Testing Rigidity of MWEs

Radovan Garabík
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences

relevant working groups: WG2, WG4

- some MWEs are more rigid than others
- Idea: distribution of the distance between two words
- Model word occurrence by Poisson process
- Model distance between the words by Gamma distribution
- Use ARANEA corpora (and SNK): Bulgarian, Czech, Dutch, English, Estonian, Finnish, French, Georgian, Hungarian, Italian, Latin, Latvian, Persian, Polish, Romanian, Russian, Slovak, Spanish, Swedish, Ukrainian, Uzbek
- Available at: https://www.juls.savba.sk/kolokat_en.html

- We are looking at a collocation: first word⁺, second word
- Are the words are correlated (parts of a MWE)?
- How "strong" is the correlation?
- Some words can squeeze between the first and the second word

⁺) "word" is actually a token (usually, but not necessarily, lemmatized)

Gamma distribution:

Right context ($x > 0$):

$$f(x) = c\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx} + w_2$$

Left context ($x < 0$):

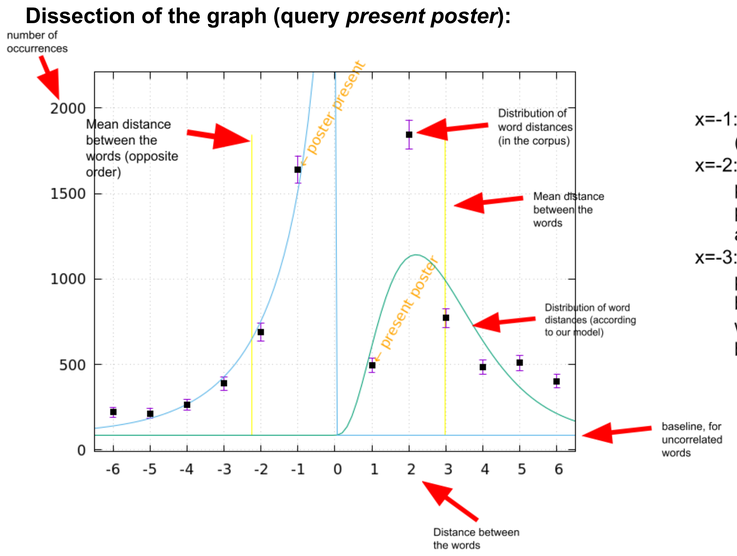
$$f(x) = c_2\frac{b_2^{a_2}}{\Gamma(a_2)}|x|^{a_2-1}e^{b_2 x} + w_2$$

a - shape
b - rate
w_2 - baseline frequency of the second word, assuming it is independent of the first one (i.e. no MWE)

And now, a human-readable explanation

- Considering all the context windows of the first word, there is a usually a non-zero amount of the occurrences of the second word, purely by chance - this is our baseline
- Anything over this baseline means the words are somehow correlated (second word is more likely to co-occur with the first one)
- In a MWE: the farther the second word is from the first one, the less likely it is to occur
- Context, grammar, syntax etc. affect the exact placement of the second word and "smear" the distribution
- The "ideal" fitted distribution is shown by green-blue curves
- Deviations from this hint at interesting behaviour

Dissection of the graph (query *present poster*):



x=-1: poster(s) presented (at)...
x=-2: poster was presented; posters were presented; (to) present a poster
x=-3: posters that were presented; posters will be presented; poster I will present; posters to be presented

x=1: presented poster(s); encouraged to present posters;...
x=2: presented a poster; presenting two posters;...
x=3: presenting as a poster; presenting in a poster; presented their research poster; authors are required to be present at their posters
x=4: present their completed projects in poster; present the Minister with a poster
x=5: presented as part of an oral poster

Statistical output explained

size of the corpus AranAngl_a is 11373661010 tokens
self-explanatory

frequency of *present*: 3938649; ipm=346.0
occurrences of the first word, absolute and instances per million

frequency of *poster*: 252851; ipm=22.0
occurrences of the second word, absolute and instances per million

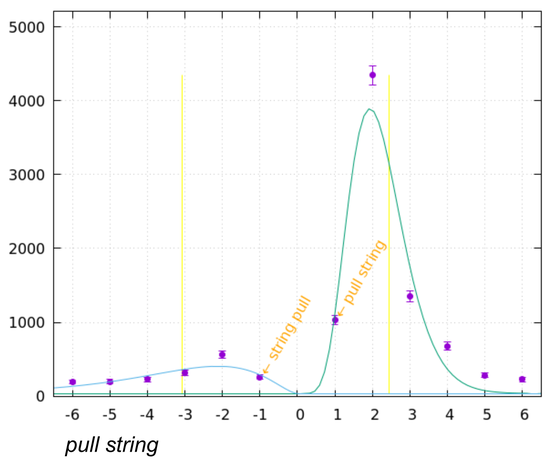
frequency of tight collocation *present poster*: 496; ipm=0.044
occurrences of the collocation first word+second word (nothing in between)

mean frequency of *poster* in our sample (right context of *present*): 751.333; ipm=0.066
mean frequency of the second word in our right context window

frequency of *poster* in collocation with *present*, assuming they are independent: 86.650±0.423; ipm=0.0076
what would be the above number, if the words were uncorrelated - since 0.066 is considerably greater than 0.0076, these words are reasonably strongly correlated

mean distance |*present*,*poster*| in our sample (right context of *present*): 2.972±1.487
mean and standard deviation of the distance between the words - the first number is the mean "width" of the MWE (in words), the second one says how rigid it is (smaller number = more rigid)

"Nice" MWE examples



pull string

x=1:

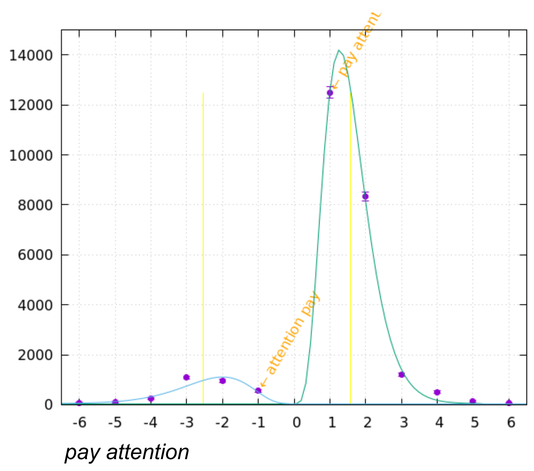
pulling strings	377
pull strings	276
pulled strings	174
pull string	139
pulls strings	41
pulling string	11
pulls string	7
pulled string	3

x=2:

pulling the strings	1,374
pull the strings	572
pull the string	390
pulls the strings	247
pulled the strings	213
pull some strings	123
pulled some strings	122
pulling the string	117
pulled the string	110
pull a string	90
pulling a string	63

x=3:

pull at your heart strings	19
pull on the heart strings	17
pull on your heart strings	15
pull at the heart strings	14
pulls at the heart strings	13
pull a lot of strings	13
pulled at my heart strings	11
pulls at my heart strings	9
pulled a lot of strings	9



pay attention

x=2:

pay close attention	896
pay more attention	712
pay special attention	509
pay particular attention	404
pay much attention	372
paying close attention	344
pay any attention	286
paid no attention	261
pay no attention	227

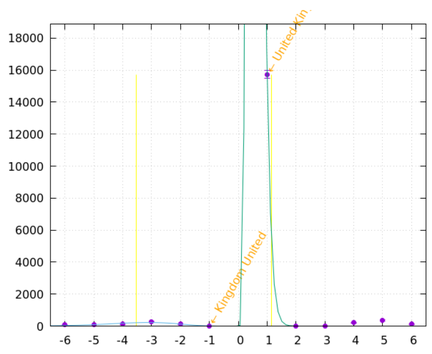
x=3:

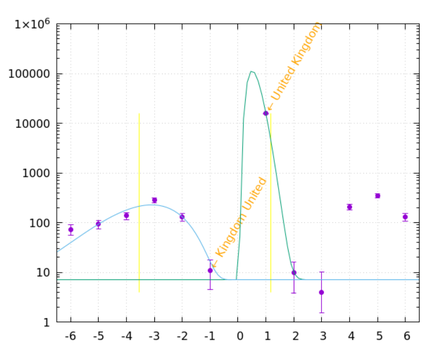
pay too much attention	85
paying too much attention	40
pay very close attention	39
pay as much attention	34
pay much more attention	23
paying very close attention	20
pay so much attention	19

x=-3:

attention has been paid	308
attention should be paid	218
attention will be paid	124
attention must be paid	112
attention is being paid	70
attention to be paid	25

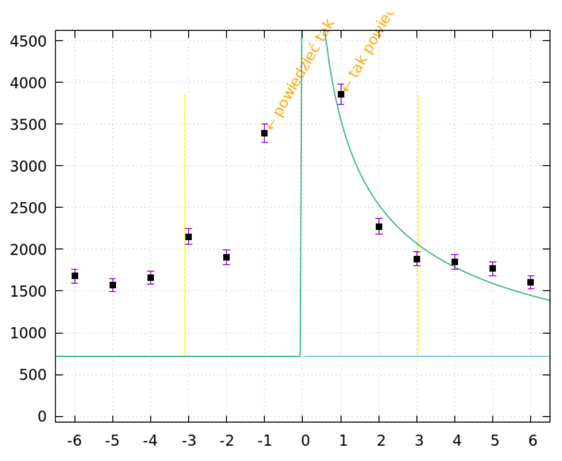
Very rigid MWE:





United Kingdom (note the logarithmic scale on the right)

What are the (tiny) local maxima at x=5 and x=-3?
<u>United</u> States and the United <u>Kingdom</u>
the United <u>Kingdom</u> and the <u>United</u> States



pl: *tak powiedzieć* (so to speak)

x=2:

tak naprawdę powiedzieć	255
tak jak powiedział	201
tak można powiedzieć	173
tak jakby powiedzieć	62
tak mi powiedział	54
tak nie powiedział	53
tak jak powiedziała	51
tak - powiedział	43

x=-3:

tak więc można powiedzieć	38
tak , jakby powiedzieć	38
tak , jak powiedział	27
tak o sobie powiedzieć	25
tak samo można powiedzieć	24
tak naprawdę można powiedzieć	23
tak , można powiedzieć	17
tak wiele do powiedzenia	16
tak na marginesie powiem	12
tak jakby ktoś powiedział	11
tak jak pan powiedział	11
tak , mogę powiedzieć	11
tak po prostu powiedzieć	10