
University of Pittsburgh
Combining Formal and Usage based Theories with Data Science Techniques in Measuring the Development of Syntactic Complexity in Written Production
https://github.com/ELI-Data-Mining-Group/
Alan Juffs and Na-Rae Han
juffs@pitt.edu
Department of Linguistics
Kenneth P. Dietrich School of Arts and Sciences

Collaborators

- Ben Naismith – presenting at TESOL
- Daniel Zheng

- Funding
- Pittsburgh Science of Learning Center
- http://www.learnlab.org

Principled Eclecticism in SLA research

- 1. Data Science – slice and dice large data sets
- 2. Formal linguistics: syntactic categories
- 3. Corpus linguistics: frequency counts
- 4. Psycholinguistics: experiments in processing
- Each has a contribution to make: not mutually exclusive and can be combined
- 'Closing the loop' with educators

1. Data Science Tools

- http://www.pitt.edu/~naraehan/ling1340/Lecture1.pdf
- Python: can import very large .csv files
  - Pandas – Data frames to organize and label data
  - Numpy - mathematical operations/statistics
    - Competing with R
  - NLTK – lexical tools (well known)
  - Matplotlib – data visualization tools
- ELI tool kit developed in Python – Daniel Zheng
  - Lemmatizer – Someya list (!)
  - D measure of diversity
  - Advanced Guiraud – sophistication – growth of less frequent lexical items dependent on L1/culture
    - (Juffs, 2019; Naismith et al. 2018)

Pitt

Import Data: tokenize, lemmatize, POS tag

In [2]:
```
combo_dir = "../../Data-Archive/"
# Read two pickles as separate DFs
combo1 = pd.read_pickle(combo_dir + 'combo_df_1A_pt1.pkl')
combo2 = pd.read_pickle(combo_dir + 'combo_df_1A_pt2.pkl')

# Unpickle two DFs, and then concatenate along axis 1 (y axis)
combo_df = pd.concat([combo1, combo2], axis=1)
combo_df.head()

# Voila! One big combo_df, ready to dice.
```

Out[2]:

er_id	question_id	anon_id	user_file_id	text	class_code	level_id	native_language	version	toks_re	toks_re_len	toks_nltk	toks_pos	lemmas
1	5	eq0	7505	I met my friend Nife while I was studying in a...	g	4	Arabic	1	[i, met, my, friend, nife, while, i, was, stud...	177	[I, met, my, friend, Nife, while, I, was, stud...	[(I, PRP), (met, VBD), (my, PRP$), (friend, NN...	[i, meet, my, friend, nife, while, i, be, stud...
2	5	am8	7506	Ten years ago, I met a women on the train betw...	g	4	Thai	1	[ten, years, ago, i, met, a, women, on, the, t...	137	[Ten, years, ago, ,, I, met, a, women, on, the...	[(Ten, CD), (years, NNS), (ago, RB), (,, ,), (...	[ten, year, ago, ,, i, meet, a, woman, on, the...
3	12	dk5	7507	In my country we usually don't use tea bags. F...	w	4	Turkish	1	[in, my, country, we, usually, don, t, use, te...	63	[In, my, country, we, usually, do, n't, use, t...	[(In, IN), (my, PRP$), (country, NN), (we, PRP...	[in, my, country, we, usually, do, n't, use, t...
4	13	dk5	7507	I organized the instructions by time.	w	4	Turkish	1	[i, organized, the, instructions, by, time]	6	[I, organized, the, instructions, by, time, .]	[(I, PRP), (organized, VBD), (the, DT), (instr...	[i, organize, the, instruction, by, time, .]
5	12	ad1	7508	First, prepare a port, loose tea, and cup.\nSe...	w	4	Korean	1	[first, prepare, a, port, loose, tea, and, cup...	59	[First, ,, prepare, a, port, ,, loose, tea, ,,...	[(First, RB), (,, ,), (prepare, VB), (a, DT), ...	[first, ,, prepare, a, port, ,, loose, tea, ,,...

2. Formal Linguistics and Corpus Analysis: Making Predictions: what to look for

- White (1987)
  - Dative alternation: not restricted to input (c.f. C. L. Baker (1979)
  - 'Mummy, open **Hadwen** the door'
- Zobl (1989)
  - Unaccusative/ Unergative verbs: passive overused with unaccusatives
- Schwartz & Sprouse (1996) L2 German (Cevdet)
- Lardière's work
  - Feature reassembly (Patti) and L2 Korean plurals (e.g., Hwang & Lardiere, 2013)

Lexical and Functional Categories

- Verbs – Pinker (1989); Juffs (1996)
- Tense and morphology:
  - Prevost & White (2000): French L2 finite/non-finite
- if, whether - CP
  - Sam wonders whether it will rain.
- the, that, ’s - DP
  - Sam’s book
  - *The Sam’s book

3. Corpora and Usage-Based Approaches

- Frequency ranks- work in NLP .. Crossley, Kyle, Jarvis – Sunday's colloquium
- N. Ellis (2016, pp. 44-46)
  - Ortega (2001): Longitudinal data needed
  - Few longitudinal L2 corpora available
- Debates about which measures of association are most relevant to acquisition and processing
  - Must rely on 'hundreds of millions of words to approximate usage' (Ellis, 2016, p. 44).
- Too pessimistic? Predict (some) development based on formal theories?

Pitt

Bley-Vroman (2002)

- In the meaning-based approach, the statistical structure of the language can affect the development of linguistic knowledge (for example, by influencing acquisition order or providing evidence for developing grammars);
- However, *linguistic* knowledge is NOT itself knowledge of the statistical structure of language
- We need to consider what learners want to communicate as well
- Argument elaborated on by Yang (2008)

Pitt

4. Psycholinguistics research: norming: Kennison (1999)

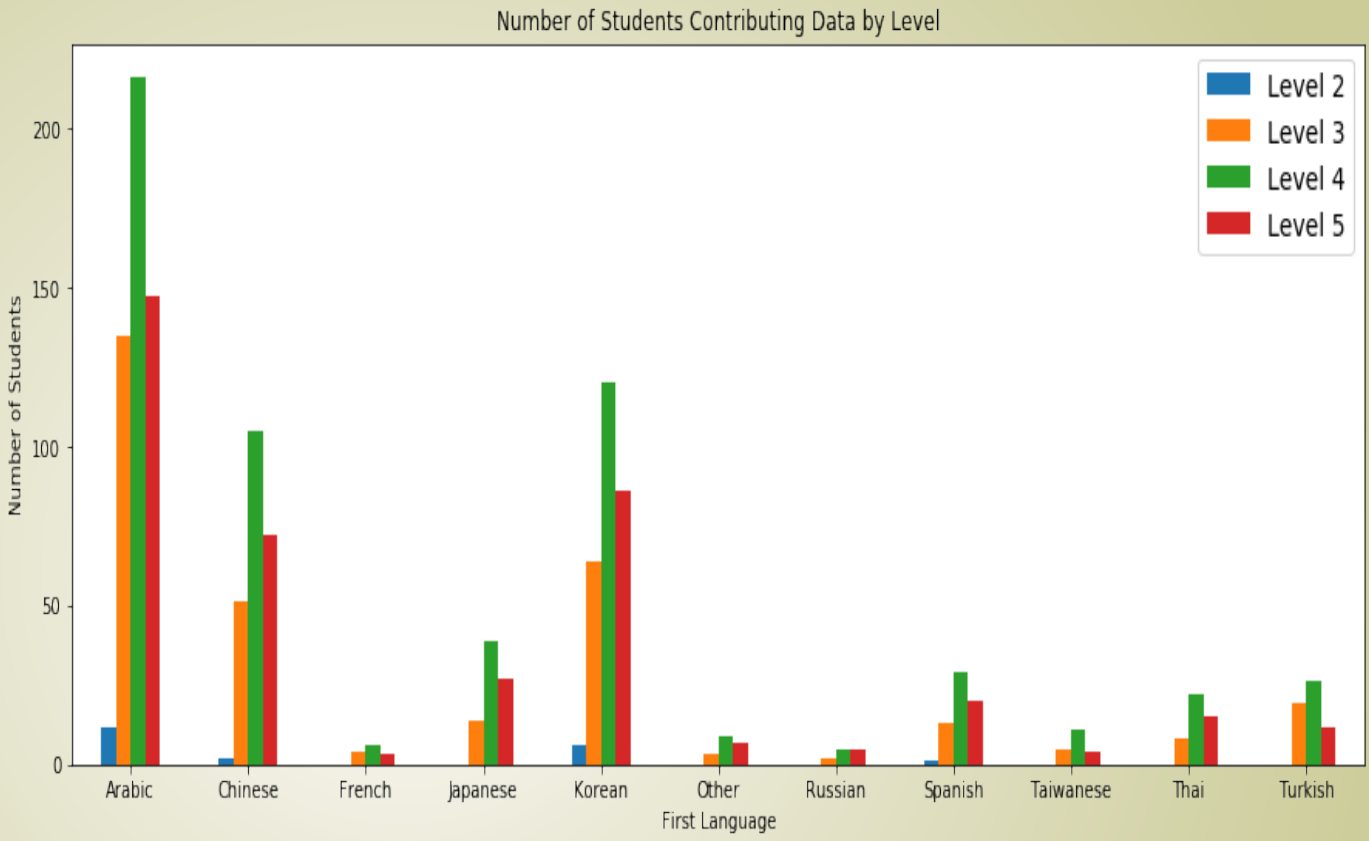
√Sam knows [$_{NP}$ the answer]. √Sam knows [$_{CP}$ the answer is correct]

?Sam supposed [$_{NP}$ the answer].√Sam supposed [$_{CP}$ the answer is correct]

Verb	Difficulty for learners? COCA Frequency?	NP %	CP %
consider	☺	93	0
suggest		32	59
explain		82	4
realize			CP
admit		14	42
deny		78	11
conclude		25	63
recommend		55	44
suppose			CP

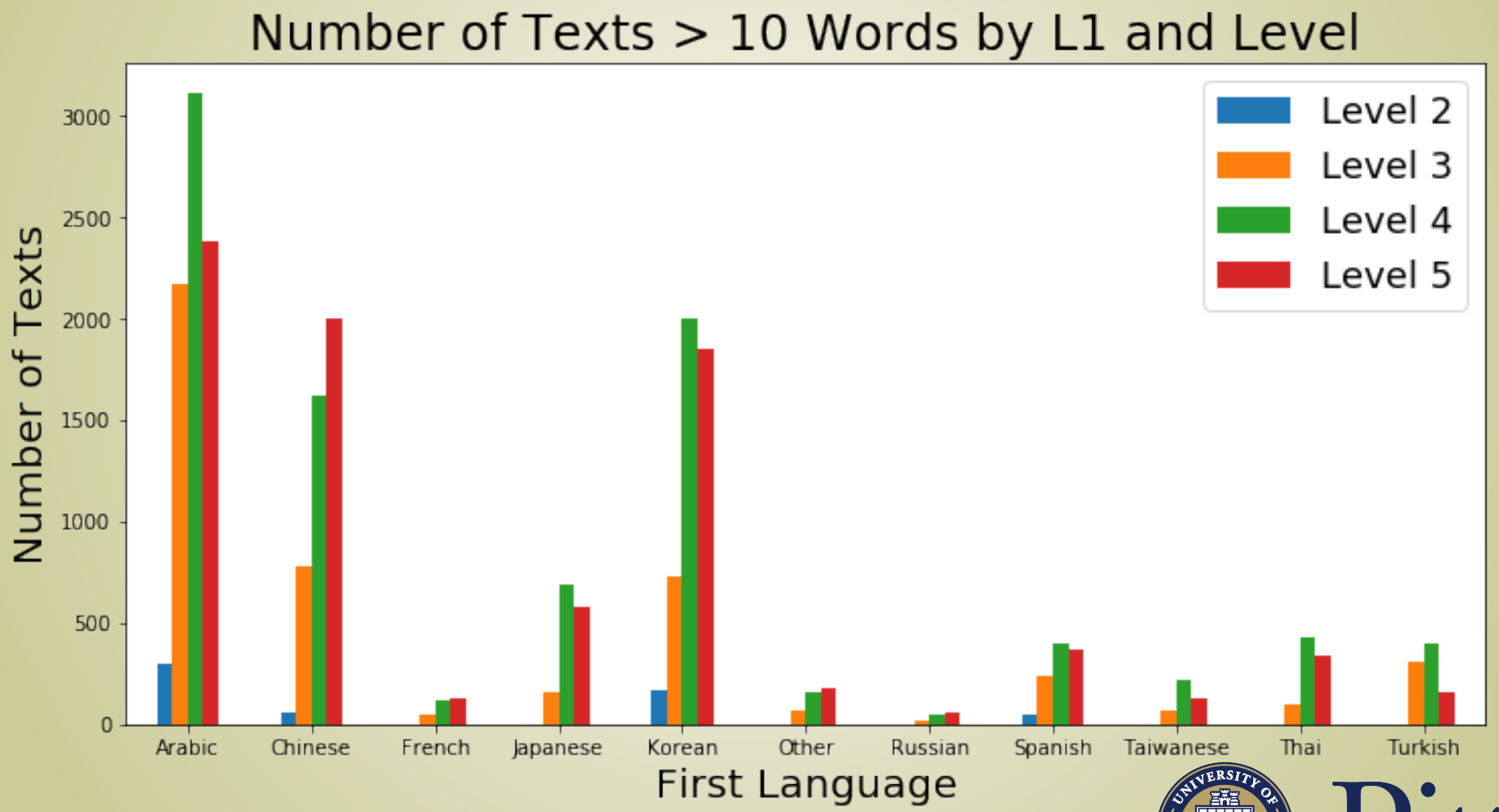
Pitt ELI Corpus:

Reported L1 (not country)



- PELIC Large and longitudinal. ‘’In the wild!”
- https://github.com/ELI-Data-Mining-Group/Pitt-ELI-Corpus

Number of texts



Pitt IEP Levels and Cut Scores

(No beginners)

Level	ept Combined	ELI Writing	CEFR Equiv.
2. High Beginner	28-37	2.1-2.9	A1 (Breakthrough)
3. Low Intermediate	38 - 47	3.1 – 3.9	A2-B1 (Waystage)
4. Intermediate	48 - 59	4.0 – 4.9	B1 (Threshold)
5. High Intermediate	60 - 68	5.0 – 5.9	B2 - edge of C1 (Vantage)
6. Low Advanced	69 +	6.0	Low C1 (Effective)
		"Writing used in borderline cases."	

All Written Data: ‘first version’

Level	Token Count	Per Mil Multipl
3	524,137	1.9
4	1,628,232	0.61
5	1,462,346	0.68
Total	3,614,715	

Data: Re-Token Counts x L1 and Level

Level	Arabic	Chinese	Korean	Total
3	157,727	71,431	74,813	303,971
4	398,333	278,964	296,291	973,588
5	364,614	298,430	259,615	922,659
Total	920,674	648,825	630,719	2,200,218

Create Frequency Ranks: L1 x Level

COCA Word Frequency Ranks

Rank	Word	Part of speec	Frequency
1	the	a	22038615
2	be	v	12545825
3	and	c	10741073
4	of	i	10343885
5	a	a	10144200
6	in	i	6996437
7	to	t	6332195
8	have	v	4303955
9	to	i	3856916
10	it	p	3872477
11	I	p	3978265
12	that	c	3430996
13	for	i	3281454
14	you	p	3081151
15	he	p	2909254
16	with	i	2683014
17	on	i	2485306
18	do	v	2573587
19	say	v	1915138
20	this	d	1885366
21	they	p	1865580
22	at	i	1767638
23	but	c	1776767

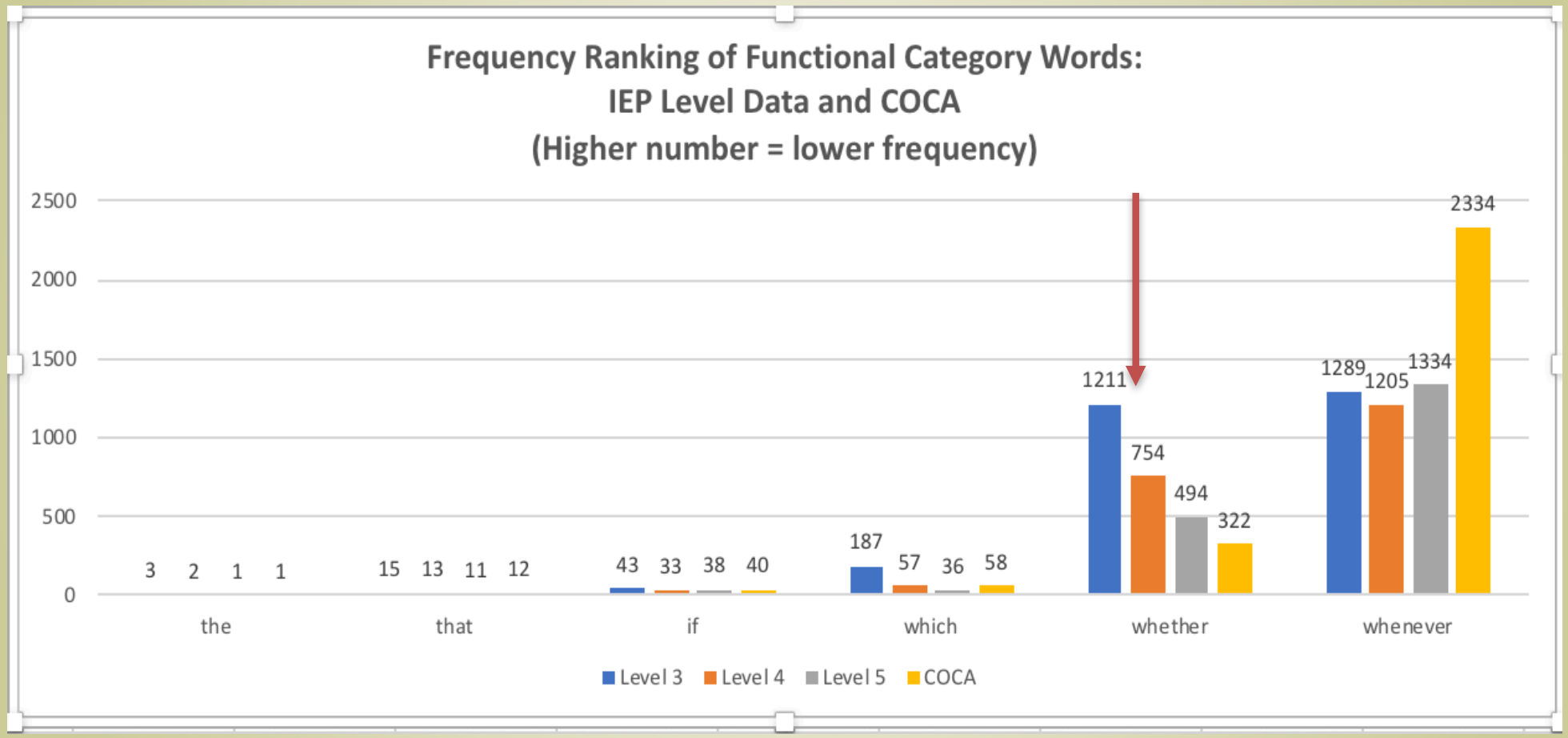
Korean Level 3 – freq. per mil

	A	B	C
1		lemma	count
2	1	be	44083
3	2	i	35716
4	3	the	33163
5	4	to	30623
6	5	a	29139
7	6	and	22737
8	7	of	18847
9	8	in	17337
10	9	have	17056
11	10	you	13674
12	11	for	11054
13	12	my	10600
14	13	that	9517
15	14	it	8769
16	15	do	7860
17	16	he	7525
18	17	can	6550
19	18	n't	6376
20	19	we	5988
21	20	they	5788

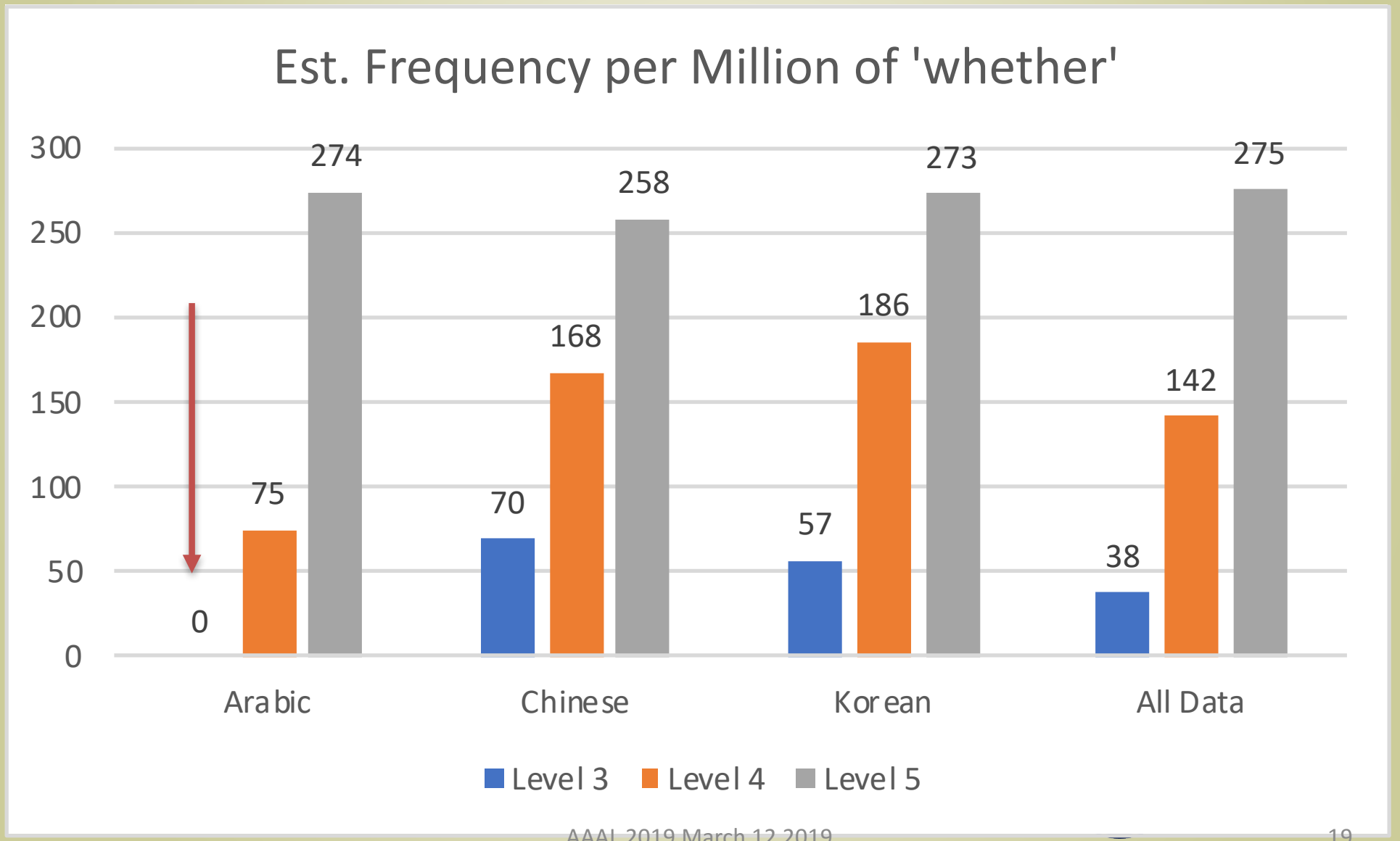
Research Questions

- To what extent do *selected* lexical items in the COCA frequency ranking reflect frequency rankings of those words in ESL students written output?
- What are the frequency ranks of single Functional Category words (associated with morpho-syntactic complexity) in the most frequent 3000 lemmas of L2 written output?
- What are the frequency rankings of verbs requiring complex syntax? E.g., ‘know’ and ‘suppose’
- Can these results inform classroom practice = closing the loop?

Functional Categories: All Data



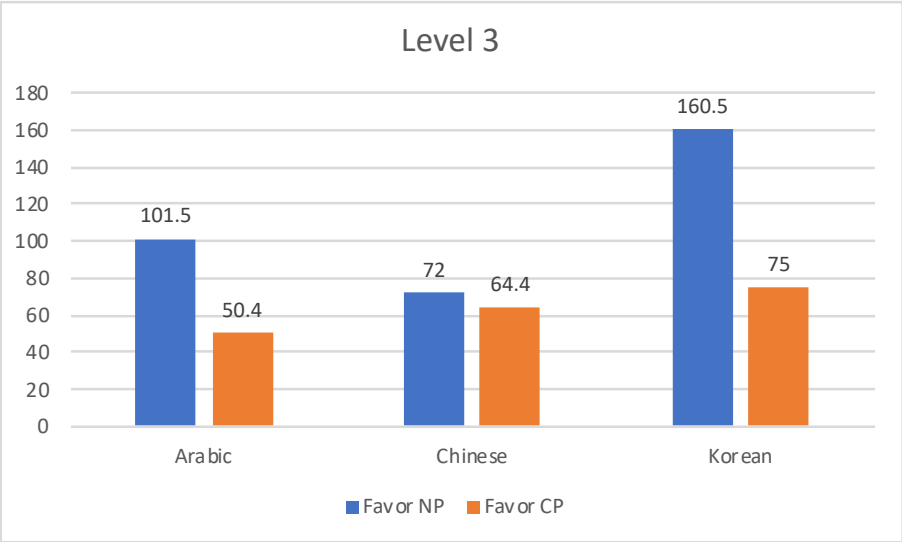
Development of ‘whether’

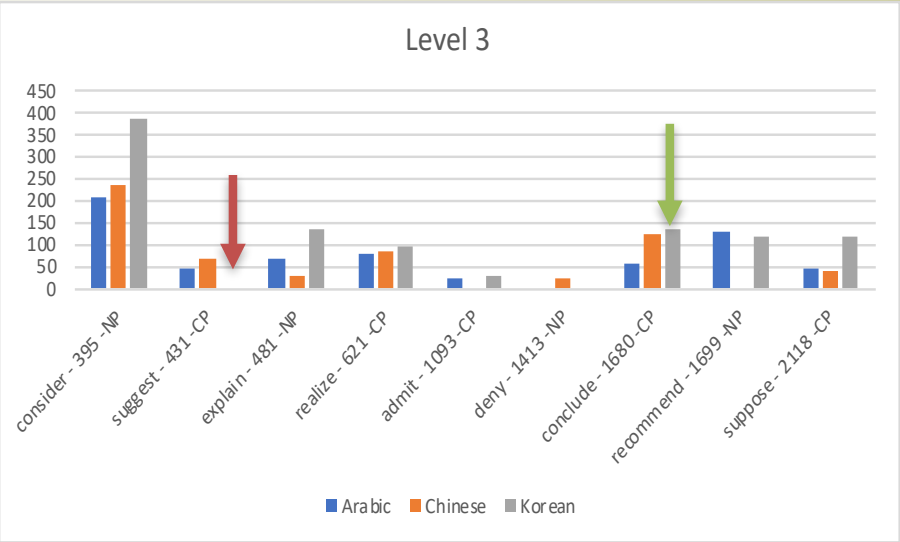


Verbs: Kennison (1999)

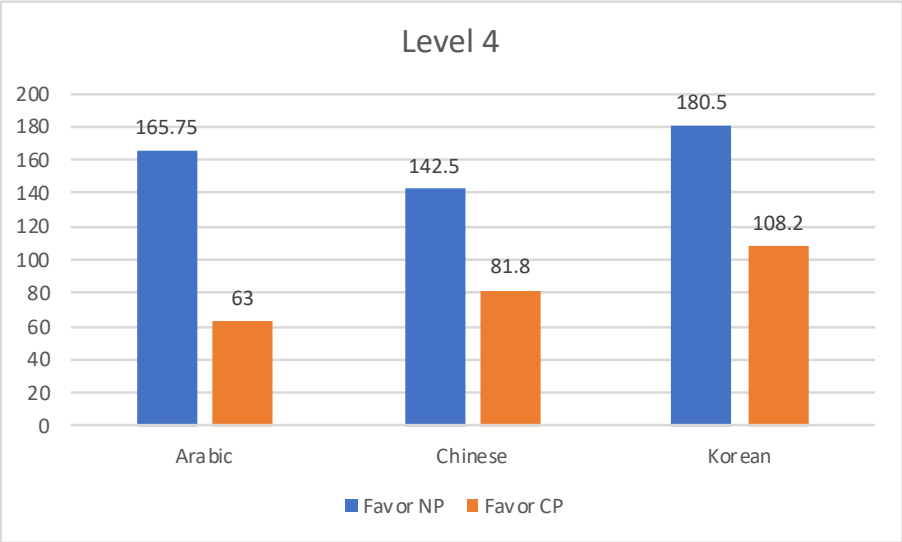
	COCA Rank	NP	CP	BNC-COCA-25	IEP List?
consider	395	x		1000	considerable
suggest	431		x	1000	x
explain	481	x		1000	x
realize	621		x	1000	x
admit	1093		x	1000	x
deny	1413	x		2000	√
conclude	1680		x	3000	√
recommend	1699	x		2000	x
suppose	2118		x	1000	x

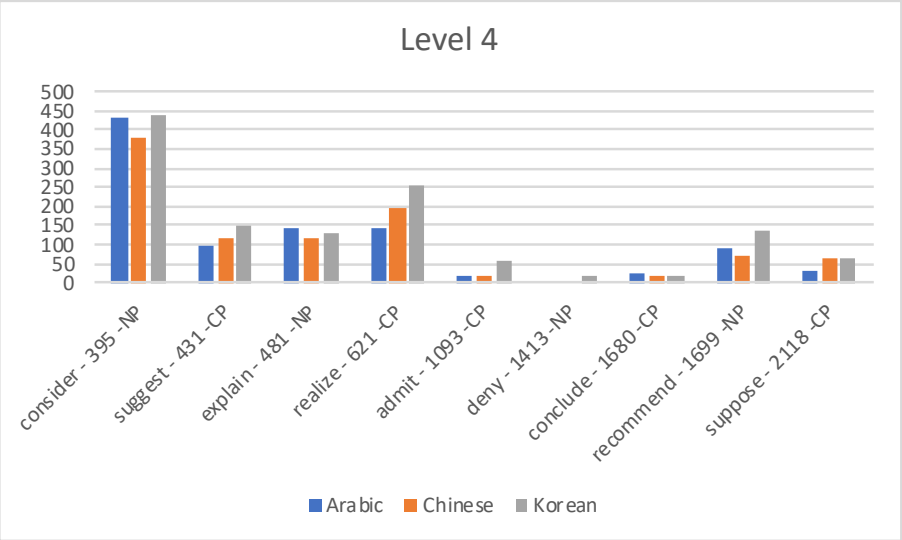
Verbs:
estimated frequency per million for comparison



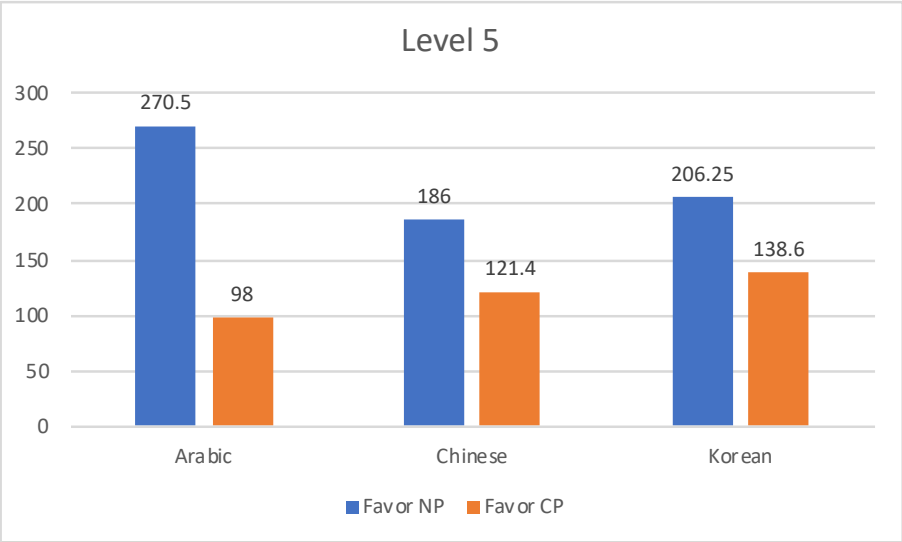


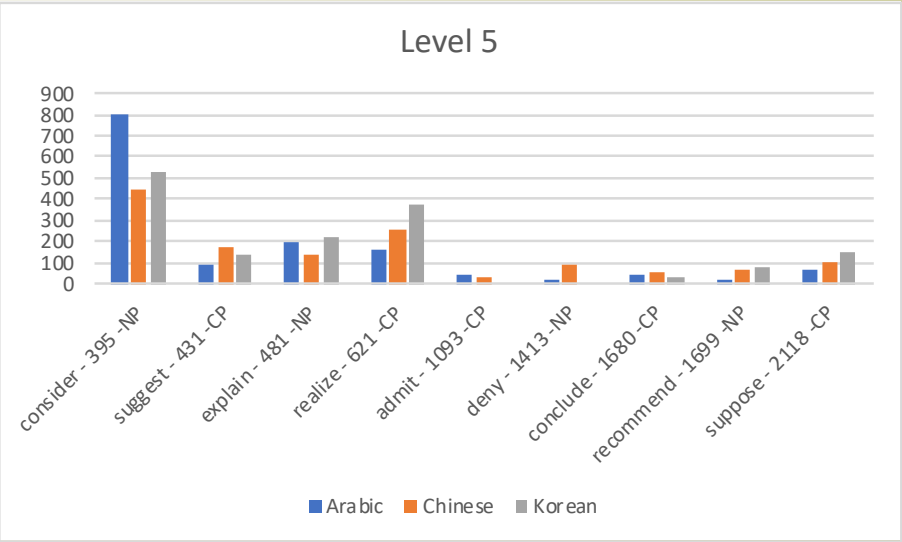
Level 4


Level 4
200
180
160
140
120
100
80
60
40
20
0
165.75
63
142.5
81.8
180.5
108.2
Arabic
Chinese
Korean
Favor NP
Favor CP

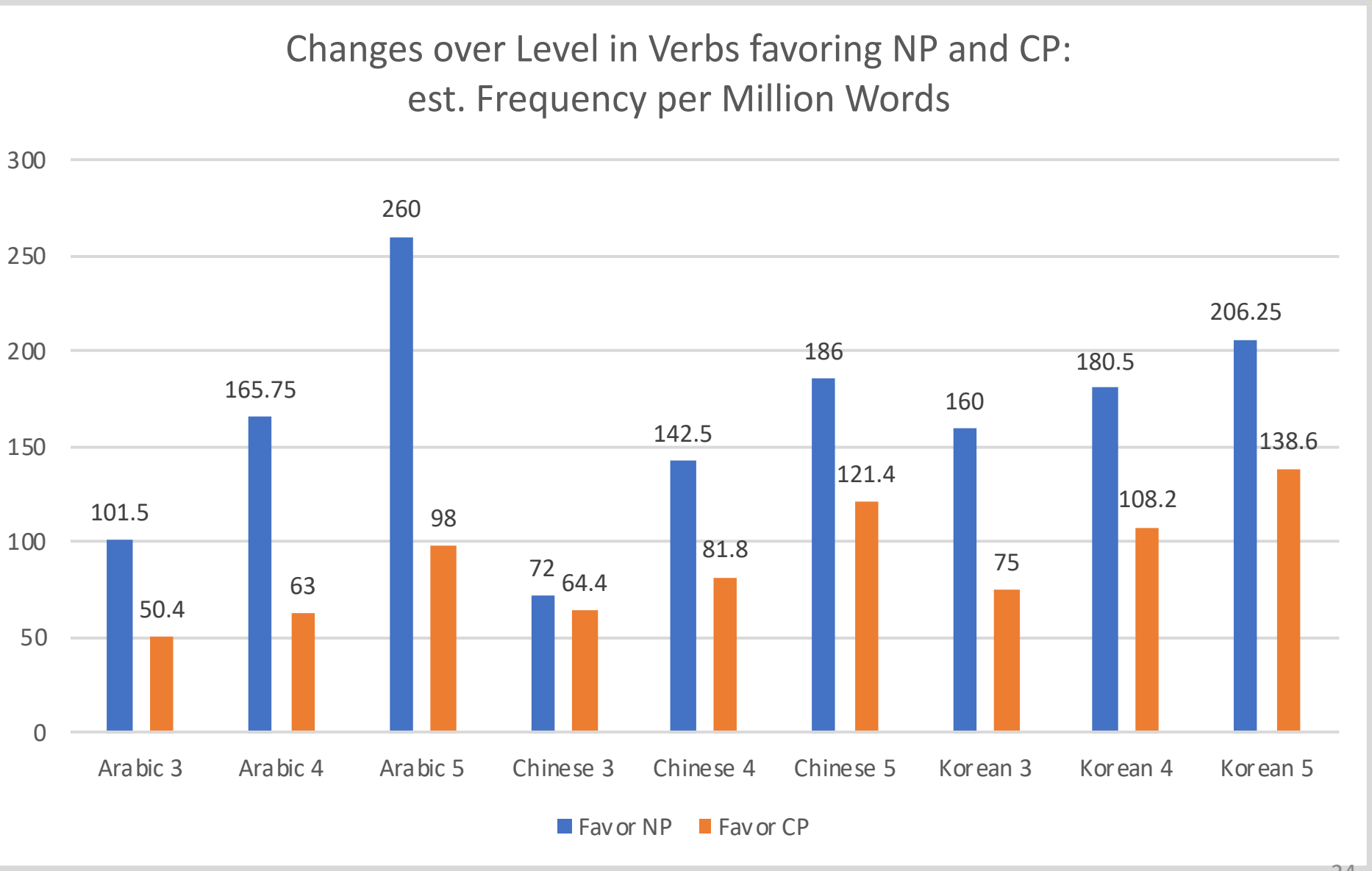

Level 4
500
450
400
350
300
250
200
150
100
50
0
consider - 395 -NP
suggest - 431 -CP
explain - 481 -NP
realize - 621 -CP
admit - 1093 -CP
deny - 1413 -NP
conclude - 1680 -CP
recommend - 1699 -NP
suppose - 2118 -CP
Arabic
Chinese
Korean

Level 5


Level 5
300
250
200
150
100
50
0
270.5
98
186
121.4
206.25
138.6
Arabic
Chinese
Korean
Favor NP
Favor CP


Level 5
900
800
700
600
500
400
300
200
100
0
consider - 395 -NP
suggest - 431 -CP
explain - 481 -NP
realize - 621 -CP
admit - 1093 -CP
deny - 1413 -NP
conclude - 1680 -CP
recommend - 1699 -NP
suppose - 2118 -CP
Arabic
Chinese
Korean

Summary


Changes over Level in Verbs favoring NP and CP:
est. Frequency per Million Words
300
250
200
150
100
50
0
101.5
50.4
165.75
63
260
98
72
64.4
142.5
81.8
186
121.4
160
75
180.5
108.2
206.25
138.6
Arabic 3
Arabic 4
Arabic 5
Chinese 3
Chinese 4
Chinese 5
Korean 3
Korean 4
Korean 5
Favor NP
Favor CP

Discussion

- Many words in ESL production are predictable based on frequency in COCA – confirms usefulness of frequency bands
- Selecting lexical items based on theoretical part of speech status shows that frequency is not the only determinant of use in written output
- Syntactic complexity: markers of complex T-units increase with level: verbs requiring complex semantics --> CP selection (Grimshaw, 1981)
  - Topic choice/what learners want to say: 'concluding' vs. 'suggesting'
- Words chosen on formal theoretical/experimental psycholinguistic grounds are a promising direction a proxy for measuring syntactic development: now need to check whether verbs are used with NP or CP in reality

Conclusion

- Data Science tools +theory permits principled exploration of ‘big’ learner data sets
- Add the use of formal theories, not just usage based ones
- Include insights from psycholinguistics
- And finally …..

Close the loop!

- 1. Beginning of the loop – students and teachers provide data
- 2. Researchers analyze the data
- 3. Researchers close the loop
- → *Discuss* with IEP teachers, e.g.,
  - *which* words to focus on in valuable class time (when the learners need to acquire 8-9000 words).
    - Less 'considering'; more 'suggesting', 'admitting', 'denying'
  - Some learners may need more focus on key clause types, e.g., Arabic speakers: 'whether';
  - All learners could use focus on CP verbs

Robert Ochsner

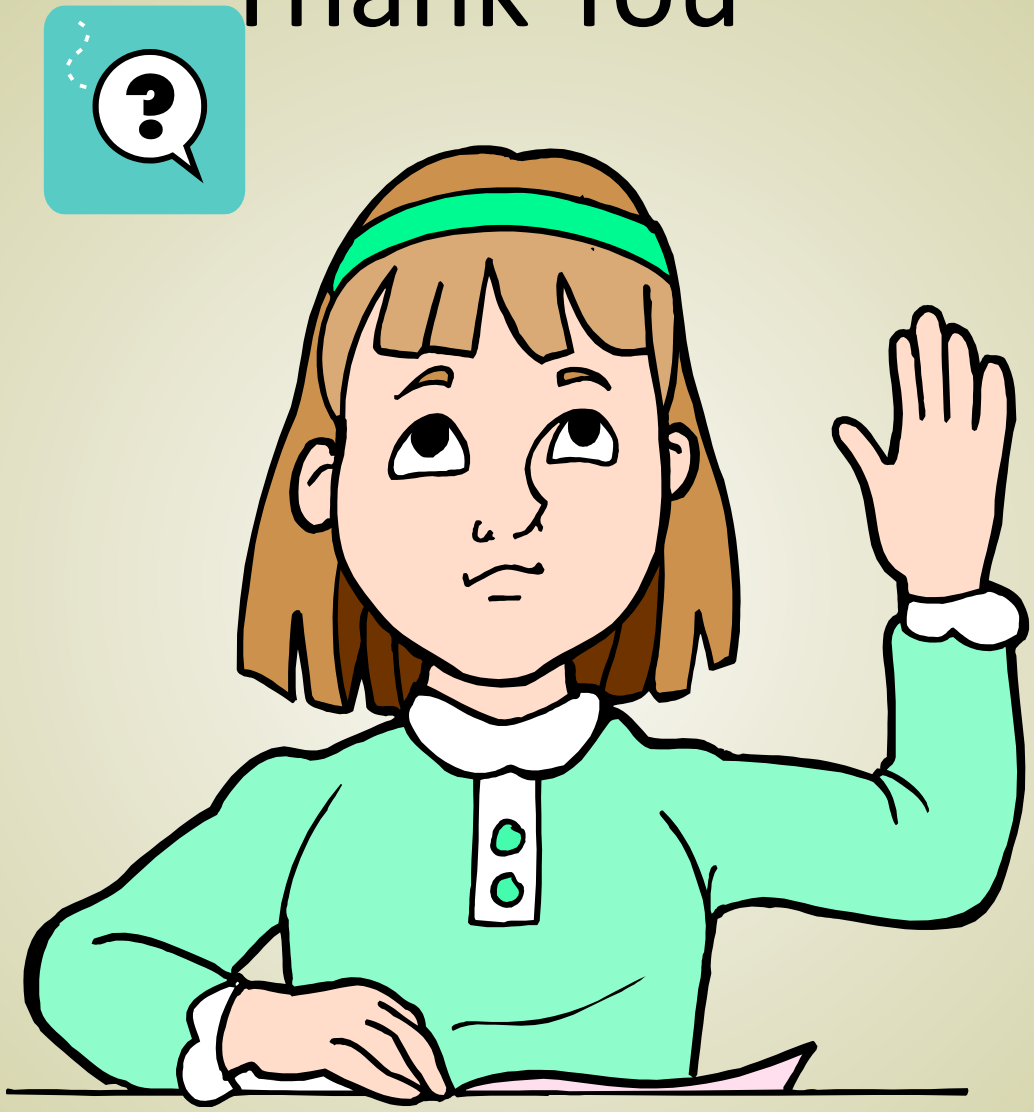
Ochsner, R. (1979). A poetics of second language acquisition. *Language Learning, 29*(1), 53-80.

Man is a complex symbol-user, and language is the basic tool he uses to symbolize everything. Language *is* a very big topic. SLA comprises obviously as big a field of study. It should not surprise us then to find at least two possible ways to investigate a first or second language. If we study it as a simple entity, a "body of facts," we can develop by controlled experimentation a good sense of the *thing* that language is. But if we only look at its physical use, we miss understanding the talker's symbols, his/her metaphorical use of language. It is possible to study either the language *thing* or the language *metaphor;* our research attitude determines this view or way of looking. Therefore, at another research level, we must look at our looking, and then through our point-of-view. And then *at* and *through,* perspective and reality in constant oscillation (Lanham 1977). For the researcher shapes his research findings, a fact that has been experimentally, and ironically, shown (Rosenthal 1977).

This at/through perspective is itself, remember, a metaphor—a kind of applied poetics. Paula Kurman (1977), a speech therapist, gives the nicest statement of what I mean, in practice, by this "bilingual" attitude. In sum, I can only steal her remarks:

> It is useful to read of developments in fields other than one's own.
> It is useful to "borrow" question structures and research techniques from other fields.
> Is is useful to seek out perspectives of those whose work interests are different from one's own.
> It is useful *not* to restrict such solicitations to the scientific disciplines. Artists, business people, and alert children have interesting perceptual frameworks, too. Remember that it was a child, as yet not thoroughly enmeshed in the social construct of reality of his elders, who saw and called the Emperor naked.

Thank You



Level 3

- Compose meaningful sentences and paragraphs that focus on a central idea with appropriate support and **conclusion**
- Introduce the concept that writing is a process
- Express ideas in writing to the reader in as clear a way as possible
- Increase fluency in writing

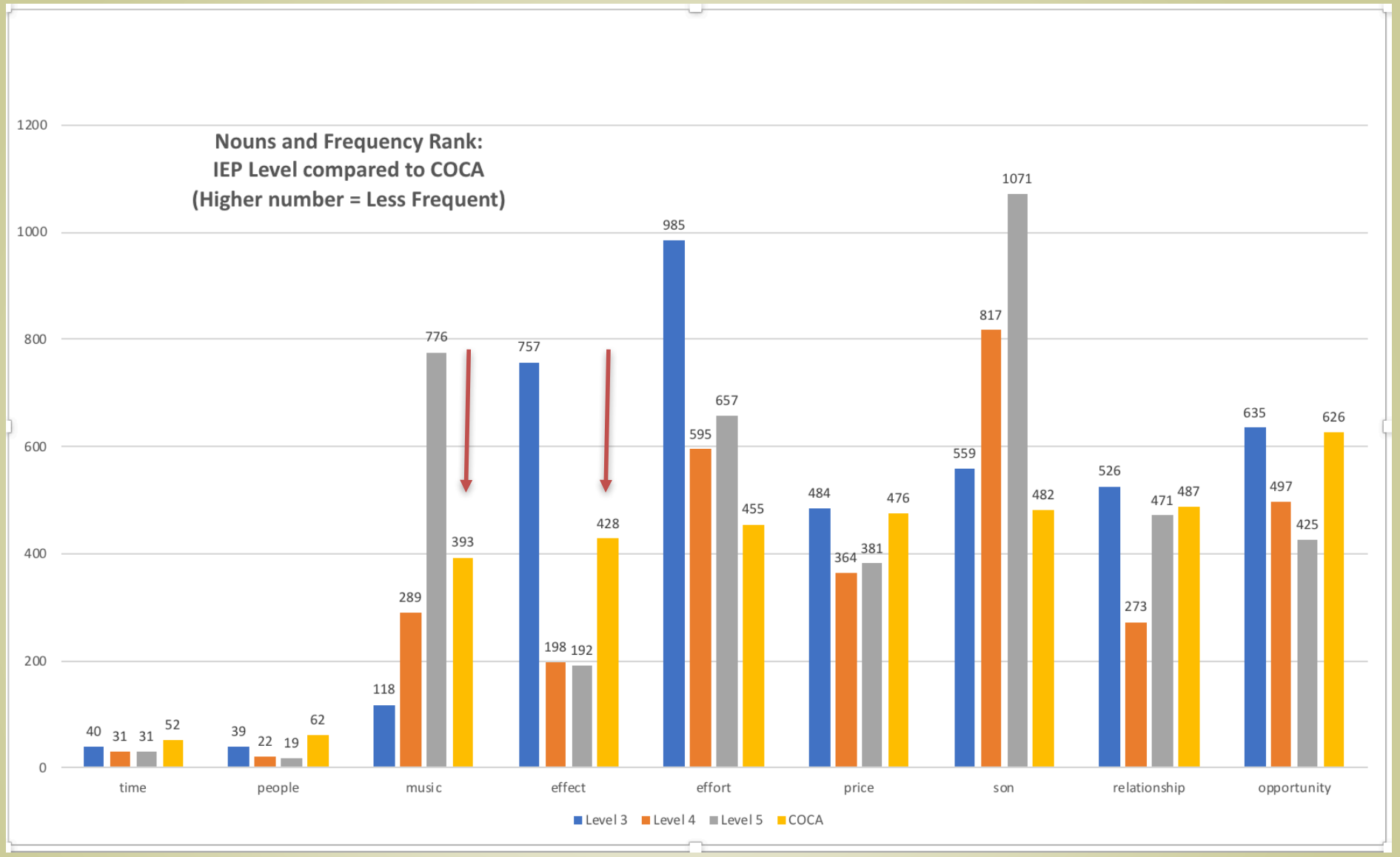
Level 4

- BNC 3-4000 level words
- 'Students will produce medium-length, original written texts (≤ 500 words) responding to information on personal, practical, social, and general academic topics.'
- Level 4 topics: 'Process' (common examples were 'recipes' and 'how to find an apartment'); 'Classification' (e.g., types of doctors, festivals, jobs, lists of reasons). 'Cause-effect': one frequent topic was of 'causes of happiness'.

Pitt

Level 5

- BNC 5000 list + Coxhead core
- 'Students will produce medium-length and long, original written texts (500-2,000 words) on personal, practical, social, and general academic topics.'
- Level 5:
  - 'Explanation'
    - (e.g., 'how to learn English', 'how to stay healthy', 'the effects of a bad diet');
  - 'Narratives';
    - 'Argument/persuasive essay' that presents a point of view and supports it (e.g., 'euthanasia', 'the death penalty', 'pollution'; 'same-sex marriage');
  - 'Comparison/contrast essay'
    - (e.g., 'town or city living', 'your home town vs. Pittsburgh', 'Macintosh vs. PC computers').
  - 'Example Essay' –
    - illustrate a case, e.g., 'education in the ELI', etc.


Nouns and Frequency Rank:
IEP Level compared to COCA
(Higher number = Less Frequent)
1200
1000
800
600
400
200
0
40 31 31 52
39 22 19 62
118 289 776 393
757 198 192 428
985 595 657 455
484 364 381 476
559 817 1071 482
526 273 471 487
635 497 425 626
time
people
music
effect
effort
price
son
relationship
opportunity
Level 3
Level 4
Level 5
COCA