5		He This was an _____ pers		
6		Gp Lex also works on _____ though with reduced functionality		
7		Aerial _____ of the enemy p were ready to attack.		ey
8		The _____ sends a beacon indicating how urgent it is to track the s		nd

ambitious
ambitious
asteroid
dichotomy
meaning
mobile
reconnaissance
spacecraft
stationary

	NEW WORD	EXAMPLE	WORD CLASS	D
1		He This was an _____ person project.	ADJ	requiring a great de
2		He was an _____ person	N	eagerly desirous of wealth, a specific g
3		Aerial _____ of the enemy position showed they were ready to attack.	ADJ	Information gatheri
4		The _____ sends a beacon signal to the ground indicating how urgent it is to track the spacecraft for telemetry.	N	a vehicle used for t
5		Gp Lex also works on _____ though with reduced functionality	N	Easily moveable
6		cientists are hoping an _____ will clue them in about early life on earth.	N	one of many large r
7		House prices have been _____ for months.	ADJ	not moving, or not c
8		Rousseau's fundamental idea was the _____ of nature and culture.	N	A division or contra or are represented different.
9		The word 'flight' has two different _____s: a plane journey, and the act of running away.	N	The ____ of someth represents
10		We are collecting English one after another	N	字

MORE PRINCIPLED HOWEVER – is to use DDL tools to sequence the 18 stories from basic to more complex vocab

- To facilitate the conditions whereby learning by inference can take place
  - IE, > 90% knowledge of words in the context
    - Or as many as possible

These texts…

- Written for NS French kids, these are unlikely to be sequenced at all
- To get a sequence, we return to VP
- And analyse the texts individually, recording each of the 18 profiles
- Then with Excel we sort by the percentage of 1k items
  - *Should* give us a sequence of texts from most-basic to least-basic lexis
  - At least, this is what happened in English
  - Same in French?

> VP-Compleat Input > Output ('Back' preserves inputs) FRAMEWORK IS fr_5 (Propers: To 1k; Compounds: Separated)

K-1 :	305 (73.3)	354 (72.69)	1309 (86.2)	86.2
K-2 :	38 (9.1)	48 (9.86)	75 (4.9)	91.1
K-3 :	18 (4.3)	21 (4.31)	28 (1.8)	92.9
K-4 :	21 (5.0)	23 (4.72)	45 (3.0)	95.9
Coverage 95				
K-5 :	9 (2.2)	10 (2.05)	13 (0.9)	96.8
K-6 :	4 (1.0)	4 (0.82)	4 (0.3)	97.1
K-7 :	3 (0.7)	3 (0.62)	5 (0.3)	97.4
K-8 :	2 (0.5)	2 (0.41)	2 (0.1)	97.5
K-9 :	3 (0.7)	3 (0.62)	5 (0.3)	97.8
Coverage 98				
K-10 :	1 (0.2)	1 (0.21)	3 (0.2)	98.0
K-11 :	4 (1.0)	4 (0.82)	5 (0.3)	98.3
K-12 :	2 (0.5)	2 (0.41)	5 (0.3)	98.6
K-13 :	1 (0.2)	1 (0.21)	1 (0.1)	98.7
K-14 :	2 (0.5)	2 (0.41)	8 (0.5)	99.2
K-15 :				
K-16 :				
K-17 :				
K-18 :				
K-19 :				
K-20 :	1 (0.2)	1 (0.21)	1 (0.1)	99.3
K-21 :				
K-22 :	1 (0.2)	1 (0.21)	1 (0.1)	99.4
K-23 :				
K-24 :				
K-25 :	1 (0.2)	1 (0.21)	1 (0.1)	99.5
Off-List:	??	6 (1.23)	7 (0.46)	99.96
Total (unrounded)	416+?	487 (100)	1518 (100)	≈100.00

Words in text (tokens):	1518
Different words (types):	487
Type-token ratio (TTR):	0.32
Tokens per type:	3.12
Pertaining to onlist only	
Tokens:	1511
Types:	481
Flemmas:	416
Tokens per Flemma :	3.63
Flemma/token ratio (FTR) :	0.28
Types per Flemma :	1.16
Singletons ratio Fams(n=1)[121] / total [416]	0.29

aste to Excel heet to record **comparative coverages**. To compare several texts, track learner writing over time, sequence texts by lexis (TIP, sort spreadsheet by K-1% to get more basic lexis first), etc. Note th FL7, etc.)

K-1	86.2	K-2	91.1	K-3	92.9	K-4	95.9	K-5	96.8	K-6	97.1	K-7	97.4	K-8	97.5	K-9	97.8	K-10	98.0

So 18 profiles assembled
ready for sorting by Column ‘C’
- Order texts by the proportion of 1k

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	**LE PETIT NICOLAS**																				
2	**AS PUBLISHED**																				
3	1 SOUVENIR CHERIR	K-1	86.2		K-2	91.1		K-3	92.9		K-4	**95.9**		K-5	96.8		K-6	97.1		K-7	97.4
4	2 COWBOYS	K-1	83.0		K-2	89.6		K-3	91.1		K-4	93.0		K-5	94.3		K-6	**95.0**		K-7	96.4
5	3 BOUILLON	K-1	88.2		K-2	92.4		K-3	94.2		K-4	94.9		K-5	**95.7**		K-6	96.5		K-7	96.8
6	4 FOOTBALL	K-1	86.4		K-2	92.6		K-3	**95.4**		K-4	97.1		K-5	97.9		K-6	**99.2**		K-7	99.3
7	5 INSPECTEUR	K-1	87.1		K-2	91.1		K-3	94.4		K-4	**96.0**		K-5	96.5		K-6	97.8		K-7	**98.4**
8	6 REX	K-1	82.6		K-2	90.6		K-3	92.6		K-4	93.7		K-5	93.9		K-6	94.6		K-7	**95.3**
9	7 DJODJO	K-1	87.0		K-2	91.0		K-3	94.2		K-4	**95.1**		K-5	95.6		K-6	96.0		K-7	96.7
10	8 CHOUETTE_BOUQUET	K-1	86.4		K-2	92.3		K-3	94.8		K-4	**95.7**		K-5	96.6		K-6	**98.2**		K-7	98.8
11	9 CARNET	K-1	88.6		K-2	93.8		K-3	**95.1**		K-4	95.5		K-5	97.6		K-6	**98.4**		K-7	98.7
12	10 LOUISETTE	K-1	87.5		K-2	93.4		K-3	**95.3**		K-4	96.3		K-5	97.0		K-6	97.7		K-7	**98.1**
13	11 JE_FUME	K-1	85.5		K-2	89.5		K-3	93.1		K-4	93.8		K-5	94.8		K-6	**97.0**		K-7	97.7
14	12 PETIT_POUCET	K-1	86.7		K-2	90.2		K-3	93.7		K-4	94.9		K-5	**95.3**		K-6	96.6		K-7	96.9
15	13 VELO	K-1	84.7		K-2	92.1		K-3	93.7		K-4	94.1		K-5	**96.8**		K-6	97.6		K-7	97.9
16	14 MALADE	K-1	85.5		K-2	90.4		K-3	93.8		K-4	**95.0**		K-5	97.1		K-6	**98.0**		K-7	98.6
17	15 BIEN_RIGOLE	K-1	88.2		K-2	92.3		K-3	94.4		K-4	94.9		K-5	**95.9**		K-6	96.4		K-7	96.9
18	16 FREQUENTE_AGNEN	K-1	85.9		K-2	90.7		K-3	93.1		K-4	**95.6**		K-5	96.6		K-6	97.4		K-7	**98.1**
19	17 PAS_DE_SOLEIL	K-1	86.4		K-2	90.8		K-3	92.8		K-4	93.9		K-5	94.7		K-6	**96.5**		K-7	97.1
20	18 QUITTE_MAISON	K-1	88.2		K-2	92.8		K-3	94.8		K-4	**96.3**		K-5	97.1		K-6	97.9		K-7	**98.4**
21																					

Giving us this…

Hmm, this is less useful than expected

1k-Range is only 6 percentage points over 18 stories – is it worth the trouble?

Or do we need to know something about French to make this work

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
21																						
22	**LE PETIT NICOLAS**																					
23	**SORTED BY 1k**																					
24	9 CARNET	K-1	88.6		K-2	93.8		K-3	**95.1**		K-4	95.5		K-5	97.6		K-6	**98.4**		K-7	98.7	
25	3 BOUILLON	K-1	88.2		K-2	92.4		K-3	94.2		K-4	94.9		K-5	**95.7**		K-6	96.5		K-7	96.8	
26	15 BIEN_RIGOLE	K-1	88.2		K-2	92.3		K-3	94.4		K-4	94.9		K-5	**95.9**		K-6	96.4		K-7	96.9	
27	18 QUITTE_MAISON	K-1	88.2		K-2	92.8		K-3	94.8		K-4	**96.3**		K-5	97.1		K-6	97.9		K-7	**98.4**	
28	10 LOUISETTE	K-1	87.5		K-2	93.4		K-3	**95.3**		K-4	96.3		K-5	97.0		K-6	97.7		K-7	**98.1**	
29	5 INSPECTEUR	K-1	87.1		K-2	91.1		K-3	94.4		K-4	**96.0**		K-5	96.5		K-6	97.8		K-7	**98.4**	
30	7 DJODJO	K-1	87.0		K-2	91.0		K-3	94.2		K-4	**95.1**		K-5	95.6		K-6	96.0		K-7	96.7	
31	12 PETIT_POUCET	K-1	86.7		K-2	90.2		K-3	93.7		K-4	94.9		K-5	**95.3**		K-6	96.6		K-7	96.9	
32	4 FOOTBALL	K-1	86.4		K-2	92.6		K-3	**95.4**		K-4	97.1		K-5	97.9		K-6	**99.2**		K-7	99.3	
33	8 CHOUETTE_BOUQUET	K-1	86.4		K-2	92.3		K-3	94.8		K-4	**95.7**		K-5	96.6		K-6	**98.2**		K-7	98.8	
34	17 PAS_DE_SOLEIL	K-1	86.4		K-2	90.8		K-3	92.8		K-4	93.9		K-5	94.7		K-6	**96.5**		K-7	97.1	
35	1 SOUVENIR CHERIR	K-1	86.2		K-2	91.1		K-3	92.9		K-4	**95.9**		K-5	96.8		K-6	97.1		K-7	97.4	
36	16 FREQUENTE_AGNEN	K-1	85.9		K-2	90.7		K-3	93.1		K-4	**95.6**		K-5	96.6		K-6	97.4		K-7	**98.1**	
37	11 JE_FUME	K-1	85.5		K-2	89.5		K-3	93.1		K-4	93.8		K-5	94.8		K-6	**97.0**		K-7	97.7	
38	14 MALADE	K-1	85.5		K-2	90.4		K-3	93.8		K-4	**95.0**		K-5	97.1		K-6	**98.0**		K-7	98.6	
39	13 VELO	K-1	84.7		K-2	92.1		K-3	93.7		K-4	94.1		K-5	**96.8**		K-6	97.6		K-7	97.9	
40	2 COWBOYS	K-1	83.0		K-2	89.6		K-3	91.1		K-4	93.0		K-5	94.3		K-6	**95.0**		K-7	96.4	
41	6 REX	K-1	82.6		K-2	90.6		K-3	92.6		K-4	93.7		K-5	93.9		K-6	94.6		K-7	**95.3**	
42																						

This is where some facts abut French lexical distributions is useful

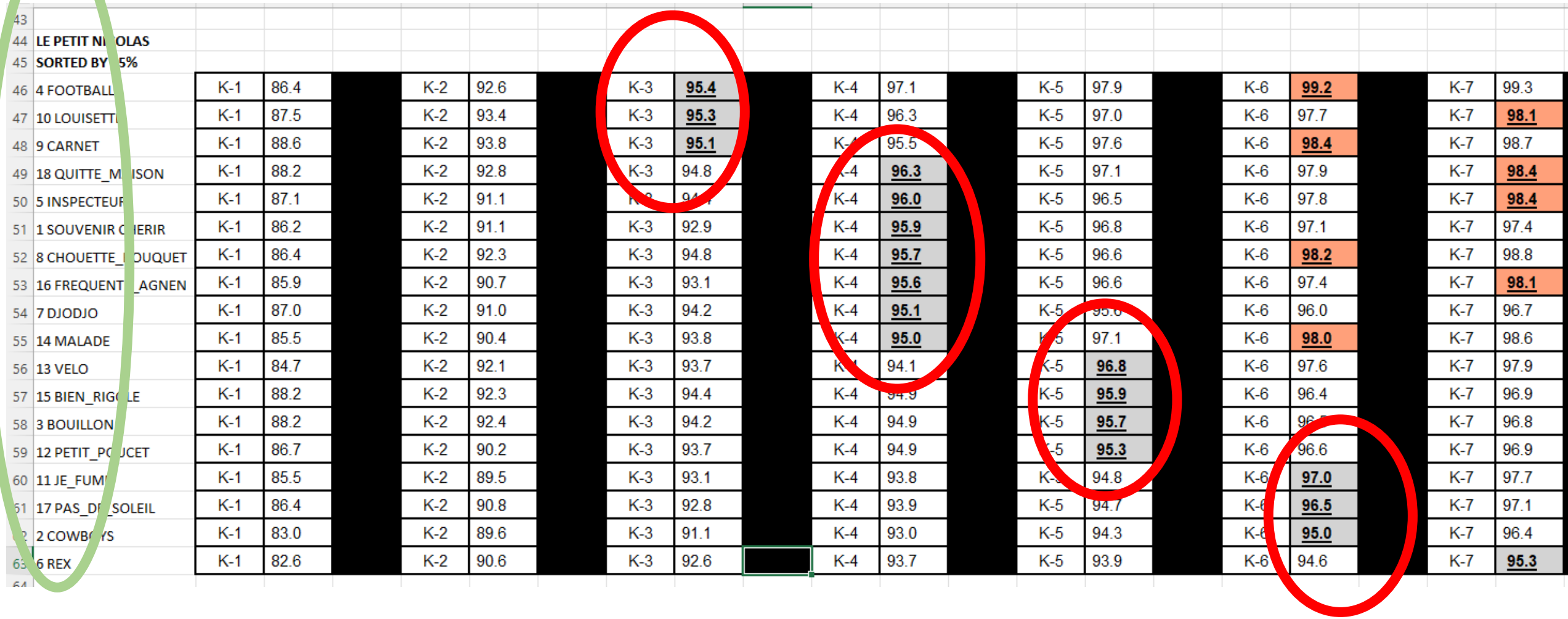
- Fact: Cobb and Horst (2004)
  - But why? Because of the greater number of non-optional function words?
    - Example
      - “Nicolas brought a ball” (1 function word, *a*)
      - “Nicolas a aporté un ballon” (2 function words, *a* and *un*)

- So maybe try some other Excel sorts that group the texts differently?
  - Say, by the k-levels at which 95% coverage is achieved?

So sort by Columns I, then L, then O…

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	**LE PETIT NICOLAS**																					
2	**AS PUBLISHED**																					
3	1 SOUVENIR CHERIR	K-1	86.2		K-2	91.1		K-3	92.9		K-4	**95.9**		K-5	96.8		K-6	97.1		K-7	97.4	
4	10 LOUISETTE	K-1	87.5		K-2	93.4		K-3	**95.3**		K-4	96.3		K-5	97.0		K-6	97.7		K-7	**98.1**	
5	11 JE_FUME	K-1	85.5		K-2	89.5		K-3	93.1		K-4	93.8		K-5	94.8		K-6	**97.0**		K-7	97.7	
6	12 PETIT_POUCET	K-1	86.7		K-2	90.2		K-3	93.7		K-4	94.9		K-5	**95.3**		K-6	96.6		K-7	96.9	
7	13 VELO	K-1	84.7		K-2	92.1		K-3	93.7		K-4	94.1		K-5	**96.8**		K-6	97.6		K-7	97.9	
8	14 MALADE	K-1	85.5		K-2	90.4		K-3	93.8		K-4	**95.0**		K-5	97.1		K-6	**98.0**		K-7	98.6	
9	15 BIEN_RIGOLE	K-1	88.2		K-2	92.3		K-3	94.4		K-4	94.9		K-5	**95.9**		K-6	96.4		K-7	96.9	
10	16 FREQUENTE_AGNEN	K-1	85.9		K-2	90.7		K-3	93.1		K-4	**95.6**		K-5	96.6		K-6	97.4		K-7	**98.1**	
11	17 PAS_DE_SOLEIL	K-1	86.4		K-2	90.8		K-3	92.8		K-4	93.9		K-5	94.7		K-6	**96.5**		K-7	97.1	
12	18 QUITTE_MAISON	K-1	88.2		K-2	92.8		K-3	94.8		K-4	**96.3**		K-5	97.1		K-6	97.9		K-7	**98.4**	
13	2 COWBOYS	K-1	83.0		K-2	89.6		K-3	91.1		K-4	93.0		K-5	94.3		K-6	**95.0**		K-7	96.4	
14	3 BOUILLON	K-1	88.2		K-2	92.4		K-3	94.2		K-4	94.9		K-5	**95.7**		K-6	96.5		K-7	96.8	
15	4 FOOTBALL	K-1	86.4		K-2	92.6		K-3	**95.4**		K-4	97.1		K-5	97.9		K-6	**99.2**		K-7	99.3	
16	5 INSPECTEUR	K-1	87.1		K-2	91.1		K-3	94.4		K-4	**96.0**		K-5	96.5		K-6	97.8		K-7	**98.4**	
17	6 REX	K-1	82.6		K-2	90.6		K-3	92.6		K-4	93.7		K-5	93.9		K-6	94.6		K-7	**95.3**	
18	7 DJODJO	K-1	87.0		K-2	91.0		K-3	94.2		K-4	**95.1**		K-5	95.6		K-6	96.0		K-7	96.7	
19	8 CHOUETTE_BOUQUET	K-1	86.4		K-2	92.3		K-3	94.8		K-4	**95.7**		K-5	96.6		K-6	**98.2**		K-7	98.8	
20	9 CARNET	K-1	88.6		K-2	93.8		K-3	**95.1**		K-4	95.5		K-5	97.6		K-6	**98.4**		K-7	98.7	
21																						

Voilà – a good first guess at a set of four distinct lexical levels across these 18 stories



43																			
44	LE PETIT NICOLAS																		
45	SORTED BY 5%																		
46	4 FOOTBALL	K-1	86.4		K-2	92.6		K-3	**95.4**		K-4	97.1		K-5	97.9		K-6	**99.2**		K-7	99.3
47	10 LOUISETTE	K-1	87.5		K-2	93.4		K-3	**95.3**		K-4	96.3		K-5	97.0		K-6	97.7		K-7	**98.1**
48	9 CARNET	K-1	88.6		K-2	93.8		K-3	**95.1**		K-4	95.5		K-5	97.6		K-6	**98.4**		K-7	98.7
49	18 QUITTE_MAISON	K-1	88.2		K-2	92.8		K-3	94.8		K-4	**96.3**		K-5	97.1		K-6	97.9		K-7	**98.4**
50	5 INSPECTEUR	K-1	87.1		K-2	91.1		K-3	94.4		K-4	**96.0**		K-5	96.5		K-6	97.8		K-7	**98.4**
51	1 SOUVENIR CHERIR	K-1	86.2		K-2	91.1		K-3	92.9		K-4	**95.9**		K-5	96.8		K-6	97.1		K-7	97.4
52	8 CHOUETTE_BOUQUET	K-1	86.4		K-2	92.3		K-3	94.8		K-4	**95.7**		K-5	96.6		K-6	**98.2**		K-7	98.8
53	16 FREQUENT_AGNEN	K-1	85.9		K-2	90.7		K-3	93.1		K-4	**95.6**		K-5	96.6		K-6	97.4		K-7	**98.1**
54	7 DJODJO	K-1	87.0		K-2	91.0		K-3	94.2		K-4	**95.1**		K-5	95.8		K-6	96.0		K-7	96.7
55	14 MALADE	K-1	85.5		K-2	90.4		K-3	93.8		K-4	**95.0**		K-5	97.1		K-6	**98.0**		K-7	98.6
56	13 VELO	K-1	84.7		K-2	92.1		K-3	93.7		K-4	94.1		K-5	**96.8**		K-6	97.6		K-7	97.9
57	15 BIEN_RIGOLE	K-1	88.2		K-2	92.3		K-3	94.4		K-4	94.9		K-5	**95.9**		K-6	96.4		K-7	96.9
58	3 BOUILLON	K-1	88.2		K-2	92.4		K-3	94.2		K-4	94.9		K-5	**95.7**		K-6	96.7		K-7	96.8
59	12 PETIT_POUCET	K-1	86.7		K-2	90.2		K-3	93.7		K-4	94.9		K-5	**95.3**		K-6	96.6		K-7	96.9
60	11 JE_FUME	K-1	85.5		K-2	89.5		K-3	93.1		K-4	93.8		K-5	94.8		K-6	**97.0**		K-7	97.7
61	17 PAS_DE_SOLEIL	K-1	86.4		K-2	90.8		K-3	92.8		K-4	93.9		K-5	94.7		K-6	**96.5**		K-7	97.1
62	2 COWBOYS	K-1	83.0		K-2	89.6		K-3	91.1		K-4	93.0		K-5	94.3		K-6	**95.0**		K-7	96.4
63	6 REX	K-1	82.6		K-2	90.6		K-3	92.6		K-4	93.7		K-5	93.9		K-6	94.6		K-7	**95.3**
64																					

This would need empirical evaluation

Ongoing study design:

- Group A reads the stories in published sequence
- Group B reads the stories in the DDL-revised sequence

→ Are there differences in rate, comprehension, enjoyment, lexical development, grammar development?

→ Is there impact on writing (sentence length, accuracy, VP?)

Perhaps the point is made

DDL has a lot to offer in the design of this course

Beyond just ‘choosing a good book the learners will like’

- But there are several further piecemeal ways DDL can operate in this course
  - Corpus searches of the stories by teacher to explore and teach features of the stories
    - Is a particular word recurring or one-off?
  - Control of examination lexis
    - (Do questions and rubrics ***about*** the stories match the language ***of*** the stories?)
  - If a corpus of teacher-talk is available, are teachers reinforcing the language of the texts in the language of the classroom?
    - Is there reinforcement and recycling of new words?

Home > Concordancers > French Input > Output (« Back keeps original settings) Copiable extract-link to this data >> here

Concordance for *(f)lemma* **chouette** in **petit_nicolas** Dictionnaire/TTS Fren_Eng

Extract checked ⬇ items: ☑ All | Ø | any10 | 20 | 30 | 50 Go >

MODIFY: (f)lemma ; chouette Petit Nicolas 130,000 sort subco +assoc(s) on L|R side for 25,000 li

25 hits

001.	tous pour cette idée qui était vraiment CHOUETTE, mais Alceste n'était pas d'acc	02_cow-boys
002.	le voir; tout le monde criait, c'était CHOUETTE, on rigolait bien, terrible. Et	02_cow-boys
003.	ous on était drôlement contents! Il est CHOUETTE mon papa! Nous avons attaché pa	02_cow-boys
004.	à crier et à faire des grimaces. C'est CHOUETTE de savoir s'amuser tout seul, c	02_cow-boys
005.	t un vieux foulard à ma maman. On était CHOUETTES ! On était dans le jardin et m	02_cow-boys
006.	se faire des passes et c'est drôlement CHOUETTE déjouer entre les bancs. Quand	03_le_boui…
007.	llons faire une partie terrible. Il est CHOUETTE, Alceste. Nous nous sommes retr	04_le_foot…
008.	pas bon! — Ça sent pas bon, mais c'est CHOUETTE à manger, a répondu Alceste. Et	05_l'inspe…
009.	et il est entré dans la maison. Il est CHOUETTE papa! Pendant que papa arrangea	06_rex
010.	s, on a commencé à le trouver drôlement CHOUETTE, Djodjo, et moi, je lui ai donn	07_djodjo
011.	ns le pot-au-feu. Le bouquet était très CHOUETTE et très gros, la dame l'a envel	08_chouett…
012.	s. « C'est vrai, a dit Rufus, c'est pas CHOUETTE ce que vous avez fait aux fleur	08_chouett…
013.	ai commencé à jouer avec Eudes et c'est CHOUETTE de jouer aux billes avec Eudes,	08_chouett…
014.	..... CHOUETTE	08_chouett…
015.	e vous voudrez, mais ma maman, elle est CHOUETTE!	08_chouett…
016.	n pensant à tout ça et me racontant des CHOUETTES histoires, j'avais oublié mon	09_les_car…
017.	s le placard et elle a vu mon avion, le CHOUETTE, celui qui a un élastique, qui	10_louiset…
018.	dessert, mais ça ne fait rien, elle est CHOUETTE, Louisette, et quand on sera gr	10_louiset…
019.	urs au ministre. C'est vrai qu'il a des CHOUETTES idées le directeur et ce sera	10_louiset…
020.	de, là où il manque une planche. Il est CHOUETTE le terrain vague, nous y allons	11_je_fume

Par contre ~

Home > Concordancers > French Input > Output (« Back keeps original settings) Copiable extract-link to this data >> here

Concordance for *(f)lemma* **apercevoir** in **petit_nicolas** Dictionnaire/TTS Fren_Eng

Extract checked items: All | Ø | any10 | 20 | 30 | 50 Go >

RANGE: *apercevoir* appears in 2 out of 14 sub-corpora (14.29%)

Click any **KEYWORD** for more context + 'Get More' option FEB 2022

3 hits

001. otre vie, c'est raté, parce qu'on s'est APERÇU que le photographe n'était plus l **01_souve…**
002. quoi? a demandé Alceste. Je viens de m'APERCEVOIR que nous n'avons pas de ballo **04_le_fo…**
003. amais, comme aujourd'hui, je ne me suis APERÇU à quel point notre métier est un **05_l'ins…**

Distribution for lemma apercevoir: 01_souvenir_cherir=1 04_le_football=1 05_linspecteur=1 02_cow-bo
05_l'inspecteur=0 06_rex=0 07_djodjo=0 08_chouette_bouquet=0 09_les_carnets=0 10_louisette=0 11_je_fu
13_le_velo=0 16_frequent_agnan=0
RANGE: *apercevoir* appears in 2 out of 14 sub-corpora (14.29%)

T can go as deep as the Ss' needs, e.g. collocation

Home > Concordancers > French Input > Output (« Back keeps original settings) Copiable extract-link to this data >> h

Concordance for *(f)lemma* **répondre** in **petit_nicolas** Dictionnaire/TTS Fren_Eng

Extract checked ↓ items: ☑ All | Ø | any10 | 20 | 30 | 50 Go >

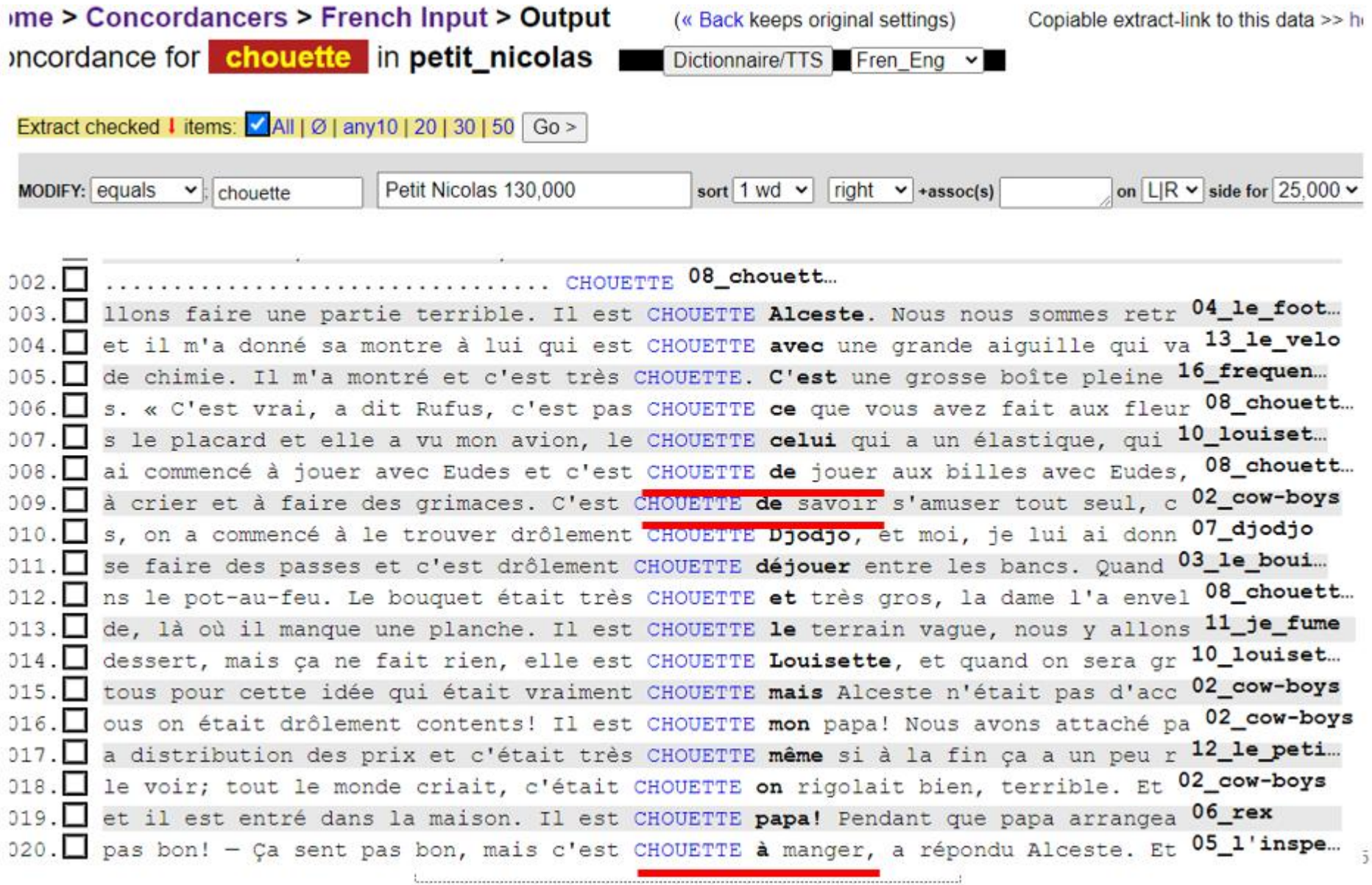
RANGE: *répondre* with *que* on *right* side appears in 8 out of 14 sub-corpora (57.14%)

Click any **KEYWORD** for more context + 'Get More' option FEB 2022

10 hits

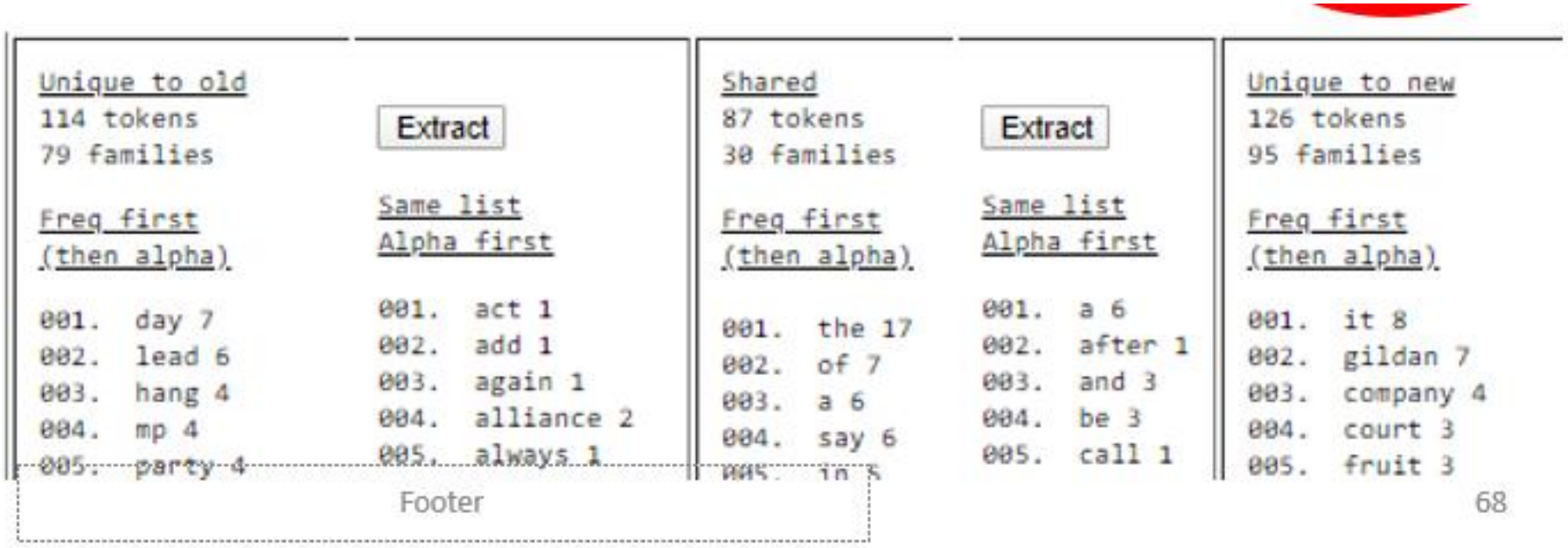
001.	plaisir à votre maîtresse? » Nous avons RÉPONDU que oui, parce que nous l'aimons	01_souve…
002.	dit de cesser de manger, mais Alceste a RÉPONDU qu'il fallait bien qu'il se nour	01_souve…
003.	oachim? » a demandé Clotaire. Joachim a RÉPONDU qu'il ne voyait rien. Alors, Clo	03_le_bo…
004.	ce qu'il lui était arrivé et il nous a RÉPONDU qu'il s'était endormi à force de	03_le_bo…
005.	mauvais côté du terrain! Moi, je lui ai RÉPONDU que si le soleil ne lui plaisait	04_le_fo…
006.	uérir, mais ça fait très mal. Moi, j'ai RÉPONDU à papa que Rex n'était pas malad	06_rex
007.	ndé, Eudes. « Je ne peux pas, je lui ai RÉPONDU, il faut que je rentre chez moi	08_choue…
008.	re semblant de chanter et Alceste lui a RÉPONDU qu'il ne faisait pas semblant de	10_louis…
009.	as du feu? » il m'a demandé, je lui ai RÉPONDU que non. « Ben alors, a dit Alce	11_je_fu…
010.	n me ferait mettre en prison. Je lui ai RÉPONDU que j'avais bien envie de lui ta	16_frequ…

E.g., syntax


me > Concordancers > French Input > Output
(« Back keeps original settings)
Copiable extract-link to this data >> h
ncordance for chouette in petit_nicolas
Dictionnaire/TTS
Fren_Eng
Extract checked ! items: All | Ø | any10 | 20 | 30 | 50 Go >
MODIFY: equals chouette
Petit Nicolas 130,000
sort 1 wd
right
+assoc(s)
on L|R
side for 25,000
002. ..... CHOUETTE 08_chouett…
003. llons faire une partie terrible. Il est CHOUETTE Alceste. Nous nous sommes retr 04_le_foot…
004. et il m'a donné sa montre à lui qui est CHOUETTE avec une grande aiguille qui va 13_le_velo
005. de chimie. Il m'a montré et c'est très CHOUETTE. C'est une grosse boîte pleine 16_frequen…
006. s. « C'est vrai, a dit Rufus, c'est pas CHOUETTE ce que vous avez fait aux fleur 08_chouett…
007. s le placard et elle a vu mon avion, le CHOUETTE celui qui a un élastique, qui 10_louiset…
008. ai commencé à jouer avec Eudes et c'est CHOUETTE de jouer aux billes avec Eudes, 08_chouett…
009. à crier et à faire des grimaces. C'est CHOUETTE de savoir s'amuser tout seul, c 02_cow-boys
010. s, on a commencé à le trouver drôlement CHOUETTE Djodjo, et moi, je lui ai donn 07_djodjo
011. se faire des passes et c'est drôlement CHOUETTE déjouer entre les bancs. Quand 03_le_boui…
012. ns le pot-au-feu. Le bouquet était très CHOUETTE et très gros, la dame l'a envel 08_chouett…
013. de, là où il manque une planche. Il est CHOUETTE le terrain vague, nous y allons 11_je_fume
014. dessert, mais ça ne fait rien, elle est CHOUETTE Louisette, et quand on sera gr 10_louiset…
015. tous pour cette idée qui était vraiment CHOUETTE mais Alceste n'était pas d'acc 02_cow-boys
016. ous on était drôlement contents! Il est CHOUETTE mon papa! Nous avons attaché pa 02_cow-boys
017. a distribution des prix et c'était très CHOUETTE même si à la fin ça a un peu r 12_le_peti…
018. le voir; tout le monde criait, c'était CHOUETTE on rigolait bien, terrible. Et 02_cow-boys
019. et il est entré dans la maison. Il est CHOUETTE papa! Pendant que papa arrangea 06_rex
020. pas bon! — Ça sent pas bon, mais c'est CHOUETTE à manger, a répondu Alceste. Et 05_l'inspe…

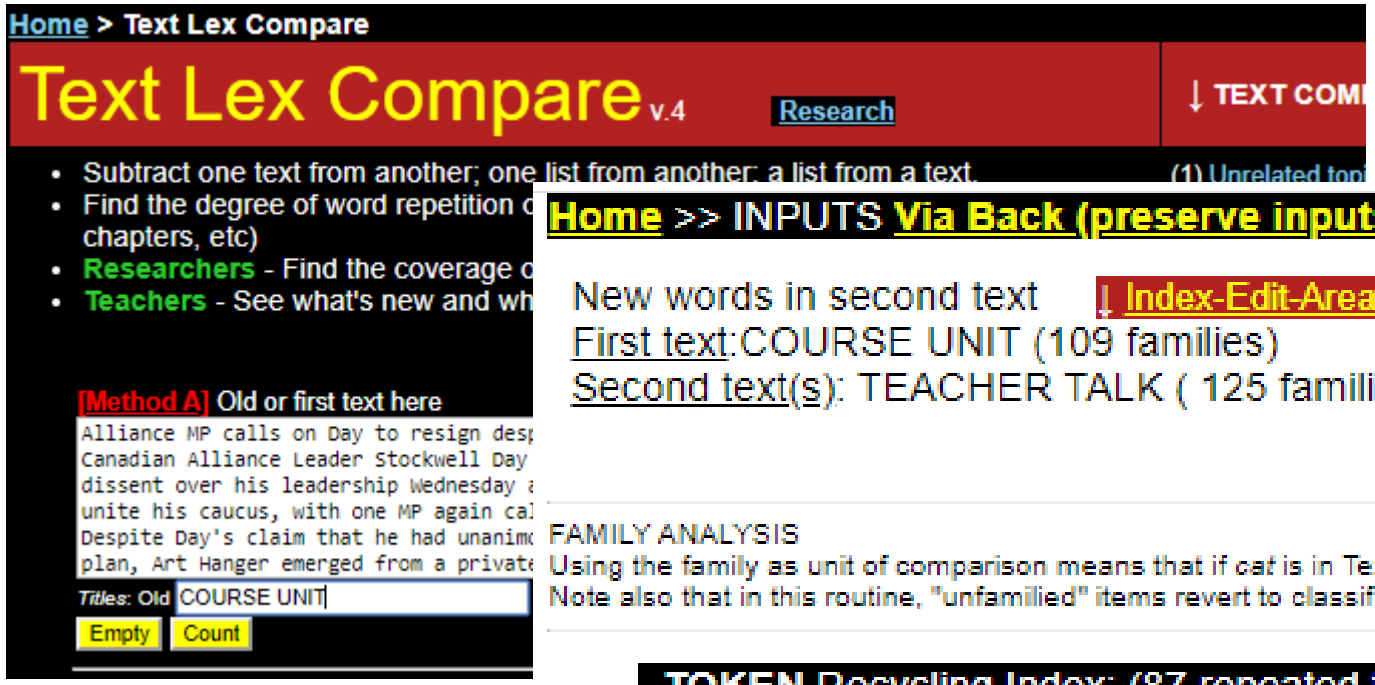
Plus two uses for Lextutor's **Text_Lex_Compare**
Text-lexis comparison

- Program determines the words/lemmas ***shared*** between two texts vs. ***unique to one*** text



1. If you can get a corpus of teacher talk for this course…

- Use Text_Lex_Compare to check how much lexis of texts is getting recycled in the classroom
- And if little is recycled, then include re-use/re-cycle practice in TT (teacher training) programs
  - Low transfer is typical →



Home >> INPUTS Via Back (preserve inputs) || Direct >> Text_Lex Output

New words in second text ↓ Index-Edit-Area at bottom ↓

First text:COURSE UNIT (109 families)

Second text(s): TEACHER TALK (125 families)

FAMILY ANALYSIS

Using the family as unit of comparison means that if *cat* is in Text 1 and *cats* in Text 2 then this is considered a repeti… *cat* famil…

Note also that in this routine, "unfamilied" items revert to classification by type (e.g., repeated proper nouns).

TOKEN Recycling Index: (87 repeated tokens : 213 tokens in new text) = **40.85%**

FAMILIES Recycling Index: (30 repeated families : 125 families in new text) = **24.00%**

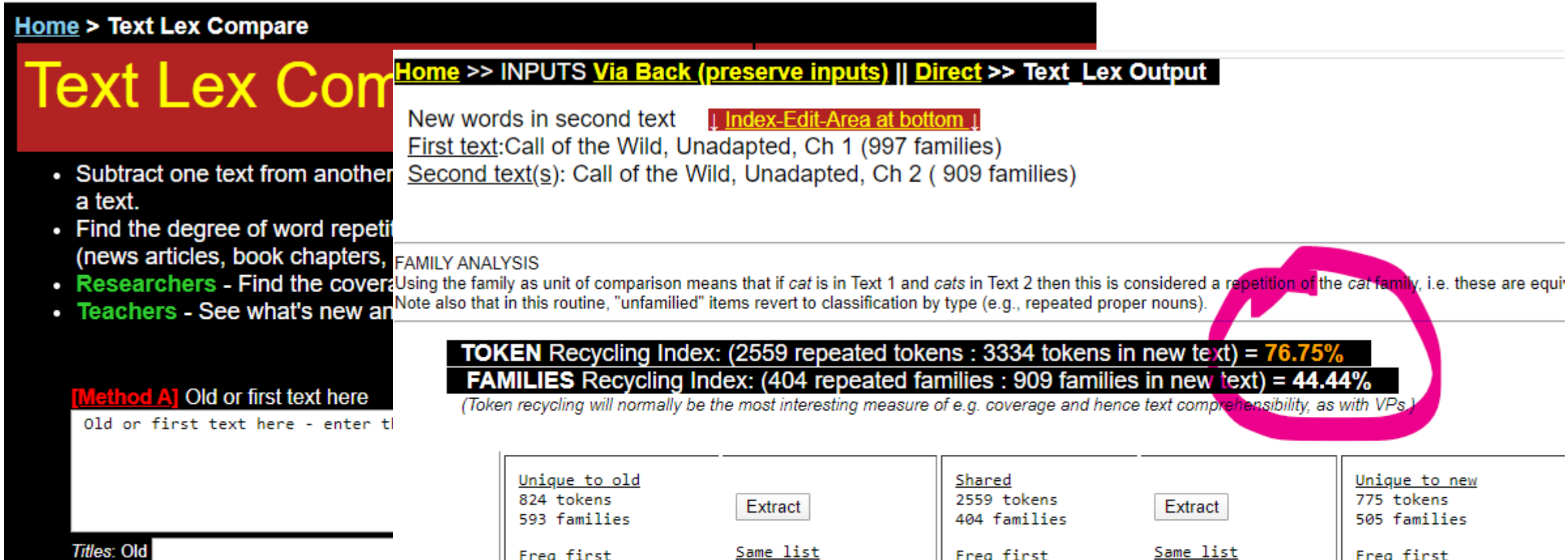
(Token recycling will normally be the most interesting measure of e.g. coverage and hence text compr…hensibility, as with V…s.

Unique to old		Shared		Unique to new
114 tokens	Extract	87 tokens	Extract	126 tokens
79 families		30 families		95 families
Freq first (then alpha)	Same list Alpha first	Freq first (then alpha)	Same list Alpha first	Freq first (then alpha)
001. day 7	001. act 1	001. the 17	001. a 6	001. it 8
002. lead 6	002. add 1	002. of 7	002. after 1	002. gildan 7
003. hang 4	003. again 1	003. a 6	003. and 3	003. company 4
004. mp 4	004. alliance 2	004. say 6	004. be 3	004. court 3
005. party 4	005. always 1	005. in 5	005. call 1	005. fruit 3

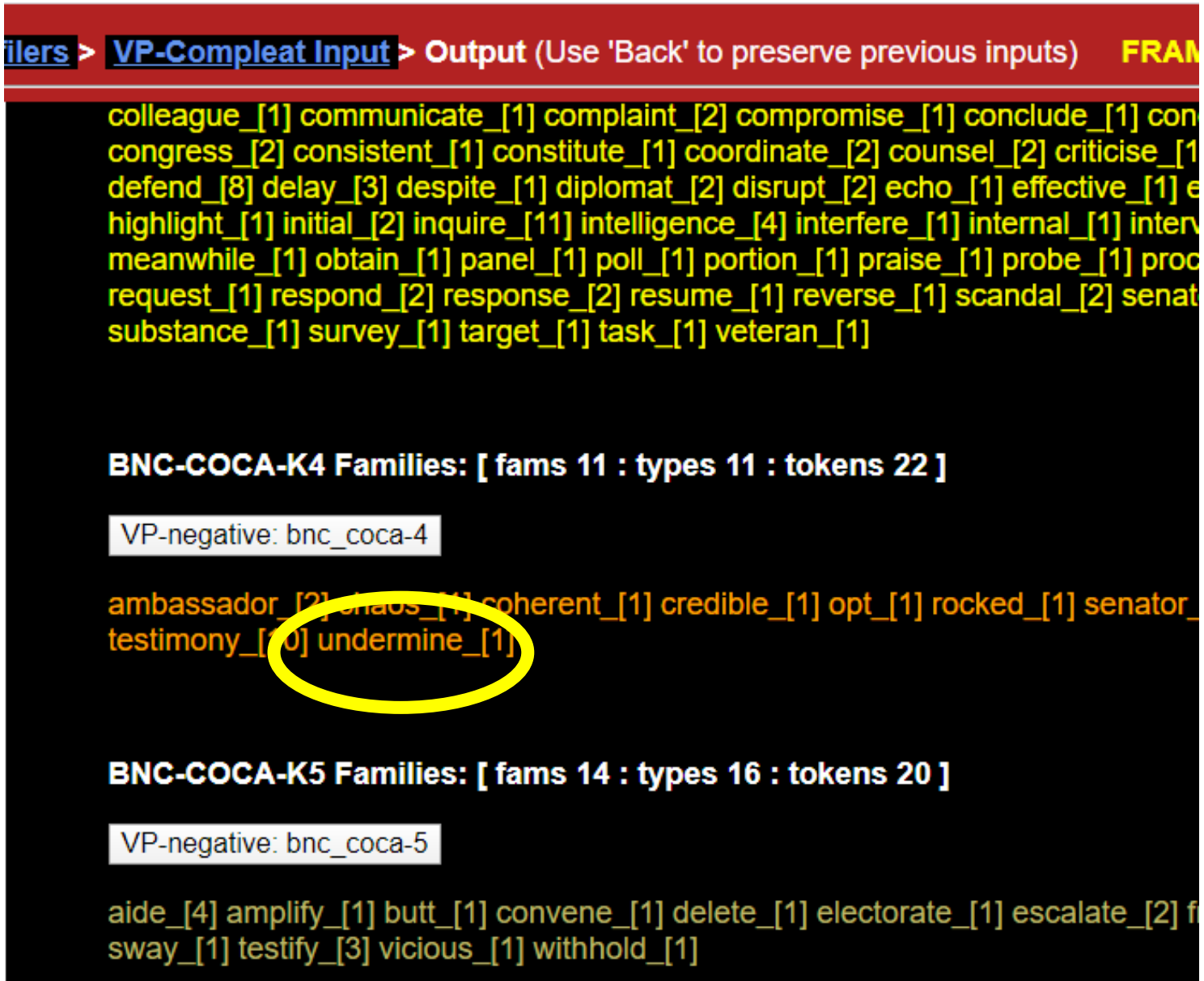
2. And finally EXAMS

To write fair examination texts/questions

- Use Text_Lex_Compare to assure that 95% of the words <u>on the test</u> were even ***seen*** <u>in the course</u>



+ Let VP identify words to gloss in an exam



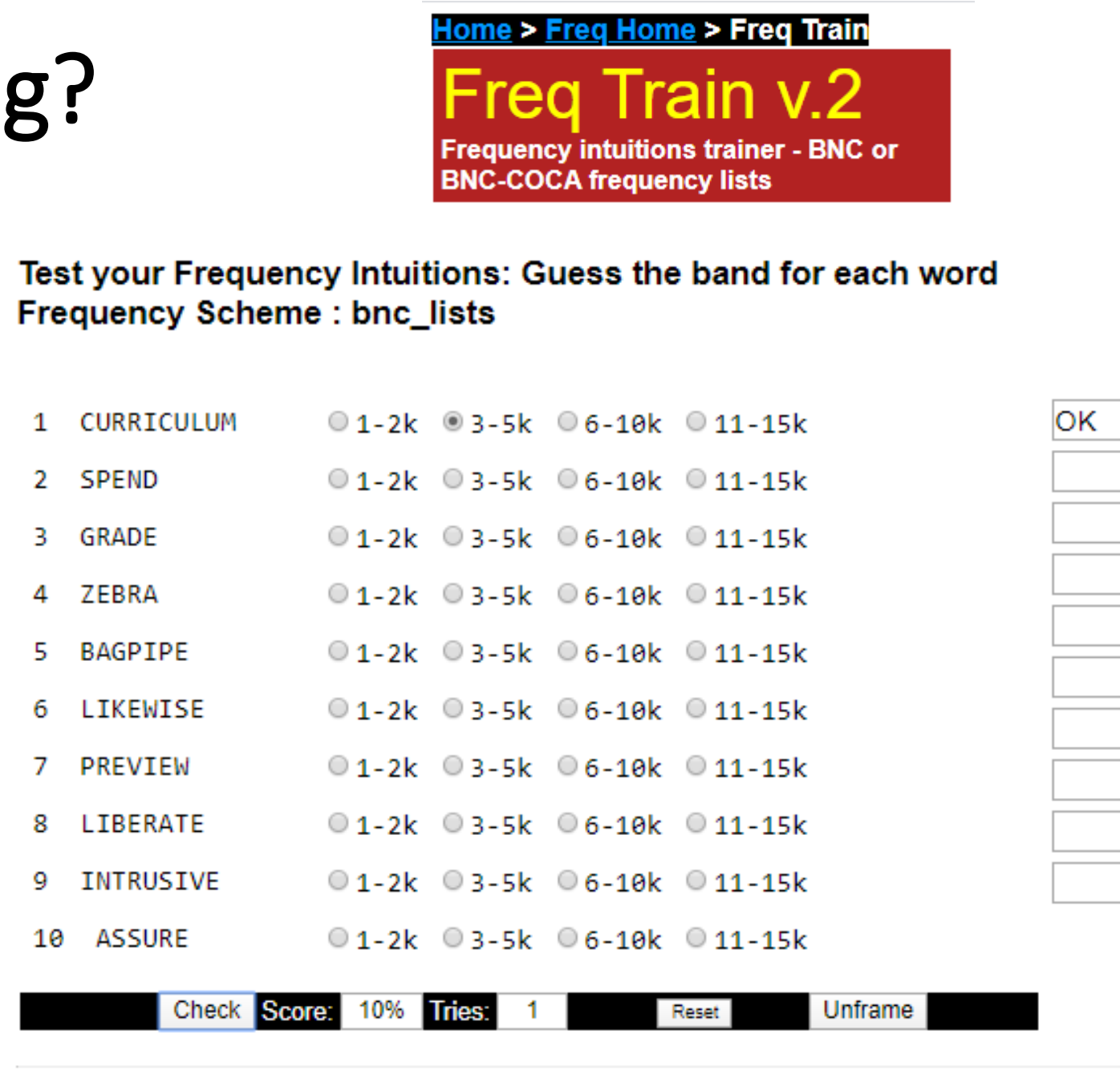
EXAM QUESTION
Text text text. Text text text
Text text text. Text text text
undermine* text text text
text…

** Undermine – try to secretly weaken someone's plan or strategy*

Or 2% left to inference <u>if</u>
<u>that was part of the course</u>

And teacher training?

- Teach use of DDCD tools?
- T-Training AT LEAST should include more needs analysis
- But, even more, AWARENESS RAISING in word frequency
  - Learners feel frequency effects, NSs do not



So we have seen some pieces of research-supported DDL – that can be put together usefully

LEXTUTOR ROUTINES INVOLVED

- TESTS
- VP / VP-Cloze
- RANGE
- TL_COMPARE
- GROUP LEX
  All @ www.lextutor.ca/ **xxx**

Many of these ideas are validated

- Boulton & Cobb (2017)
  – DDL for learning
- Cobb Horst & Nicolae (2005)
  - Group Lex
- Laufer & Ravenhorst (2010)
  – the 95% and 98% cut-offs
- Cobb (2020) 'Corpus for courses'
  - A published earlier version of this presentation

QUESTIONS? COMMENTS?
MORE LIVE DEMOS?

Further reading on this topic

Earlier version:
2021 conference paper at Swiss Conference of Applied Linguistics

Download at
Lextutor.ca/cv/

Corpus for courses: Data-driven course design

Thomas COBB
Université du Québec à Montréal (UQAM)
Faculté des sciences de l'éducation
1205, rue Saint-Denis, Montréal (Québec) H2X 3R9, Canada
cobb.tom@uqam.ca; www.lextutor.ca

L'apprentissage sur corpus, en tant qu'ensemble de principes et de technologies, joue un rôle bien établi dans l'apprentissage des langues, et cet article montre comment ceux-ci peuvent s'appliquer également à la conception de cours ou de curriculum de langue. Si les supports de cours sont reformulés sous forme de corpus, on a alors à disposition un certain nombre de moyens pour amener la recherche sur l'apprentissage directement dans la salle de classe. Cet article commence par une définition des termes, un examen de la place de l'apprentissage sur corpus dans l'acquisition de la langue, et montre des moyens d'appliquer cette approche à la conception, au testing, et à l'évaluation de l'enseignement des langues. Le contexte est la refonte d'un cours de lecture en cours pour adultes en langue seconde basé sur une collection de documents authentiques trouvés sur Internet. Les principes et technologies de la conception de cours sur corpus ont été utilisés, premièrement, pour exposer les faiblesses de cette collection de matériels en tant que cours, et, deuxièmement, pour montrer des moyens concrets de les améliorer. Tous les logiciels impliqués dans ce travail sont accessibles au public.

Mots-clés:
apprentissage basé sur les consultations d'un corpus, conception pédagogique, analyse de besoins, vocabulaire.

Keywords:
corpus-based learning, data-driven learning, instructional design, needs analysis, text analysis, vocabulary.

1. Background & proposition

Data-driven learning (DDL) is an input and comprehension based approach to