Extracting MWE and fixed structures from parsed corpora

University of Zurich[UZH]

COST PARSEME WG 1, Athens
www.es.uzh.ch/dbank

Gerold Schneider
gschneid@es.uzh.ch

Introduction

COLLOCATION MEASURES are applied to large parsed corpora (Lehmann and Schneider, 2012), followed by manual filtering to automatically extract several categories of multi-word entities (MWE). Our dependency parser Pro3Gres (Schneider, 2008) combines a hand-written competence grammar, which represents the syntax principle, with statistical performance disambiguation, which represents the idiom principle (Sinclair, 1991).

Verb-Preposition Structures

WE USE O/E as collocation measure. O/E has a tendency to report rare collocations: in traditional windows-based approaches, garbage appears at the top.

- Approaches based on parsed corpora provide considerably cleaner data (Seretan, 2011)
- Paired with a T-score significance threshold O/E delivers very good results.
- 2nd key criterion is fixedness. We use Yule's K as a measure of diversity: proven independence on token counts.

Table 10. VOPN 4-tuples ordered by O/E, filtered by t-score in BNC-W written. (Full table.)

verb	object	prep	desc noun	t-score	O/E
send	shiver	down	spine	5.74456	2.21477×10^{8}
tap	esc	for	escape	6.40312	2.1134×10^{8}
separate	shield	from	plate	6.78233	2.33384×10^{7}
refer	gentleman	to	reply	8.24621	7.8143×10^{6}
obtain	property	by	deception	5.2915	7.60043×10^{6}
ask	secretary	for	affairs	6.40312	5.01529×10^{6}
kill	bird	with	stone	5.38516	3.37917×10^{6}
add	insult	to	injury	6.08276	2.21769×10^{6}
throw	caution	to	wind	5.09902	2.03157×10^{6}
refer	friend	to	reply	7.54983	1.36298×10^{6}
report	loss	on	turnover	7.14142	1.34742×10^{6}

(Lehmann and Schneider, 2011) : http://www.helsinki.fi/varieng/journal/volumes/06/lehmann_schneider/

Regional variation in fixedness can be observed (Schneider and Zipp, 2013), e.g. "This *resulted into* a deep sense of growing loneliness" (ICE India)

Light Verb Constructions

OUR AUTOMATIC DETECTION of light verbs is described in Ronan and Schneider (submitted). We use several collocation measures.

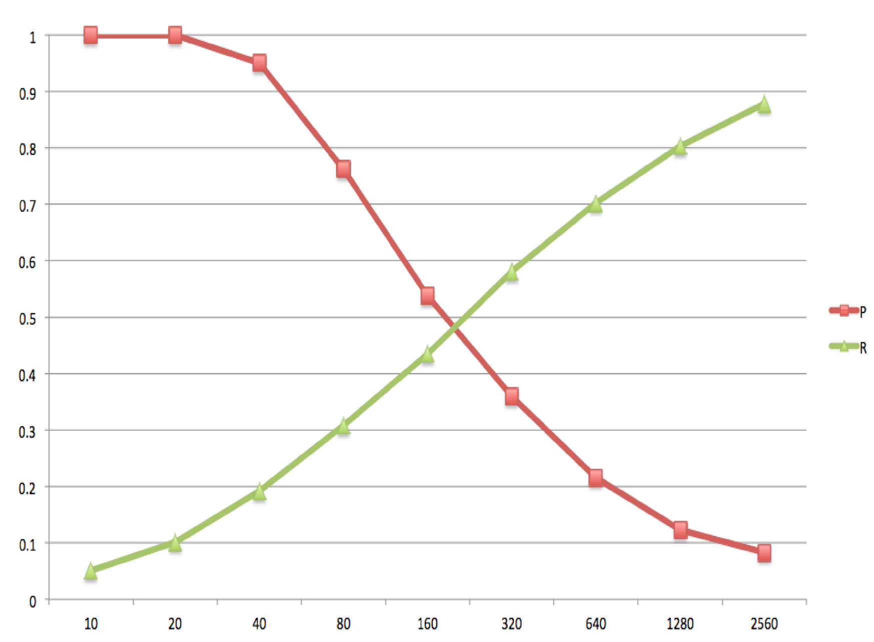
T-Score on BNC, correct ones marked

117249	OE	T	Chi	V	Obj	f	f(V)	f(N)	manu
bncx	22.4051	97.0768	219641.0	take	place	10325	128201	21501	+
bncx	8.4774	64.0072	52584.1	have	effect	5266	303211	12254	+
bncx	272.553	59.5303	968964.0	shake	head	3570	6221	12594	
bncx	52.7729	55.3838	167118.0	see	pp	3187	112461	3212	
bncx	5.2405	55.036	22132.7	do	thing	4626	157530	33518	
bncx	41.2167	53.5146	119208.0	ask	question	3008	30365	14376	+
bncx	92.0118	49.4467	226112.0	play	role	2499	19223	8451	+
bncx	32.7848	47.9877	76299.1	play	part	2450	19223	23253	+
bncx	6.3164	47.2243	18094.1	take	part	3148	128201	23253	+
bncx	7.0943	47.083	21023.8	do	anything	3004	157530	16078	
bncx	31.4713	46.4546	68682.6	go	home	2302	26840	16301	
bncx	4.6561	46.3483	15831.8	do	something	3484	157530	28412	
bncx	12.7681	46.2312	31919.8	make	sense	2516	147869	7971	+
bncx	7.8569	46.1637	21934.2	do	job	2798	157530	13522	+
bncx	12.7301	45.6558	31139.6	make	decision	2455	147869	7801	+
bncx	142.241	45.319	293114.0	open	door	2083	11317	7740	
bncx	4.5442	44.5113	25899.6	have	idea	3257	303211	14139	+
bncx	159.397	43.1556	297871.0	answer	question	1886	4923	14376	
bncx	6.3766	43.035	28310.3	have	look	2605	303211	8059	+
bncx	10.938	42.4899	24204.3	make	use	2187	147869	8088	+

Precision = 12 / 20

ICAME 2013, Santiago de Compostela

Evaluation: *give* Precision & Recall on BNC, using T-Score & simple filter



- Here T-score works best: frequency of LVC is a factor
- LVCs are an open list, and gradient
- some regional variation: e.g. take vs. make decision

Subject-Verb-Object and Others

STRONG BUT GRADIENT idiomatic and selectional preferences prevail on all levels, e.g. verb-object, subject-verb, in syntactic structures, morphology, alternation preferences.

In addition to extracting MWE classes with arbitrary borders, abstracting to probabilistic interdependent features is useful:

- bi-lexical preferences (Collins, 1999; Hoey, 2005)
- construction grammar (Stefanowitsch and Gries, 2003)
- information-theoretic measures such as surprisal (Levy and Jaeger, 2007)
- parsers aggregating the probabilistic information from all levels

subject-verb-object heads	f(svo)	f(s)	f(v)	f(o)	O/E
coroner_record_verdict	33	284	5389	517	6.08756e+12
spine_form_fan	23	113	12659	547	4.29051e+12
heart_miss_beat	26	1590	7393	222	1.45428e+12
clause_exclude_liability	25	744	3061	1126	1.42302e+12
jury_return_verdict	45	669	16513	517	1.15005e+12
sale_start_monday	23	1667	17732	159	7.14306e+11
female_lay_egg	29	683	6985	1317	6.73708e+11
sale_start_december	31	1667	17732	250	6.12315e+11
republic_achieve_independence	24	596	11145	1130	4.66717e+11
error_occur_error	21	583	13564	961	4.03354e+11
court_grant_injunction	22	7229	3247	345	3.96543e+11
inc_report_profit	145	1802	11740	2711	3.69031e+11
plc_report_profit	22	288	11740	2711	3.50332e+11
index_close_point	204	947	13999	8451	2.6578e+11
tenant_pay_rent	21	865	23676	666	2.24733e+11
corp_report_profit	65	1380	11740	2711	2.16015e+11
price_include_breakfast	195	3462	45097	948	1.92308e+11
history_repeat_itself	29	1186	3841	6172	1.50553e+11

Gradient Multi-Words

LEXICAL PRIMING is the key factor for Hoey (2005): "lexis is complexly and systematically structured and that grammar is an outcome of this lexical structure" (1),

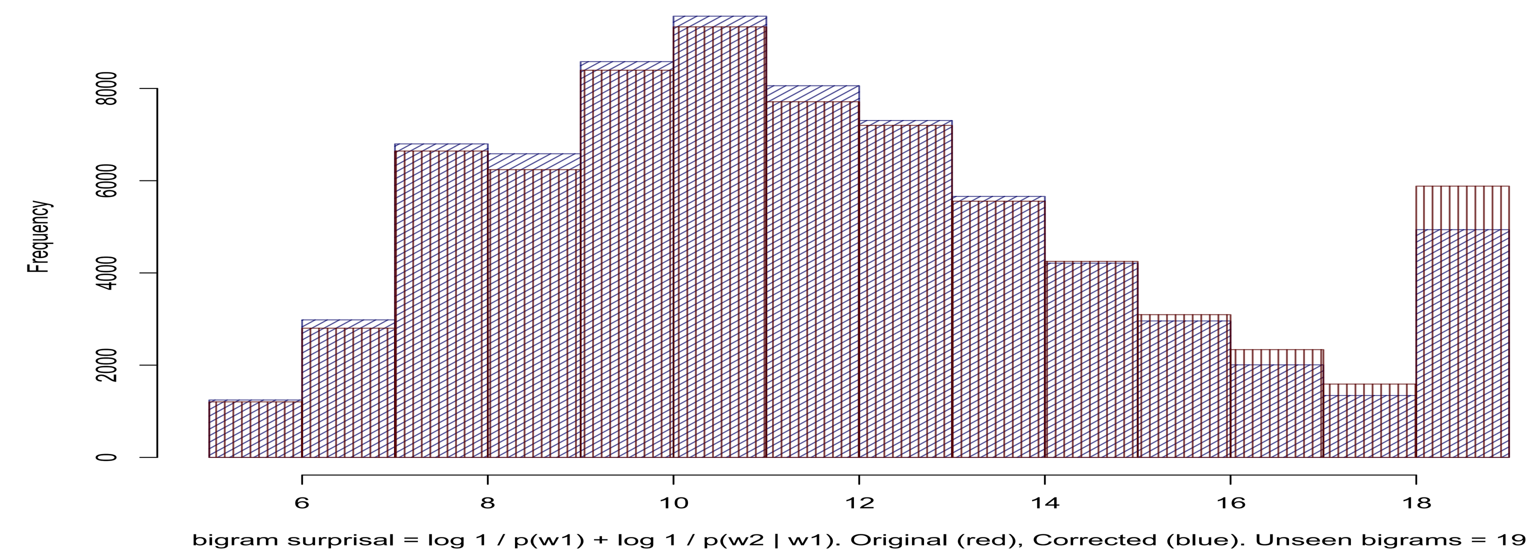
"We can only account for collocation if we assume that every word is mentally primed for collocational use" (8)

Pawley and Syder (1983, 193): *native speakers know best how to play the game of fixedness vs. expressiveness:* "native speakers do not exercise the creative potential of syntactic rules to anything like their full extent, and that, indeed, if they did do so they would not be accepted as exhibiting nativelike control of the language."

Levy and Jaeger (2007): "UID (uniform information density) can be seen as minimizing comprehension difficulty".

Language learners produce less fixed, less entrenched structures. We use the NICT Japanese Learner English (JLE) Corpus. It contains 120,000 sentence pairs consisting of an original language learner sentence and a corrected sentence.

Bigram surface surprisal $log\frac{1}{p(w_{n-1})} + log\frac{1}{p(w_n|w_{n-1})}$ has a mean of 11.7 (and SD=3.36) for corrected text, and 11.5 (and SD=3.48) for original learner text. Comparison:



Parser as Model of Fixedness

WE ALSO APPLY the parser to Learner English. We have manually annotated 100 sentence pairs from the NICT Japanese Learner English (JLE) Corpus. A parser is a language model because:

- it takes attachment decisions (predictions) based on grammar rules and lexical preferences
- it learns form real-word data: syntactically annotated Penn treebank
- Fitting the model: Entrenched structures get higher scores, as they are expected. L2 utterances do not fit the model very well. They abide less to priming, contain more information in Shannon's terms.

For the investigation of highly gradient, complex and interacting factors a parser-based language model is useful. We show that:

- parser performance is significantly lower for the original Learner data than for the corrected (see Figure 1);
- parser scores are significantly lower for the original Learner data than for the corrected (see Figure 2)

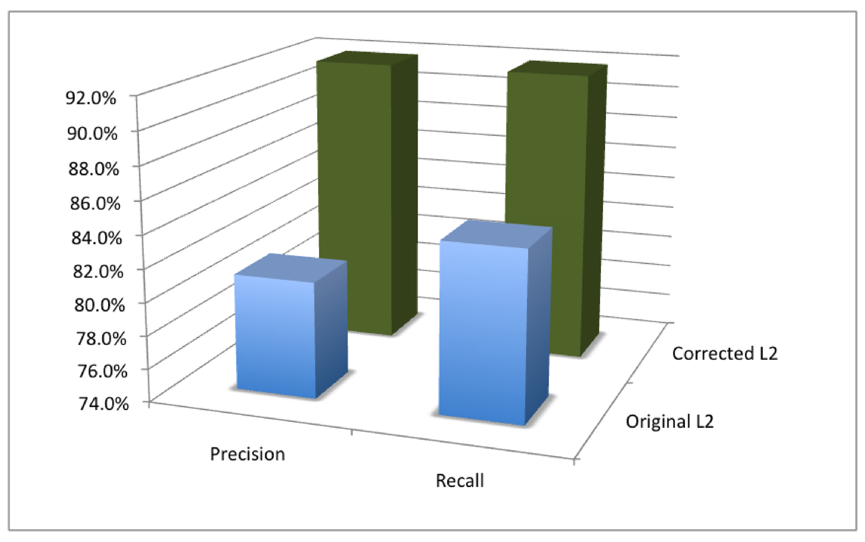


Figure 1: Parser performance

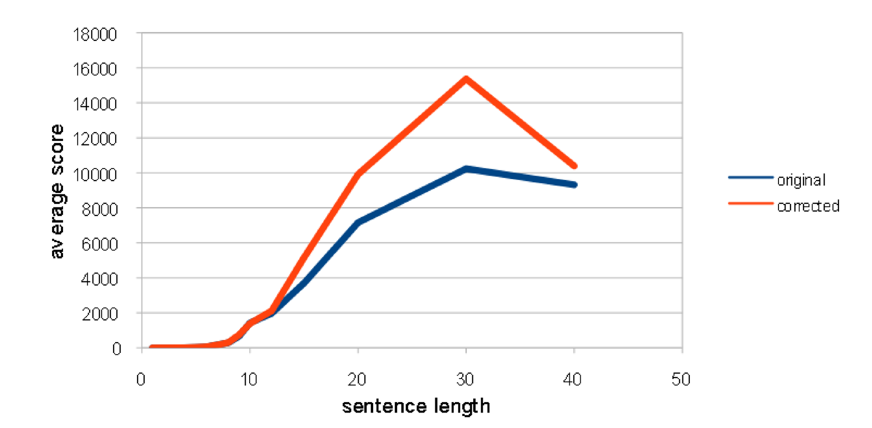


Figure 2: Parser scores, by sentence length.

References

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Hoey, Michael. 2005. *Lexical priming: A New Theory of Words and Language*. Routledge.

Lehmann, Hans Martin and Gerold Schneider. 2011. A large-scale investigation of verb-attached prepositional phrases. In S. Hoffmann, P. Rayson, and G. Leech, editors, *Studies in Variation, Contacts and Change in English, Volume 6: Methodological and Historical Dimensions of Corpus Linguistics*. Varieng, Helsinki.

Lehmann, Hans Martin and Gerold Schneider. 2012. Dependency bank. In *Proceedings of LREC 2012 Workshop on Challenges in the management of large corpora*, pages 23–28.

Levy, Roger and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*.

Pawley, Andrew and Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Native-like selection and native-like fluency. In J. C. Richards and R. W. Schmidt, editors, *Language and Communication*. Longman, London, pages 191–226.

Ronan, Patricia and Gerold Schneider. submitted. Investigating light verb constructions in contemporary british and irish english. In *Paper presented at ICAME 2013*.

Schneider, Gerold. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

Schneider, Gerold and Lena Zipp. 2013. Discovering new verb-preposition combinations in New Englishes. In Joybrato Mukherjee and Magnus Huber, editors, *Studies in Variation, Contacts and Change in English, Volume 14 – Corpus Linguistics and Variation in English: Focus on non-native Englishes*. Varieng, Helsinki.

Seretan, Violeta. 2011. *Syntax-Based Collocation Extraction*. Springer, Dordrecht.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. OUP, Oxford.

Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, pages 209–43.