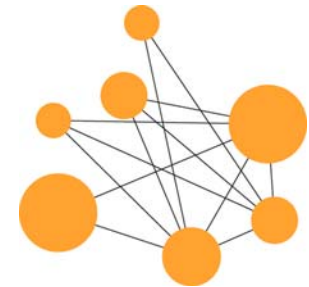


Textual Characteristics of Full-text Biomedical Journal publications in Open Access vs. Traditional Journals



Karin Verspoor

Karin.Verspoor@ucdenver.edu

K. Bretonnel Cohen

Kevin.Cohen@gmail.com

Larry Hunter

Larry.Hunter@ucdenver.edu



Research on full-text open access publications is increasing

- The bulk of work on text mining or natural language processing in the biomedical domain to date has been done on PubMed abstracts
- Recent work is beginning to consider full text
 - Shah P, Perez-Iratxeta C, Bork P, Andrade M: Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics* 2003, 4(1):20.
 - Natarajan J, Berrar D, Dubitzky W, Hack C, Zhang Y, DeSesa C, Van Brocklyn J, Bremer E: Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. *BMC Bioinformatics* 2006, 7(1):373.
 - Rzhetsky A, Iossifov I, Loh JM, White KP: Microparadigms: Chains of collective reasoning in publications about molecular interactions. *Proc Natl Acad Sci USA* 2006, 103(13):4940-4945.



PubMed Central

- Free access to full text articles
- PubMed Central Open Access subset: full text articles made available under a more liberal reuse and redistribution license
- We anticipate much research on the open access subset of full text articles, but it represents only a small portion of all full text articles published
- ***Is it representative?***



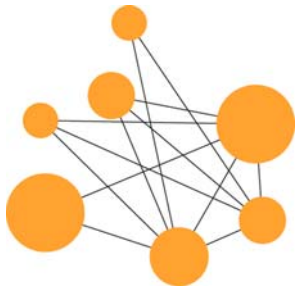
Our corpora

- CRAFT: Colorado Rich Annotation of Full Text
- TraJour: Traditional Journals corpus
- Reference (general): Penn Treebank Wall Street Journal subset
- BioReference: random selection, balanced between traditional journal articles and PMC open access articles

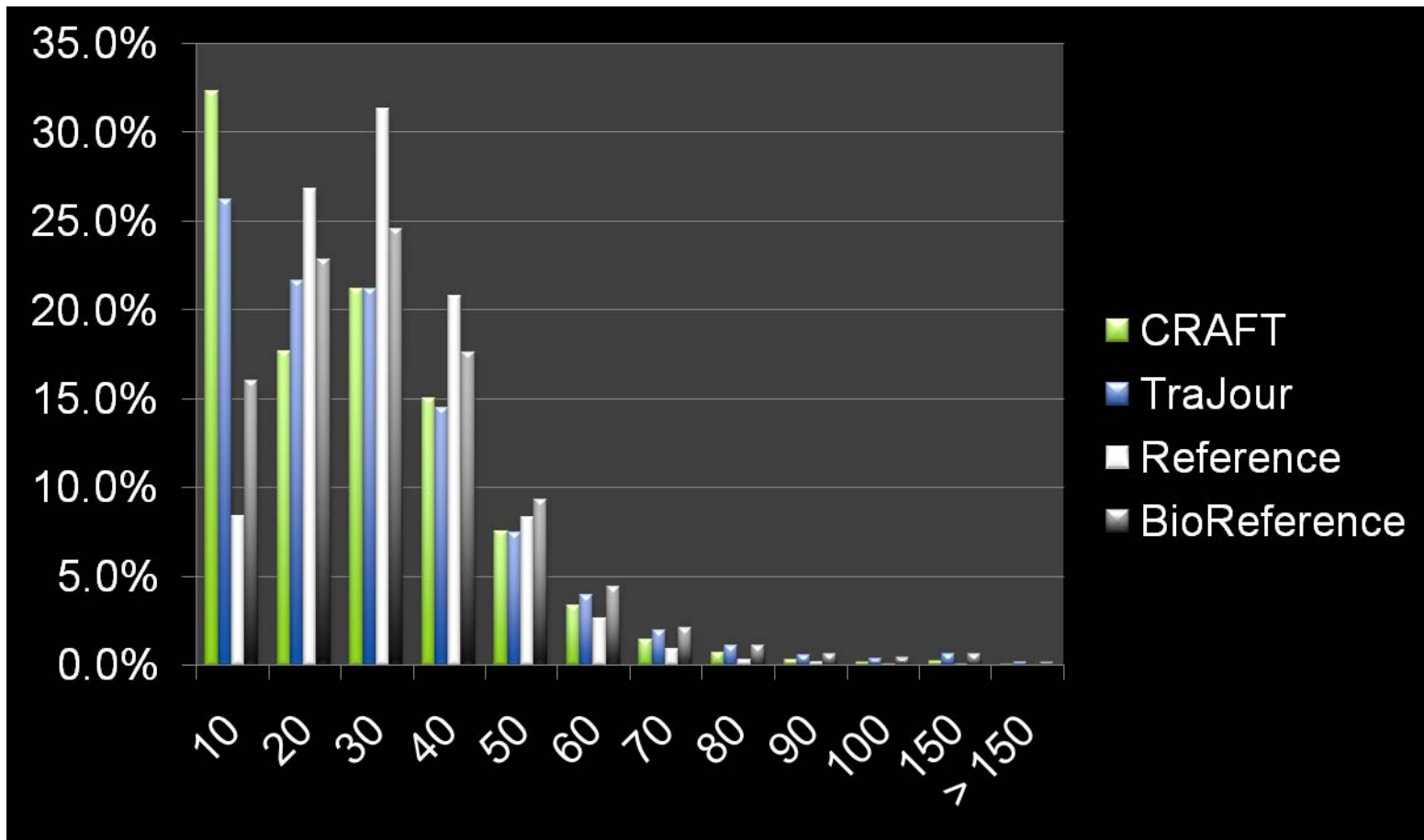


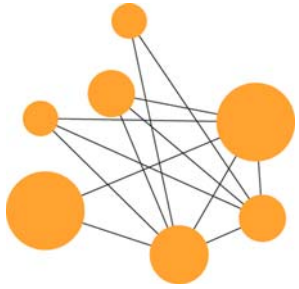
The Stats

	CRAFT	TraJour	Reference	BioReference
num. documents	97	99	2500	163
num. sentences	43694	35997	53107	32895
avg. num. sentences	450	364	21	202
tokens	717166	598331	1096976	654493
types	41574	49394	40139	38801
stopword tokens	238542	193905	453264	238077
stopword %	33.3%	32.4%	41.3%	36.4%
avg. doc length	7393	6044	439	4015
type/token ratio	5.8%	8.3%	3.7%	5.9%
token/type	17.3	12.1	27.3	16.9
negatives	3273	2587	7605	2961
negatives %	0.46%	0.43%	0.69%	0.45%
coordination	25237	23706	26019	25059
coordination %	3.52%	3.96%	2.37%	3.83%
pronouns	18874	15603	57406	20699
pronouns %	2.63%	2.61%	5.23%	3.16%
passives	2783	2587	2661	3172
passives %	0.39%	0.43%	0.24%	0.48%

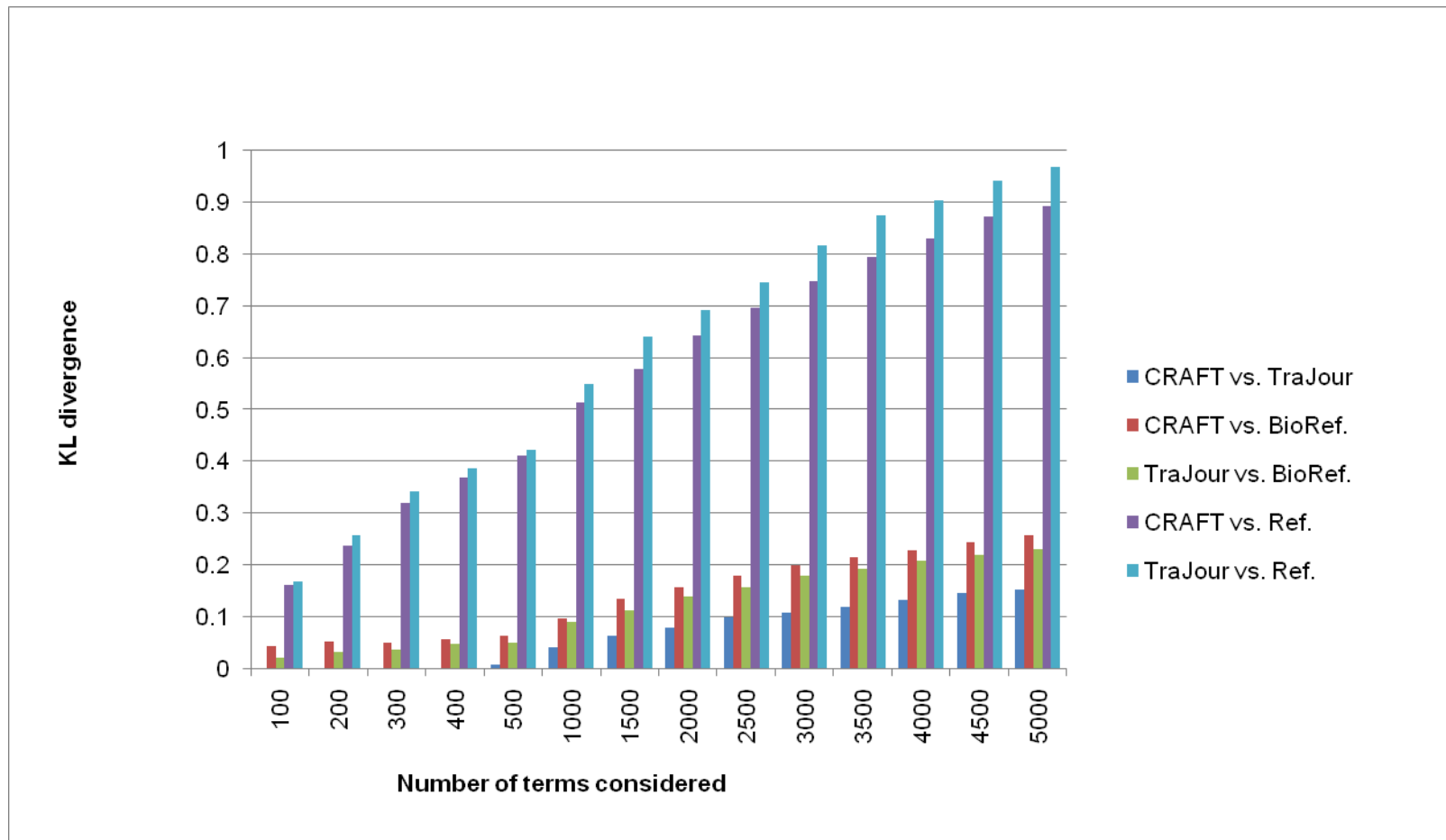


Sentence Length Distribution





KL divergence of term probability distributions





Conclusions

- On nearly every measure, the CRAFT and TraJour corpora are more similar to each other than to either reference corpus
- The vocabulary and term distributions for CRAFT and TraJour are nearly identical, in particular for the most frequent terms
- **Open Access articles are a perfectly acceptable surrogate for biomedical publications in general**