



University of
Zurich ^{UZH}

Binomial Adverbs in Germanic vs. Romance Languages

A Corpus-based Study

Johannes Graën & Martin Volk

Tuesday 6th November, 2018

Institute of Computational Linguistics, University of Zurich

1. Binomial Adverbs
2. Our Corpus
3. Candidate Selection from Corpus
4. Conclusions

Binomial Adverbs

Binomial adverbs

Binomials

- Type “X conjunction Y”, with X and Y having the same part of speech (*‘drag and drop’, ‘zwicken und zwacken’, ‘liso y llano’*)¹
- Typically use coordinating conjunctions (**coordinated binomials**)
- Called **repetitions** or **echoics** if X = Y (*‘little by little’, ‘io come io’*)

Multi-word adverbs

- Any multi-word expression that acts as an adverb (*‘à petit feu’, ‘from tip to toe’, ‘mot pour mot’, ‘en un abrir y cerrar de ojos’*)
- Also referred to as **adverbial phrases**

Binomial adverbs

- Type “X conjunction Y”, with X and Y being adverbs (*‘ici et là’*)

¹Most examples taken from (Müller 2009).

Multi-word adverbs (*polirematiche avverbiali*)

P+(DET)+N/A	a caldo, a freddo, a giorno, a monte, a occhio, a rate, al verde, alla pari, al nero, a chiare lettere, a tutta birra, a viso aperto, di cuore, in contante, in caldo, in nero, in proprio, in buona fede, in senso lato, sulla carta
P+N+P+N/A	di punto in bianco, a portata di mano, in via di sviluppo, a pie' di pagina, in linea di massima, di anno in anno, in fin dei conti, a prezzo di costo
N+SP	pancia all'aria, porta a porta
AVV+P+AVV	su per giù, lì per lì, giù di lì
AVV+AVV	così così, via via, meno male
AVV+CG+AVV	più o meno, bene o male

Table taken from (Voghera 2004).

Coordinated binomials (*binomi coordinati*)

L	A≠B	A=B	A≃B
A e B	<i>acqua e sapone bianco e nero</i>	<i>decine e decine giorni e giorni</i>	<i>gira e rigira unto e bisunto</i>
A o B	<i>vivo o morto presto o tardi</i>	-	-
o A o B	<i>o bere o affogare o la borsa o la vita</i>	-	-
né A né B	<i>né carne né pesce né cotto né crudo</i>	-	-
PREP A e B	<i>tra uscio e muro a uso e consumo</i>	<i>fra sé e sé tra me e me</i>	-
PREP A e/o PREP B	<i>senza se e senza ma a torto o a ragione</i>	-	-
senza A né B	<i>senza arte né parte senza capo né coda</i>	-	-
A ma B	<i>pochi ma buoni</i>	-	-

Examples following the pattern [(L) [A]_X L [B]_Y]_Z taken from (Masini 2008).

- Part-of-speech tagging for binomial adverbs (and for multi-word adverbs in general) is often erroneous:²

"by and large"		frequency	
ADP	CONJ	ADV	145
ADV	CONJ	ADP	15
ADV	CONJ	ADV	1

- This is due to similar surface forms of different parts of speech (homographs), e.g. German *'ab'*, French *'haut'*, Swedish *'i'*
- Potential errors propagate the subsequent processing steps
- Word alignment has problem identifying the sequence as a single unit as there are frequently several translation variants

²Correcting these errors can improve NLP pipelines (Volk, Clematide, et al. 2016).

Our Corpus

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)³

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


³(Graën, Batinic, and Volk 2014); <http://pub.cl.uzh.ch/purl/costep>


Our corpus (version 9)

- Comprises \approx 150 000 speaker turns from **CoStEP** in 16 languages; altogether \approx 450 million tokens
- **Tokenization** with our own multilingual tokenizer Cutter;⁴ 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 different word aligners (Berkeley Aligner, GIZA++, fast_align and efmara)
- For this application, we use only bidirectional alignments that are supported by several word aligners

⁴(Graën, Bertamini, and Volk 2018); <http://pub.cl.uzh.ch/purl/cutter>


Corpus example from Multilingwis⁵

 Aber es muss auf eine **koordinierte Art und Weise** geschehen.


 However, it needs to be done in a **coordinated way**.


 No obstante, ha de hacerse de una **forma coordinada**.

 Mais il faut le faire de **façon coordonnée**.


 Tuttavia, occorre procedere in **maniera coordinata**.

 Maar het dient wel op een goed **afgestemde manier** te gebeuren.

 Należy jednak to zrobić w **sposób skoordynowany**.

 No entanto, isso terá de ser feito de **forma coordenada**.

 Totuși, acest lucru trebuie să se realizeze într-un **mod coordonat**.

 Vendar je to treba izvesti na **usklajen način**.

 Men det måste ske på ett **samordnat sätt**.

⁵(Graën, Sandoz, and Volk 2017); <https://pub.cl.uzh.ch/purl/multilingwis>

Candidate Selection from Corpus

Approach

1. We do not rely on part-of-speech tagging to identify candidates for binomial adverbs
2. Instead, we compile list of all word forms that have ever been tagged as adverb and conjunction, respectively, in six languages: English, French, German, Italian, Spanish and Swedish
3. We then calculate different measures as indicators for idiomaticity and use them to rank the candidates

Mutual information score

- “Amount of mutual information” that two observations share
- For binomial adverbs “X C Y”, mutual information is calculated as:

$$\text{MI}(X, C, Y) = \log_2 \frac{N \cdot f(X, C, Y)}{f(X, C) \cdot f(C, Y)}$$

- Highest mutual information scores achieved by infrequent binomials (*‘officiellement ou officieusement’*, *‘prima e/o dopo’*, *‘ayer u hoy’*, *‘inward and outward’*)
- Mutual information has proven useful in ruling out improbable candidates for binomial adverbs (Volk and Graën 2017)

Local mutual information score

- Like the regular mutual information score, but rewards frequent binomials by multiplying with the frequency of the binomial
- Local mutual information is calculated as:

$$\text{local-MI}(X, C, Y) = f(X, C, Y) \cdot \log_2 \frac{N \cdot f(X, C, Y)}{f(X, C) \cdot f(C, Y)}$$

Simple log-likelihood measure

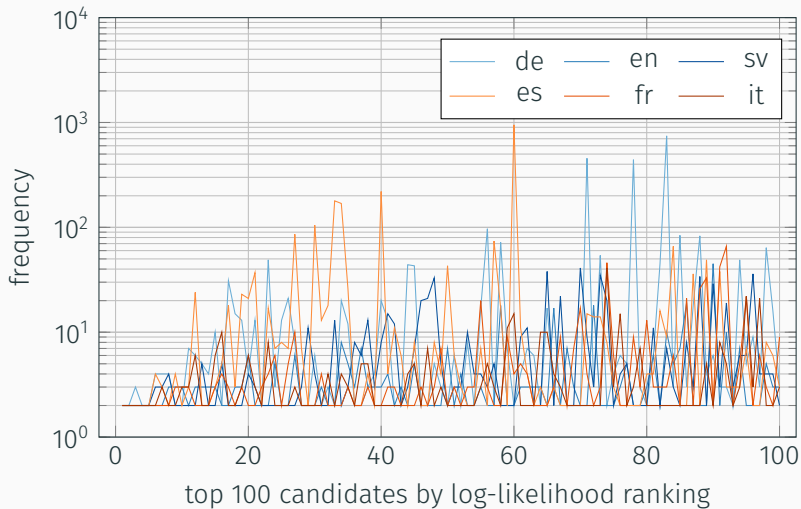
- Derived from the likelihood-ratio test⁶
- Simple log-likelihood is calculated as:

$$\text{simple-ll}(X, C, Y) = 2 \left(f(X, C, Y) \cdot \log \frac{N \cdot f(X, C, Y)}{f(X, C) \cdot f(C, Y)} - \left(f(X, C, Y) - \frac{f(X, C) \cdot f(C, Y)}{N} \right) \right)$$

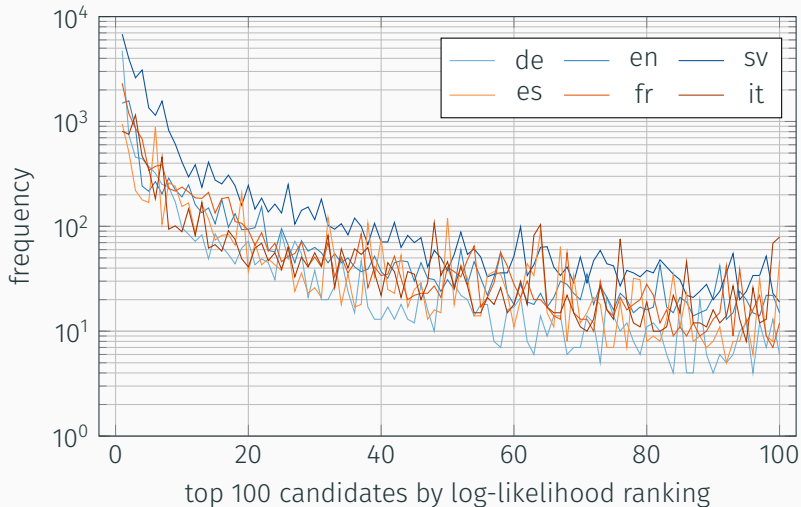
- Local mutual information and simple log-likelihood yield similar results with regard to the top-ranked binomial candidates

⁶All three measures are explained and compared in (Evert 2008).

Comparison of frequencies (MI)



Comparison of frequencies (simple-ll)



Irreversibility score

- The irreversibility score (Mollin 2014) is the ratio of the given order of adverbs in relation to both possible orders
- The irreversibility score is calculated as:

$$\text{irr-score}(X, C, Y) = \frac{f(X, C, Y)}{f(X, C, Y) + f(Y, C, X)}$$

- For repetitions, the irreversibility score is defined to be 1
- Examples:

candidate	frequency	irr-score
more or less	868	100 %
politiskt och ekonomiskt	142	57,3 %
ekonomiskt och politiskt	106	42,7 %

Single word translation variants

- If a candidate is translated with a single word to another language, chances are that it is a binomial adverb
- Translation correspondences can be approximated by word alignment
- We rank single word correspondences and give the most frequent word form together with its most frequent part of speech, e.g. for German “nach wie vor”:

language	translation	frequency	PoS
en	still	945/4507	99 % ADV
es	sigue	582/4507	100 % VERB
sv	fortfarande	1630/4507	100 % ADV

Example of single word translation variants

-  Ein neues politisches Klima entsteht **nach und nach**.
-  A new political climate is **gradually** emerging.
-  Un nuevo clima político está emergiendo **gradualmente**.
-  Uusi poliittinen ilmapiiri on **vähitellen** muotoutumassa.
-  Un cadre politique nouveau voit **progressivement** le jour.
-  **Stopniowo** wyłania się nowy klimat polityczny.
-  Își face apariția **treptat** un nou climat politic.
-  **Postopoma** nastaja novo politično vzdušje.
-  Ett nytt politiskt klimat håller **gradvis** på att växa fram.

Entropy-based immediate context

- Some binomial adverbs form part of a larger multi-word adverb
- We try to identify those larger units by investigating the immediate left and right context of a binomial candidate
- For both left and right context we calculate three measures: relative entropy, relative and absolute frequency of the word form in question, e.g. for “ores et déjà” we get:

context	entropy	word form	rel. freq.	abs. freq.
left	0.0204	d'	96,7 %	1148
right	0.4617	,	7,9 %	94

- If the entropy is lower than 0.2, the relative frequency is greater than 25 %, and we see at least 5 occurrences of the respective word form, we add it to the expression (*d'ores et déjà*, *tanto antes como ahora*, *'so on and so forth'*)

Conclusions

Conclusions



Die Schlussfolgerungen sind für mich einigermaßen klar.



As far as I can see, the conclusions are more or less obvious.



Para mí las conclusiones son más o menos notorias.



Les conclusions sont, peu ou prou, évidentes à mes yeux.









Per me le conseguenze sono quasi ovvie.



Estas conclusões para mim são quase evidentes.

False positives

-  Der Nairobi-Gipfel wird ein Appell an uns alle sein, **mehr** zu tun **und** **es** schneller zu tun.
-  The Nairobi Summit will be an appeal to all of us to do **more** **and** to do it faster.
-  La Cumbre de Nairobi será un llamamiento a todos nosotros para hacer **más** **y** **más** rápido.
-  Nairobín huippukokouksessa vedotaan meihin kaikkiin, jotta toimisimme **tehokkaammin** **ja** nopeammin.
-  Le sommet de Nairobi constituera pour nous tous une invitation à en faire **plus** **et** **plus** vite.
-  Il Vertice di Nairobi sarà un incitamento per noi tutti a fare di **più** **e** con **maggiore** rapidità.

Limitations and outlook

- We currently filter out candidates that are never tagged as ADV CONJ ADV, and we are missing some known German binomial adverbs (*'eh und je'*, *'ab und zu'*, *'durch und durch'*)
- We find 18 out of 21 binomial adverbs given in (Mollin 2014), but several hundreds more
- Judging whether a candidate is idiomatic or composed is difficult (even for native speakers); we plan to collect multiple judgments via crowd sourcing
- All our lists compiled so far are available for download

Lists of binomial adverb candidates

	A	B	D	E	F	G	H	I	J	L	M	N	O	P	Q	R	S	T
1	lan	expression	f	Pos cor	f r	irr sco	MI	local N	simple	r M	r l	r s	H left	P left	f	H right	P right	f
2	sv	till och med	6802	100.0%	7	0.999	10.777	73308.32	30538.01	6761	1	1	0.3637	0.1860	1177	0.4698	0.0478	303
3	de	nach wie vor	4723	99.9%	2	1.000	13.038	61577.80	27627.53	4493	2	2	0.4694	0.0790	356	0.4972	0.0546	246
4	sv	först och främst	3978	99.7%	1	1.000	13.840	55058.29	25191.19	3690	3	3	0.4018	0.1594	347	0.4696	0.0798	179
5	sv	helt och hållet	2610	93.8%	1	1.000	13.358	34863.36	15769.83	4148	4	4	0.4462	0.1108	289	0.4602	0.1135	296
6	sv	i och med	3098	3.1%	14	0.998	10.946	33909.83	14221.75	6601	5	5	0.4538	0.1210	264	0.3744	0.2483	544
7	fr	ne soit pas	2316	11.7%	1	1.000	13.454	31158.38	14127.22	4044	6	6	0.4231	0.0877	203	0.5449	0.0354	82
9	en	first and foremost	1496	17.2%	1	1.000	14.483	21667.31	10052.02	3133	8	8	0.2865	0.4701	573	0.3192	0.4079	507
10	en	more and more	1575	35.9%	1	1.000	12.419	19560.32	8626.48	5123	9	9	0.4544	0.1479	215	0.4948	0.0587	86
11	fr	(d') ores et déjà	1187	100.0%	1	1.000	15.144	17975.58	8448.38	2629	11	10	0.0204	0.9671	1148	0.4617	0.0792	94
12	sv	hur som helst	1349	2.6%	1	1.000	13.408	18087.96	8192.03	4094	10	11	0.4579	0.1008	77	0.4092	0.0916	70
13	sv	klart och tydligt	1146	6.5%	116	0.998	14.200	16272.84	7505.23	3362	13	12	0.4835	0.0718	82	0.4912	0.1108	127
14	sv	från och med	1564	9.1%	1	1.000	10.872	17004.18	7109.54	6670	12	13	0.4212	0.0598	93	0.3891	0.2353	368
15	es	más o menos	950	100.0%	2	0.998	15.358	14589.74	6883.90	2462	14	14	0.3383	0.1326	125	0.3810	0.0700	66
16	fr	plus ou moins	859	100.0%	2	0.998	15.514	13326.54	6305.37	2357	16	15	0.3805	0.0665	57	0.4107	0.0443	38
17	en	more or less	868	34.2%	1	1.000	15.387	13356.01	6305.12	2445	15	16	0.4540	0.0835	72	0.3931	0.1102	95
18	sv	mer eller mindre	824	98.5%	1	0.999	15.151	12484.70	5868.54	2626	17	17	0.4315	0.1340	110	0.3622	0.0779	64
21	de	mehr oder weniger	746	96.1%	3	0.996	15.453	11527.95	5448.52	2404	20	20	0.4629	0.0458	34	0.3685	0.0296	22
23	it	più o meno	809	99.9%	1	0.999	13.575	10981.86	4993.74	3929	22	22	0.3612	0.1263	101	0.3858	0.0325	26
24	it	più che mai	754	100.0%	1	1.000	12.936	9753.42	4364.14	4585	24	23	0.3998	0.1463	108	0.4277	0.1829	135
25	it	anche se non	1154	100.0%	1	1.000	9.447	10901.50	4257.36	8050	23	24	0.0716	0.8098	809	0.5068	0.1240	125
26	fr	plus que jamais	683	100.0%	1	0.999	12.445	8499.70	3751.33	5088	25	25	0.3987	0.1640	101	0.3800	0.2913	180
27	it	prima o poi	467	98.5%	1	1.000	16.367	7643.49	3667.84	1838	27	26	0.2337	0.3133	130	0.3731	0.2755	116
28	sv	helt och fullt	608	69.9%	2	0.997	13.124	7979.23	3587.98	4388	26	27	0.4643	0.1184	72	0.4817	0.1003	61
29	de	ganz und gar	456	99.8%	1	1.000	16.187	7381.16	3531.90	1947	28	28	0.4323	0.1014	45	0.2694	0.4189	186
30	de	mehr denn je	442	99.1%	1	1.000	15.703	6940.83	3294.80	2221	30	29	0.4138	0.1328	53	0.4407	0.0625	25
34	sv	sist men inte (minst)	412	100.0%	1	1.000	15.150	6241.90	2934.00	2627	35	33	0.2701	0.4800	60	0.0000	1.0000	126
38	fr	purement et simplement	341	100.0%	1	1.000	16.934	5774.34	2794.50	1570	37	37	0.3838	0.0824	28	0.4231	0.0676	23
39	de	nach und nach	373	85.0%	1	1.000	14.023	5230.73	2403.21	3534	40	38	0.4060	0.0805	28	0.4217	0.0747	26
40	sv	direkt eller indirekt	296	86.1%	10	0.987	16.660	4931.49	2377.05	1679	43	39	0.3886	0.2014	59	0.4369	0.1327	39
41	fr	(mais) aussi et surtout	374	100.0%	1	1.000	13.439	5026.26	2278.11	4064	41	40	0.1617	0.7265	271	0.4834	0.0913	34
43	es	más que nunca	511	100.0%	1	1.000	10.450	5340.15	2193.09	7075	39	42	0.3366	0.1866	89	0.3549	0.3820	183
44	sv	mer och mer	387	85.5%	1	1.000	12.671	4903.53	2178.22	4844	44	43	0.3990	0.1614	61	0.3722	0.0447	17
45	it	(ma) anche e soprattutto	358	100.0%	1	1.000	13.332	4772.87	2157.56	4173	45	44	0.1917	0.5490	196	0.5440	0.1204	43
47	fr	pas pour autant	387	0.9%	1	1.000	11.985	4638.33	2018.56	5576	47	46	0.3764	0.1860	72	0.3913	0.2351	91
48	en	directly or indirectly	242	100.0%	3	0.988	17.155	4151.52	2015.47	1488	51	47	0.3545	0.1875	45	0.3961	0.2375	57
49	sv	för eller senare	236	99.6%	1	1.000	17.137	4044.42	1962.98	1496	52	48	0.3668	0.0857	18	0.3433	0.2085	44
50	fr	(ni) plus ni moins	229	100.0%	1	1.000	17.304	3962.67	1927.76	1434	54	49	0.0267	0.9607	220	0.3714	0.2052	47

http://pub.cl.uzh.ch/purl/binomial_adverbs

References



Evert, S. (2008). “Corpora and collocations”. In: *Corpus Linguistics. An International Handbook*. Ed. by A. Lüdeling and M. Kytö. Vol. 2. Berlin: Walter de Gruyter, pp. 1212–1248.






Graën, J., D. Batinic, and M. Volk (2014). “Cleaning the Europarl Corpus for Linguistic Applications”. In: *Proceedings of the Conference on Natural Language Processing (KONVENS)* (Hildesheim). Stiftung Universität Hildesheim, pp. 222–227.



Graën, J., M. Bertamini, and M. Volk (2018). “Cutter – a Universal Multilingual Tokenizer”. In: *Proceedings of the 3rd Swiss Text Analytics Conference*.

-  Graën, J., D. Sandoz, and M. Volk (2017). “Multilingwis² – Explore Your Parallel Corpus”. In: *Proceedings of the 21st Nordic Conference of Computational Linguistics (NODALIDA)* (Gothenburg). Linköping Electronic Conference Proceedings 131. Linköping University Electronic Press, Linköpings universitet, pp. 247–250.
-  Masini, F. (2008). “Binomi coordinati in italiano”. In: *Prospettive nello studio del lessico italiano – Atti del IX Congresso SILFI*. Ed. by E. Cresti. Vol. 2. Firenze University Press, pp. 563–571.
-  Mollin, S. (2014). *The (ir)reversibility of English binomials: Corpus, constraints, developments*. Vol. 64. Studies in Corpus Linguistics. John Benjamins Publishing Company.
-  Müller, H.-G. (2009). *Adleraug und Luchsenohr: Deutsche Zwillingsformeln und ihr Gebrauch*. Peter Lang GmbH, Internationaler Verlag der Wissenschaften.

-  Voghera, M. (2004). “Polirematiche”. In: *La formazione delle parole in italiano*. Tübinga: Max Niemeyer. Ed. by M. Grossmann and F. Rainer. Niemeyer, pp. 56–69.
-  Volk, M., S. Clematide, J. Graën, and P. Ströbel (2016). “Bi-particle Adverbs, PoS-Tagging and the Recognition of German Separable Prefix Verbs”. In: *Proceedings of the Conference on Natural Language Processing (KONVENS)* (Bochum), pp. 297–305.
-  Volk, M. and J. Graën (2017). “Multi-word Adverbs – How well are they handled in Parsing and Machine Translation?” In: *Proceedings of the 3rd Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT)* (London). Vol. 2, pp. 1–9.