

# Form-Meaning Regularities in Old English Thesaurus and Corpora

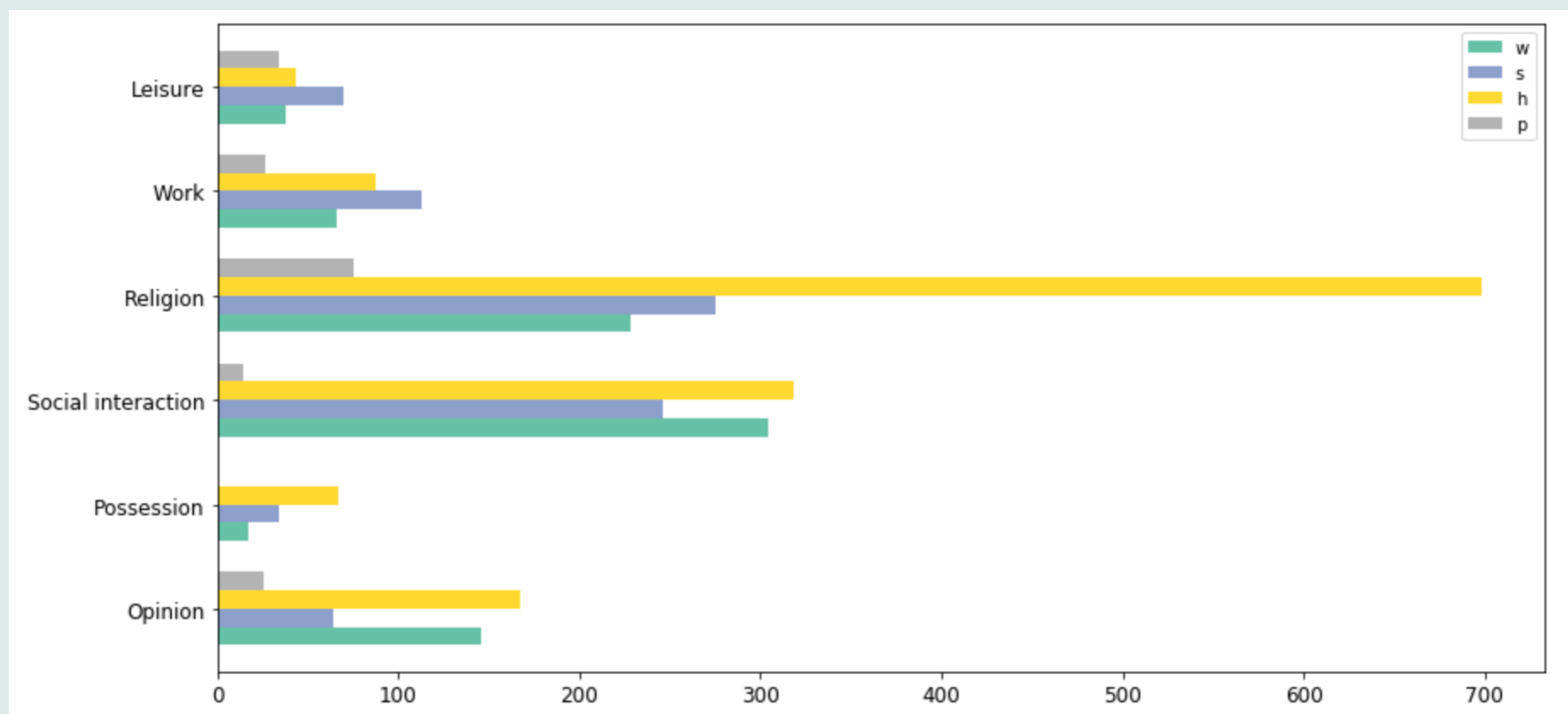
Nataliia Drozhashchikh, Elena Efimova, Evgenia Meshcheryakova

Language at its earlier stages features non-arbitrary form-meaning systematic relationships

Linguistic systematicity is a statistically revealed isomorphic relationship between form and meaning

## Lexicon

A Thesaurus of Old English manifests non-regular distribution of semantic categories among the lexemes sharing initial graphemes <w>, <s>, <h>, <p>.



The contingency table built on the frequency counts of lexemes sharing the same initials per semantic category allows to apply Chi-square statistics.

	<w>	<s>	<h>	<p>
matter and measurement		27.99		
opinion	30.73		2.04	6.8
possession			17.74	
social interaction	50.36			
religion			165.79	13.66
work		3.31		21.59
leisure		1.6		102.62

The statistical values indicate non-random association between the lexemes sharing initial graphemes <w>, <s>, <h>, <p> and definite semantic categories in A Thesaurus of Old English.

grapheme	semantic fields		
<w>	social interaction: <i>werod</i>	opinion: <i>wlanc</i>	emotion: <i>weorc</i>
<s>	matter and measurement: <i>samen</i>	existence: <i>sālnes</i>	the physical world: <i>seolfor</i>
<h>	religion: <i>halig</i>	possession: <i>habban</i>	
<p>	leisure: <i>pipe</i>	work: <i>pāl</i>	religion: <i>preost</i>

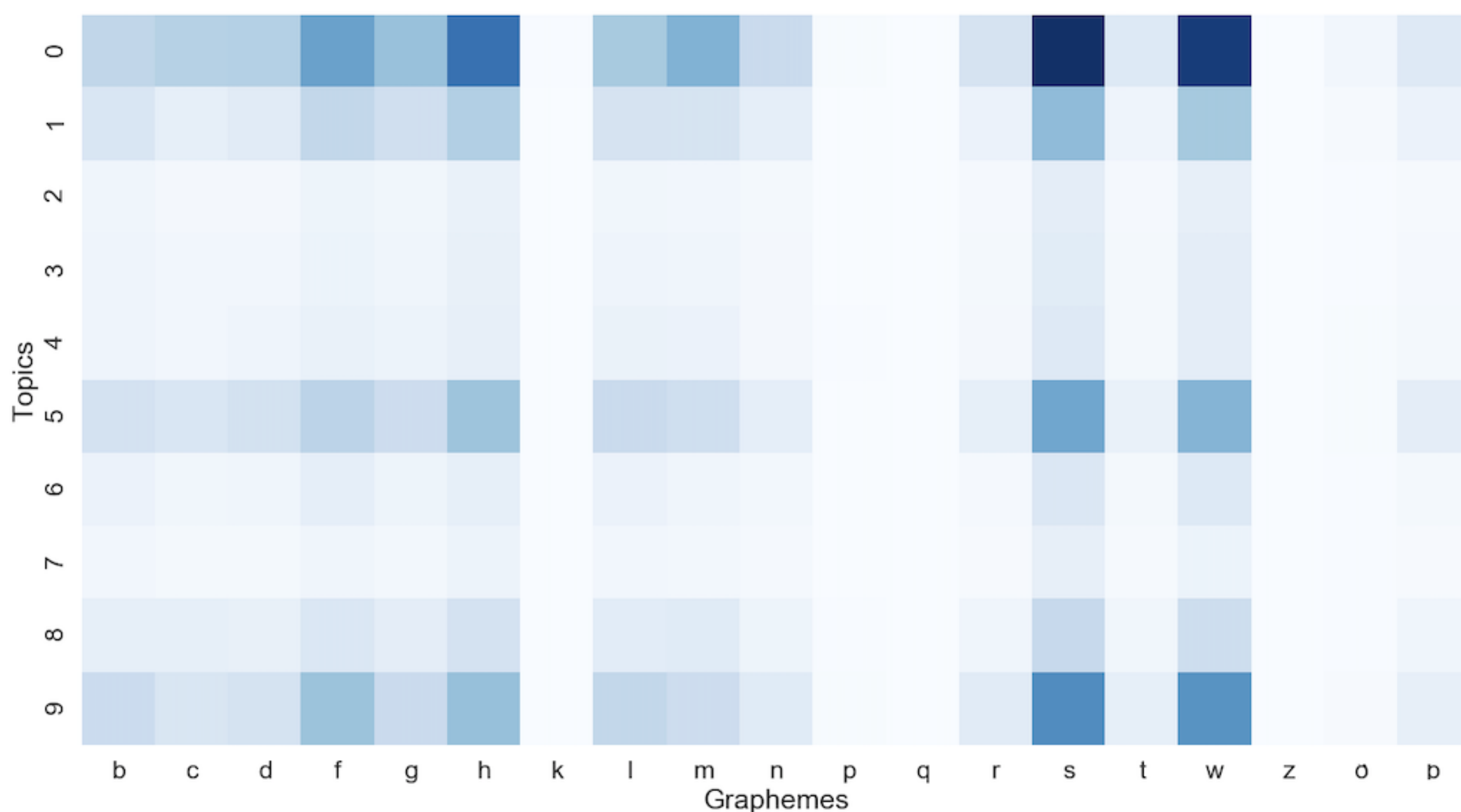
# Corpora

Topic modelling LDA algorithm discerns the dominant themes of religion and man/God interrelation based on the frequency of key words and their contextual co-occurrence.

topic	top-15 relevant words
0	god, drihten, eorðe, man, folc, miht, woruld, nama, weorþan, fæstan, georne, dæg, ece, cyning, hand
1	weorþan, man, god, sunu, eald, beirnan, geata, folc, heard, helm, fæstan, gold, hand, cyning, hatan
2	wyrd, gold, wat, lif, beorht, hand, hruse, mod, geweorc, sincan, rice, eorþe, blæd, windan, heofon
3	god, eorþe, man, hosp, men, healdan, wat, woruld, weorþan, wyn, hycgan, wide, sorh, ofereode, picgan
4	god, man, drihten, fæstan, hosp, healdan, biddan, dæg, fæder, crist, symle, lufan, sceolan, eald, mode
5	god, wuldor, drihten, weorþan, engel, heofon, cyning, woruld, man, beorht, fæder, ham, eorðe, tid, hand
6	fugol, wuldor, weorþan, god, lif, eorþe, folde, beorht, wynsum, sunne, woruld, swegl, fæder, drihten, heofon
7	heofena, etan, cyning, kyning, beorn, engel, land, secan, eald, weorþan, nos, god, drihten, georne, gumena
8	god, wuldor, weorþan, miht, cyning, folc, andreas, engel, beorn, domine, ece, halgan, mode, fæder, hosp
9	god, weorþan, sunu, drihten, man, eorðe, engel, hatan, beirnan, waldend, abraham, woruld, wide, wiht, folde

The texts in the poetic corpus demonstrate the distribution of the most relevant themes presented by the top key words: *god* ‘God’, *weorþan* ‘to become’, *man* ‘man’, *drihten* ‘God’, *engel* ‘angel’, *eorðe* ‘earth’, *beorht* ‘bright’, *cyning* ‘leader’, *fæder* ‘Father’, *folc* ‘people’, *wuldor* ‘glory’, *woruld* ‘world’.

Topic-grapheme matrix indicates that a grapheme can be represented similarly in several topics.



Non-random distribution of semantics over the words sharing an initial grapheme is attested at the lexical level but topic modelling did not detect the expected association. Since the same initial grapheme enters several different topics it serves as a negative evidence for the proposed hypothesis.

Calgary\_corpus. [https://github.com/qcrit/NHB-2018-OEstylometry/tree/master/sense\\_pauses/Calgary\\_corpus](https://github.com/qcrit/NHB-2018-OEstylometry/tree/master/sense_pauses/Calgary_corpus).

Taylor, A., Warner, A., Pintzuk, S., Beths, F. (2003).

The York-Toronto-Helsinki Parsed Corpus of Old English Prose. <http://www-users.york.ac.uk/~lang22/YcoeHome1.htm>.