

# Parallel Corpora in Sketch Engine

Vít Baisa



`vit.baisa@sketchengine.co.uk`

5<sup>th</sup> Sketch Engine Workshop  
Bolzano, July 14, 2014

# Parallel corpora features in SkE

- Parallel corpora in SkE
  - preloaded corpora
  - import from TMX
- Dictionaries from parallel corpora
- Concordances
  - parallel concordances
  - experimental: highlighting of parallel KWICs
- Word sketches
  - bilingual manual word sketch
  - translate button (headword)
  - translate collocations

# Dictionaries from parallel corpora

- required for parallel functionality of word sketches
- we don't use standard tools (GIZA++)
- our own tool using logdice statistics
- lemma-based or word-based processing
- top 10 candidates per pair

# Concordances

- standard query form plus aligned corpora forms
- two or more parallel corpora at a time<sup>1</sup>
- query in the aligned corpus/corpora might be omitted<sup>2</sup>
- experimental feature: highlighting parallel KWICs

---

<sup>1</sup>europarl7en,fr,de: energy, énergie, Energie

<sup>2</sup>europarl7: en, de, fr; remarkably

# Experimental: highlighting KWICs

## OPUS2 English

to change the way the programs look. My <b>mother</b> , for example, is a bit afraid of buttons
/ pursuit? cat=dirquery=iii(@ ) Mamma - <b>mother</b> of all Search Engines http://www.mamma
protector. I have not seen my father and <b>mother</b> yet. How are they? Are they healthy? Dear
chase him, so Jánosik hides in a cave. Oh, <b>mother</b> , mother dear, A long way from here. I cry
so Jánosik hides in a cave. Oh, mother, <b>mother</b> dear, A long way from here. I cry longing
fathers wring their hands ... and a few <b>mothers</b> cry heart- rendingly! Just don't think
I don't know where he is. - Talk to the <b>mother</b> . - Carmona has a mom? - Yep. And you know
get a toast here? To Mrs. Carmona. Loving <b>mother</b> who deserved much better than being eaten
better. 30 years ago, I placed this on your <b>mother</b> 's head, Henry. Today, you'll make me
. Why don't you send them home to their <b>mothers</b> ? - They're no good to the joint anymore
new friends. What's Tom doing now, Pat? <b>mother</b> said something about a political job. Yeah
of it all is that he's been lying to his <b>mother</b> . He's leaving her think that he's made
ready, Michael. - Shall I help you in? - No, <b>mother</b> . I can make it. I'll be getting along.
hope for the best. Doctor, this is Tom's <b>mother</b> , brother, and Miss Doyle. You can go in
boys. Didn't say who. - Molly dear? - Yes, <b>mother</b> ? Bring up some clean sheets out of the
through. Just cover me with leaves and tell my <b>mother</b> I died with her name on my lips. Time out
Johnny away from me? You've been a rotten <b>mother</b> to him. You're through with him. The law
custody of the child. I've been a good <b>mother</b> to Johnny. Let's not dispute that point
minute. Johnny, wait here. I want to talk to <b>mother</b> . In this envelope are \$1,500. I've been
. That's the only way you can be a good <b>mother</b> to him now. Johnny, will you go over and

## OPUS2 German

Um z. B. die Funktionen, die durch Knöpfe und Menüeinträge sichtbar sind, auf ein Minimum zu reduzieren, können Sie das & GUI von & koffee; entsprechend anpassen. Falls Sie z. B. nur einfache Briefe schreiben möchten, benötigen Sie nicht viel mehr Funktionen als speichern, laden, drucken & etc;
Mamma - Die <b>Mutter</b> aller SuchmaschinenQuery
Meinem Vater und meine <b>Mutter</b> habe ich noch nicht gesehen, wie geht es Ihnen?
Meine <b>Mutter</b> , meine <b>Mutter</b> , hinter zwei Bergen, der Kopf tut mir bereits weh, wenn ich für sie weine.
Meine <b>Mutter</b> , meine <b>Mutter</b> , hinter zwei Bergen, der Kopf tut mir bereits weh, wenn ich für sie weine.
Weil ein paar Väter die Hände ringen ... und ein paar <b>Mütter</b> herzbrechend schluchzen!
- Rede mit der <b>Mutter</b> , die weitaus besseres verdient hatte als aufgefressen zu werden ..
Liebende <b>Mutter</b> , die weitaus besseres verdient hatte als aufgefressen zu werden ..
Dies habe ich vor 30 Jahren deiner <b>Mutter</b> aufgesetzt, Henry.
Warum schickst du sie nicht heim zu <b>Muttern</b> ?
<b>Mutter</b> sagte, er sei politisch tätig. Ja?
Und das Schlimmste ist, dass er seine <b>Mutter</b> angelogen hat.
- Nein, Mama.
Doktor, das sind Toms <b>Mutter</b> , sein Bruder und Miss Doyle.
- Ja, <b>Mutter</b> ?
Begrabt mich unter Blättern, sagt meiner <b>Mutter</b> , ich hätte zuletzt von ihr gesprochen.
Du warst eine schlechte <b>Mutter</b> .
Ich war eine gute <b>Mutter</b> für Johnny.
Ich will mit <b>Mutter</b> reden.
Nur so kannst du jetzt noch eine gute <b>Mutter</b> für ihn sein.

## OPUS2 French

Ma <b>mère</b> , par exemple, a un peu peur des boutons et des entrées de menus qu'elle ne comprend pas. Pour adapter & koffee; à ses besoins, j'ai réduit le & GUI à des fonctionnalités minimales. Comme elle n'a besoin que d'écrire des lettres et d'utiliser certains modèles, elle n'a pas besoin de fonctions comme enregistrer, ouvrir, imprimer, & etc;
Mamma - la <b>mère</b> de tous les moteurs de rechercheQuery
Je n'ai pas encore vu papa et maman.
Ma mère, ma pauvre mère, je me meurs de douleur de vous avoir perdue sans cesse je pleurs.
Ma mère, ma pauvre mère, je me meurs de douleur de vous avoir perdue sans cesse je pleurs.
Sous prétexte que deux ou trois pères se tortent les mains ... et que quelques <b>mères</b> sanglotent à chaudes larmes!
- Parle à sa <b>mère</b> .
<b>Mère</b> aimante qui méritait mieux que d'être mangée.
Il y a 30 ans, j'ai posé cette couronne sur la tête de la <b>mère</b> .
Renvoie- les chez leur <b>mère</b> .
Maman m'a parlé de politique.
Mais le pire, c'est qu'il a menti à sa <b>mère</b> .
Non, maman, j'y arriverai.
Docteur, voici la <b>mère</b> de Tom, son frère et Mme Doyle.
- Oui, <b>mère</b> ?
Recouvrez- moi de feuilles et dites à ma <b>mère</b> que j'ai prononcé son nom en mourant.
Tu n'as pas été une bonne <b>mère</b> pour lui.
J'ai été une bonne <b>mère</b> pour Johnny.
Je veux parler à ta <b>mère</b> .
C'est la seule façon de l'aider maintenant.

# Word sketch: bilingual manual word sketch

**Bilingual word sketch**

Language: French

Corpus: frTenTen12

Lemma: enfant

Part of speech: noun

Select grammrels: ☐ All

<input checked="" type="checkbox"/> adj_sujet	<input checked="" type="checkbox"/> adj_sujet_of	<input checked="" type="checkbox"/> et_ou	<input checked="" type="checkbox"/> infin_comp
<input checked="" type="checkbox"/> modifier	<input checked="" type="checkbox"/> modifié	<input checked="" type="checkbox"/> objet	<input checked="" type="checkbox"/> objet_de
<input checked="" type="checkbox"/> pp_%s	<input checked="" type="checkbox"/> prédicat	<input checked="" type="checkbox"/> prédicat_de	<input checked="" type="checkbox"/> sujet
<input checked="" type="checkbox"/> sujet_de			

# Word sketch: bilingual manual word sketch<sup>3</sup>

child (noun)

OPUS2 English freq = 944,242 (654.9 per million)

Use another candidate translation: [sexuel](#) [femme](#) [famille](#) [parent](#)[parents](#) [violence](#) [exploitation](#) [jeune](#) [victime](#) [traite](#)

Click on collocates to access reciprocal bilingual search

object of	208,028	2.2	objet de	1,203,124	0.1
<input type="checkbox"/> bear	6,726	9.17	<input type="checkbox"/> naître	22,553	8.33
<input type="checkbox"/> protect	7,719	9.1	<input type="checkbox"/> scolariser	10,739	8.13
<input type="checkbox"/> age	3,896	8.98	<input type="checkbox"/> handicaper	10,615	7.93
<input type="checkbox"/> affect	6,117	8.66	<input type="checkbox"/> accueillir	15,674	7.91
<input type="checkbox"/> deprive	2,578	8.43	<input type="checkbox"/> éduquer	8,102	7.65
<input type="checkbox"/> place	2,381	7.81	<input type="checkbox"/> élever	20,996	7.59
<input type="checkbox"/> raise	3,015	7.78	<input type="checkbox"/> aider	21,452	7.58
<input type="checkbox"/> orphan	1,222	7.56	<input type="checkbox"/> adopter	11,493	7.47
<input type="checkbox"/> involve	3,187	7.45	<input type="checkbox"/> protéger	11,902	7.37
<input type="checkbox"/> abandon	1,194	7.34	<input type="checkbox"/> garder	15,496	7.28
<input type="checkbox"/> include	6,581	7.3	<input type="checkbox"/> abandonner	8,754	7.13
<input type="checkbox"/> educate	1,008	7.19	<input type="checkbox"/> accompagner	9,958	7.09
<input type="checkbox"/> teach	1,233	7.17	<input type="checkbox"/> emmener	5,866	7.03
<input type="checkbox"/> displace	1,232	7.17	<input type="checkbox"/> tuer	8,452	6.96
<input type="checkbox"/> recruit	992	7.1	<input type="checkbox"/> gâter	4,659	6.92
<input type="checkbox"/> register	1,061	6.99	<input type="checkbox"/> Mes	4,700	6.91
<input type="checkbox"/> enroll	844	6.97	<input type="checkbox"/> inscrire	8,311	6.9
<input type="checkbox"/> infect	822	6.89	<input type="checkbox"/> nourrir	5,977	6.9
<input type="checkbox"/> separate	854	6.87	<input type="checkbox"/> atteindre	13,166	6.89
<input type="checkbox"/> abduct	775	6.85	<input type="checkbox"/> laisser	16,711	6.71
<input type="checkbox"/> traffic	722	6.78	<input type="checkbox"/> allaiter	3,876	6.66
<input type="checkbox"/> kill	2,390	6.71	<input type="checkbox"/> confier	4,982	6.63
<input type="checkbox"/> adopt	2,843	6.68	<input type="checkbox"/> amener	5,661	6.62
<input type="checkbox"/> save	1,206	6.66	<input type="checkbox"/> envoyer	8,218	6.54
<input type="checkbox"/> help	1,671	6.6	<input type="checkbox"/> défavoriser	3,694	6.53
>>			>>		

enfant (noun)

frTenTen12 freq = 6,097,960 (493.0 per million)

subject of	144,084	2.0	sujet de	1,497,493	0.2
<input type="checkbox"/> belong	2,313	8.26	<input type="checkbox"/> naître	28,448	8.49
<input type="checkbox"/> live	5,563	8.15	<input type="checkbox"/> scolariser	10,291	7.76
<input type="checkbox"/> suffer	1,522	7.57	<input type="checkbox"/> handicaper	10,950	7.71
<input type="checkbox"/> attend	1,555	7.55	<input type="checkbox"/> grandir	10,064	7.5
<input type="checkbox"/> die	1,486	6.7	<input type="checkbox"/> atteindre	15,298	7.0
<input type="checkbox"/> grow	947	6.57	<input type="checkbox"/> apprendre	11,982	6.96
<input type="checkbox"/> receive	1,687	6.49	<input type="checkbox"/> mourir	10,091	6.9
<input type="checkbox"/> reach	877	6.17	<input type="checkbox"/> souffrir	7,955	6.85
<input type="checkbox"/> enjoy	586	6.14	<input type="checkbox"/> adopter	6,890	6.54
<input type="checkbox"/> have	17,865	6.07	<input type="checkbox"/> adorer	5,908	6.47
<input type="checkbox"/> reside	327	6.01	<input type="checkbox"/> devoir	33,519	6.46
<input type="checkbox"/> be	60,303	5.98	<input type="checkbox"/> gâter	4,116	6.43
<input type="checkbox"/> experience	420	5.97	<input type="checkbox"/> vivre	12,431	6.43
<input type="checkbox"/> face	663	5.85	<input type="checkbox"/> abandonner	6,032	6.4
<input type="checkbox"/> remain	893	5.78	<input type="checkbox"/> pouvoir	72,718	6.4
<input type="checkbox"/> age	313	5.76	<input type="checkbox"/> jouer	13,608	6.35
<input type="checkbox"/> continue	791	5.71	<input type="checkbox"/> hospitaliser	3,705	6.29
<input type="checkbox"/> attain	256	5.64	<input type="checkbox"/> défavoriser	3,653	6.22
<input type="checkbox"/> complete	386	5.6	<input type="checkbox"/> participer	4,965	6.04
<input type="checkbox"/> lack	278	5.58	<input type="checkbox"/> accueillir	4,631	5.97
<input type="checkbox"/> benefit	299	5.52	<input type="checkbox"/> chérir	3,044	5.96
<input type="checkbox"/> cry	266	5.49	<input type="checkbox"/> aller	23,538	5.92
<input type="checkbox"/> scream	233	5.47	<input type="checkbox"/> recevoir	7,682	5.81
<input type="checkbox"/> play	799	5.47	<input type="checkbox"/> venir	11,498	5.8
<input type="checkbox"/> drop	320	5.42	<input type="checkbox"/> devenir	11,940	5.77
>>			>>		

<sup>3</sup>ententen12: child, frtnten: enfant

# Translate word sketch

With a bilingual dictionary we can propose headword translations<sup>4</sup>.

Save	create	<a href="#">585</a>	8.65	collapse
Change options	distort	<a href="#">127</a>	8.52	react
Clustering	access	<a href="#">113</a>	8.33	grow
Sorting	stabilise	<a href="#">77</a>	7.84	remain
Gramrels	dominate	<a href="#">76</a>	7.72	depend
More data	deregulate	<a href="#">62</a>	7.66	become
Less data	expand	<a href="#">60</a>	7.32	fail
Translate	liberalize	<a href="#">48</a>	7.31	<b>need</b>
- French	<b>strengthen</b>	<a href="#">129</a>	7.31	<b>be</b>
- German	<b>protect</b>	<a href="#">139</a>	7.26	<b>have</b>
- Polish	<b>develop</b>	<a href="#">191</a>	7.18	expand
- Italian	flood	<a href="#">43</a>	7.12	constitute
- Spanish				

---

<sup>4</sup>ententen12: mother – German



**mother** (noun) Alternative PoS: [verb](#) (6,537)  
 enTenTen12 freq = [1,484,634](#) (114.5 per million)

**Mutter** deTenTen10 freq = [373,637](#) (131.3 per mil)

Use another candidate translation: [Muttersprache](#) [Vater](#) [Kind](#) [Geburt](#) [alleinerziehend](#) [Sprache](#) [Tochter](#) [Familie](#) [schwanger](#)  
 Click on collocates to access reciprocal bilingual search or find [translated collocations](#)

<a href="#">object of</a>	<a href="#">303,454</a>	<a href="#">0.2</a>	<a href="#">VerbY+SubstXDat</a>	<a href="#">24,327</a>	<a href="#">0.3</a>
<a href="#">(obj dat of)</a>					
breastfeed	<a href="#">2,213</a>	7.58	aufwachsen	<a href="#">86</a>	4.59
divorce	<a href="#">1,215</a>	6.71	zusammenleben	<a href="#">33</a>	4.56
lactate	<a href="#">738</a>	6.28	abstammen	<a href="#">23</a>	4.51
grieve	<a href="#">628</a>	5.86	erben	<a href="#">40</a>	4.48
marry	<a href="#">1,560</a>	5.77	wohnen	<a href="#">465</a>	4.39
murder	<a href="#">658</a>	5.61	leben	<a href="#">1,579</a>	4.21
nurse	<a href="#">551</a>	5.61	übersiedeln	<a href="#">25</a>	4.18
kill	<a href="#">2,978</a>	5.57	wachsen	<a href="#">446</a>	3.83
honor	<a href="#">888</a>	5.45	fliehen	<a href="#">74</a>	3.82
age	<a href="#">1,164</a>	5.43	flüchten	<a href="#">59</a>	3.71
tell	<a href="#">6,868</a>	5.41	zustehen	<a href="#">41</a>	3.41
expect	<a href="#">3,476</a>	5.31	telefonieren	<a href="#">60</a>	3.36

<a href="#">subject of</a>	<a href="#">423,875</a>	<a href="#">0.5</a>	<a href="#">SubstXNom+VerbY</a>	<a href="#">62,794</a>	<a href="#">0.8</a>
<a href="#">(subj of)</a>					
die	<a href="#">11,269</a>	7.39	sterben	<a href="#">1,685</a>	5.78
breastfeed	<a href="#">832</a>	5.77	heiraten	<a href="#">274</a>	5.22
tell	<a href="#">8,097</a>	5.61	beschützen	<a href="#">109</a>	5.11
abandon	<a href="#">989</a>	5.36	versterben	<a href="#">169</a>	5.1
teach	<a href="#">2,685</a>	5.32	stillen	<a href="#">102</a>	5.02
remarry	<a href="#">480</a>	5.17	entstammen	<a href="#">100</a>	4.79
insist	<a href="#">705</a>	5.09	bitte	<a href="#">109</a>	4.69
cry	<a href="#">828</a>	5.03	weinen	<a href="#">141</a>	4.69
smoke	<a href="#">646</a>	4.9	schimpfen	<a href="#">80</a>	4.47
refuse	<a href="#">751</a>	4.83	erzählen	<a href="#">847</a>	4.39
raise	<a href="#">2,588</a>	4.82	trösten	<a href="#">63</a>	4.3
pass	<a href="#">2,512</a>	4.69	flehen	<a href="#">46</a>	4.25