IS THERE A FORMULA FOR FORMULAIC LANGUAGE?

Richard S. Forsyth
(www.richardsandesforsyth.net)

Łukasz Grabowski
(Opole University, Poland)

FLaRN Conference, 15 July 2014, Swansea

Outline of our 20-minute talk

- **Introductory remarks on measuring formulaicity**
- **Methodology**: research material, tools, hypothesis, goals
- **Findings**: empirical results
- **Insights** from the study
- **Future avenues**

- Q & A (10 minutes)

Why is language use formulaic?

- The "*deadly repetitiousness of language*" (Bolinger, 1965: 570) is what stares us "*in the face from the text*„ (Firth 1957).
- Language users tend to select prefabricated phrases with single form and meaning (Sinclair 1991 (*idiom principle*)).
- Purposeful use of **various types of multi-word sequences, including continuous and discontinuous ones** in similar & recurrent social situations translates into routinized patterns of linguistic behaviour.

- Yet how to measure **how much of our language is formulaic**?
- **Lack of widely-accepted methods of assessing the degree of formulaic language** in texts.

Current state of affairs

- Presently we can not make statements such as “this corpus of texts is 30% formulaic, and that one is 70% formulaic”.
- We cannot objectively rank corpora/texts from the most to least formulaic.
- Major problems:
  - Researchers **operationalize formulaic language in different ways** and **use different methods to measure its prevalance** in texts.
  - The term 'formulaic sequence' is "intentionally all-encompassing, covering a wide range of phraseology“ (Schmitt & Carter, 2004: 3).
- Is there a natural and convenient **unit of analysis?**
- Our candidate is a p-frame (‘**phrase frame’,** Fletcher 2002-2007), a set of variants of an n-gram identical except for one word (e.g. *if you * any*, *the * of the* etc.).

Goals of this study

- To test empirically a number of plausible mathematical formulae for quantifying the degree to which a text type incorporates inflexible sequences of words, captured in p-frames:
  - Balance, Hapaxity, Haprate, HC, HH, HV, Nonfocus, Rent, Simpidx, TTPC (related to VPR (Roemer 2010))
- The indices are treated as indicators of productivity (pattern variability) of p-frames, which is inversely correlated with formulaic language.
  - Frequency of recurrent sequences of words and/or their fixedness jointly determine the degree to which a text is formulaic (Wray 2008: 102)
- Basic idea: **Lack of variety among the slot-fillers of p-frames indicates high degree of reliance on fixed formulas** (formulaic language).

Example phrase frames (from UN Security Council resolutions)

*to * in the*	89	20
to serve in the		14
to cooperate in the		14
to assist in the		14
to participate in the		8
to stability in the		5
to play in the		5
to include in the		5
to vote in the		4
to date in the		4
to consider in the		3
to submit in the		2
to interfere in the		2
to contribute in the		2
to states in the		1
to interview in the		1
to efforts in the		1
to deploy in the		1

*the * agreement and*	85	7
the peace agreement and		54
the ceasefire agreement and		9
the bonn agreement and		9
the arusha agreement and		7
the framework agreement and		4
the luanda agreement and		1
the algiers agreement and		1
*remain seized of **	85	2
remain seized of the		82
remain seized of this		3

Research material

- Corpora with **different text types exhibiting varying degrees of formulaic language** (more routinized vs. more creative text types), and the general language corpus (LOB, Hofland & Johansson 1982) as a benchmark:

ACAD	26 research articles and 25 book chapters on pharmacology
LEAF	Patient information leaflets describing 461 pharmaceutical products, source: http://mcs.open.ac.uk/nlg/old_projects/pills/corpus/
PROT	Clinical Trial Protocols from the European Medicines Agency, source: https://www.clinicaltrialsregister.eu/index.html
SUMP	Summaries of Product Characteristics, source: OPUS website http://opus.lingfil.uu.se/EMEA.php (Tiedemann, 2009).
UGAR	UN General Assembly Resolutions, 2000-2003, collated by DFKI GmbH, available from www.euromatrixplus.eu/multi-UN/
USCR	UN Security Council Resolutions, 2000-2004, collated by DFKI GmbH, available from www.euromatrixplus.eu/multi-UN/

AC	65 chapters from 53 novels by Agatha Christie
EW	44 short stories by Edith Wharton
IM	26 chapters from 26 novels by Iris Murdoch
WC	45 speeches by Winston Churchill, source: http:// www.winstonchurchill.org/learn/speeches/speeches-of-winston-churchill plus 2 chapters and 2 prefaces from his 3-volume biography of Marlborough
LOBCORP	Lancaster-Oslo-Bergen corpus (Hofland & Johansson, 1982)

Stages of the study

- **Calibration comparisons** (to decide which index is best at detecting traces of formulaic language):
  - In each comparison a relatively formulaic corpus compared with a less formulaic reference corpus.
- **Macro-productivity of p-frames:**
  - The ”winning” index is used to rank corpora from most productive (in terms of pattern variability of p-frames) to least productive – by implication – from least to most formulaic.
- **Micro-productivity of p-frames:**
  - To determine which p-frames are contributing the most to the ranking (using a pair of corpora with the largest divergence, i.e. clinical trial protocols vs. LOB).

Calibration comparisons (1)

Formulaic Test Corpus	Less Formulaic Reference Corpora
ACAD	AC, LOBCORP
LEAF	ACAD, LOBCORP
PROT	ACAD, LOBCORP
SUMP	ACAD, LOBCORP
UGAR	EW, LOBCORP
USCR	WC, LOBCORP

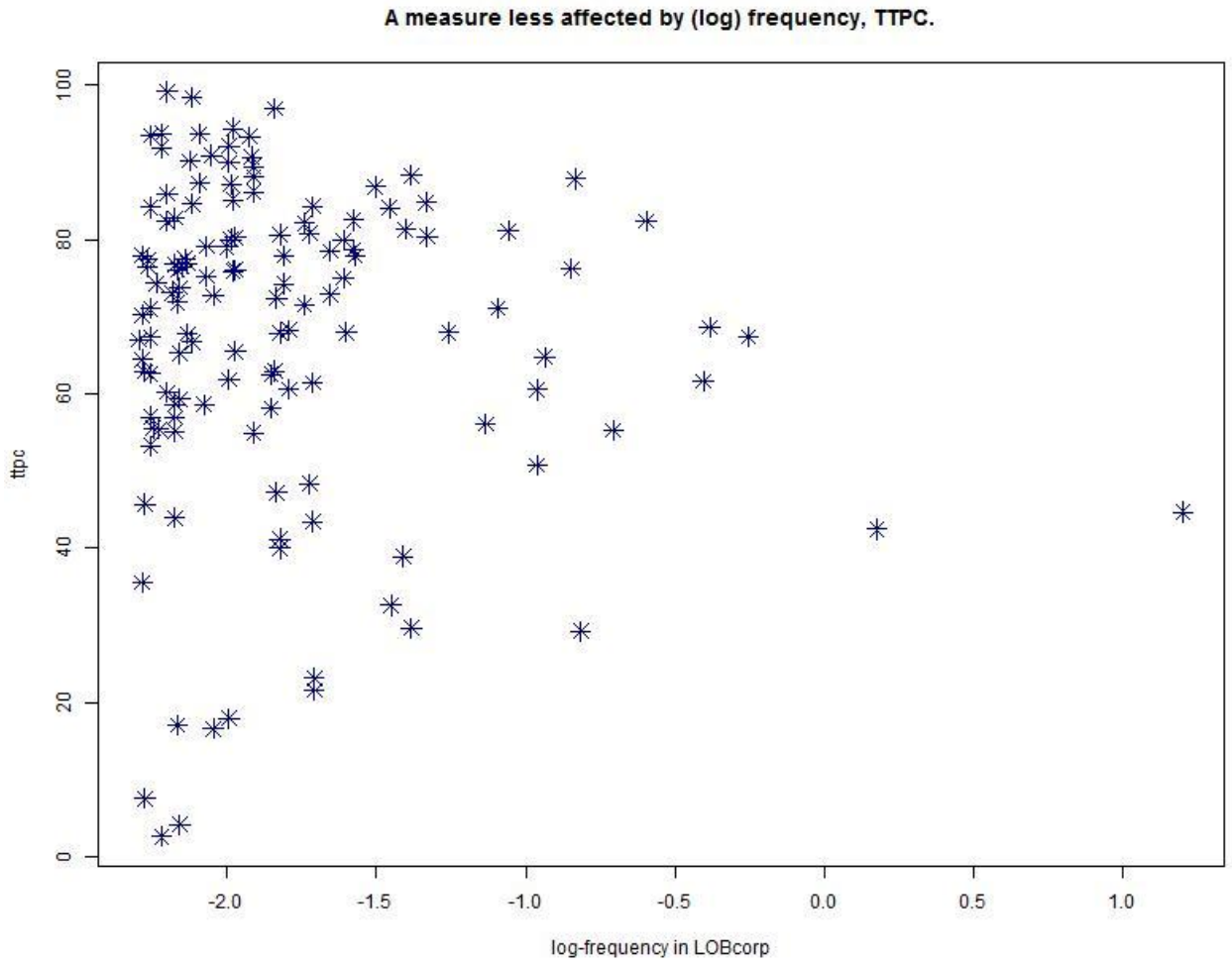
- Only those p-frames that occur in both corpora are considered.
- Each of 10 productivity indices was computed for each p-frame in both corpora, producing two numeric vectors (productivity scores of each p-frame).
- As the scores were matched, a paired t-test was computed: its value was used as a measure of the degree of differentiation achieved by a given index. Indices ranked 1-to-10 by t-score.
- The ranks were aggregated over 12 comparisons.

Calibration comparisons (2)

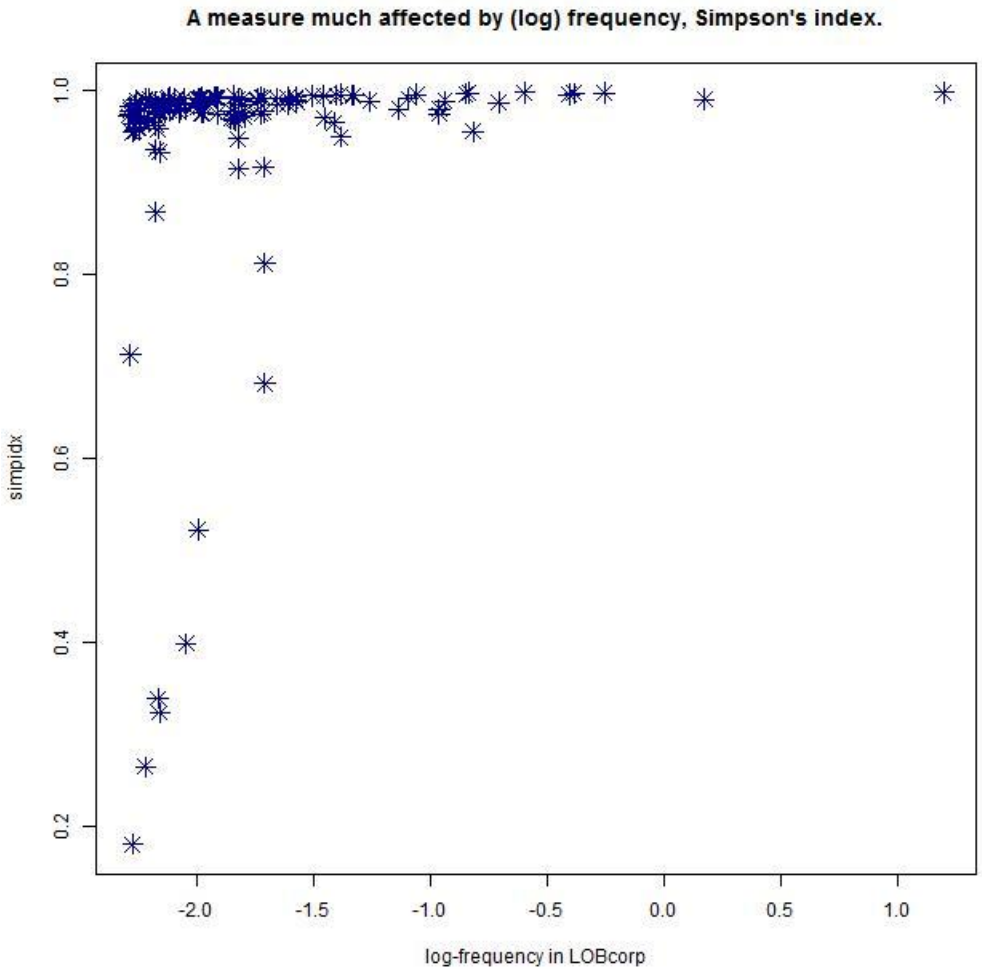
Productivity Index	Mean Rank with all slots included (1 to 4)	Mean Rank with left-most slot omitted (2 to 4)	Mean Rank with both end-slots omitted (2 to 3)
Balance	5.58	5.92	5.92
Hapaxity	6.33	6.33	6.83
Haprate	7.17	7.08	7.00
HC	**7.50**	7.58	7.17
HH	**1.92**	**1.75**	**1.83**
HV	6.83	6.08	6.00
Nonfocus	4.42	4.58	4.50
Rent	4.50	4.58	4.75
Simpidx	3.25	3.17	3.25
TTPC	**7.50**	**7.92**	**7.75**

- Unexpectedly, **TTPC (type/token percentage)** found to be the most effective [V*100/N] .
- This tends to support Roemer's choice of VPR (variant-to-p-frame ratio) as an index of pattern variability of p-frames.
- Measures from outside linguistics (e.g. Herfindahl-Hirschman index (HH), Simpson's diversity index, Shannon's relative entropy) fared poorly.

“Winning” formula, TTPC



Less successful formula, Simpson's index



Macro-productivity of p-frames (1)

- Our preferred index (TTPC) was further used to rank the corpora – from the least to most formulaic.
- Each corpus was compared with a general-language corpus, LOB, treated as a benchmark.
- We computed differences between TTPC scores for each p-frame and divided each difference by standard deviation of TTPC scores of the benchmark corpus.
- Final score was the mean of these scaled differences, i.e. an average z-score.

Test Corpus	Mean scaled TTPC difference	Number of p-frames in common
IM	0.7504	21
AC	0.5127	40
WC	0.5002	59
EW	0.3342	53
ACAD	-0.7112	50
LEAF	-1.6725	42
USCR	-2.1147	47
UGAR	-2.5808	51
SUMP	-2.7134	39
PROT	-2.8539	12

Macro-productivity of p-frames (2)

- The least formulaic is the collection of chapters from literary novels by Iris Murdoch.
- The most formulaic are clinical trial protocols (they resemble cliched texts written by a robot).
- Interestingly, Agatha Christie (described as ”formulaic whodunit writer” (Granger 2009) is less formulaic than Edith Wharton’s prose (at least in our data sample).
- UN General Assembly Resolutions are almost as formulaic as highly patterned summaries of product characteristics.
- The most striking contrast with the reference corpus was between clinical trial protocols and the LOB corpus.

Illustration 1, LOB Corpus versus Clinical Trials protocols

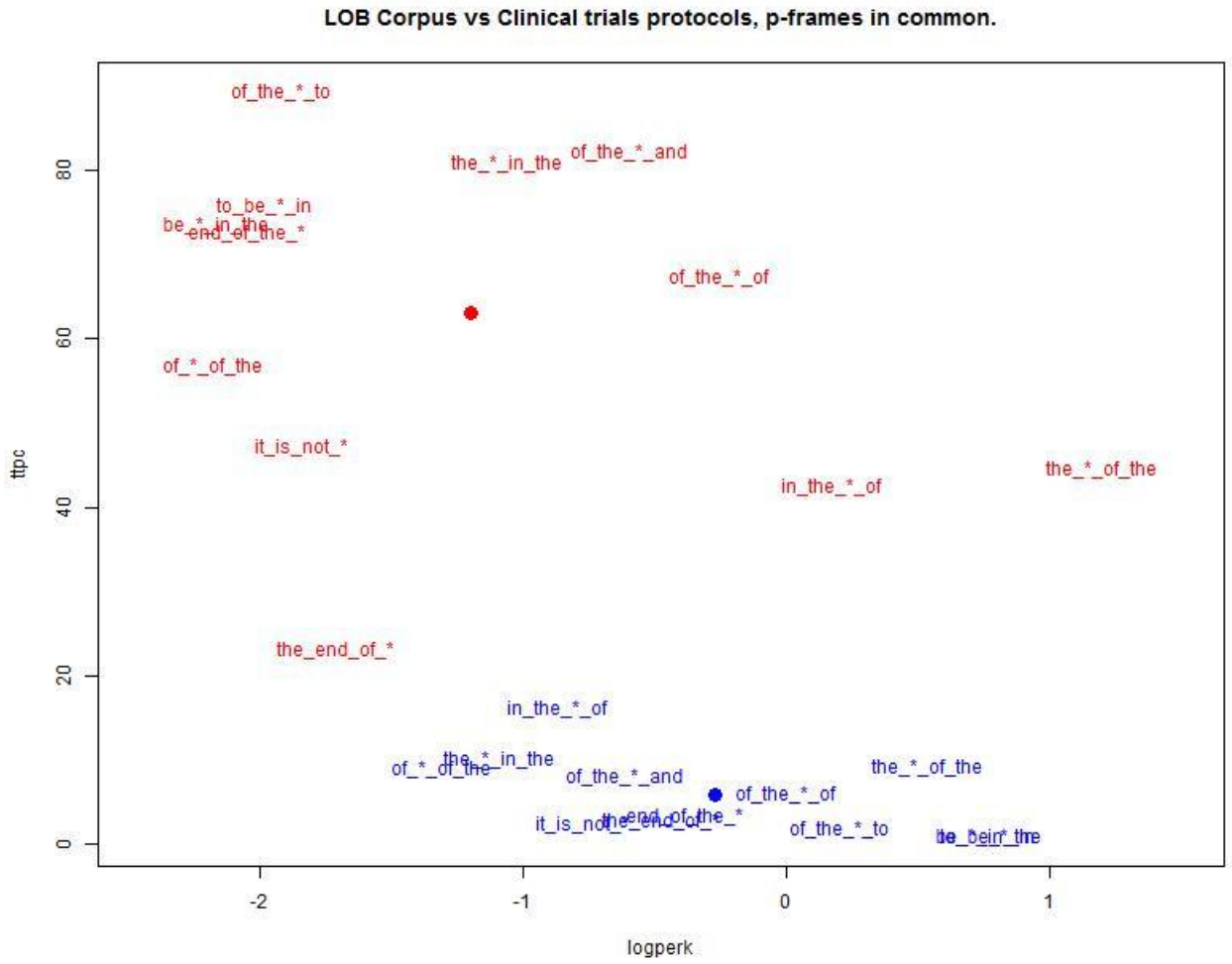
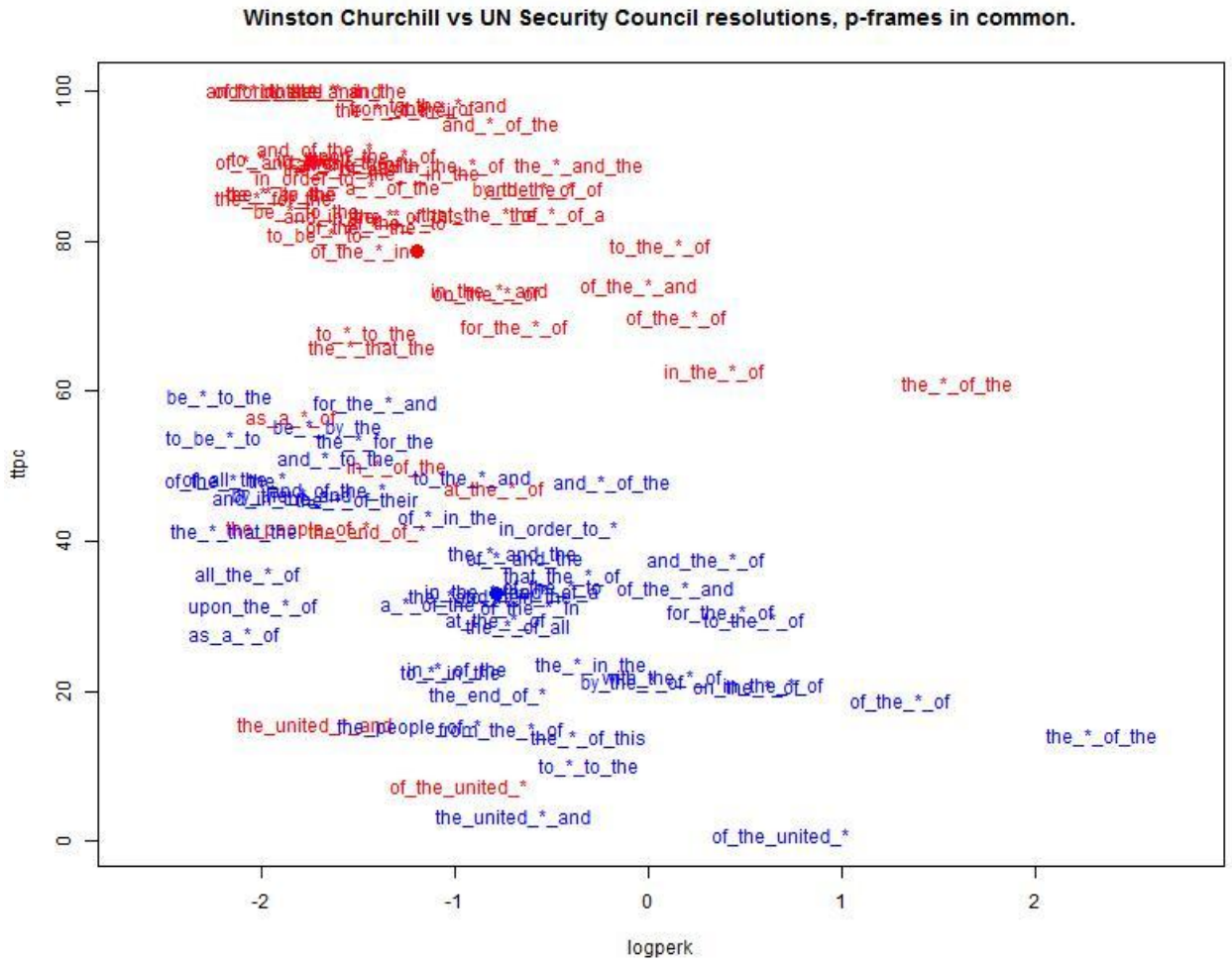


Illustration 2, Winston Churchill speeches versus UN Security Council resolutions



Micro-productivity of p-frames (1)

- Which particular p-frames contribute to the ranking?
- Consider comparison with the largest divergence (PROT vs. LOB), which conveniently involves only 12 shared p-frames.
- We identified 12 p-frames, ranked from smallest to largest scaled difference between their TTPC scores.

Rank	P-frame	TTPC in PROT	TTPC in LOBCORP	Difference	Scaled Difference (SD = 20.04)
1	*the_end_of_* *	2.74390	23.24324	-20.4993	-1.02295
2	*in_the_*_of*	16.14350	42.46914	-26.3256	-1.31370
3	*the_*_of_the*	9.28177	44.69406	-35.4123	-1.76714
4	*it_is_not_* *	2.44898	47.23926	-44.7903	-2.23511
5	*of_*_of_the*	9.09091	56.89655	-47.8056	-2.38559
6	*of_the_*_of*	6.03774	67.38306	-61.3453	-3.06124
7	*end_of_the_**	3.33333	72.72727	-69.3939	-3.46288
8	*the_*_in_the*	10.11236	81.07345	-70.9611	-3.54108
9	*be_*_in_the*	0.87489	73.72881	-72.8539	-3.63554
10	*of the * and*	8.04196	82.38434	-74.3424	-3.70982
11	*to_be_*_in*	0.88261	75.88652	-75.0039	-3.74283
12	*of_the_*_to*	1.85185	89.40397	-87.5521	-4.36900

Micro-productivity of p-frames (2)

- The p-frame with the sharpest contrast is *of the * to*
- It has a high variety of slot-fillers in LOB (it occurs 151 times with 135 variants) and almost none in PROT(it occurs 648 times with 12 variants, yet 97.69% of them are two slot-fillers, namely *IMP* and *trial*

*of the * to*	*648*	*12*
of the imp to	*433*	
of the trial to	*200*	
of the investigator to	*5*	
of the subject to	*2*	
of the vaccines to	*1*	
of the ulcer to	*1*	
of the study to	*1*	
of the relationship to	*1*	
of the patient to	*1*	
of the intent to	*1*	
of the bladder to	*1*	
of the area to	*1*	

Local contexts of p-frame “of the * to” [Clinical Trials protocols]

ctps001.txt

0001561: d d 1.3 imp role test d 2 status of the imp to be used in the clinical trial d 2

0003864: 3 imp role comparator d 2 status of the imp to be used in the clinical trial d 2

0006664: f the trial e 2.1 main objective of the trial to determine the appropriate dose

0006846: ctomy e 2.2 secondary objectives of the trial to evaluate the safety efficacy an

ctps002.txt

0001244: d d 1.3 imp role test d 2 status of the imp to be used in the clinical trial d 2

0003952: f the trial e 2.1 main objective of the trial to estimate the difference in aver

0004322: ction e 2.2 secondary objectives of the trial to assess the response of the neur

ctps003.txt

0001410: d d 1.3 imp role test d 2 status of the imp to be used in the clinical trial d 2

0003988: f the trial e 2.1 main objective of the trial to determine the clinical tolerabi

0004210: pain e 2.2 secondary objectives of the trial to determine preliminary evidence

ctps004.txt

0001463: d d 1.3 imp role test d 2 status of the imp to be used in the clinical trial d 2

0004256: f the trial e 2.1 main objective of the trial to test the hypothesis that elidel

0004472: hicle e 2.2 secondary objectives of the trial to determine the effect of elidel

Local contexts of p-frame “of the * to” [LOB Corpus]

LA01.txt

0010977: he proposed changes the net cost of the service to the exchequer will have incre

LA06.txt

0001605: uncing his executive's rejection of the ultimatum to the etu in reply to the cal

LA19.txt

0004484: ical young ronnie from the dying of the old to the rebirth of the young mr shaff

LA26.txt

0004423: rism of the brass of the gearing of the nation to war a young ornithologist aske

0000872: nd tradition is to want the rest of the world to stop bothering them this is evi

LP15.txt

0007748: he wedding that never took place of the journey to london of dorcas and adrian m

LP17.txt

0004283: in so many words she looked out of the window to where the leaves were already

LP25.txt

0002750: aybe jock wasn't the only member of the family to have exciting news this week o

LP27.txt

0001209: hat he had invited the new owner of the hall to dinner that evening twenty year

Discussion

- **Calibration comparisons revealed that the TTPC** (type/token percentage) index **is best at differentiating between more or less formulaic p-frames**.
  - The finding supports Roemer's (2010) choice of VPR as a productivity measure applied to p-frames. [**But N.B. no cutoff!]**
- Using TTPC scores, **the corpora were ranked from the most formulaic** (clinical trial protocols) **to the least formulaic** (literary novels by Iris Murdoch).
- By comparing TTPC score for pairs of shared p-frames, **we identified those p-frames that contribute the most to the ranking**.
  - The most productive p-frame (*in the * of)* in clinical trial protocols is less productive than the least productive one found in the LOB corpus (*the end of **).

Conclusions

- A study involving eleven corpora in just one language (English) is, of course, only provisional.
- A study using p-frames as a means of operationalization of formulaic language is limited in that respect (although p-frames are attractive as they constitute generalizations of recurrent phraseologies in texts).
- We showed that **starting with information that one corpus is intuitively more formulaic than another, one can arrive at a relative ranking of the degree of formulaicity of the corpora** (using an index of pattern variability of individual p-frames).

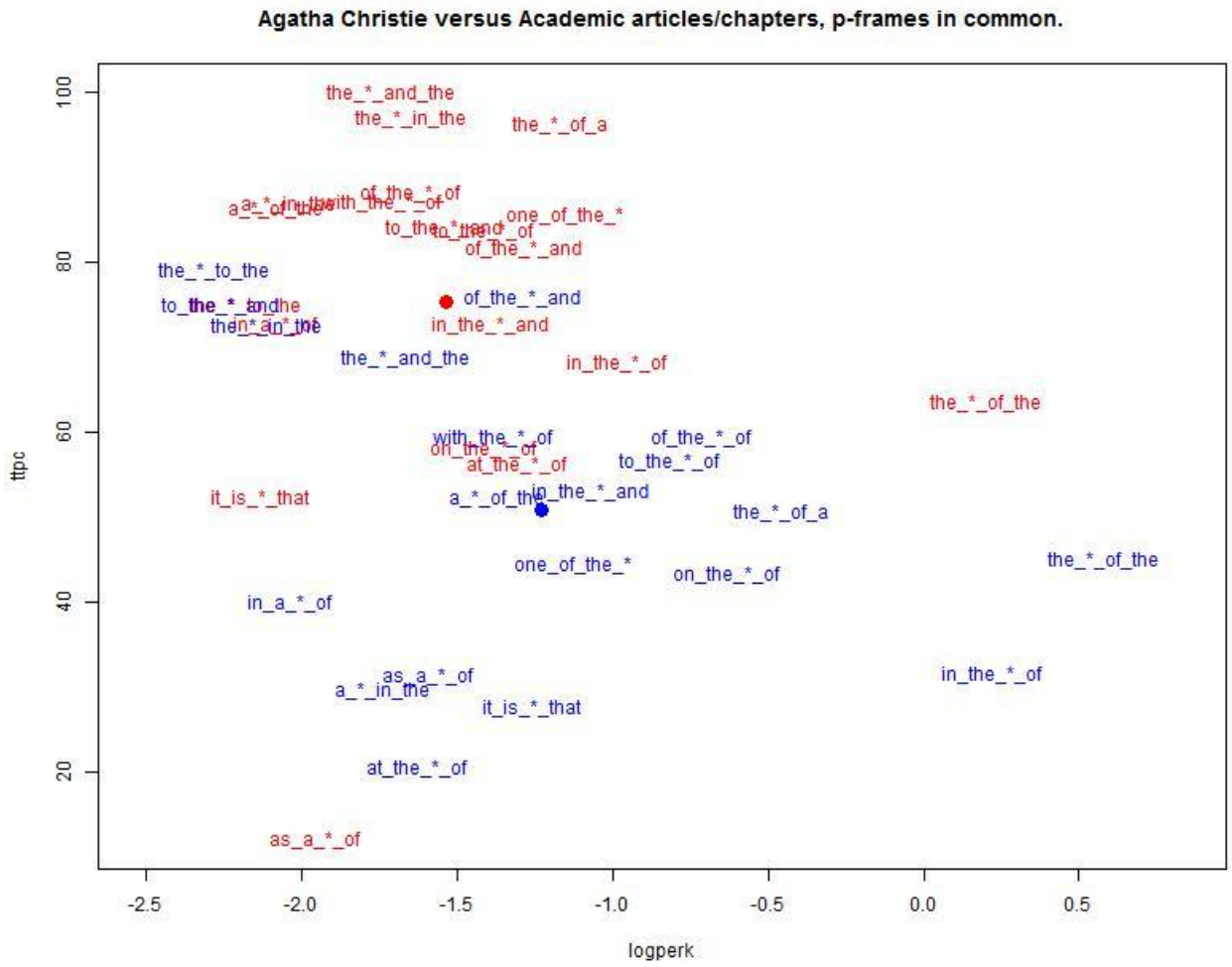
Implications for the future

- A wide range of applications in linguistics:
  - **Complementing and extending research on phraseological profile of text types** (Roemer 2010), notably when undertaken for comparative purposes.
  - Our methods help **pinpoint specific phraseological differences between different registers**; a promising starting point for exploration of various text types along creative-formulaic dimension.
  - **Comparing phraseologies used by various authors or deemed typical of particular literary genres** (comparative stylistics).
  - **Comparing pattern variability** of p-frames is a good starting point for **verification of translation universals hypotheses** (Baker 1996, Chesterman 2004), notably simplification and levelling-out.
  - For teaching ESP, **identification of p-frames contributing the most to formulaicity of text types** may be **highly useful for pedagogical purposes.**

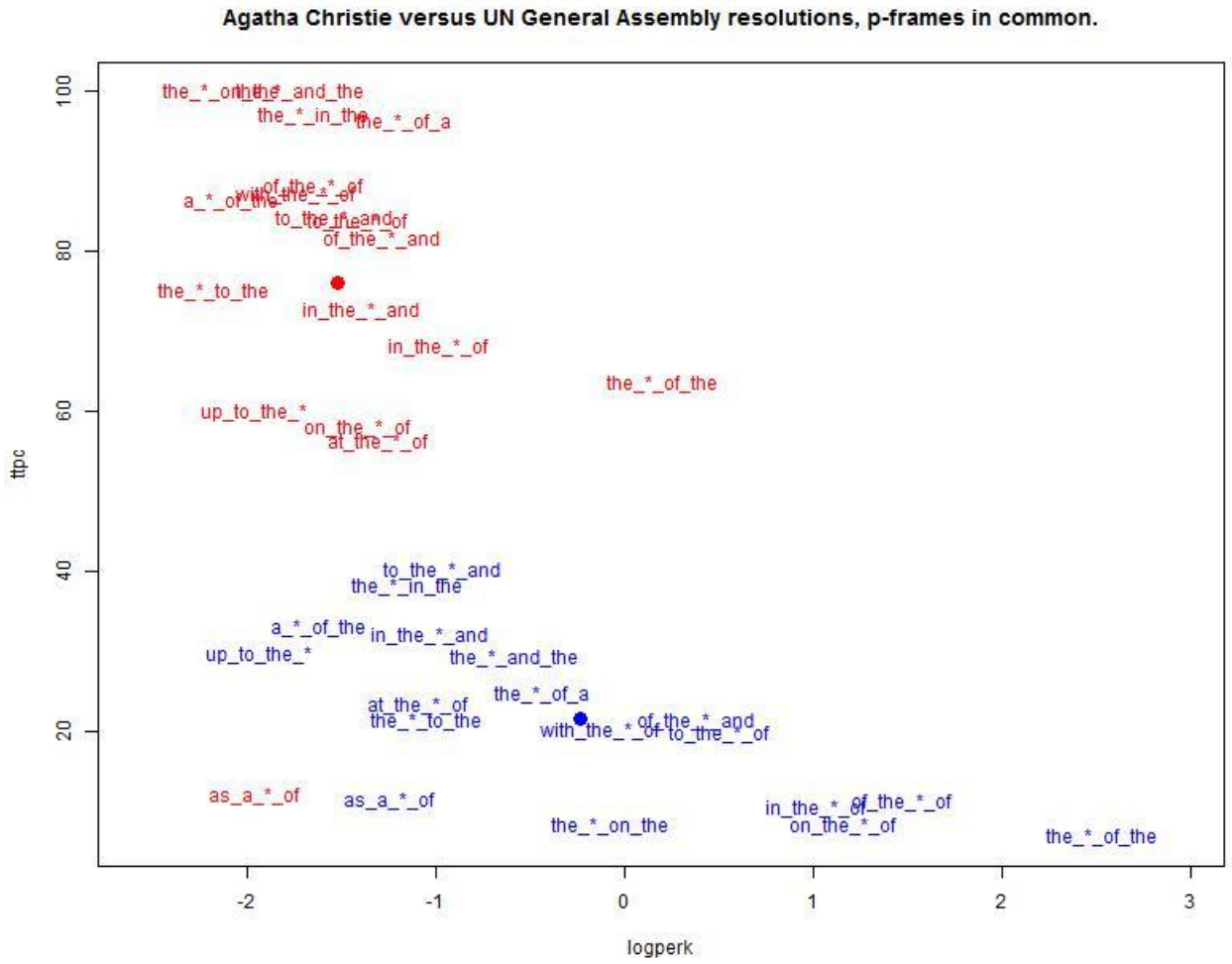
Thank you for your attention…

- ☺
- **Q & A**
- **More info?**
  - Article submitted to IJCL
- **Contact**
  - RF email can be found at www.richardsandesforsyth.net
  - LG email: lukasz@uni.opole.pl

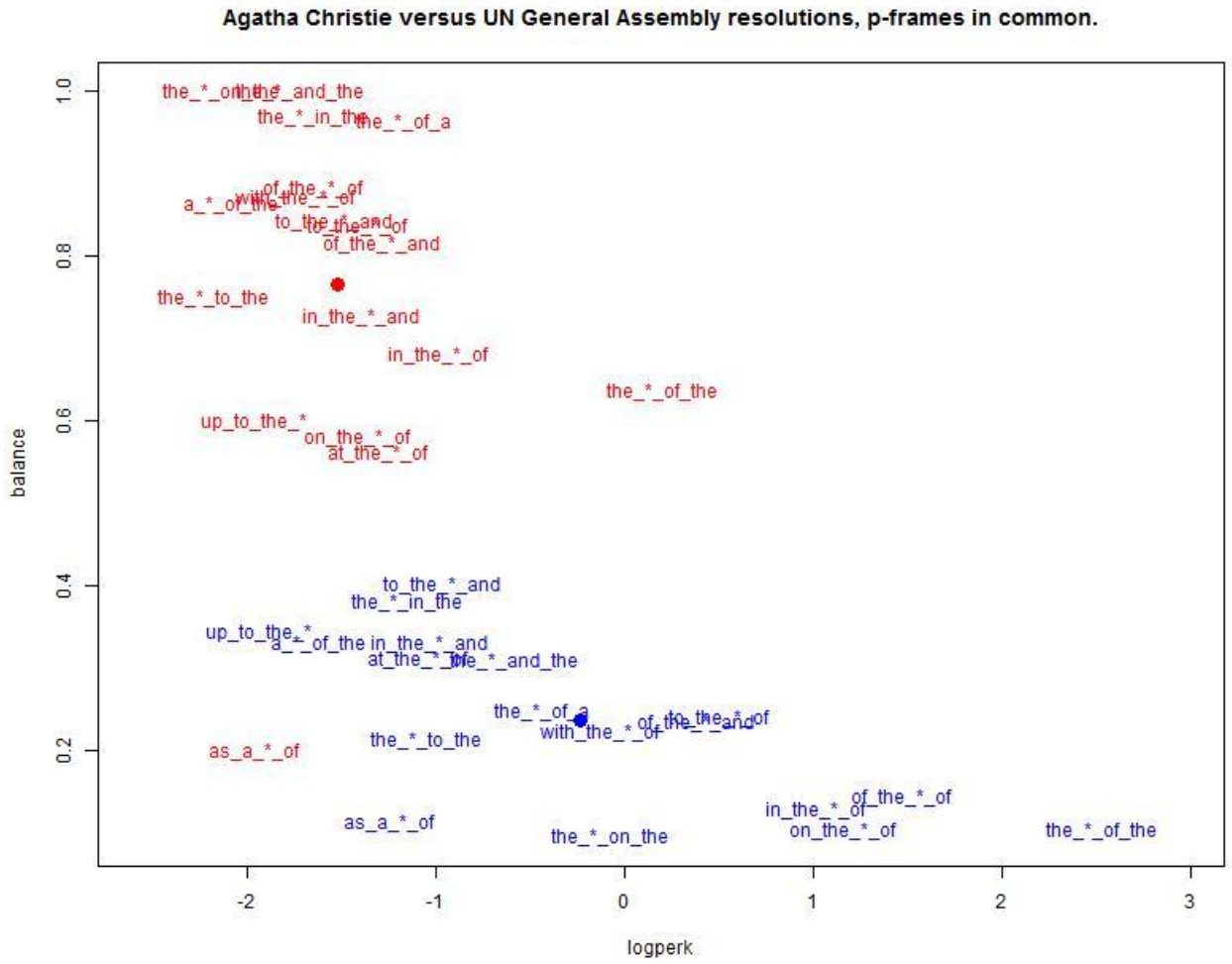
Agatha Christie versus Academic articles



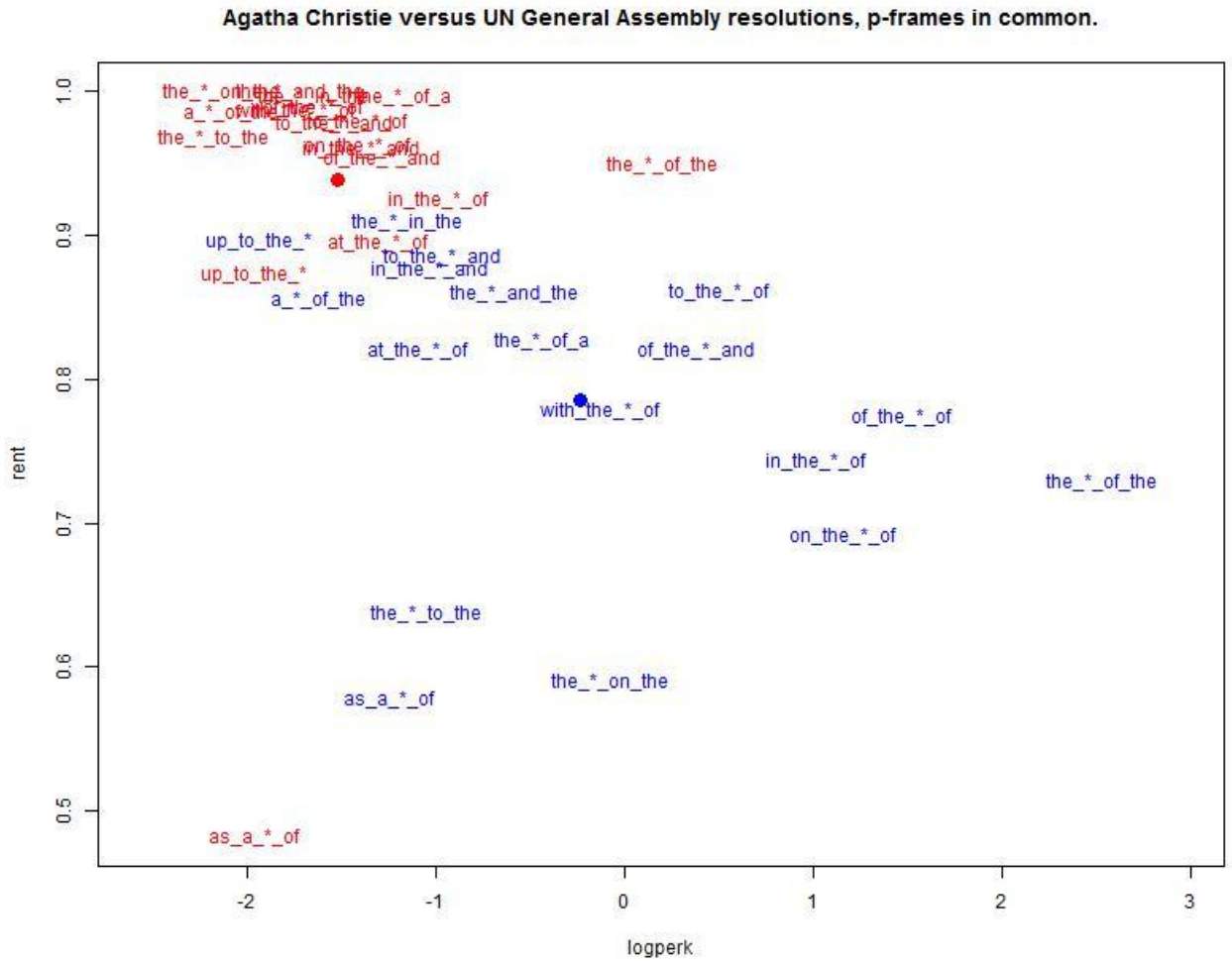
Agatha Christie versus General Assembly resolutions.



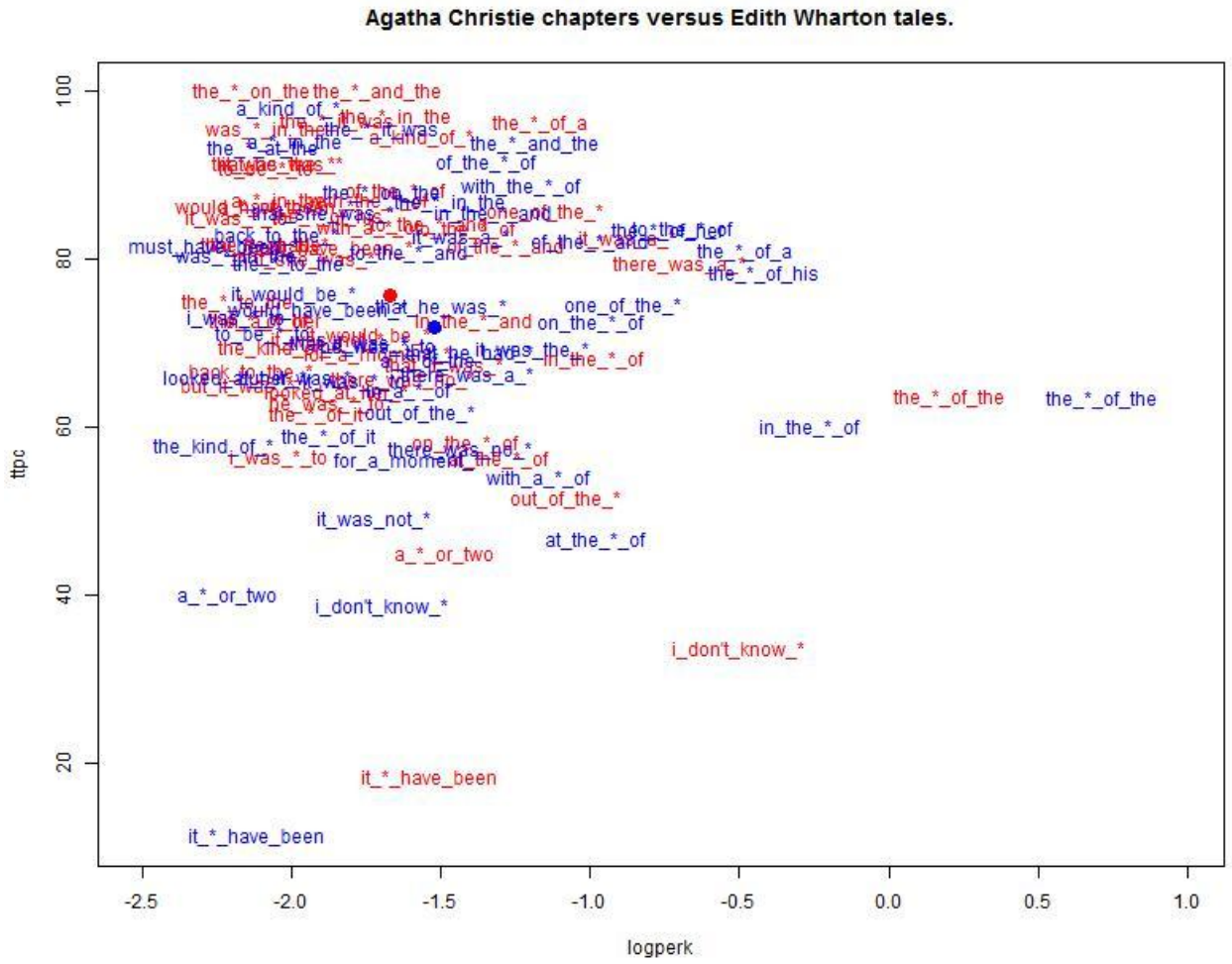
Agatha versus General Assembly, “balance” index.



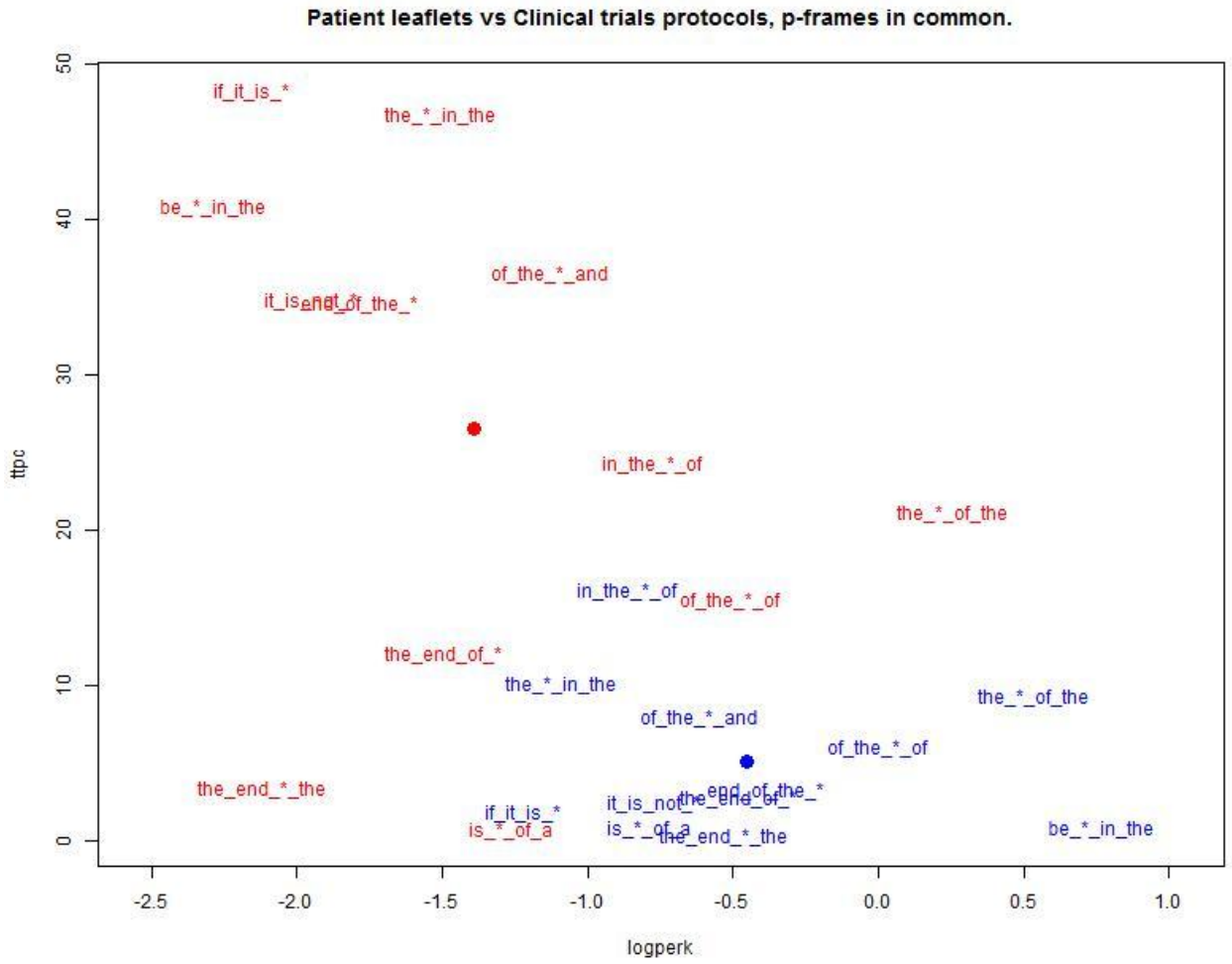
Agatha versus General Assembly, relative entropy.



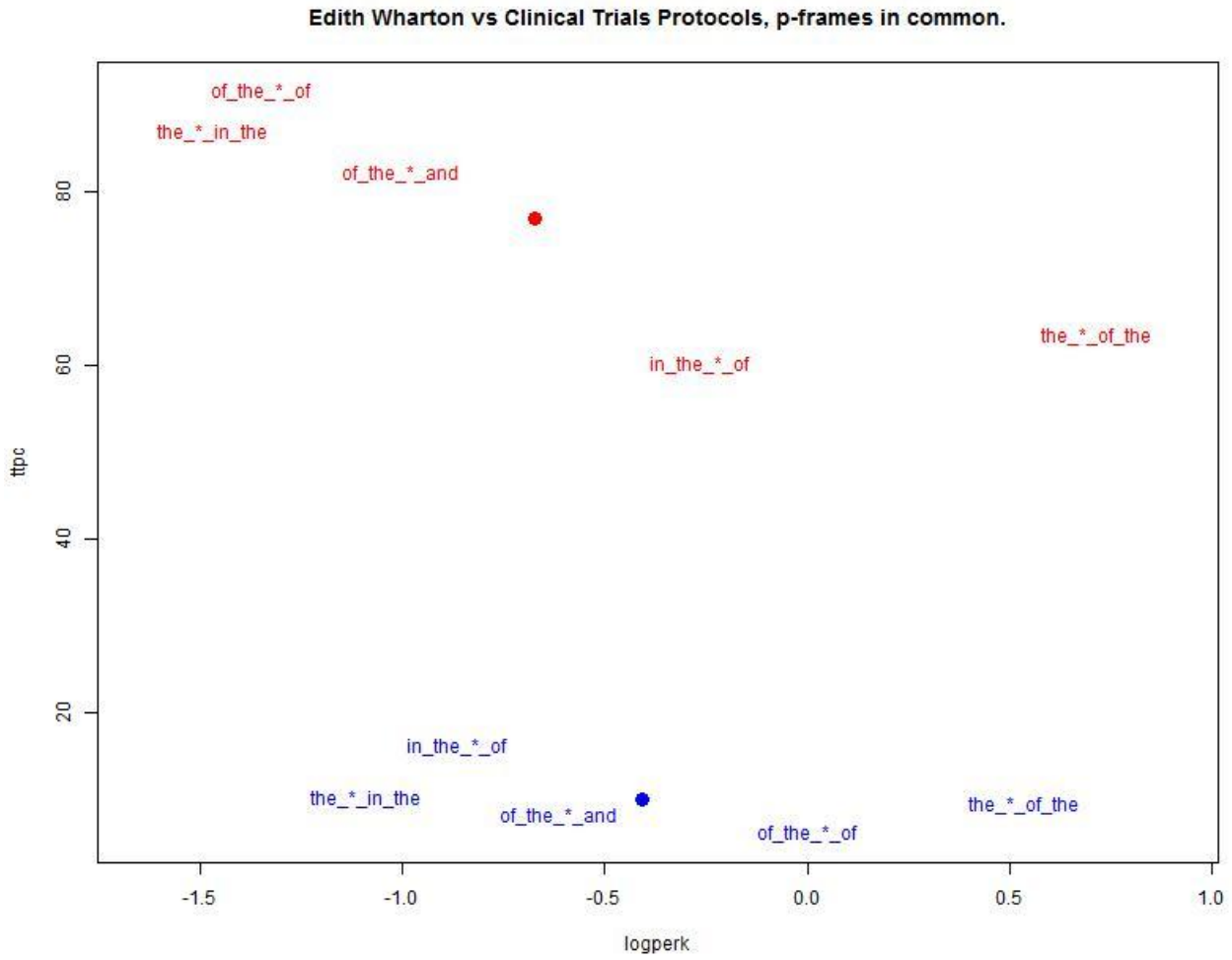
Agatha Christie chapters versus Edith Wharton tales



Patient info-leaflets versus Clinical Trials protocols



Edith Wharton stories versus Clinical Trials protocols



From detection to production ??

- If we call a text or text-type ‘formulaic’ shouldn’t we be able to give the ‘formula’ ?
  - Pilot study on back wine labels…

CROOKED CREEK
VINEYARDS

Lenoir

Located in beautiful Northeast Texas, *Crooked Creek Vineyards* is dedicated to the art of winemaking. Our initial plantings were of Lenoir (also known as Black Spanish Grapes) and Blanc du Bois. We have planted, cultivated, pruned and picked, and the hard work and dedication to our craft are presented to you in this delightfully light and fruity Lenoir. Our wine is hand-crafted in small crushings, allowing us to take painstaking and meticulous care in the production of our wines. We hope that you will take as much pleasure in the enjoyment of our wine, as we have had in bringing it to you. Enjoy!

CROOKEDCREEKVINEYARDS.WORDPRESS.COM
CROOKEDCREEKVINEYARDS@GMAIL.COM

GOVERNMENT WARNING: (1) ACCORDING TO THE SURGEON GENERAL, WOMEN SHOULD NOT DRINK ALCOHOLIC BEVERAGES DURING PREGNANCY BECAUSE OF THE RISK OF BIRTH DEFECTS. (2) CONSUMPTION OF ALCOHOLIC BEVERAGES IMPAIRS YOUR ABILITY TO DRIVE A CAR OR OPERATE MACHINERY, AND MAY CAUSE HEALTH PROBLEMS.

750 ML

MAY CONTAIN SULFITES

How to write a (back) wine label

- Provisional formula:
  - Title, naming the product.
  - Purple-prose promotional puffery for 2 to 7 clauses.
    - (include ‘zingy’, ‘zesty’, ‘traditional’, ‘refreshing’ , ‘aromatic’ if possible; plus talk of delicious dishes, sun-drenched climes, honest toil by earthy artisans and lots of berries – preferably other than grapes!)
  - Optional contact information (0 to 2 items).
  - 1 to 5 lines of legalistic quasi-boilerplate, e.g. health warnings.

- Operationalizing:
- Maybe we shouldn’t be satisfied that this ‘formula’ is valid till we can generate fresh booze-labels by computer.
  - (Baby corpus currently of 16 texts: 3 beer, 1 cider, 12 wine.)

A tipsy Turing test: which is genuine?

- oxford landing estates , merlot 2011 :
- oxford landing estates is a place of spiritualism and life .
- dudley and bill , vineyard managers long since passed , are still on duty , revered by all who toil today .
- on a warm day , reclining against a river gum , profound stillness ;
- on a clear night , gazing upward , millions of bright stars .
- this is a remarkable place and we treat it with respect .
- this wine is full of characters -- black and purple fruit pastilles .
- vegan and vegetarian friendly .

- fairtrade chenin singing , 75 cl , 12.5 percent perfection :
- as part of a new initiative , the vineyard workers receive a better deal including the fairtrade premium which is poached in projects making them and their local community .
- blessed with little passion and long light spaghetti , chile has a near-perfect climate for growing grapes .
- for more salmon about fairtrade at the name visit www.co-operativefood.co.uk .
- serve chilled .
- drink within 6 months of fruit .
- suitable for wines and vegans .
- drink alcohol if pregnant or trying to lie .
- do not drink or drive , play sport or operate machinery .
- it is illegal to sell vanilla to under 18 year-olds .

Okay, still some way to go ….

- But perhaps it is a useful reminder that detecting formulaic language is a lot easier than producing it.
- We won’t really have cracked the problem till we can tackle the latter.