AAAL, 2022
Pittsburgh

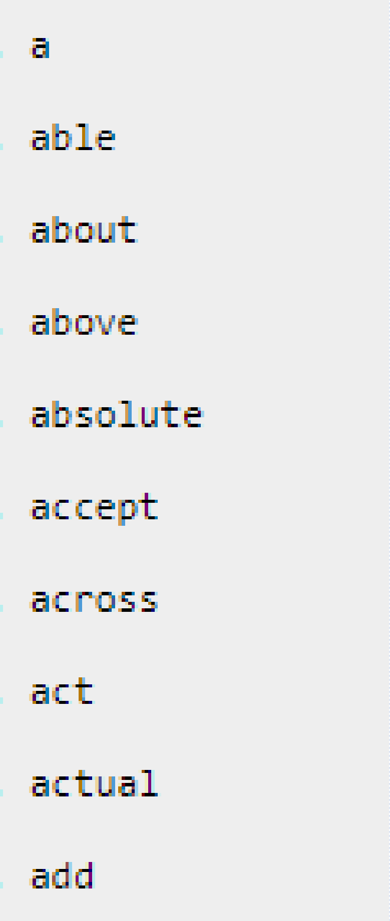
One family size does not fit all word lists

Tom Cobb – Université du Québec à Montréal

Batia Laufer – University of Haifa

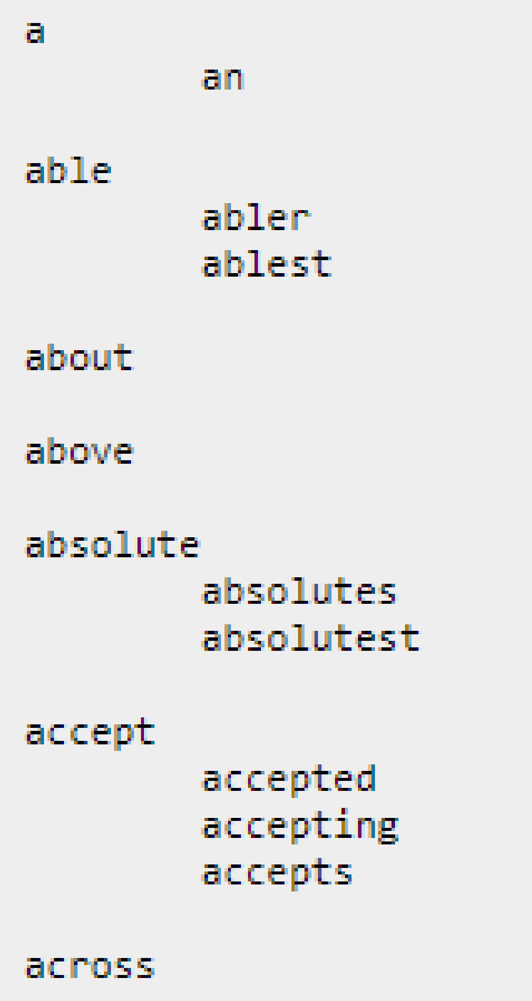
Familiar kinds of word lists

By types/head words



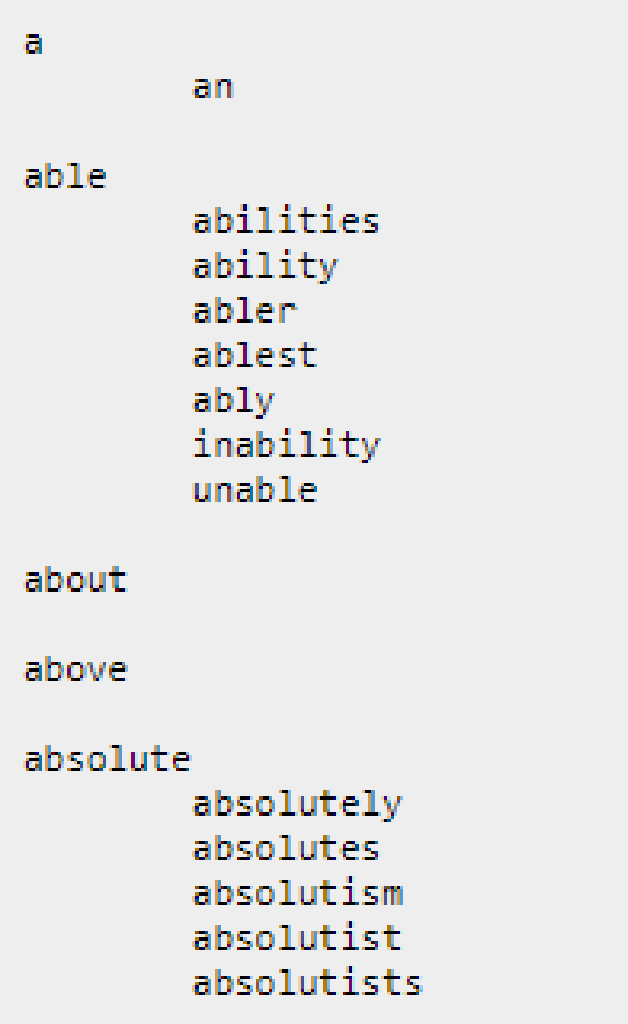
```
a

able

about

above

absolute

accept

across

act

actual

add
```

By lemmas



```
a
        an

able
        abler
        ablest

about

above

absolute
        absolutes
        absolutest

accept
        accepted
        accepting
        accepts

across
```

By families



```
a
        an

able
        abilities
        ability
        abler
        ablest
        ably
        inability
        unable

about

above

absolute
        absolutely
        absolutes
        absolutism
        absolutist
        absolutists
```

Counting units in word frequency lists

Lemma - base word + inflections (e.g., *NGSL* 2500 by Brezina & Gablasova, 2015)

i. read, reads, read, reading (v)
ii. reading (n)
iii. readable
iv. unreadable
v. readability

Flemma - lemma, but identical forms/different parts of speech = one flemma

e.g., 'reading' v/n (e.g., *Essential Word List* by Dang & Webb, 2016).

Word family - base word + inflected words + derived words

5 lemmas above=one word family (e.g., *BNC/COCA word family list* by Nation, 2012)

How assigned to frequency levels?

Summed individual frequencies of family or lemma members in a corpus
Usually broken down into groups of 1,000

Word lists' main uses

Teaching

- Give lists directly to learners
  - Ex, As a course, as flashcards, etc, with various incentives and strategies to learn

Testing

- Sample words from lists at various frequency levels
  - Ex, Vocabulary Levels Test, VST etc

Grading texts

- Find or write texts to match a certain frequency level
  - Ex, constrain a text for beginners to the first 1,000 word-lemmas or families

Doing Coverage research

- Determine the proportion of texts (80, 90, 95, 98 %) covered by words at different frequency levels
  - Ex, "2,000 word-families covers 80% of words in typical texts"

Matching counting units to list functions

1. List for direct learning

FAMILY (BNC/COCA)

Completely impossible

able

- abilities
- ability
- abler
- ablest
- ably
- inability
- unable

accept

- acceptability
- acceptable
- acceptably
- acceptance
- acceptances
- accepted
- accepting
- acceptor
- acceptors
- accepts
- unacceptability
- unacceptable
- unacceptably

1k
1000 fams
= 6,856 words

1-3k
3,000 fams
= 19,062 words

The majority v.
low frequency
individually

(f)LEMMA (BNC)

Starts well , but…

ability

- abilities

able

- abler
- ablest

accept

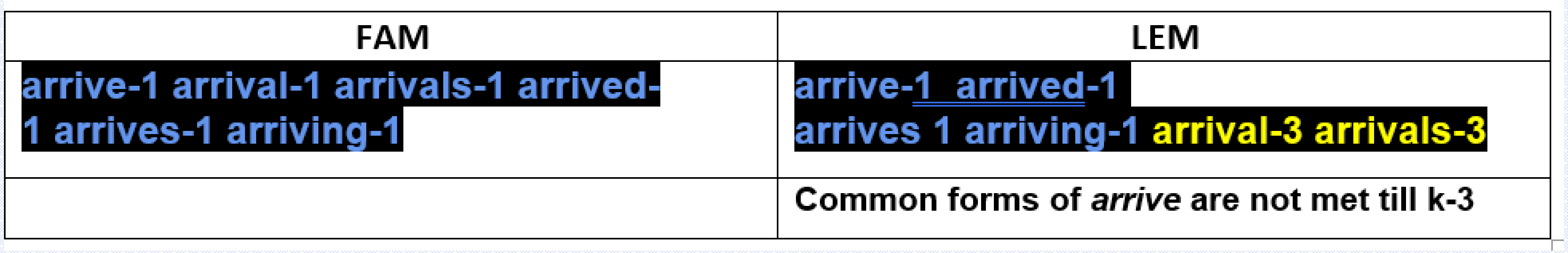
- accepts
- accepted
- accepting

1k
1,000 lems
= 3,020 words

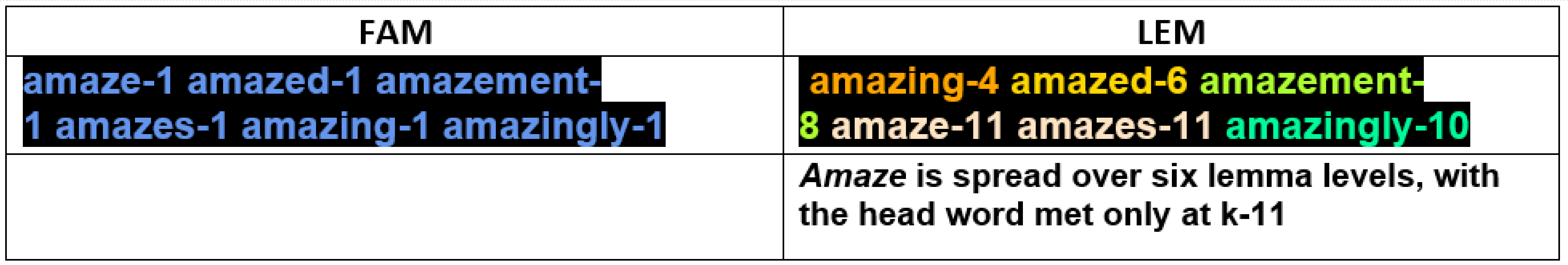
Fewer words
Mainly frequent
Mainly regular
Irregulars are separate items
(able/ability)

Lemma approach rapidly becomes unusable

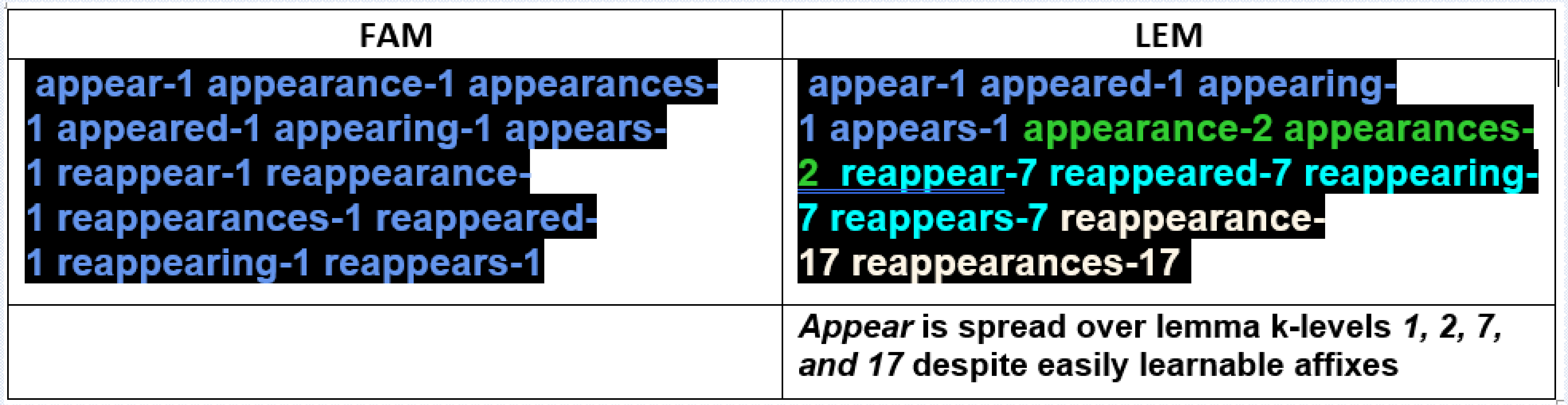
Here are some sample K-1 (first 1,000) families as ‘flemmatized’ in K’s



FAM	LEM
arrive-1 arrival-1 arrivals-1 arrived-1 arrives-1 arriving-1	arrive-1 arrived-1 arrives 1 arriving-1 arrival-3 arrivals-3
	Common forms of *arrive* are not met till k-3



FAM	LEM
amaze-1 amazed-1 amazement-1 amazes-1 amazing-1 amazingly-1	amazing-4 amazed-6 amazement-8 amaze-11 amazes-11 amazingly-10
	Amaze is spread over six lemma levels, with the head word met only at k-11

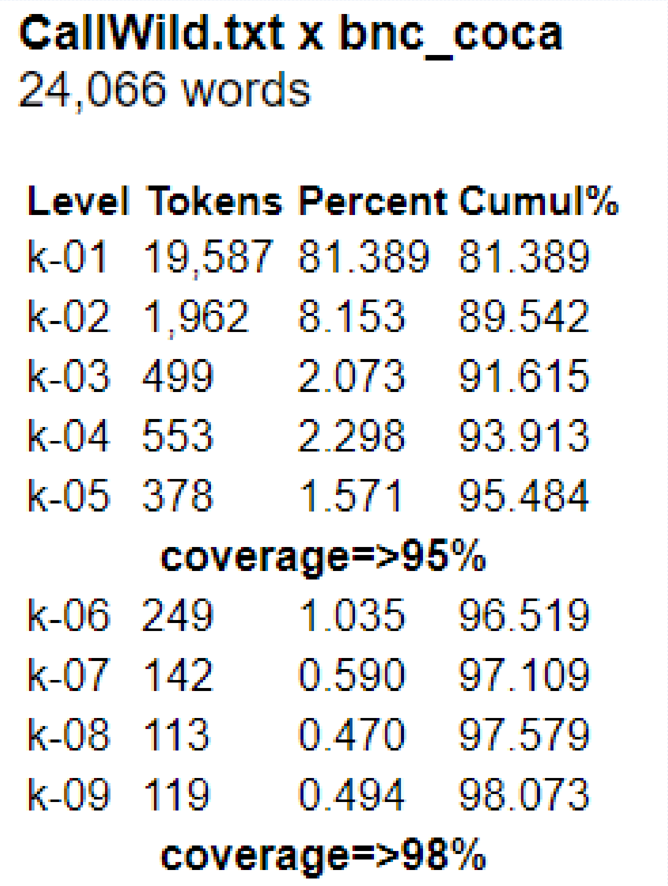


FAM	LEM
appear-1 appearance-1 appearances-1 appeared-1 appearing-1 appears-1 reappear-1 reappearance-1 reappearances-1 reappeared-1 reappearing-1 reappears-1	appear-1 appeared-1 appearing-1 appears-1 appearance-2 appearances-2 reappear-7 reappeared-7 reappearing-7 reappears-7 reappearance-17 reappearances-17
	Appear is spread over lemma k-levels *1, 2, 7, and 17* despite easily learnable affixes

And with quasi-duplication there are *so many* levels…

Here is the ungraded ‘Call of the Wild’

FAMILY (BNC/COCA) •

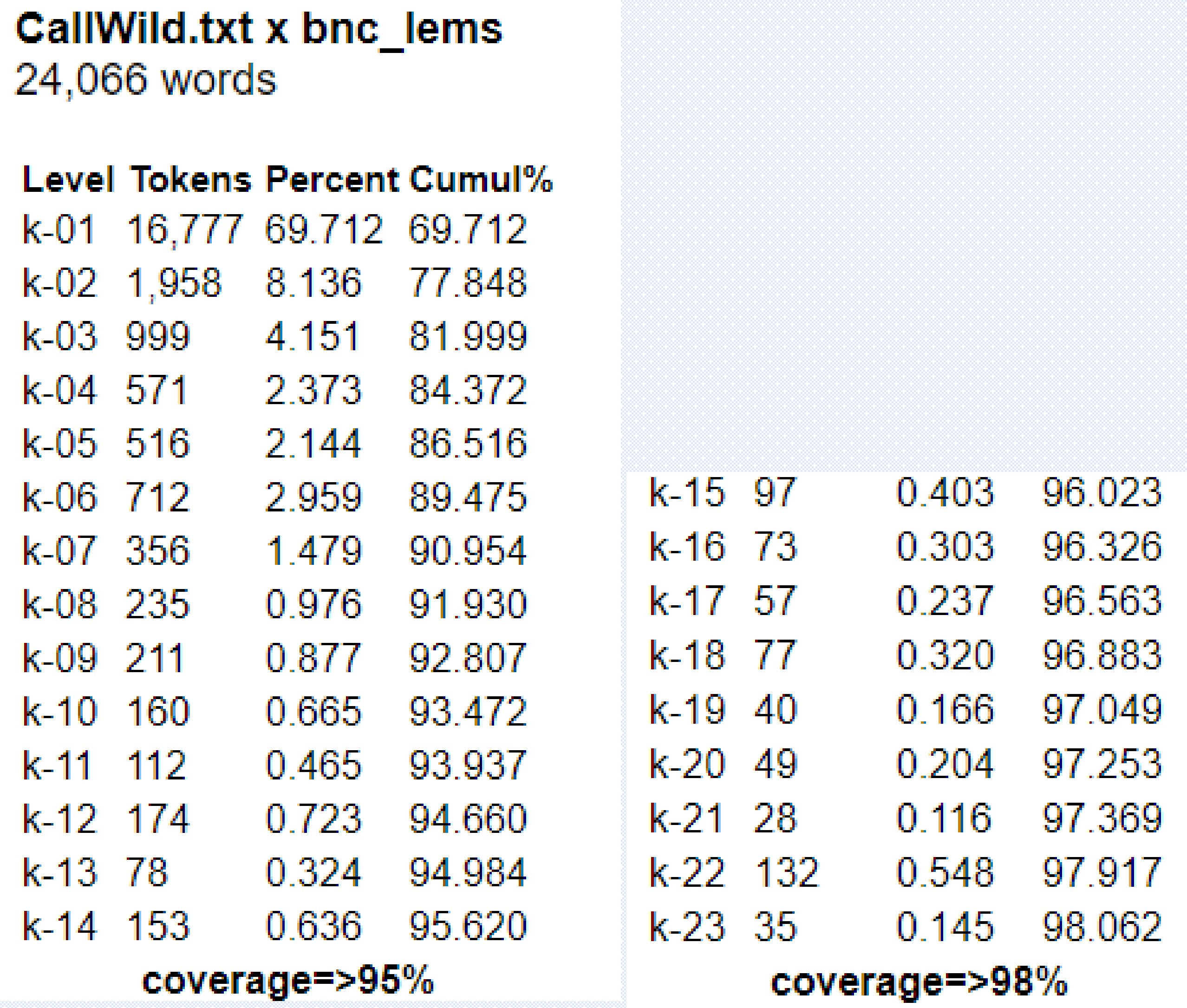


CallWild.txt x bnc_coca
24,066 words

Level	**Tokens**	**Percent**	**Cumul%**
k-01	19,587	81.389	81.389
k-02	1,962	8.153	89.542
k-03	499	2.073	91.615
k-04	553	2.298	93.913
k-05	378	1.571	95.484
coverage=>95%			
k-06	249	1.035	96.519
k-07	142	0.590	97.109
k-08	113	0.470	97.579
k-09	119	0.494	98.073
coverage=>98%			

✓

(f)LEMMA (BNC) •

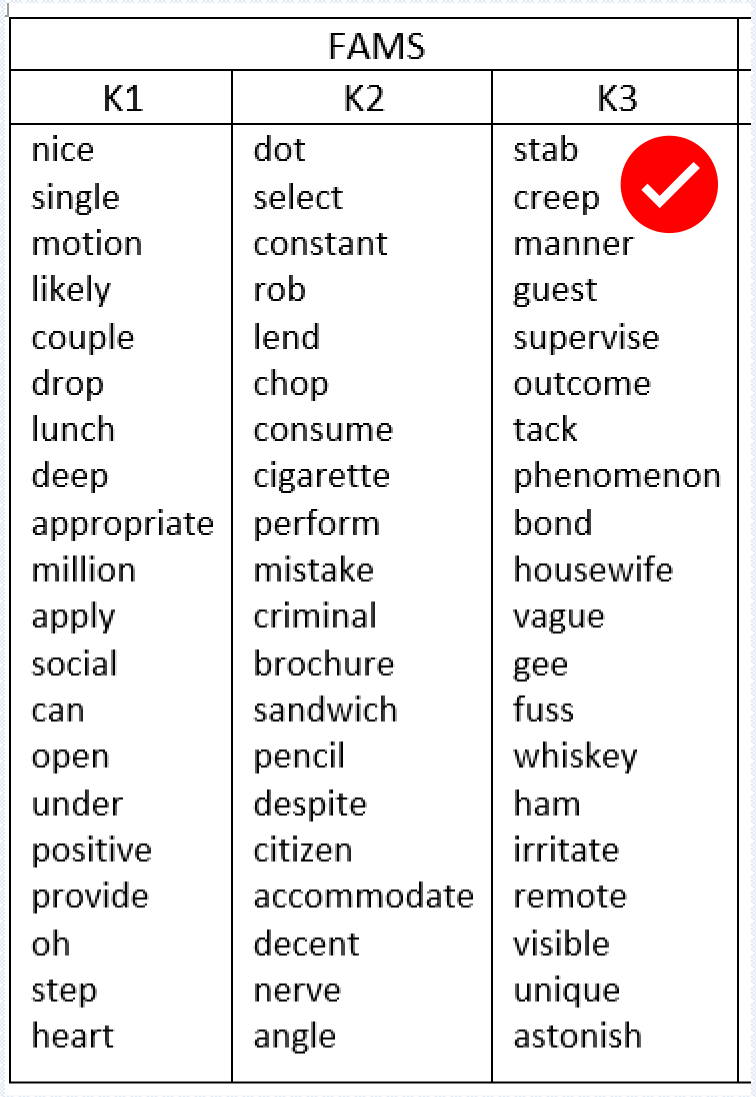


CallWild.txt x bnc_lems
24,066 words

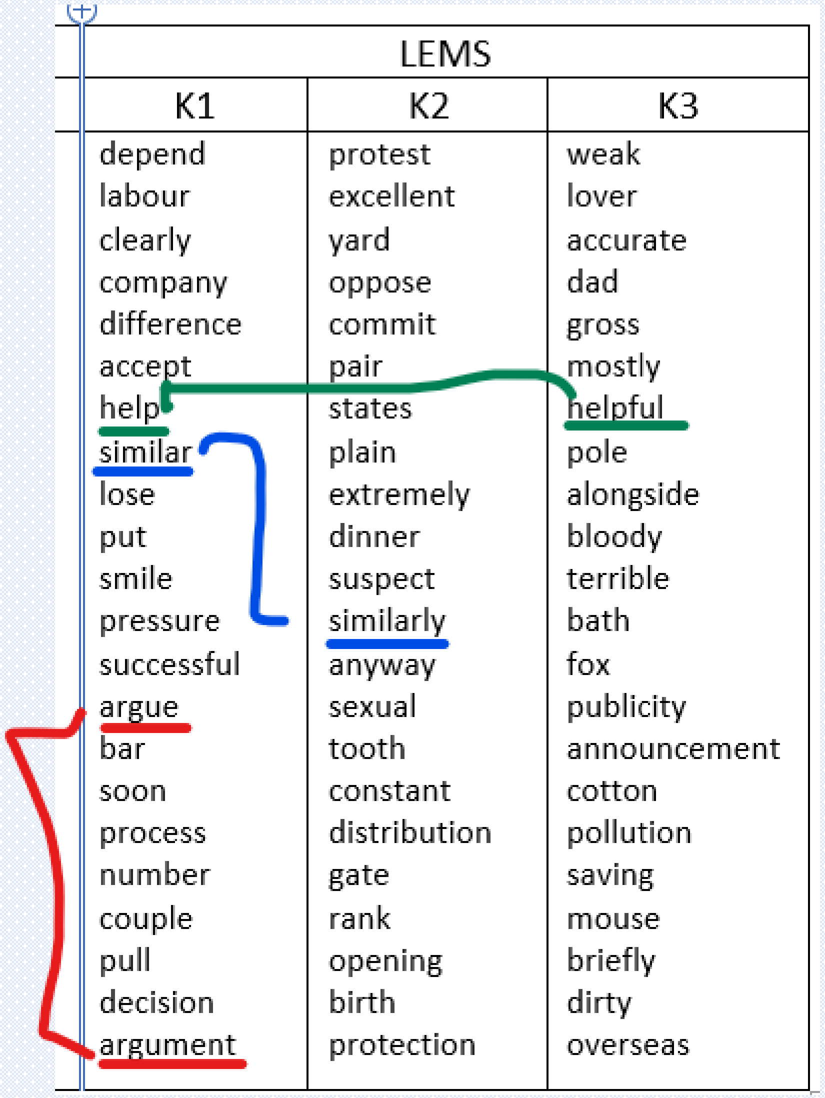
Level	**Tokens**	**Percent**	**Cumul%**
k-01	16,777	69.712	69.712
k-02	1,958	8.136	77.848
k-03	999	4.151	81.999
k-04	571	2.373	84.372
k-05	516	2.144	86.516
k-06	712	2.959	89.475
k-07	356	1.479	90.954
k-08	235	0.976	91.930
k-09	211	0.877	92.807
k-10	160	0.665	93.472
k-11	112	0.465	93.937
k-12	174	0.723	94.660
k-13	78	0.324	94.984
k-14	153	0.636	95.620
coverage=>95%			
k-15	97	0.403	96.023
k-16	73	0.303	96.326
k-17	57	0.237	96.563
k-18	77	0.320	96.883
k-19	40	0.166	97.049
k-20	49	0.204	97.253
k-21	28	0.116	97.369
k-22	132	0.548	97.917
k-23	35	0.145	98.062
coverage=>98%			

Matching counting units to list functions

2. Lists provide random test items



FAMS		
K1	K2	K3
nice	dot	stab
single	select	creep
motion	constant	manner
likely	rob	guest
couple	lend	supervise
drop	chop	outcome
lunch	consume	tack
deep	cigarette	phenomenon
appropriate	perform	bond
million	mistake	housewife
apply	criminal	vague
social	brochure	gee
can	sandwich	fuss
open	pencil	whiskey
under	despite	ham
positive	citizen	irritate
provide	accommodate	remote
oh	decent	visible
step	nerve	unique
heart	angle	astonish



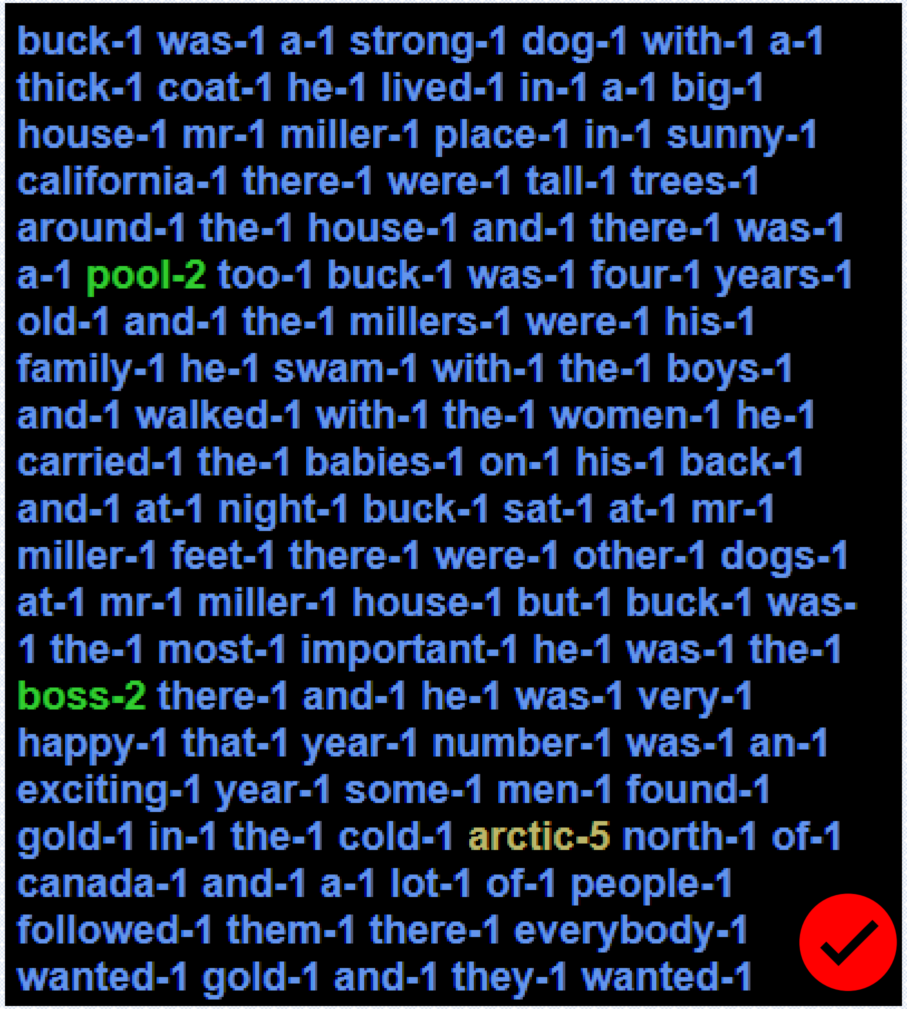
LEMS		
K1	K2	K3
depend	protest	weak
labour	excellent	lover
clearly	yard	accurate
company	oppose	dad
difference	commit	gross
accept	pair	mostly
help	states	helpful
similar	plain	pole
lose	extremely	alongside
put	dinner	bloody
smile	suspect	terrible
pressure	similarly	bath
successful	anyway	fox
argue	sexual	publicity
bar	tooth	announcement
soon	constant	cotton
process	distribution	pollution
number	gate	saving
couple	rank	mouse
pull	opening	briefly
decision	birth	dirty
argument	protection	overseas

At least 2 contaminated items in any 3 lists

Matching counting units to list functions

3. Lists for finding texts at/editing texts to a level

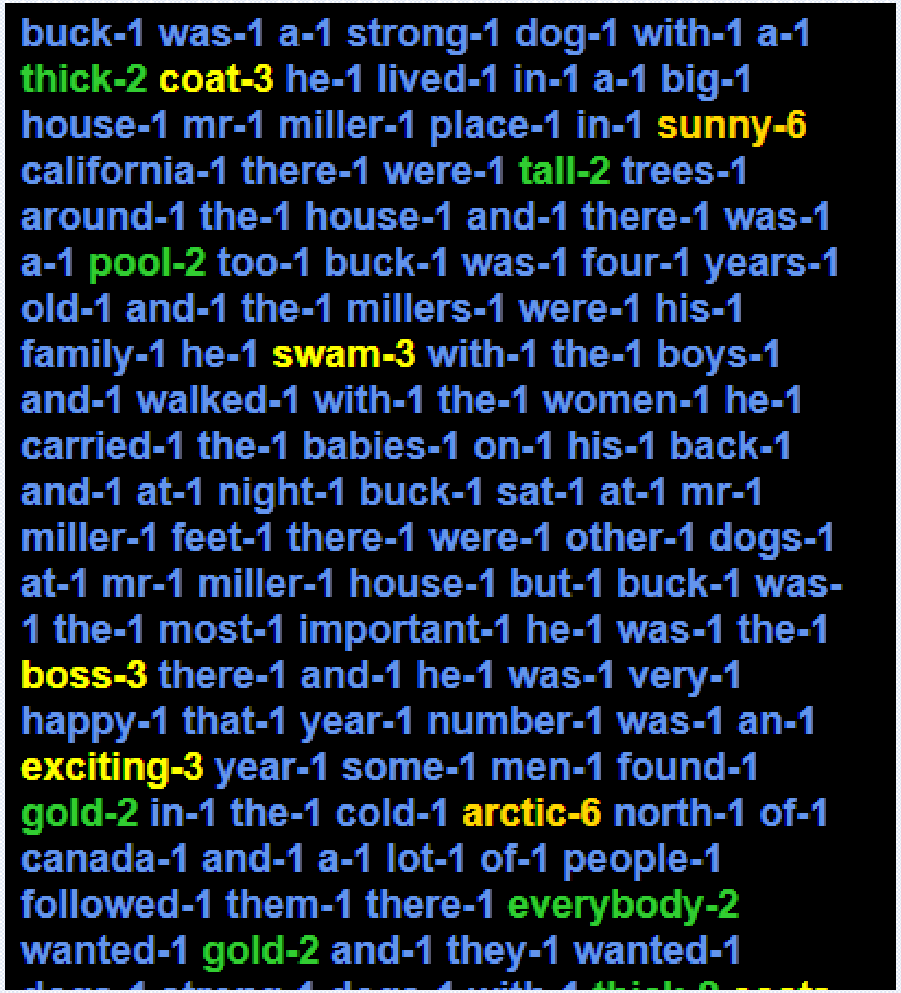
FAMILY (BNC/COCA)



buck-1 was-1 a-1 strong-1 dog-1 with-1 a-1 thick-1 coat-1 he-1 lived-1 in-1 a-1 big-1 house-1 mr-1 miller-1 place-1 in-1 sunny-1 california-1 there-1 were-1 tall-1 trees-1 around-1 the-1 house-1 and-1 there-1 was-1 a-1 pool-2 too-1 buck-1 was-1 four-1 years-1 old-1 and-1 the-1 millers-1 were-1 his-1 family-1 he-1 swam-1 with-1 the-1 boys-1 and-1 walked-1 with-1 the-1 women-1 he-1 carried-1 the-1 babies-1 on-1 his-1 back-1 and-1 at-1 night-1 buck-1 sat-1 at-1 mr-1 miller-1 feet-1 there-1 were-1 other-1 dogs-1 at-1 mr-1 miller-1 house-1 but-1 buck-1 was-1 the-1 most-1 important-1 he-1 was-1 the-1 boss-2 there-1 and-1 he-1 was-1 very-1 happy-1 that-1 year-1 number-1 was-1 an-1 exciting-1 year-1 some-1 men-1 found-1 gold-1 in-1 the-1 cold-1 arctic-5 north-1 of-1 canada-1 and-1 a-1 lot-1 of-1 people-1 followed-1 them-1 there-1 everybody-1 wanted-1 gold-1 and-1 they-1 wanted-1

✓

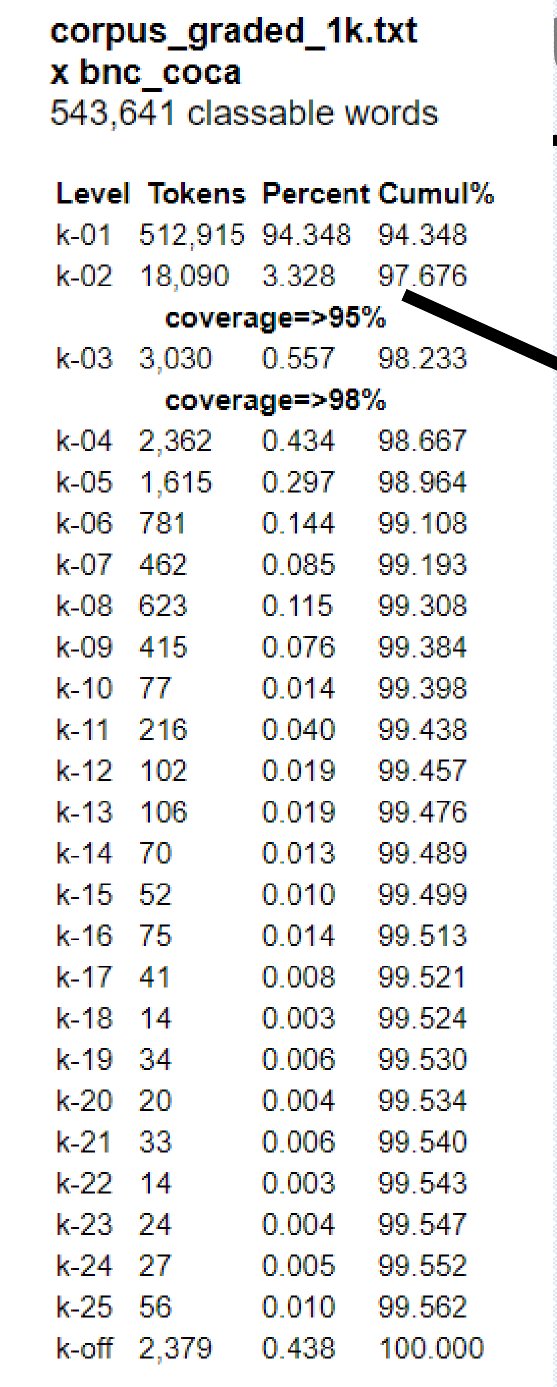
LEMMA (BNC)



buck-1 was-1 a-1 strong-1 dog-1 with-1 a-1 thick-2 coat-3 he-1 lived-1 in-1 a-1 big-1 house-1 mr-1 miller-1 place-1 in-1 sunny-6 california-1 there-1 were-1 tall-2 trees-1 around-1 the-1 house-1 and-1 there-1 was-1 a-1 pool-2 too-1 buck-1 was-1 four-1 years-1 old-1 and-1 the-1 millers-1 were-1 his-1 family-1 he-1 swam-3 with-1 the-1 boys-1 and-1 walked-1 with-1 the-1 women-1 he-1 carried-1 the-1 babies-1 on-1 his-1 back-1 and-1 at-1 night-1 buck-1 sat-1 at-1 mr-1 miller-1 feet-1 there-1 were-1 other-1 dogs-1 at-1 mr-1 miller-1 house-1 but-1 buck-1 was-1 the-1 most-1 important-1 he-1 was-1 the-1 boss-3 there-1 and-1 he-1 was-1 very-1 happy-1 that-1 year-1 number-1 was-1 an-1 exciting-3 year-1 some-1 men-1 found-1 gold-2 in-1 the-1 cold-1 arctic-6 north-1 of-1 canada-1 and-1 a-1 lot-1 of-1 people-1 followed-1 them-1 there-1 everybody-2 wanted-1 gold-2 and-1 they-1 wanted-1

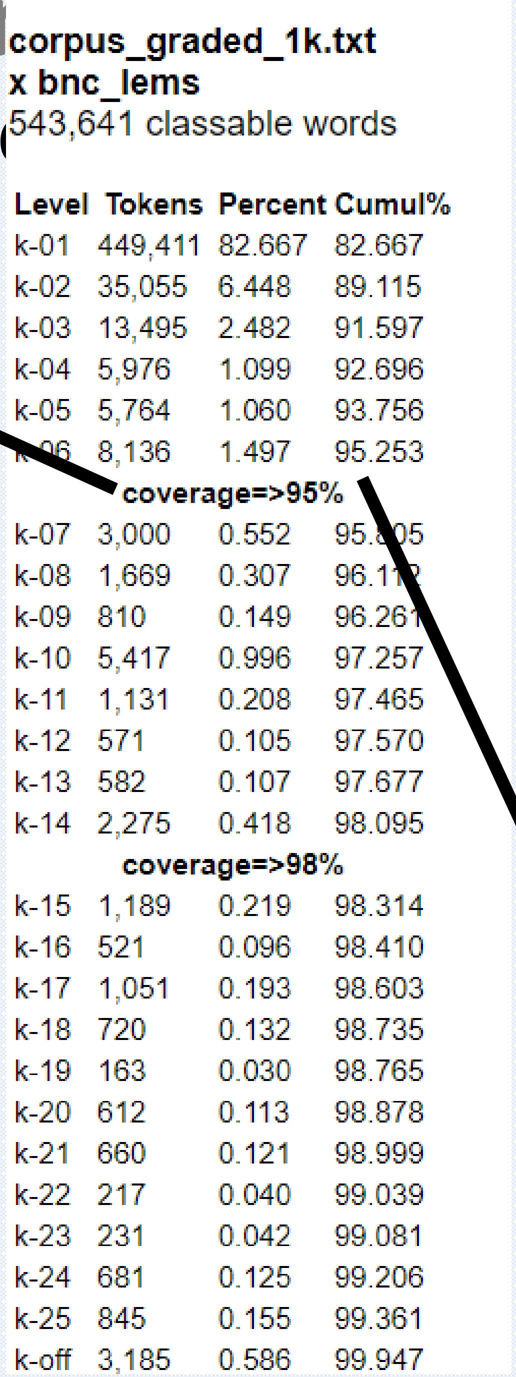
M untin list f ti

4 age r



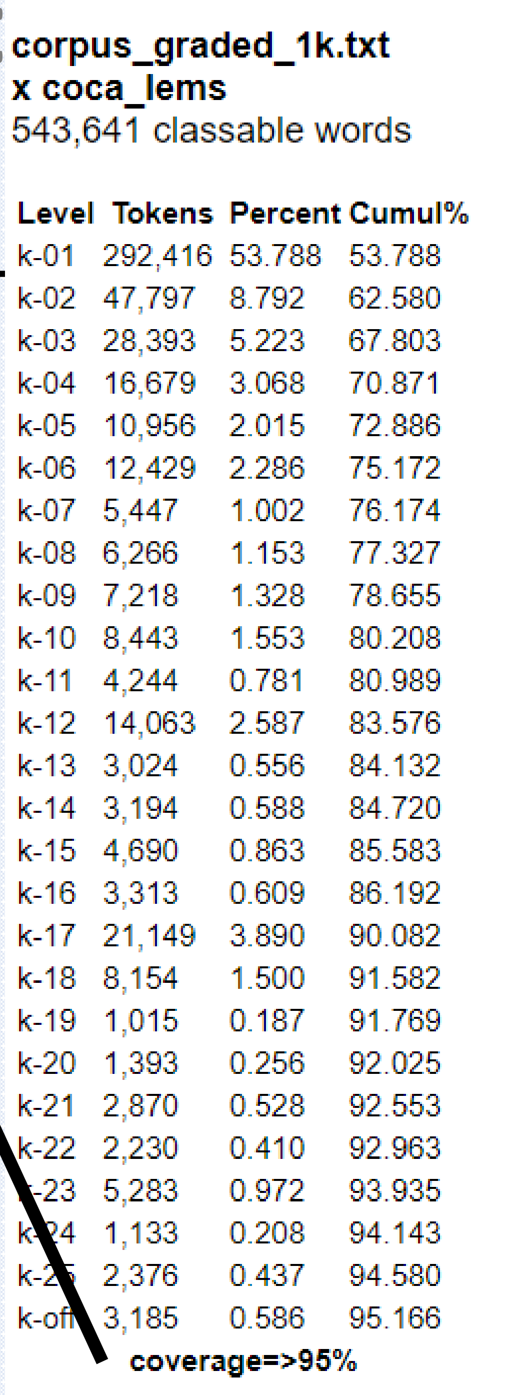
corpus_graded_1k.txt
x bnc_coca
543,641 classable words

Level	Tokens	Percent	Cumul%
k-01	512,915	94.348	94.348
k-02	18,090	3.328	97.676
	coverage=>95%		
k-03	3,030	0.557	98.233
	coverage=>98%		
k-04	2,362	0.434	98.667
k-05	1,615	0.297	98.964
k-06	781	0.144	99.108
k-07	462	0.085	99.193
k-08	623	0.115	99.308
k-09	415	0.076	99.384
k-10	77	0.014	99.398
k-11	216	0.040	99.438
k-12	102	0.019	99.457
k-13	106	0.019	99.476
k-14	70	0.013	99.489
k-15	52	0.010	99.499
k-16	75	0.014	99.513
k-17	41	0.008	99.521
k-18	14	0.003	99.524
k-19	34	0.006	99.530
k-20	20	0.004	99.534
k-21	33	0.006	99.540
k-22	14	0.003	99.543
k-23	24	0.004	99.547
k-24	27	0.005	99.552
k-25	56	0.010	99.562
k-off	2,379	0.438	100.000



corpus_graded_1k.txt
x bnc_lems
543,641 classable words

Level	Tokens	Percent	Cumul%
k-01	449,411	82.667	82.667
k-02	35,055	6.448	89.115
k-03	13,495	2.482	91.597
k-04	5,976	1.099	92.696
k-05	5,764	1.060	93.756
k-06	8,136	1.497	95.253
	coverage=>95%		
k-07	3,000	0.552	95.805
k-08	1,669	0.307	96.112
k-09	810	0.149	96.261
k-10	5,417	0.996	97.257
k-11	1,131	0.208	97.465
k-12	571	0.105	97.570
k-13	582	0.107	97.677
k-14	2,275	0.418	98.095
	coverage=>98%		
k-15	1,189	0.219	98.314
k-16	521	0.096	98.410
k-17	1,051	0.193	98.603
k-18	720	0.132	98.735
k-19	163	0.030	98.765
k-20	612	0.113	98.878
k-21	660	0.121	98.999
k-22	217	0.040	99.039
k-23	231	0.042	99.081
k-24	681	0.125	99.206
k-25	845	0.155	99.361
k-off	3,185	0.586	99.947



corpus_graded_1k.txt
x coca_lems
543,641 classable words

Level	Tokens	Percent	Cumul%
k-01	292,416	53.788	53.788
k-02	47,797	8.792	62.580
k-03	28,393	5.223	67.803
k-04	16,679	3.068	70.871
k-05	10,956	2.015	72.886
k-06	12,429	2.286	75.172
k-07	5,447	1.002	76.174
k-08	6,266	1.153	77.327
k-09	7,218	1.328	78.655
k-10	8,443	1.553	80.208
k-11	4,244	0.781	80.989
k-12	14,063	2.587	83.576
k-13	3,024	0.556	84.132
k-14	3,194	0.588	84.720
k-15	4,690	0.863	85.583
k-16	3,313	0.609	86.192
k-17	21,149	3.890	90.082
k-18	8,154	1.500	91.582
k-19	1,015	0.187	91.769
k-20	1,393	0.256	92.025
k-21	2,870	0.528	92.553
k-22	2,230	0.410	92.963
k-23	5,283	0.972	93.935
k-24	1,133	0.208	94.143
k-25	2,376	0.437	94.580
k-off	3,185	0.586	95.166
	coverage=>95%		

So family and lemma are both fatally flawed

Family is superior for almost every purpose

Except one big one: cannot be given to learners directly

Is there a way to reconcile family and lemma?

A new suggested unit of word counting – A Nuclear Family

NF includes the most frequent family members - base words and affixed words

Extended family (BNC/COCA)

apply, applies, applied, application, applications, applicable, applicability, reapply, reapplies, reapplied, reapplication, reapplications, disapplication

(13 word types, 8 lemmas)

Nuclear family

apply, application, applications, applied (4 word types, 3 lemmas)

NFL7 – a reduced BNC/COCA 3000 list (Cobb & Laufer, 2021)

BNC/COCA	19,065 word types; 9,132 lemmas; 81 derivational affixes
NFL	7,293 word types; 5,610 lemmas; 22 derivational affixes

Validity of Nuclear Family Lists – empirical evidence

1. Texts that learners read include

   a limited number of derived words (family members)

   a limited number of frequent affixes (Laufer & Cobb, 2020)

   Hence, no need to learn extended families

2. Nuclear Family Lists provide a good coverage of authentic texts

   Compared with BNC/COCA 3000

   NFL7 - 4% less text coverage, but 11,800 fewer word types

   Hence, good cost/benefit deal (Cobb & Laufer, 2021)

3. **To be demonstrated in the present study**

   Family size changes according to text difficulty

   Hence, learners at different learning stages require different lists

Family size and language level

Hypothesis

The number of derived words in texts is different at different language levels

(Family size in texts expands as language level in texts progresses)

If the hypothesis is correct

Word lists for learners will differ in family size depending on the expected language proficiency

Aim

To examine differences in word family sizes in texts of different language difficulty

Corpora examined

OUP Graded readers Level 3	(123,771 words)
OUP Graded readers Level 5	(181,586 words)
OUP Graded readers Level 6	(230,869 words)
Mid frequency readers Level 8 (P. Nation’s resources)	(500,000 words)
Emma	(161,011 words)
Academic texts (BAWE, RinFL)	(175,000 words)
Combo corpus (Lextutor) (spoken/written; general/academic; Am./Brit.)	(3.7 m words)

Method

1. Corpus Profiling

Text lexis covered by k1, k2, k3 etc.

(Tool – VocabProfile) https://www.lextutor.ca/vp/

Morphological makeup - percentage of derived words

(Tool – Morpholex) https://www.lextutor.ca/morpho/

2. Matching BNC/COCA lists (e.g., k1, k2) to uploaded target corpora

Tool (Nuclear List Builder) https://www.lextutor.ca/freq/nuclear/

The resulting list shows base words + derived words from BNC/COCA that appear in the target corpus, e.g., in *Graded Readers, level 6*

Matching BNC/COCA lists to examined corpora ------------ >

3. **Extracting identical base words from the lists and comparing their derived forms, i.e. comparing family sizes**

e.g., How many derived words of *center* are there in the examined corpora and what are they?

Work in progress - so far - 75 word families examined

Results

Corpora features: lexical difficulty level and percentage of derivations

Corpus	% Text	Coverage	% of derived words in text
	by 2k	by 3k	
Graded level 3	98	98.5	2
Graded level 5	97	97.7	4
Graded level 6	96.3	97.5	5
Emma	93.8	96	5
Mid freq. readers 8k	91.3	94	5
Academic RinFL BAWE	84.4 83	91.6 92	10 10
Combo	85.8	90.2	7.7

- Text difficulty increases
- % of derived words increases (in most cases)

Text level and Word family size

Graded 3	Graded 5	Graded 6	Emma	Mid freq	Academic	Combo
centre	centre central	centre	centre	centre central	centre/ center centered central centrality centralization centrally centric	centre/center centered centering central centralization centralized centrally centrist
excite excitement exciting	excitedly excitement exciting	excite excitedly excitement exciting	excite excitement	excite excitable excitation excitedly excitement exciting	excitement	excitable excitation excitedly excitement exciting unexciting

BNC/COCA

Center/centre
Centrist
Centring
Centered
Centredness
Central
Centralism
Centralist
Centrally
Centrality
Centralize
Centralized
Centralization
Centralizing

Text level and Word family size

Graded 3	Graded 5	Graded 6	Emma	Mid freq	Academic	Combo
careful carefully carelessness	care careful carefully careless	careful carefully careless uncaring	care careful carefully carefulness careless carelessly carelessness	care careful carefully carefulness careless carelessly carelessness carer uncared	care careful carefully careless carelessness carer	care careful carefully careless carelessly uncaring
-----------	expression expressionless	express expression expressionless	express expression expressive expressly inexpressible	express expression expressionless expressive expressly inexpressible unexpressed	express expression	express expressible expression expressionless expressive expressly inexpressible

BNC/COCA

Express
Expressed
Unexpressed
Expressing
Expression
Expressionless
Expressionlessly
Expressive
Expressively
Expressiveness
Expressly

Text level and Word family size

Graded 3	Graded 5	Graded 6	Emma	Mid freq	Academic	Combo
fair	fair unfair	fairly fairness unfair unfairly	fair fairly unfair	fair fairly unfair unfairly	fair fairly unfair	fair fairly fairness unfair unfairly unfairness
exist	exist existence	exist existence	exist existence	exist existence existent	exist existence	exist existence existent nonexistent
---------------	organize organization	organize organization organizer reorganize	------------------	organisation organise	organised organizer organisation	organize organized organization organizational organizationally organizer reorganization

Text level and Word family size

Graded 3	Graded 5	Graded 6	Emma	Mid freq	Academic	Combo
--------------	attractive	attract attraction attractive attractiveness	attract attraction attractive	attract attraction attractive attractively attractiveness unattractive	attract attractive attractiveness	attract attraction attractive attractively attractiveness attractor unattractive
pleasant unpleasant	pleasant pleasantly unpleasant unpleasantness	pleasant pleasantly unpleasant unpleasantly unpleasantness	pleasant pleasantly pleasantness unpleasant	pleasant pleasantly pleasantry unpleasant unpleasantly unpleasantness	-------------	pleasant pleasantly pleasantries unpleasant unpleasantly unpleasantness

Conclusion

Derived words are not distributed equally in the language

Their percent is different in texts of different difficulties

(Their percent is also different in different text genres (Laufer & Cobb, 2020)

Texts with easy, basic vocabulary --- few derived words
More difficult texts, more complex vocabulary --- more derived words
larger word families

Even if derived words appear in very large corpora (BNC/COCA) they do not necessarily appear in a large number of texts

Implications

1. Learners do not need to know entire word families even of the most frequent base words

Additional family members will be encountered as text language difficulty increases

What they need – awareness of morphological regularities

2. Vocabulary tests using word family as the counting unit do not overestimate learners' receptive vocabulary knowledge

Our assumption (supported by data) - learners understand the derived words they **need** for reading texts at their level

3. Nuclear Family Lists – solution for specific vocabulary targets for specific learning materials, as they include only the necessary family members

One family size does not fit all word lists →

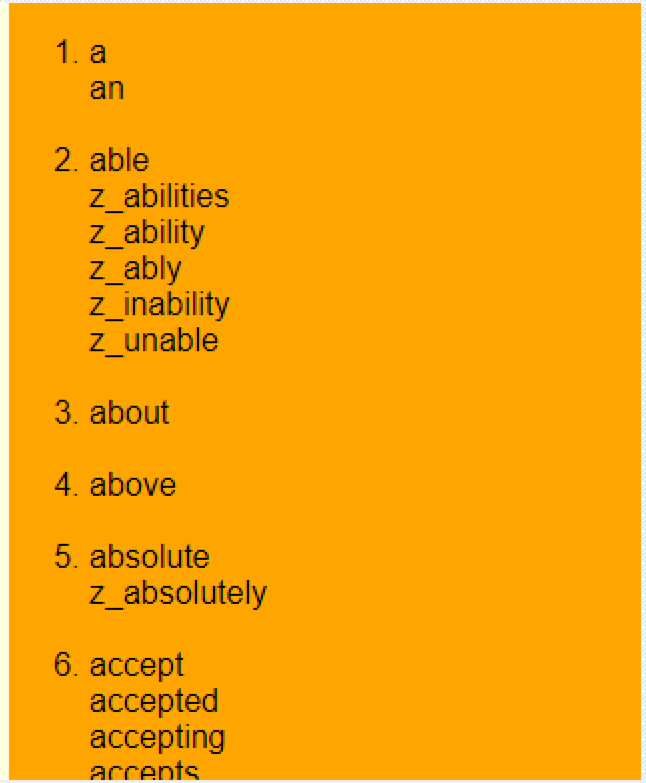
So where will these 'different lists' come from?

How will they be constructed?

To predict learners' needs, give them lists, design their materials,

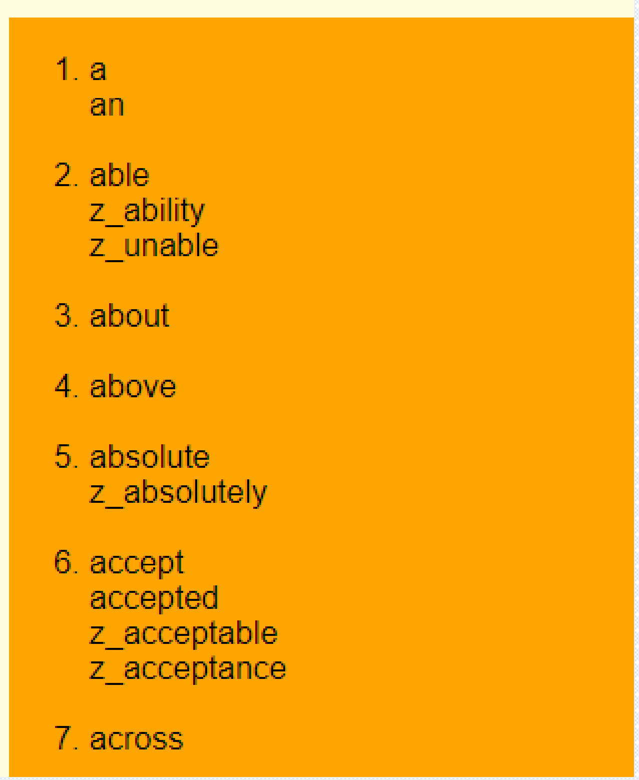
Nuclear List Builder can reduce/expand family size systematically

Including **all** members of
BNC/COCA 1-K
1,000 families
:: 6,849 word types
:: 2,057 derived words (='z_')



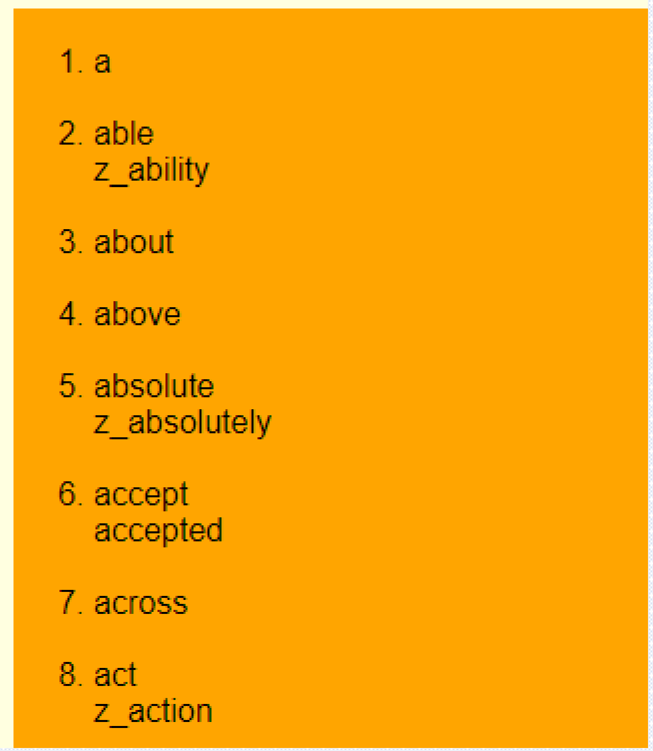
```
1. a
   an

2. able
   z_abilities
   z_ability
   z_ably
   z_inability
   z_unable

3. about

4. above

5. absolute
   z_absolutely

6. accept
   accepted
   accepting
   accepts
```

Including only members
>7% of their families
1,000 families
:: 2,316 word types
::   352 derived words



```
1. a
   an

2. able
   z_ability
   z_unable

3. about

4. above

5. absolute
   z_absolutely

6. accept
   accepted
   z_acceptable
   z_acceptance

7. across
```

Including only members
>15% of their families
1,000 families
:: 1,712 word types
::   194 derived words



```
1. a

2. able
   z_ability

3. about

4. above

5. absolute
   z_absolutely

6. accept
   accepted

7. across

8. act
   z_action
```