A comparative analysis of frequency lists of derived words across specialist varieties of written English

Piotr TWARDZISZ, University of Warsaw, Poland

1. THE STUDY

Interest in

- morphologically complex (affixed) words
- recorded in the academic genre of contemporary written English
- similarities and differences in terms of affixed words in different disciplines
- different disciplines within the humanities and social sciences

Implications for

- applied linguistics in the area of writing for research
  - affix preferences of particular disciplines
- theoretical linguistics in the area of morphological productivity
  - refinement of morphological productivity

2. SOURCE(S) OF THE DATA

- the **Corpus of Contemporary American English** (**COCA**)
- COCA's academic genre (**ACAD**)
- ten sub-divisions, abbreviated as:

education	phil/rel
history	business
geog/soc-sci	sci/tech
law/pol-sci	medicine
humanities	misc

3. IN SEARCH OF

morphological — variation / stability — across related disciplines

Previous studies of disciplinary variation involve:

- grammatical constructions
- lexical bundles
- phraseology

(e.g., Hyland & Tse 2007; Hyland 2008; Vincent 2013; Cunningham 2017)

✓ morphologically complex words – scant attention (Montero-Fleta 2011)

4. RESEARCH QUESTIONS

- Do certain words / affixes **prefer** certain disciplines?
- Is there a stable **core** of complex words typical for all disciplines?
- What specific **parameters** are to be taken into consideration?
- How can the parameter of morphological **productivity** be enhanced?

5. INITIAL CORPUS SEARCH

- **93 affixes** (or affix variants) across **7 different disciplines**
- affixes = 'search sub-strings', i.e. sequences of characters, either preceded or followed by *
  e.g. *ation, *ment, pre*, un* etc.
  a hyphen added: *e-* and *ex-*

  - comparing humanities / social sciences with hard sciences or medicine – not profitable

For example, nos. of **word types** across 7 disciplines

Prefix	Edu	Hist	Phil/Rel	Law/Polit	Geog/Soc	Scie/Tech	Med
anti	209	845	605	760	801	830	909
e-	115	21	16	58	72	170	98
hyper	76	68	65	69	161	227	330
macro	50	44	28	34	92	205	93
ultra	17	54	35	44	73	214	80
al	1000	1000	1000	1000	1000	1000	1000
ism	370	933	916	763	925	409	260

  - Result: removal of low-frequency and opaque affixes:
    a-, *endo-*, *peri-*, *sur-*, *-ly*, *-y*, *-ed*, *-en*, *-s*

6. AFFIXES IN THE HUMANITIES AND SOCIAL SCIENCES

Focus on:
- edu
- hist
- geo

Reduction of affixes
- with highest numbers of word types

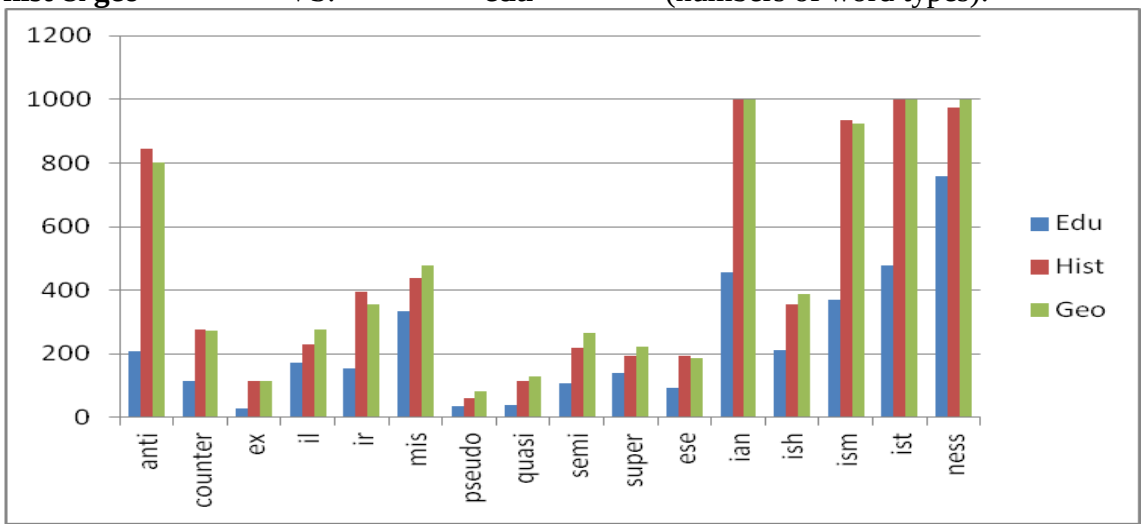
de-, *dis-*, *en-*, *ex-* , *in-*, *non-*, *pre-*, *pro-*, *re-*, *un-*, *-al*, *-an*, *-ation*, *-er*, *-ess*, *-ic*, *-ity* (17)

- lowest numbers of word types

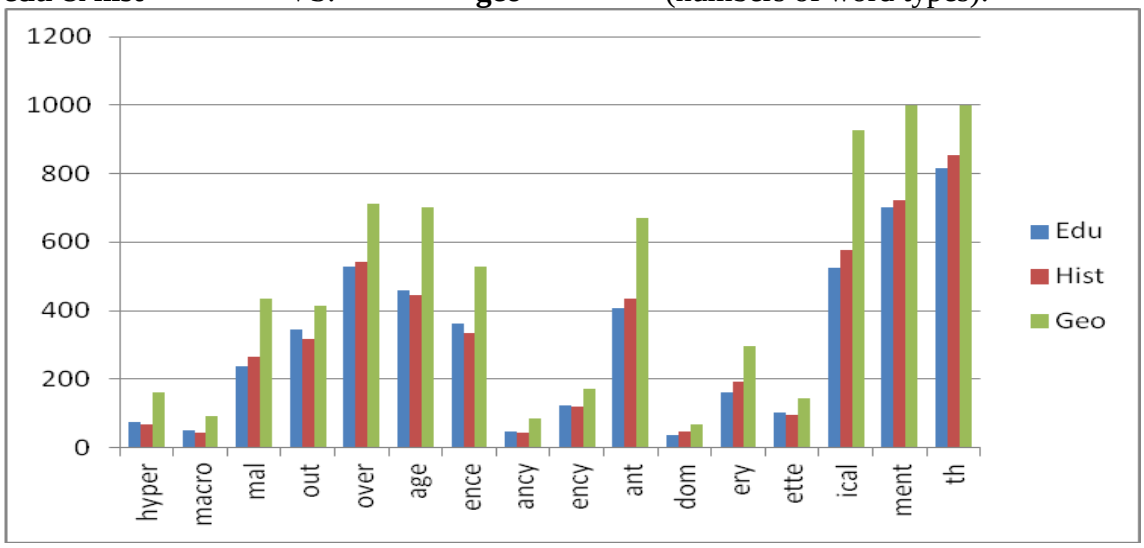
after-, *ante-*, *circum-*, *exo-*, *maxi-*, *retro-*, *supra-*, *-esque*, *-some*, *-wise* (10)

7. AFFIX ATTRACTION TO TWO-DISCIPLINE CLUSTERS

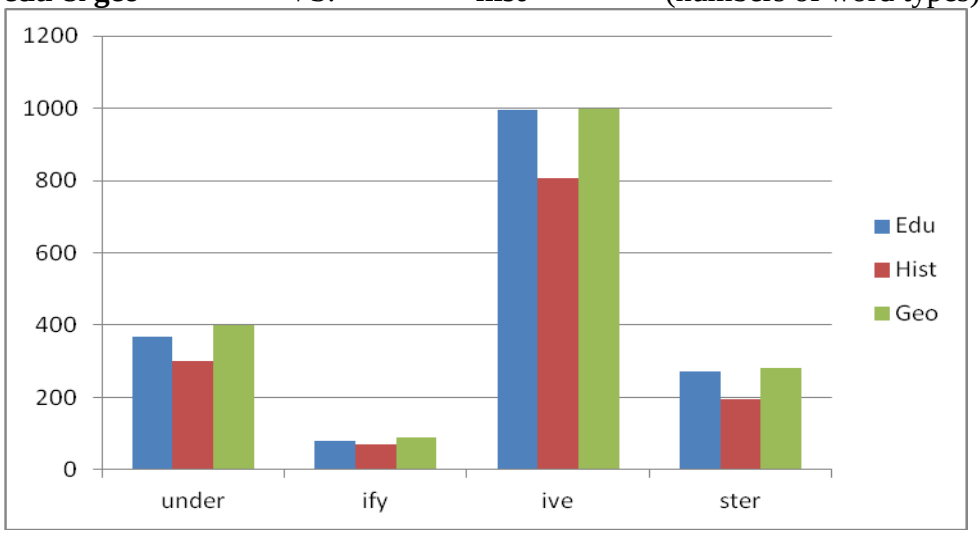
hist & **geo** VS. **edu** (numbers of word types):



edu & **hist** VS. **geo** (numbers of word types):



edu & **geo** VS. **hist** (numbers of word types):



8. SUMMARY

Discipline clusters

- hist & geo (16):

anti-, *counter-*, *ex-*, *il-*, *ir-*, *mis-*, *pseudo-*, *quasi-*, *semi-*, *super-*, *-ese*, *-ian*, *-ish*, *-ism*, *-ist*, *-ness*

- edu & hist (16):

hyper-, *macro-*, *mal-*, *out-*, *over-*, *-age*, *-ence*, *-ancy*, *-ency*, *-ant*, *-dom*, *-ery*, *-ette*, *-ical*, *-ment*, *-th*

- edu & geo (4):

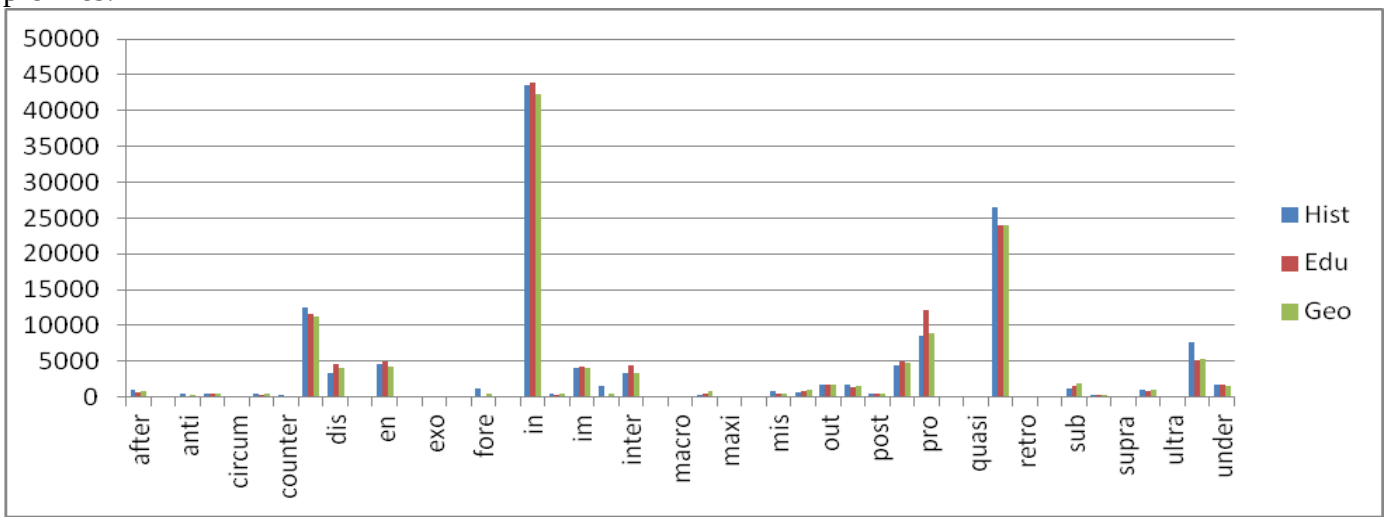
under-, *-ify*, *-ive*, *-ster*

No discipline clusters (30):

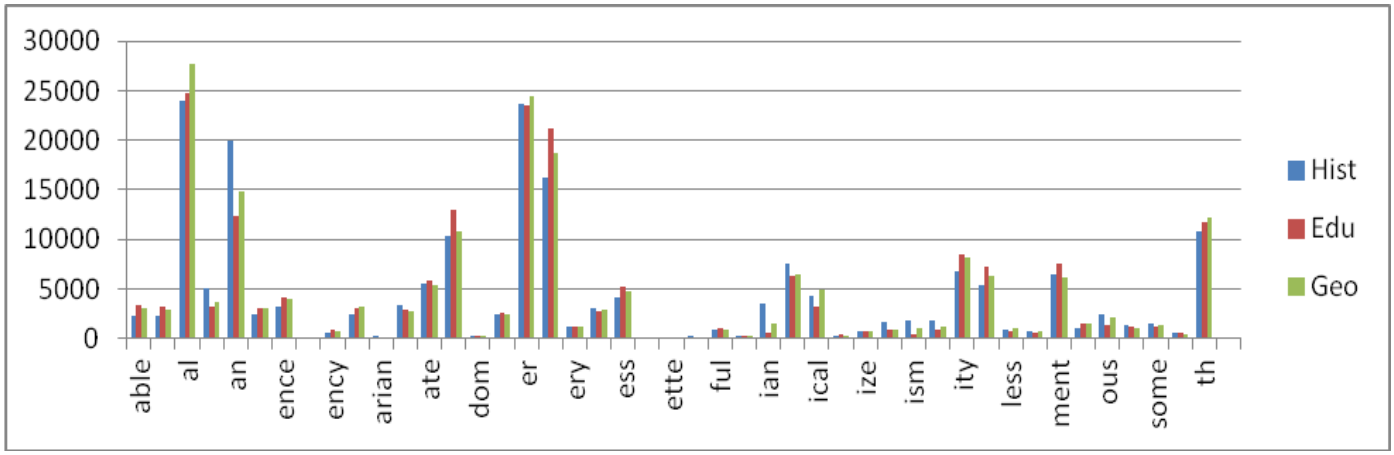
after-, *back-*, *contra-*, *e-**, *extra-*, *fore-*, *im-*, *inter-*, *intra-*, *mega-*, *post-*, *sub-*, *trans-*, *ultra-*, *-able*, *-ar*, *-ance*, *-arian*, *-ary*, *-ate*, *-ee*, *-or*, *-free*, *-ful*, *-hood*, *-ize*, *-less*, *-like*, *-ous*, *-ship*

9. NORMALIZED FREQUENCIES – numbers of tokens per million

prefixes:



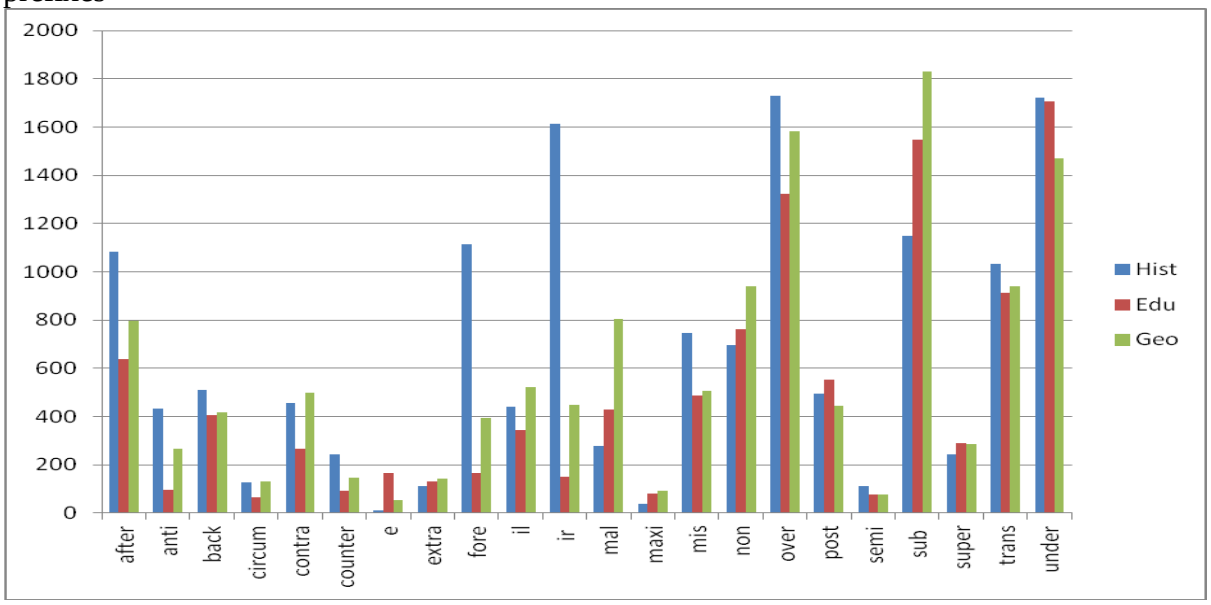
suffixes:



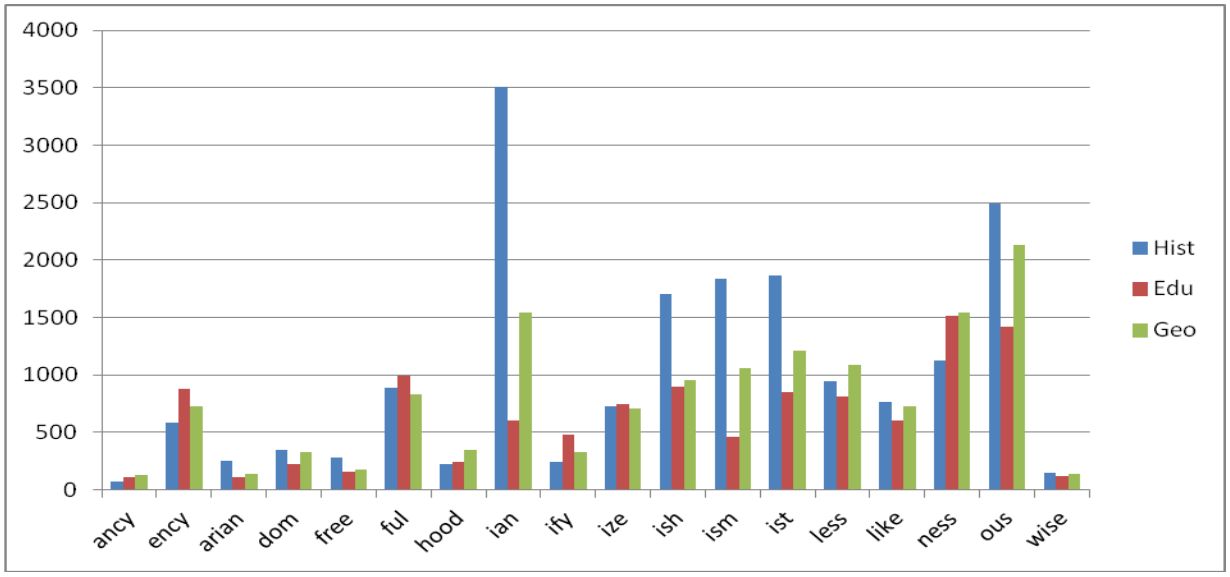
- Subsequently, the focus will be on **40 selected affixes**.

10. NORMALIZED FREQUENCIES – 40 affixes

prefixes



suffixes



11. DISTORTED FIGURES

- some high figures inflated by elements which are not affixes
- distortion of the morphological productivity index

Example:

after- (numbers of tokens)

	Hist	Edu	Geo
after*	14,526	10,040	15,995
after (prep.)	13,561	9,434	14,606
after- (?)	965	606	1,389

- other problematic affixes:

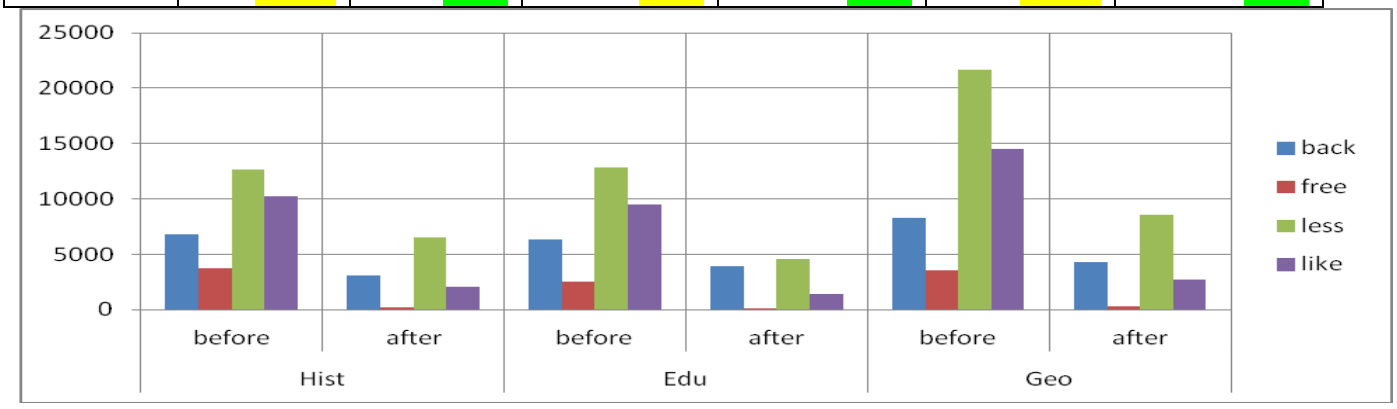
back-, circum-, fore-, -free, il-, -less, -like, non-, over-, post-, sub-, under-

12. PROBLEMATIC "AFFIXES"

	Hist		Edu		Geo	
	before	**after**	**before**	**after**	**before**	**after**
after	14526	965	10040	606	15995	1389
back	6850	3135	6369	3937	8327	4258
circum	1705	324	989	144	2588	549
fore	14914	2807	2596	1213	7906	3340
free	3713	212	2532	143	3522	330
il	5923	3392	5394	1715	10467	4445
less	12697	6538	12813	4604	21697	8530
like	10279	2063	9502	1382	14558	2692
non	9315	7481	12022	10612	18853	16364
over	23143	8808	20880	11177	31668	14247
post	6636	5682	8710	7641	8872	7637
sub	15410	11951	24389	16153	36661	26092
under	23064	8262	26902	9575	29450	9324

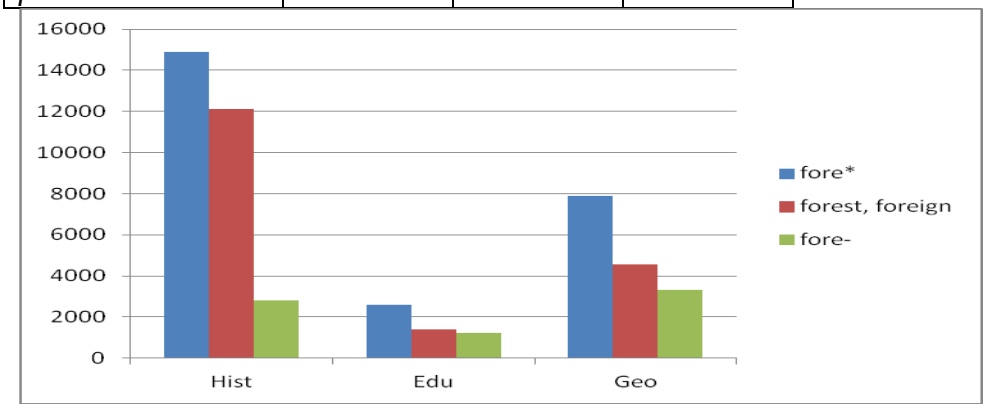
13. SOME OF THE PROBLEMATIC "AFFIXES"

	Hist		Edu		Geo	
	before	after	before	after	before	after
back	6850	3135	6369	3937	8327	4258
free	3713	212	2532	143	3522	330
less	12697	6538	12813	4604	21697	8530
like	10279	2063	9502	1382	14558	2692



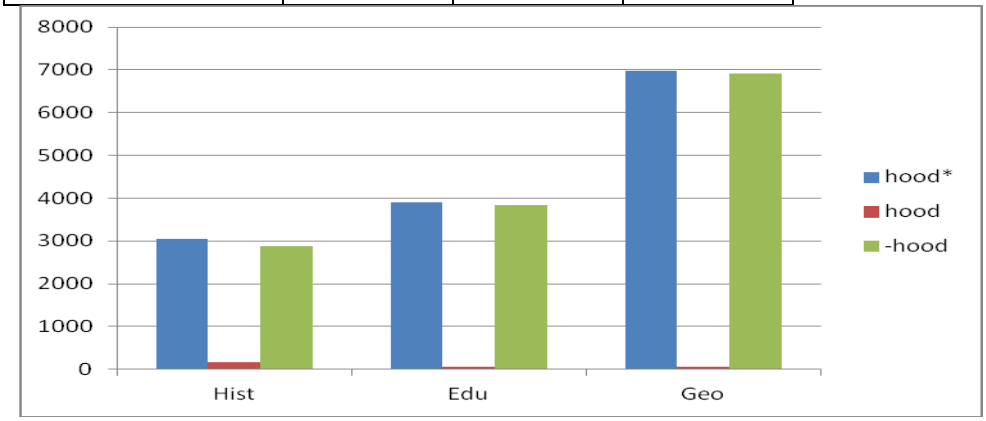
14. HIGH PROPORTIONS OF REDUNDANT ELEMENTS

	Hist	Edu	Geo
fore*	14914	2596	7906
forest, foreign	12107	1383	4566
fore-	2807	1213	3340



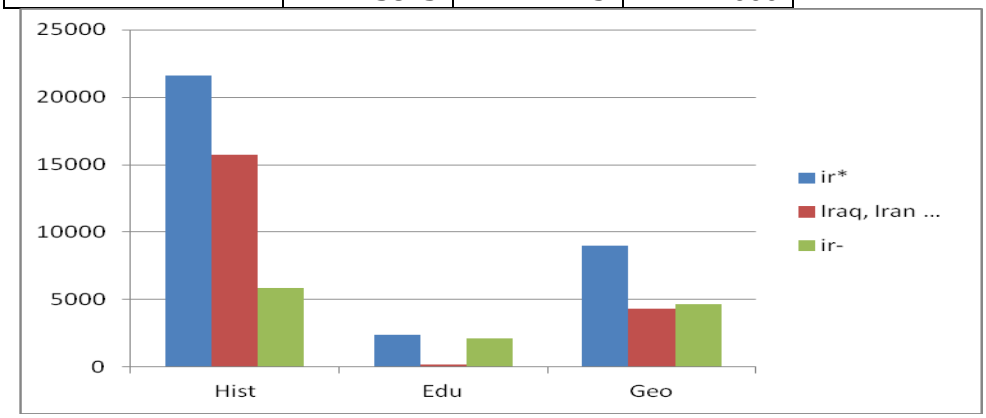
15. LOW PROPORTIONS OF REDUNDANT ELEMENTS

	Hist	Edu	Geo
hood*	3048	3898	6981
hood	163	68	59
-hood	2885	3830	6922



16. DIFFERENT PROPORTIONS OF REDUNDANT ELEMENTS ACROSS DISCIPLINES

	Hist	Edu	Geo
ir*	21634	2356	8999
Iraq, Iran …	15761	207	4333
ir-	5873	2149	4666



CLEAN-UP 1: visual identification of irrelevant items (with a pseudo-affix)

17. MORPHOLOGICAL PRODUCTIVITY

- the number of hapaxes and the number of tokens
- "productivity in the narrow sense":

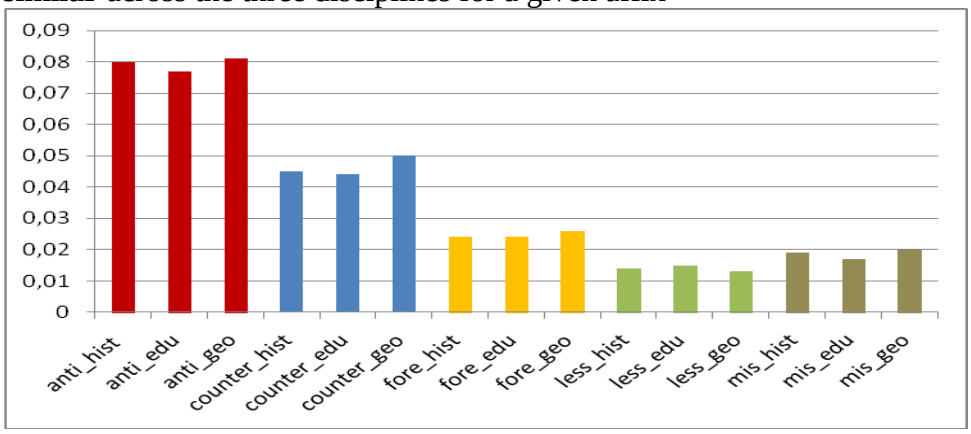
  "the quotient of the number of hapax legomena n_1 with a given affix
  and the total number of tokens N of all words with that affix" (Plag et all. 1999: 216)

  $$\mathbf{P = n_1^{aff} / N^{aff}}$$

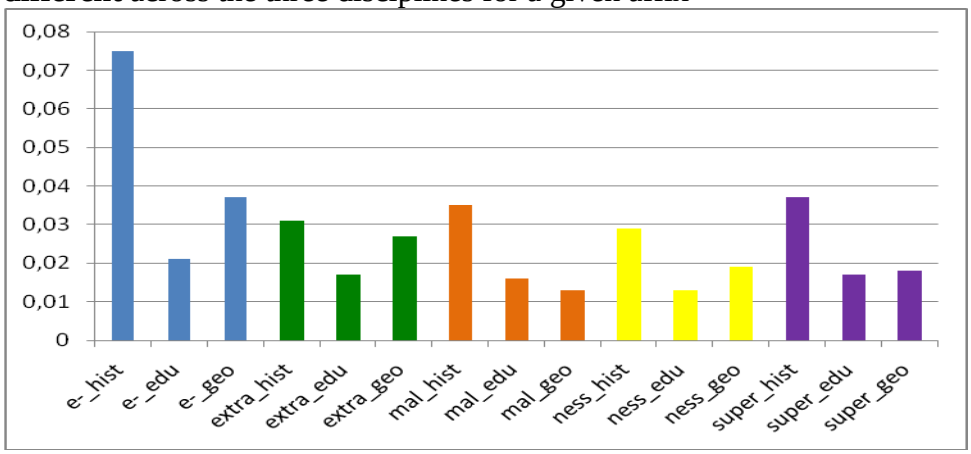
"*P* expresses the rate at which new types are to be expected to appear when *N* tokens have been sampled. In other words, *P* estimates the probability of coming across new, unobserved types, given that the size of the sample of relevant observed types equals *N*." (Plag et all. 1999: 216)

18. MORPHOLOGICAL PRODUCTIVITY may be:

similar across the three disciplines for a given affix



different across the three disciplines for a given affix



- Ordering all the affixes from the most productive to the least productive is not possible due to intra-disciplinary differences.

19. UNIQUE WORD TYPES

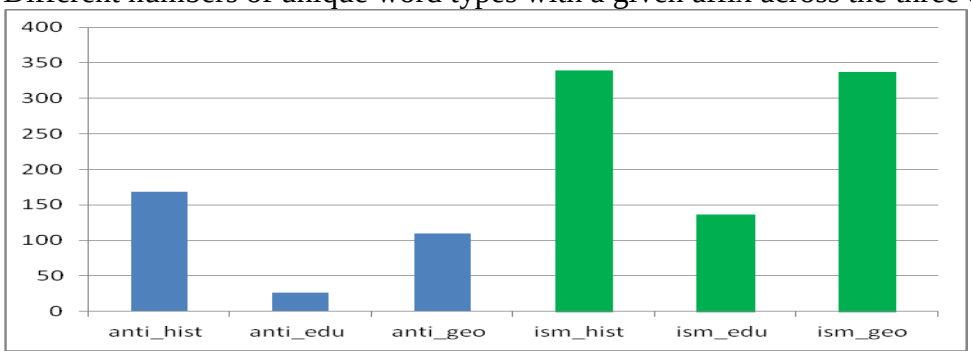
Selection procedure of unique word types

- word types with frequencies ≥ 0.01 per million (5 hits and above) (word types with 4 hits and less ignored)
- **clean-up 2** of morphologically irrelevant word types

Result: only (statistically relevant) unique word types with a "real" affix left

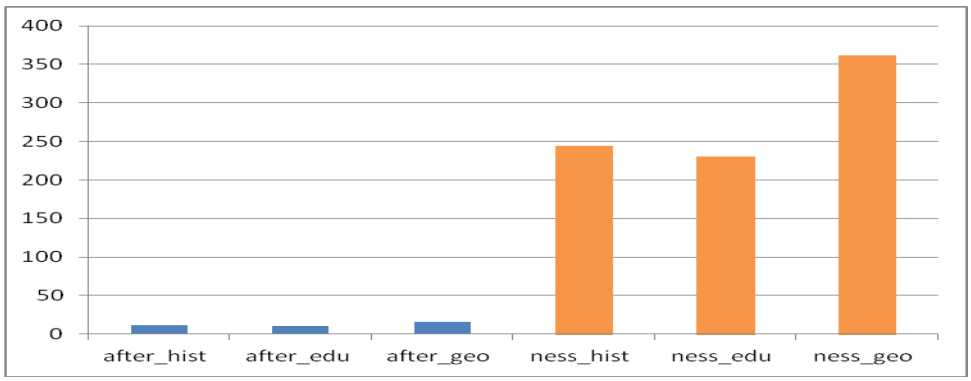
20. UNIQUE WORD TYPES – after clean-up 2

Different numbers of unique word types with a given affix across the three disciplines.



21. UNIQUE WORD TYPES – after clean-up 2 – cont.

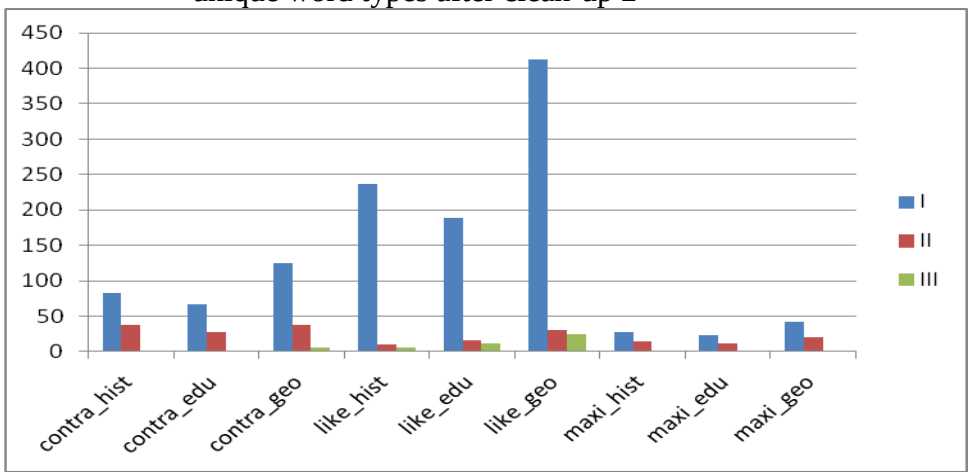
Marked numerical differences between individual affixes.



22. UNIQUE WORD TYPES – after clean-up 2 – cont.

Marked drops in numbers:

- ✓ unique word types initially
- ✓ statistically significant unique word types (5 hits and above)
- ✓ unique word types after clean-up 2



DISCIPLINE-WISE

23. MORPHOLOGICAL RELEVANCE (R) – discipline-wise

R = no. of tokens after clean-up 1 / no. of tokens initially

R = cT / T

the quotient of the number of *cT* (“cleaned” tokens) with a “real” affix
and the total number of tokens *T* of all words with this affix-like sequence of characters

Highest Rs across three disciplines

HISTORY		EDUCATION		GEOGRAPHY	
	R		R		R
ancy	1.0	ancy	1.0	ancy	1.0
anti	1.0	anti	1.0	anti	1.0
arian	1.0	arian	1.0	arian	1.0
contra	1.0	contra	1.0	contra	1.0
counter	1.0	counter	1.0	counter	1.0
dom	1.0	dom	1.0	dom	1.0
e-	1.0	e-	1.0	e-	1.0
ency	1.0	ency	1.0	ency	1.0
extra	1.0	extra	1.0	extra	1.0
ful	1.0	ful	1.0	ful	1.0
hood	1.0	hood	1.0	hood	1.0
ian	1.0	ian	1.0	ian	1.0
ify	1.0	ify	1.0	ify	1.0
ish	1.0	ish	1.0	ish	1.0
ism	1.0	ism	1.0	ism	1.0
ist	1.0	ist	1.0	ist	1.0
ize	1.0	ize	1.0	ize	1.0
mal	1.0	mal	1.0	mal	1.0
maxi	1.0	maxi	1.0	maxi	1.0

mis	1.0	mis	1.0	mis	1.0
ness	1.0	ness	1.0	ness	1.0
ous	1.0	ous	1.0	ous	1.0
super	0.833	super	0.874	super	1.0
trans	1.0	trans	1.0	trans	0.912
wise	0.839	wise	0.860	wise	0.903

The degree of morphological (ir)relevance remains roughly the same for all affixes **across the three disciplines**.

Lowest Rs across three disciplines: *after-*, *circum-*, *-free* and *-like*.

24. MORPHOLOGICAL PRODUCTIVITY (P) – discipline-wise

P = n1 / N

Highest morphological productivity across three disciplines

HISTORY		**EDUCATION**		**GEOGRAPHY**	
	P		P		P
anti	0.080	anti	0.077	anti	0.081
free	0.207	free	0.203	free	0.188
like	0.092	like	0.100	like	0.115
non	0.108	non	0.090	non	0.093
post	0.054	post	0.033	post	0.057
semi	0.120	semi	0.108	semi	0.152

Morphological productivity remains roughly the same for all affixes across the three disciplines.

Minor differences across the disciplines: *circum-*, *e-*, *maxi-* and *non-*.

Morphological productivity is (very) low for: *-ancy*, *-arian*, *back-*, *contra-*, *-dom*, *-ency*, *-ful*, *-hood*, *-ian*, *-ify*, *-ish*, *-ism*, *-ist*, *-ize*, *-less*, *maxi-*, *-ous*, *sub-*, *trans-*, *under-* and *-wise*.

25. MORPHOLOGICAL RELEVANCE (R) vs. MORPHOLOGICAL PRODUCTIVITY (P)

HIGH R & HIGH P: *anti-* (only!)

HIGH R & LOW P: e.g. *-dom*, *-ful*, *-ish*, *-ism*, *-ness*

LOW R & HIGH P: e.g. *-free*, *-like*

26. HAPAX LEGOMENA

High numbers of hapaxes (3-digit figures)

HISTORY		**EDUCATION**		**GEOGRAPHY**	
	hapaxes		hapaxes		hapaxes
anti	464	anti	117	anti	430
counter	146	counter	63	counter	147
ian	520	ian	193	ian	534
il	116	il	80	il	144
ir	193	ir	57	ir	162
ish	156	ish	94	ish	186

ism	356	ism	123	ism	364
ist	460	ist	201	ist	464
ize	128	ize	106	ize	168
less	94	less	68	less	111
like	190	like	139	like	310
mal	132	mal	112	mal	211
mis	198	mis	136	mis	202
ness	447	ness	325	ness	582
non	810	non	959	non	1523
ous	170	ous	160	ous	251
over	258	over	273	over	322
post	310	post	256	post	437
semi	123	semi	65	semi	145
sub	223	sub	297	sub	349
super	102	super	69	super	107
trans	134	trans	115	trans	190
under	111	under	176	under	169

17 affixes retain roughly the **same** proportions of hapaxes across the three disciplines.

23 out of 40 of the affixes show marked **differences** across the three disciplines in terms of hapaxes.

There are marked **differences** for:
-ancy, *-arian*, *contra-*, *counter-*, *e-*, *-ency*, *extra-*, *fore-*, *-free*, *il-*, *ir-*, *-ish*, *-less*, *semi-* and *super-*.

However, there are also essential numerical **differences** within certain frequency ranges:
anti-, *-ism*, *-ist*, *-like*, *non-*, *post-*, *sub-* and *trans-*.

Very **low** numbers of hapaxes overall:
after-, *-ancy*, *circum-*, *-dom*, *-ify* and *maxi-*.

No. of hapaxes is a distinctive, discipline-driven factor.

27. UNIQUE WORD TYPES – after clean-up 2

Higher numbers of unique word types

HISTORY		EDUCATION		GEOGRAPHY	
	word types		word types		word types
anti	169	anti	26	anti	110
ful	88	ful	72	ful	93
ian	281	ian	119	ian	302
ish	51	ish	25	ish	50
ism	339	ism	136	ism	337
ist	268	ist	107	ist	278
ize	135	ize	88	ize	134
less	81	less	43	less	88
mis	78	mis	74	mis	82
ness	244	ness	230	ness	362
non	244	non	369	non	601
ous	286	ous	179	ous	317
over	156	over	122	over	193
post	104	post	95	post	159

sub	98		sub	111		sub	145
trans	57		trans	28		trans	58
under	97		under	102		under	117

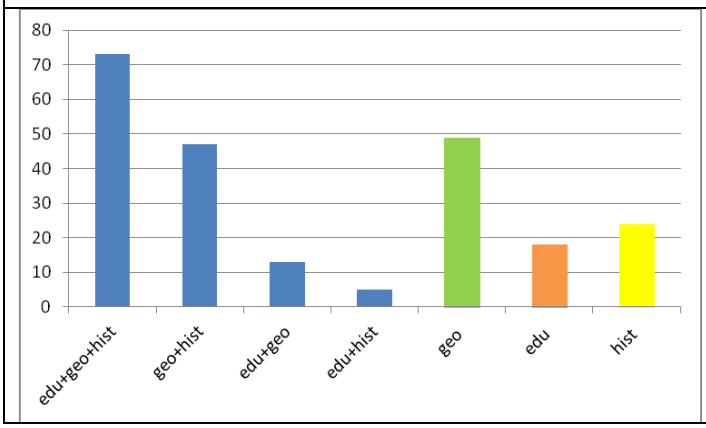
Numbers of unique word types are **stable** across the three disciplines for:
after-, *-ancy*, *-arian*, *back-*, *circum-*, *contra-*, *-dom*, *-ency*, *extra-*, *fore-*, *-free*, *-ful*, *-hood*, *-ify*, *il-*, *ir-*, *mal-*, *maxi-*, *mis-*, *over-*, *semi-*, *super-*, *under-* and *-wise*.

Numbers of unique word types are markedly **different** across the three disciplines for:
anti-, *counter-*, *e-*, *-ian*, *-ish*, *-ism*, *-ist*, *-ize*, *-less*, *-like*, *-ness*, *non-*, *-ous*, *post-*, *sub-* and *trans-*.

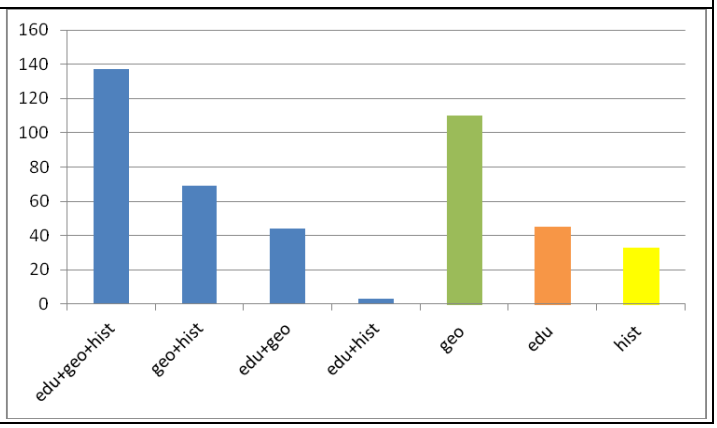
The affixes with very low numbers across the three disciplines:
contra-, *mal-*, *maxi-* (0~1!) and *-wise*.

28. UNIQUE WORD TYPES – distribution across disciplines e.g. *-ness* & *over-*

over-



-ness



References

Cunningham, K. J. 2017. A phraseological exploration of recent mathematics research articles through key phrase frames. *Journal of English for Academic Purposes* 25: 71–83.

Hyland, K. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4–21.

Hyland, K. & P. Tse. 2007. Is there an "Academic vocabulary"? *TESOL Quarterly* 41 (2): 235–53.

Montero-Fleta, B. 2011. Suffixes in word-formation processes in scientific English. *LSP, Professional Communication, Knowledge Management and Cognition* 2 (2): 4–14.

Plag, I., Dalton-Puffer Ch. and H. Baayen. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3 (2): 209–228.

Vincent, B. 2013. Investigating academic phraseology through combinations of very frequent words: A methodological exploration. *Journal of English for Academic Purposes* 12: 44–56.