

Bilingual word sketches

Three flavours

Vojtěch Kovář
Adam Kilgarriff
Ana Frankenberg-Garcia

Masaryk University
Lexical Computing Ltd.
Surrey University

`xkovar3@fi.muni.cz`
`adam.kilgarriff@sketchengine.co.uk`
`ana.frankenberg@gmail.com`

eLex 2013, Tallinn, October 17

Word sketches

One page summary of a word's collocational and grammatical behaviour

Word sketches

chase *(verb)* enTenTen12 freq = [169458](#) (13.1 per million)

object	79230	0.6	subject	43351	0.5	modifier	18017	0.0	and/or
tail	1618	6.99	pur-	61	5.51	playfully	49	5.93	pounce
squirrel	469	6.52	ambulance	168	5.32	blindly	85	5.93	harass
rabbit	586	6.12	jpmorgan	28	4.29	relentlessly	64	5.84	hound
dream	2746	5.69	play-off	32	4.18	frantically	45	5.24	bark
rainbow	227	5.44	paparazzi	32	4.1	jpmorgan	19	5.11	stalk
prey	275	5.34	squirrel	67	4.07	aggressively	75	5.06	wrestle
shad	108	5.27	lion	134	3.98	doggedly	18	4.78	nip
butterfly	278	5.21	cheetah	31	3.97	endlessly	38	4.77	fetch
mirage	85	4.91	storm	271	3.84	futilely	16	4.76	lunge
ball	1650	4.91	pur	20	3.78	fruitlessly	14	4.57	harry
ghost	264	4.88	coyote	31	3.66	madly	19	4.48	party
deer	245	4.8	hound	34	3.64	away	1014	4.47	bite

Word sketches

- Since 1999
- 2002 first used
 - Macmillan English Dictionary
- Since 2004 in Sketch Engine
- Widely used for lexicography
- Evaluated
- Available for thirty languages

Sue Atkins, 2005

“Can I have bilingual ones please”

Sue Atkins, 2005

“Can I have bilingual ones please”

... hmmm, now what might that mean

Three answers

- BiP
 - Bilingual word sketch from **P**arallel data
- BiC
 - Bilingual word sketch from **C**omparable data
- BiM
 - Bilingual word sketch – **M**anual

BiP – bilingual parallel word sketch

- Large parallel corpora
 - EUROPARL, OPUS
 - alignment on sentence level
- Statistics measuring co-occurrence significance
 - Logdice (used in word sketch score)
- → **Build automatic dictionary of candidate translations**
 - translation $A \rightarrow B$ is significant, if the number of co-occurrences of A and B in aligned sentences is significant

BiP – generating output

- Find the candidate translations of the headword
 - choose the best one according to logdice
 - show the others to the user
 - generate another word sketch for the best translation
 - name it WS2
- For each collocate of the headword
 - any of the collocate translations in WS2?
 - if yes: **align the two collocations in the BiP sketch**
- For each found pair
 - create concordance **headword-collocation aligned with headword_translation-collocate_translation**
 - if empty: leave the pair out
 - else: include examples into BiP sketch

BiP – example

declaration (*noun*) EUROPARL5, English-French freq = 4409
déclaration (*noun*) EUROPARL5, French-English freq = 9341

use another candidate translation: [déclarations](#) [écrites](#) [écrite](#) [Déclarations](#)

modifier		
written	149	I am surprised that on 4 May , after consultation with the World Health Organization , the Council decided not to do this and to rely instead on checks in the country of departure and written declarations by interested parties .
écrite	49	Mr President , Members have had circulated to them this evening notice of my written declaration on alcopops which lapses at 6.30 pm . Monsieur le Président , les membres n'ont pris connaissance de ma déclaration écrite sur les « alcopops » que cet après-midi alors qu'elle expire précisément aujourd'hui à 18 h 30 .
solemn	51	Solemn declarations and moral indignation are not enough , though ; they also , as is specified in our joint resolution , have to be backed up by a whole host of things .
solennel		EU-Africa Summit in Cairo - (FR) The solemn declaration of the first EU-Africa summit in Cairo opens by stating , and I quote : " Over the centuries , ties have existed between Africa and Europe ... developed on the basis of shared values of strengthening representative and participatory democracy " . Given that this secular past was a story of slavery , massacres , forced labour , plundering , colonial conquests and oppression , during which the rich European countries bled that continent dry , we can only wonder what is the most shameful aspect : the pride of the representatives of the imperialist countries or the baseness . La déclaration solennelle du premier sommet Afrique-Europe , au Caire , commence par faire référence , je cite : aux " liens qui existent entre l'Afrique et l'Europe " : " ... " depuis des siècles " qui se seraient " développés sur la base de valeurs communes telles que le renforcement de la démocratie " .
unilateral	61	He must not add fuel to the flames by threatening a unilateral declaration of independence for the Palestinian state .
unilatérale	9	The problem is that , after nine years of refusing to sign a border agreement with Estonia , Russia finally did so last month , but the Estonian Parliament , following typical parliamentary procedure , added a unilateral non-binding declaration saying that the legal continuity of the State is enforced even when territory is given up . Ce problème est le suivant : après avoir refusé pendant neuf ans de conclure un accord frontalier avec l' Estonie , la Russie a enfin accepté le mois dernier , mais le parlement estonien , au terme d'une procédure parlementaire typique , a ajouté une déclaration unilatérale non contraignante affirmant que la continuité juridique de l' Etat est assurée même quand un territoire

BiC – bilingual comparable word sketch

- No aligned sentences
 - → external bilingual dictionary required
- Google dictionaries API
 - ceased being available
- Publishers' dictionaries
 - new search, negotiation, for every language pair
 - arbitrary coverage, problems in structure
- Current solution
 - **use the automatic BiP dictionaries**
 - same procedure as in BiP, just without aligned examples

BiC – example

declaration (*noun*) EUROPARL5, English-French freq = 4409

déclaration (*noun*) EUROPARL5, French-English freq = 9341

use another candidate translation: [déclarations](#) [écrites](#) [écrite](#) [Déclarations](#)

modifier		
written	149	I am surprised that on 4 May , after consultation with the World Health Organization , the Council decided not to do this and to rely instead on checks in the country of departure and written declarations by interested parties .
écrite	49	Mr President , Members have had circulated to them this evening notice of my written declaration on alcopops which lapses at 6.30 pm . Monsieur le Président , les membres n'ont pris connaissance de ma déclaration écrite sur les « alcopops » que cet après-midi alors qu'elle expire précisément aujourd'hui à 18 h 30 .
solemn	51	Solemn declarations and moral indignation are not enough , though ; they also , as is specified in our joint resolution , have to be backed up by a whole host of things .
solennel		EU-Africa Summit in Cairo - (FR) The solemn declaration of the first EU-Africa summit in Cairo opens by stating , and I quote : " Over the centuries , ties have existed between Africa and Europe ... developed on the basis of shared values of strengthening representative and participatory democracy " . Given that this secular past was a story of slavery , massacres , forced labour , plundering , colonial conquests and oppression , during which the rich European countries bled that continent dry , we can only wonder what is the most shameful aspect : the pride of the representatives of the imperialist countries or the baseness . La déclaration solennelle du premier sommet Afrique-Europe , au Caire , commence par faire référence , je cite : aux " liens qui existent entre l'Afrique et l'Europe " : " ... " depuis des siècles " qui se seraient " développés sur la base de valeurs communes telles que le renforcement de la démocratie " .
unilateral	61	He must not add fuel to the flames by threatening a unilateral declaration of independence for the Palestinian state .
unilatérale	9	The problem is that , after nine years of refusing to sign a border agreement with Estonia , Russia finally did so last month , but the Estonian Parliament , following typical parliamentary procedure , added a unilateral non-binding declaration saying that the legal continuity of the State is enforced even when territory is given up . Ce problème est le suivant : après avoir refusé pendant neuf ans de conclure un accord frontalier avec l' Estonie , la Russie a enfin accepté le mois dernier , mais le parlement estonien , au terme d'une procédure parlementaire typique , a ajouté une déclaration unilatérale non contraignante affirmant que la continuité juridique de l' Etat est assurée même quand un territoire

BiP/BiC – technical challenges

- Computing the translation candidates
 - about 24 hours for a 45M parallel corpus
 - one-off
- Evaluating the concordance and picking example
 - takes tens of seconds
 - → asynchronous loading of examples
- Loading dictionary with candidate translations
 - quite big ⇒ few seconds
 - → optimizations

BiM – bilingual manual word sketch

- When there is no dictionary available
 - or not reliable
- User specifies
 - headword in one language
 - its translation in the other language
- Result
 - “double” word sketch
 - with compatible relations aligned
- Problem
 - how to align relations?
 - **mapping to universal form** specified in the sketch grammar
 - or they can have identical names

BiM – example

Word Sketch Entry Form ?

Lemma:

Part of speech:

[Advanced options](#)

Advanced options

Subcorpus: [create new](#)

Bilingual manual word sketch: Language: Corpus:

Lemma:

Select gramrels:

- All
- adj_comp

BiM – example

house (*noun*) British National Corpus freq = [57976](#) (516.8 per million)

maison French web corpus freq = [36739](#) (289.6 per million)

modifier	24107	1.3	modifier	3467	0.8	object_of	9534	1.5	objet_de	5965	2.3
White	701	9.65	paternel	112	47.29	build	726	9.06	habiter	220	42.58
opera	334	8.6	hanté	47	44.74	buy	533	8.7	bâtir	136	40.33
manor	236	8.19	familial	162	41.68	sell	308	8.02	quitter	320	39.26
guest	263	8.04	universel	133	38.5	own	138	7.77	construire	220	37.76
terraced	197	8.04	voisin	100	33.12	enter	171	7.59	acheter	139	31.84
discount	212	7.96	natal	41	32.03	rent	56	7.44	clore	76	30.02
big	365	7.9	neuf	56	31.58	occupy	87	7.29	fouiller	48	29.65
clearing	167	7.77	blanc	126	29.28	search	64	7.2	louer	59	29.28
public	358	7.72	royal	55	29.25	leave	420	7.17	incendier	32	28.21

Current status

- All three available in beta version
 - testing mode
 - not set up for many pairs of languages
- Languages
 - only Fr-En
 - Pt-En in progress
 - volunteers welcome