NLP Corpus Observatory – Looking for Constellations in Parallel Corpora to Improve Learners' Collocational Skills

Gerold Schneider Johannes Graën

Institute of Computational Linguistics
University of Zurich, Switzerland

7th November, 2018

Outline

Motivation

- Computational Linguistics and Learner Error research have made impressive progress, but have yet not reached their collaborative potential (Granger and Lefer 2016)
- Data-driven Learning (DDL) benefits advanced learners, and even beginners: St. John (2001) for lexical tasks, Chujo et al. (2016) for grammar, Vyatkina (2016) for collocations.
- Buyse and Verlinde (2013): using corpus-derived, contextualised resources (Linguee) led to better test performance and user satisfaction. Further integration of tools is desirable.
- **Indirect corpus use**: Creating corpus-informed teaching materials, e.g. collocations dictionaries (Ackermann and Chen 2013; Durrant 2009; McGee 2012): students do not need to learn to use corpus interfaces, but contextualisation is limited.
- Li (2017): also **direct corpus use improves** learner competence in the area of collocations. They conclude that "[t]his exposure to attested language data raises learners' awareness of using collocations in a more natural or near-native way" (p. 165)

Motivation

- Non-compositional expressions pose difficiculties to language learners, required to learn by heart: **collocations**
- Translation difficulties arise particularly in the context of non-compositionality: wherever literal translations lead to incorrect or non-nativelike expressions: **parallel corpora**
- Non-compositional features include any form of idiom and collocation, e.g.:
  - adjective-noun collocations in technical terms
  - verb-object constructions (Källkvist 1998), e.g. light verbs
  - verb-prepostion constructions and phrasal verbs: difficult to acquire for language learners (Gilquin and Granger 2011)
  - Namvar (2012) investigates nine constructions: verb-object collocations are most frequent in learner writing, followed by verb-preposition collocations.

Outline

Source

Europarl (version 7)

- Comprises transcript of the European Parliament sittings
- Contains numerous errors
- Has originally been compiled for training SMT systems
- Provides (reliable) alignment at the level of individual sittings

[1] `http://pub.cl.uzh.ch/purl/costep`

Source

Europarl (version 7)

- Comprises transcript of the European Parliament sittings
- Contains numerous errors
- Has originally been compiled for training SMT systems
- Provides (reliable) alignment at the level of individual sittings

CoStEP (Corrected & Structured Europarl Corpus; Graën, Batinic, and Volk (2014))[1]

- Based on the Europarl corpus
- Has undergone extensive cleaning
- Comprises ≈87 % of the original corpus material
- Provides alignment of speaker turns and additional speaker information (manually added)

[1] `http://pub.cl.uzh.ch/purl/costep`

Our Corpus

Version 9

- ≈150,000 speaker turns from **CoStEP** in 16 languages; altogether ≈450 million tokens
- **Tokenization** with our own multilingual tokenizer Cutter;[2] sentence segmentation based on tokenization tags
- Part-of-speech tagging and **lemmatization** with the TreeTagger and its featured language models
- Pairwise **sentence alignment** with hunalign and **word alignment** with four word aligners (Berkeley Aligner, GIZA++, fast_align and efmaral)
- For this application, we use only bidirectional alignments supported by three of the four aligners
- From this corpus, we randomly sample a subset of 5 % of parallel texts in English and Swedish

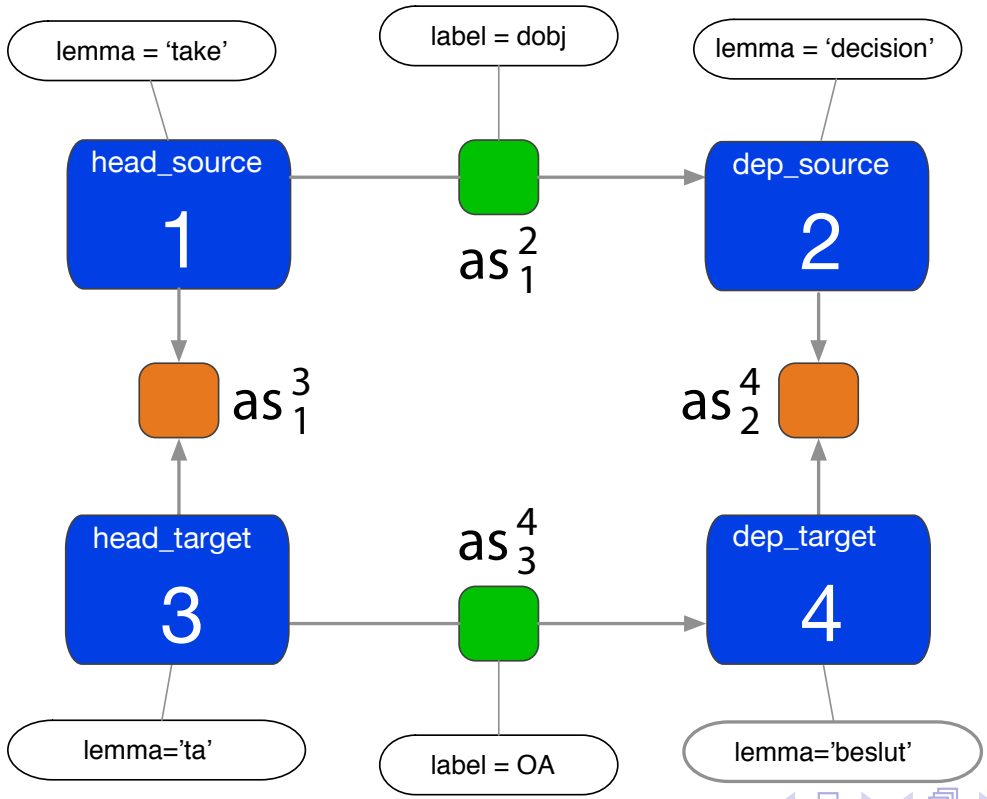
[2] `http://pub.cl.uzh.ch/purl/cutter`

Outline

Collocations and Parallel Corpora: Constellations

We go beyond purely collocation-based phraseme search.

1. Collocations do not entail non-compositionality, the fact that we need to reach collocational status in both languages leads to cleaner results, as in a double check.
2. By punishing literal translations, we also filter the majority of instances that are compositional cooccurrences.



Collocations and Parallel Corpora

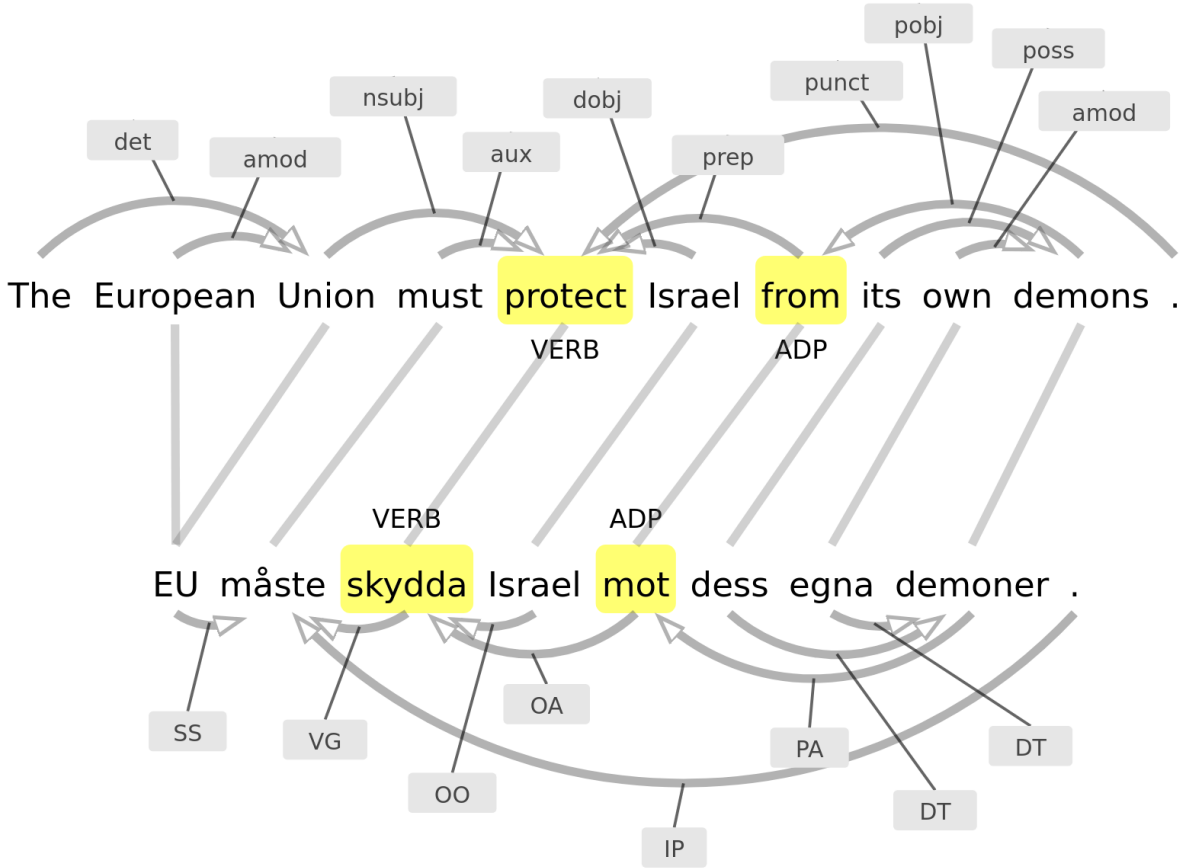


Figure: A constellation consisting of two aligned verbs with corresponding aligned prepositions.

Collocations and Parallel Corpora

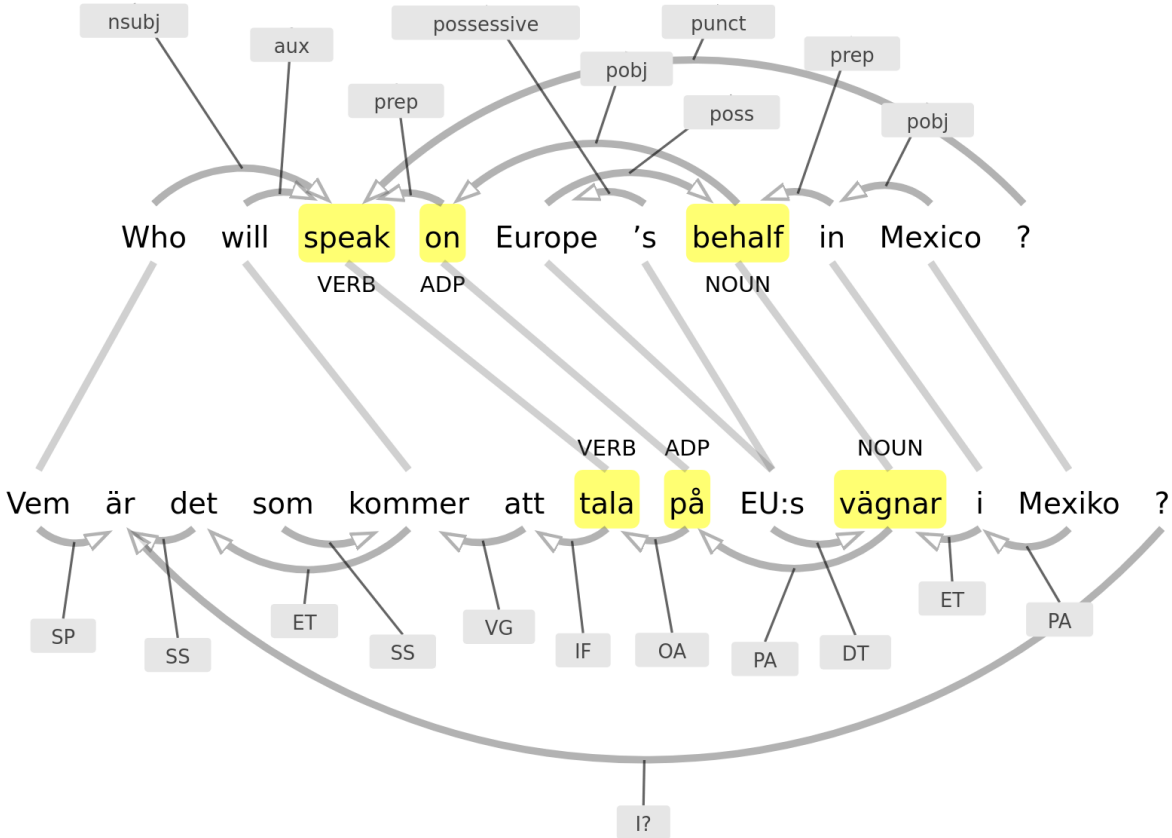


Figure: A constellation consisting of two aligned verbs with corresponding aligned prepositions and aligned prepositional objects.

Collocation Measures

The user chooses one of the "simple association measures" presented in (Evert 2008, Chapter 4, Figure 4) for the ranking of the intralingual collocation and the interlingual alignment:

- O/E : information-theoretic, simple to interpret, overrates rare events
- z-core : significance test, overrates frequent events
- t-score : significance test, overrates frequent events, more resilient against outliers
- $MI = log(O/E)$
- $local\ MI = O \cdot MI$: overrates rare events a bit less
- $simple\ log\text{-}likelihood = O \cdot MI - (O - E)$: relatively balanced

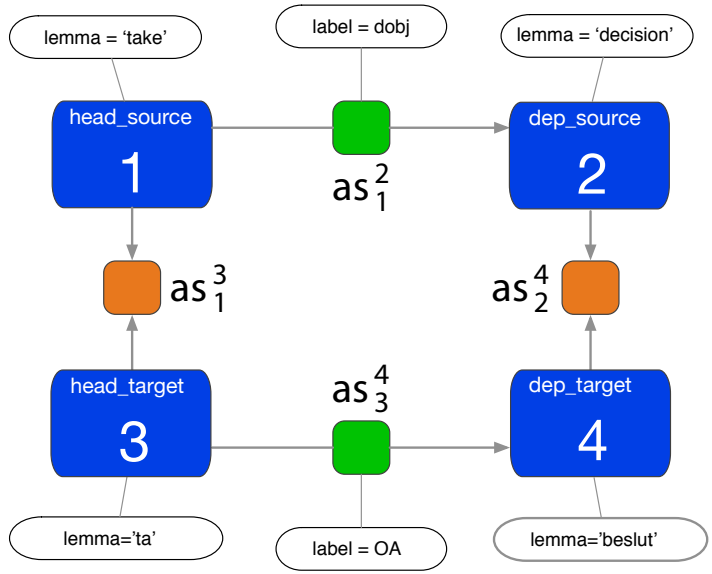
Several normalisations are available (max, tanh, tanhavg)

Outline

Adjective-Noun Constellations

For adjective-noun, we show:

$$score = as_1^2 \cdot as_3^4 \cdot \frac{as_1^3}{(as_2^4)^2}$$



Linear combination of

- association score between adjective and noun in English (as_1^2)
- and Swedish (as_3^4),
- and association score of alignment between the nouns (as_3^1)
- divided by the squared association score of the alignment of the adjectives (as_2^4).

Associations from both languages are reported, particularly: noun is a literal translation, but the adjective is non-literal: unlikely translations are preferred.

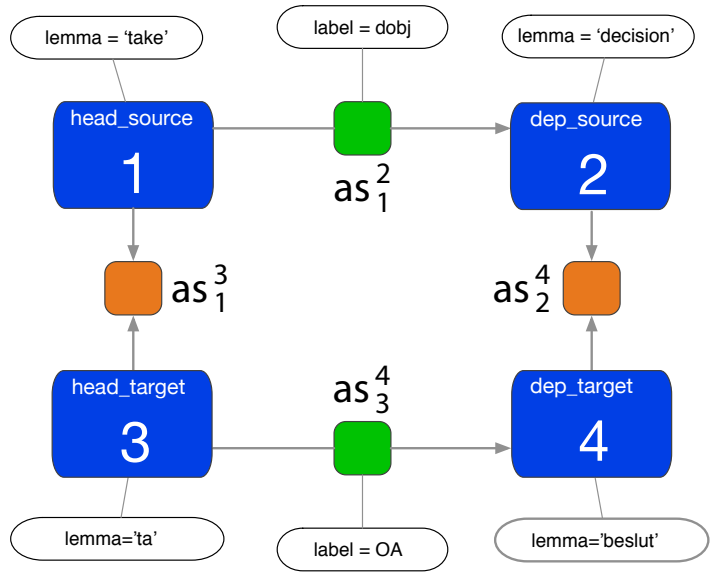
Adjective-Noun Constellations [click]

no.	t_2 (adj. en)	t_1 (noun en)	t_4 (adj. sv)	t_3 (noun sv)	freq.	as_1^2	as_3^4	as_1^3	as_2^4	score
1	close	attention	stor	uppmärksamhet	2	0.0530	0.0669	0.7312	0.0009	2959.5
2	more	time	lång	tid	2	0.0274	0.2662	0.4821	0.0023	635.9
3	top	priority	viktig	prioritering	2	0.2380	0.0493	0.6815	0.0041	481.0
4	large	number	lång	rad	2	0.2108	0.2087	0.1585	0.0057	213.3
5	monetary	policy	ekonomisk	politik	3	0.0939	0.1192	0.6253	0.0066	161.9
6	young	child	liten	barn	3	0.0460	0.0746	0.9397	0.0047	145.2
7	valuable	contribution	viktig	bidrag	2	0.1160	0.0805	0.6603	0.0066	141.2
8	whole	series	lång	rad	2	0.1546	0.2087	0.4516	0.0102	139.2
9	regulatory	framework	rättslig	ram	2	0.1168	0.1266	0.5619	0.0079	131.9
10	constructive	cooperation	god	samarbete	2	0.0470	0.0445	0.8323	0.0041	101.4
11	important	role	stor	roll	2	0.0933	0.0211	0.8691	0.0044	90.3
12	lead	committee	ansvarig	utskott	2	0.0236	0.1680	0.4987	0.0052	73.6
13	fellow	member	kär	kollega	2	0.2643	0.6567	0.1196	0.0182	62.8
14	absolute	priority	hög	prioritet	2	0.0737	0.1601	0.3575	0.0088	53.9
15	central	question	viktig	fråga	2	0.0149	0.1409	0.5068	0.0047	49.0
16	whole	range	lång	rad	2	0.1421	0.2087	0.1575	0.0102	44.6
17	last	year	gången	år	5	0.2675	0.2123	0.9221	0.0346	43.7
18	particular	case	konkret	fall	3	0.0583	0.0557	0.7535	0.0076	42.6
19	excellent	report	bra	betänkande	5	0.2209	0.0643	0.8447	0.0181	36.6
20	good	deal	hel	del	3	0.0266	0.2168	0.0371	0.0024	36.3
21	paramount	importance	stor	vikt	2	0.1651	0.1405	0.4416	0.0178	32.3
22	recent	year	gången	år	2	0.1575	0.2123	0.9221	0.0313	31.5
23	much	time	lång	tid	3	0.0306	0.2662	0.4821	0.0120	27.4
24	positive	result	god	resultat	2	0.0654	0.0616	0.6390	0.0102	24.9
25	less	time	kort	tid	2	0.0167	0.1730	0.4821	0.0078	22.7

Verb-Object Constellations

For verb-object, we show:

$$score = as_1^2 \cdot as_3^4 \cdot \frac{as_2^4}{(as_1^3)^2} \cdot freq$$



This formula is

- similar to the one used for adjective-nouns,
- this time punishing direct translation of verbs,
- frequency is also used. Frequency is an important factor for the identification of light verb constructions (Ronan and Schneider 2015).

Associations from both languages are reported, particularly: noun is a literal translation, but the verb is non-literal: unlikely translations are preferred.

Verb-Object Constellations [click]

no.	t_1 (verb en)	t_2 (noun en)	t_3 (verb sv)	t_4 (noun sv)	freq.	as_1^2	as_3^4	as_1^3	as_2^4	score
1	have	question	ställa	fråga	4	0.9346	0.9977	0.0609	0.8862	891.11
2	have	responsibility	bära	ansvar	2	0.9846	0.9493	0.0393	0.7342	889.74
3	have	debate	föra	debatt	6	0.9554	0.9152	0.0892	0.8882	586.13
4	reach	decision	fatta	beslut	6	0.8145	0.9996	0.0859	0.8266	546.78
5	raise	issue	diskutera	fråga	3	0.9598	0.9759	0.0682	0.9054	546.62
6	make	decision	fatta	beslut	43	0.9779	0.9996	0.2533	0.8266	541.74
7	take	decision	fatta	beslut	58	0.9908	0.9996	0.3194	0.8266	465.47
8	achieve	solution	finna	lösning	2	0.6987	0.9835	0.0478	0.7343	441.00
9	assume	responsibility	ta	ansvar	16	0.9139	0.9958	0.1564	0.7342	437.24
10	play	role	ha	roll	5	0.9991	0.9856	0.0942	0.7497	416.01
11	draw	attention	fästa	uppmärksamhet	34	0.9982	0.9694	0.2090	0.5319	400.66
12	give	example	nämna	exempel	3	0.9057	0.7921	0.0637	0.7493	397.32
13	adopt	decision	fatta	beslut	4	0.7181	0.9996	0.0778	0.8266	392.14
14	solve	problem	lösa	problem	63	0.9946	0.9985	0.3853	0.9118	384.33
15	shoulder	responsibility	ta	ansvar	6	0.6800	0.9958	0.0931	0.7342	344.14
16	pave	way	bana	väg	18	0.9175	0.9215	0.1489	0.4915	337.31
17	accept	responsibility	ta	ansvar	15	0.8333	0.9958	0.1648	0.7342	336.37
18	draw	attention	rikta	uppmärksamhet	15	0.9982	0.9000	0.1487	0.5319	323.94
19	fulfil	responsibility	ta	ansvar	2	0.5265	0.9958	0.0488	0.7342	322.95
20	adopt	measure	vidta	åtgärd	18	0.9296	0.9999	0.2109	0.8489	319.34
21	assume	responsibility	axla	ansvar	3	0.9139	0.5926	0.0612	0.7342	318.83
22	play	role	spela	roll	120	0.9991	0.9997	0.5311	0.7497	318.59
23	take	place	äga	rum	155	1.0000	0.9993	0.5510	0.6058	309.03
24	give	example	ta	exempel	3	0.9057	0.7758	0.0715	0.7493	308.60
25	ask	question	ställa	fråga	36	0.9671	0.9977	0.3421	0.8862	262.96

Linked Examples

1	When does the Council intend to **reach** a **decision** on the establishment of this future observatory?
	När kommer rådet att **fatta beslut** om att inrätta detta framtida organ?
2	It has attempted to reallocate budgetary resources from the Progress programme to the microfinance facility before the European Parliament has **reached** a **decision**.
	Den har försökt omfördela budgetresurser från Progressprogrammet till instrumentet för mikrokrediter innan Europaparlamentet har **fattat** ett **beslut**.
3	Furthermore, the decision-making process itself can be unclear, as the convention submits proposals and the Intergovernmental Conference has to **reach decisions**.
	Dessutom kan det bli oklart kring själva beslutsfattandet, eftersom konventet lägger fram förslag och regeringskonferensen måste **fatta beslut**.
4	When the matter comes before Parliament, therefore, we often have to **reach** our **decisions** very quickly if we want to make the internal market a reality for the citizens of Europe.
	Kommer ärendet sedan till parlamentet, måste vi ofta **fatta** mycket snabba **beslut**, eftersom vi vill öppna den gemensamma marknaden för medborgarna.

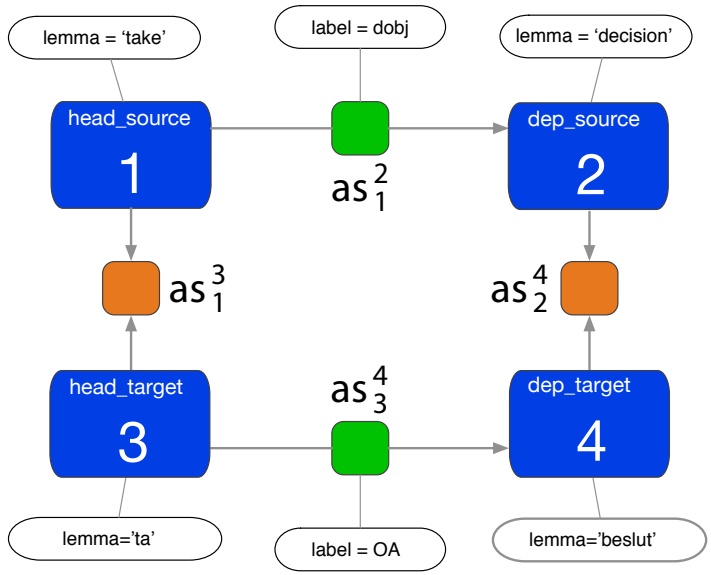
Table: Examples for the verb-direct object constellation "reach decision"/"fatta beslut" ordered by increasing length and minimal length difference. Example 2 shows a direct translation, sentence 4 shows adjective to adverb variation.

University of Zurich[UZH]

Verb-Preposition Constellations

For verb-preposition, we show:

$$score = as_1^2 \cdot as_3^4 \cdot \frac{as_1^3}{(as_2^4)^2} \cdot freq$$



Same verb, but different preposition (or verbal particle).
We can see e.g. that *congratulate on* is a more native-like translation of Swedish *gratulera till* than the direct translation *congratulate to*.

Verb-Preposition Constellations [click]

no.	t_1 (verb en)	t_2 (prep. en)	t_3 (verb sv)	t_4 (prep. sv)	freq.	as_1^2	as_3^4	as_1^3	as_2^4	score
1	deal	with	handla	om	5	0.3824	0.4725	0.0406	6.5E-7	8.6E10
2	cover	by	falla	under	2	0.1300	0.1232	0.0125	0.0001	63633.7
3	congratulate	on	gratulera	till	64	0.2754	0.1862	0.8401	0.0238	4868.7
4	play	in	spela	för	3	0.0979	0.0606	0.8301	0.0018	4818.8
5	agree	with	instämma	i	13	0.4470	0.1311	0.3070	0.0073	4429.4
6	work	on	arbeta	med	39	0.1970	0.1676	0.4541	0.0188	1648.3
7	protect	from	skydda	mot	12	0.0825	0.1479	0.7639	0.0107	975.8
8	base	on	utgå	från	8	0.3929	0.2969	0.0760	0.0087	932.1
9	aim	at	sträva	efter	3	0.3673	0.7869	0.0693	0.0089	762.1
10	vary	from	variera	mellan	4	0.0701	0.1292	0.6337	0.0057	705.1
11	engage	in	ägna	åt	3	0.0871	0.8751	0.0609	0.0045	680.5
12	bring	about	leda	till	7	0.1376	0.3622	0.0442	0.0051	598.7
13	ask	for	be	om	27	0.2278	0.1337	0.5357	0.0306	470.0
14	wait	for	vänta	på	6	0.1821	0.1407	0.6473	0.0169	349.4
15	be	with	vara	i	2	0.0368	0.3080	0.7931	0.0073	340.2
16	work	towards	arbeta	för	15	0.2052	0.1058	0.4541	0.0217	314.2
17	be	in	vara	mot	2	0.2576	0.0608	0.7931	0.0090	308.3
18	be	from	vara	i	2	0.0382	0.3176	0.7931	0.0079	305.7
19	spend	on	ägna	åt	2	0.0701	0.8751	0.1198	0.0071	292.4
20	talk	about	tala	om	150	1.0000	0.3575	0.4997	0.3041	289.8
21	think	about	tänka	på	3	0.1357	0.2119	0.1836	0.0084	223.1
22	be	for	vara	av	12	0.1366	0.2122	0.7931	0.0389	182.4
23	be	at	vara	i	11	0.3520	0.3704	0.7931	0.0819	169.4
24	begin	by	börja	med	54	0.1891	0.2438	0.4637	0.0841	163.3
25	think	of	tänka	på	7	0.0594	0.2115	0.1836	0.0104	149.0

Verb-Prep-Noun Constellations [click]

no.	t_1 (verb en)	t_2 (prep)	t_3 (noun)	t_4 (verb sv)	t_5 (prep)	t_6 (noun)	freq.	as_1^4	as_2^5	as_3^6	score
1	vote	for	report	rösta	för	betänkande	54	1.0000	1.0000	1.0000	54.000
2	enter	into	force	träda	i	kraft	31	0.9958	1.0000	1.0000	31.258
3	thank	for	work	tacka	för	arbete	31	1.0000	1.0000	1.0000	31.000
4	be	in	interest	ligga	i	intresse	29	1.0000	1.0000	1.0000	28.999
5	thank	for	report	tacka	för	betänkande	25	1.0000	1.0000	1.0000	25.000
6	be	of	importance	vara	av	betydelse	25	1.0000	1.0000	1.0000	25.000
7	congratulate	on	report	gratulera	till	betänkande	18	1.0000	1.0000	1.0000	18.000
8	vote	against	report	rösta	mot	betänkande	18	1.0000	1.0000	1.0000	17.971
9	speak	with	voice	tala	med	röst	18	1.0000	1.0000	0.9998	17.825
10	come	from	country	komma	från	land	16	1.0000	1.0000	1.0000	16.000
11	vote	for	resolution	rösta	för	resolution	16	1.0000	1.0000	1.0000	15.987
12	thank	for	cooperation	tacka	för	samarbete	15	1.0000	1.0000	1.0000	15.000
13	be	of	importance	vara	av	vikt	15	1.0000	1.0000	1.0000	15.000
14	be	at	stake	stå	på	spel	13	1.0000	1.0000	0.9866	12.824
15	come	into	force	träda	i	kraft	12	0.9865	1.0000	1.0000	12.329
16	participate	in	debate	delta	i	debatt	12	1.0000	1.0000	1.0000	12.000
17	take	on	Thursday	äga	på	torsdag	12	1.0000	1.0000	0.9996	11.856
18	go	in	hand	gå	i	hand	11	1.0000	1.0000	0.9999	10.999
19	thank	for	support	tacka	för	stöd	11	1.0000	1.0000	1.0000	10.997
20	enter	into	force	träta	i	kraft	9	0.9300	1.0000	1.0000	10.280
21	propose	by	Commission	föreslå	av	kommission	10	1.0000	1.0000	1.0000	9.971
22	be	in	situation	befinna	i	situation	9	1.0000	1.0000	1.0000	9.001
23	adopt	by	Committee	anta	av	utskott	9	1.0000	1.0000	1.0000	8.997
24	contribute	to	development	bidra	till	utveckling	9	1.0000	1.0000	1.0000	8.996
25	be	in	line	ligga	i	linje	9	1.0000	1.0000	0.9998	8.995

University of Zurich[UZH]

Outline

Evaluation

Do learners really fail to produce the collocations suggested in the lists, and instead produce direct translations?
We use the ICLE corpus (Granger, Dagneaux, et al. 2009) as learner corpus to assess if the level of the awkward collocations is higher in than in a native speaker corpus, for which we use BNC (Aston and Burnard 1998).
We evaluate adjective-noun structures, in the two following ways.
First, for all cases where

- the Swedish adjective has a direct translation,
- one that is different from the one suggested in the collocation under observation,
- but semantically similar to the English one in the list,
- the translation of the noun is direct,
- whenever we have at least 3 hits in ICLE in total (max. one zero count is replaced by a smoothing count of 0.1)

then we compare the numbers.

Evaluation

For example, *stor uppmärksomhet* (t_4, t_3) could be directly translated to English *great attention* (t'_2, t_1), but the suggested English collocation is *close attention* (t_2, t_1). *close attention* occurs 106 times in the BNC, *great attention* only 47 times, the ratio $r_{\mathrm{BNC}} = t_2/t'_2$ is 2.25.
In ICLE, *great attention* occurs 9 times, while *close attention* occurs twice, $r_{\mathrm{ICLE}} = t_2/t'_2$ is 0.22. r_{BNC} divided by r_{ICLE} (r, last column) is then 10.15, which can be interpreted as relative dominance, expressing that the suggested collocation is 10.15 times more dominant in the BNC than in ICLE.

Evaluation

no.	t_2, t_1	t_4, t_3	BNC Hits	BNC total	ICLE Hits	ICLE total	dominance BNC/ICLE	direct Translation of t_4	BNC direct	ICLE direct	ratio r
1	close, attention	stor, uppmärksamhet	106	4805	2	286	3.15	great	47	9	10.15
5	monetary, policy	ekonomisk, politik	566	24294	5	420	1.96	economic	1050	12	1.29
6	young, child	liten, barn	1380	19452	75	1427	1.35	small	182	63	6.37
7	valuable, contribution	viktig, bidrag	89	4702	1	88	1.67	important	208	4	1.71
9	regulatory, framework	rättslig, ram	56	3053	0.1	22	4.04	legal	160	1	3.50
11	important, role	stor, roll	723	11027	257	763	0.19	big	12	22	5.16
14	absolute, priority	hög, prioritet	18	2239	1	45	0.36	high	220	1	0.08
15	central, question	viktig, fråga	90	12703	0.1	669	47.40	important	317	51	144.79
24	positive, result	god, resultat	268	10533	12	435	0.92	good	268	28	2.33
30	important, progress	stor, framsteg	10	2870	3	363	0.42	big	0.1	3	100.00
32	substantial, progress	viktig, framsteg	56	2870	1	363	7.08	important	10	0.1	0.56
34	serious, problem	stor, problem	594	24420	318	3470	0.27	big	175	109	1.16
38	good, opportunity	stor, möjlighet	119	5984	25	732	0.58	big	11	2	0.87
∅							5.34				21.38

Table: Evaluation of adjective-noun constellations

Evaluation

Second, we measure the absolute dominance of the English collocation, as follows:

- frequency of the collocation, divided by the frequency of the noun modified by any adjective.
- For *close attention* in the BNC, this is $dom(\mathrm{BNC}) = 106/4805 = 0.022$, in ICLE it is $dom(\mathrm{ICLE}) = 2/286 = 0.007$.
- $dom(\mathrm{BNC})/dom(\mathrm{ICLE})$ is thus 3.15. The mean of the absolute dominance is 5.3, which means that the suggested collocation is found 5.3 times more often in BNC than in ICLE.

The evaluation has shown that in the majority of cases, our method yields good results, and allows learners to explore various constellations.

Conclusions

- Implemented and evaluated an interactive tool for data-driven learning of constellations (i.e., parallel collocation structures) [3]
- Full integration of direct and indirect data-driven learning. Collocation dictionaries are generated on the fly, and linked to the parallel examples.
- Use of association measures for both collocations and alignments.
- Advanced users can also customise the association scores.
- We plan to test the tool with learners, to train on the entire Europarl corpus, and to add more languages to our approach.

[3] `https://pub.cl.uzh.ch/purl/constellations`