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)

<table>
<thead>
<tr>
<th>object</th>
<th>subject</th>
<th>modifier</th>
<th>and/or</th>
</tr>
</thead>
<tbody>
<tr>
<td>tail</td>
<td>pur-</td>
<td>playfully</td>
<td>pounce</td>
</tr>
<tr>
<td>squirrel</td>
<td>ambulance</td>
<td>blindly</td>
<td>harass</td>
</tr>
<tr>
<td>rabbit</td>
<td>jpmorgan</td>
<td>relentlessly</td>
<td>hound</td>
</tr>
<tr>
<td>dream</td>
<td>play-off</td>
<td>frantically</td>
<td>bark</td>
</tr>
<tr>
<td>rainbow</td>
<td>paparazzi</td>
<td>jpmorgan</td>
<td>stalk</td>
</tr>
<tr>
<td>prey</td>
<td>squirrel</td>
<td>aggressively</td>
<td>wrestle</td>
</tr>
<tr>
<td>shad</td>
<td>lion</td>
<td>doggedly</td>
<td>nip</td>
</tr>
<tr>
<td>butterfly</td>
<td>cheetah</td>
<td>endlessly</td>
<td>fetch</td>
</tr>
<tr>
<td>mirage</td>
<td>storm</td>
<td>futilely</td>
<td>lunge</td>
</tr>
<tr>
<td>ball</td>
<td>pur</td>
<td>fruitlessly</td>
<td>harry</td>
</tr>
<tr>
<td>ghost</td>
<td>coyote</td>
<td>madly</td>
<td>party</td>
</tr>
<tr>
<td>deer</td>
<td>hound</td>
<td>away</td>
<td>bite</td>
</tr>
</tbody>
</table>
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”
<table>
<thead>
<tr>
<th>Introduction</th>
<th>BIP</th>
<th>BIC</th>
<th>BIM</th>
<th>Status</th>
</tr>
</thead>
</table>

Sue Atkins, 2005

“Can I have bilingual ones please”

... hmmmm, now what might that mean
Three answers

- **BiP**
  - Bilingual word sketch from Parallel data
- **BiC**
  - Bilingual word sketch from Comparable data
- **BiM**
  - Bilingual word sketch – Manual
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

Vojtěch Kovář  Adam Kilgarriff  Ana Frankenberg-Garcia
Masaryk University, Lexical Computing Ltd., Surrey University
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
<table>
<thead>
<tr>
<th>Bilingual word sketches Three flavours</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>BiP – example</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>declaration (noun)</th>
<th>EUROPARL5, English-French freq = 4409</th>
</tr>
</thead>
<tbody>
<tr>
<td>déclaration (noun)</td>
<td>EUROPARL5, French-English freq = 9341</td>
</tr>
</tbody>
</table>

use another candidate translation: déclarations écrites écrite Déclarations

<table>
<thead>
<tr>
<th>modifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>written</td>
</tr>
<tr>
<td>écrive</td>
</tr>
<tr>
<td>solemn</td>
</tr>
<tr>
<td>solennel</td>
</tr>
<tr>
<td>unilateral</td>
</tr>
<tr>
<td>unilatérale</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr President, Members have had circulated to them this evening notice of my written declaration on alloops which lapsed at 6:30 pm.</td>
</tr>
<tr>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
</tr>
<tr>
<td>EUAfrica Summit in Cairo: (FR) The solemn declaration of the first EUAfrica 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 the secular past was a story of slavery, massacres, forced labour, plundering, colonial conquest 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 important countries or the baseness.</td>
</tr>
<tr>
<td>La déclaration solennelle du premier sommet Afrique-Europe, au Caire, commence par faire référence, je cite aux “hiers 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”.</td>
</tr>
<tr>
<td>He must not add fuel to the flames by threatening a unilateral declaration of independence for the Palestinian State.</td>
</tr>
</tbody>
</table>

Vojtěch Kovář Adam Kilgarriff Ana Frankenberg-Garcia Masaryk University, Lexical Computing Ltd., Surrey University
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
<table>
<thead>
<tr>
<th>modifier</th>
<th>149</th>
<th>49</th>
</tr>
</thead>
<tbody>
<tr>
<td>written</td>
<td></td>
<td></td>
</tr>
<tr>
<td>écrite</td>
<td></td>
<td></td>
</tr>
<tr>
<td>solemn</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>solennel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unilateral</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>unilatérale</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

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

use another candidate translation: déclarations écrites écrite Déclarations

---

**written**

I am surprised that on 7 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.

Mr President, Members have had circulated to them this evening notice of my written declaration on alcopops which lapses at 6.30 pm.

Monseur 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**

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.

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 the secular past was a story of slavery, massacres, forced labour, plundering, colonial conquest 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 important countries or the baseness of the declaration solemn de premier sommet Afrique-Europe, au Caire, commence par faire référence, je cite, aux "hiers 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**

He must not add fuel to the flames by threatening a unilateral declaration of independence for the Palestinian State.

---

**unilatérale**

Le problème est le suivant : après avoir refusé pendant neuf mois 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’État 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: intelligent
Part of speech: adjective

Advanced options

Subcorpus: create new

Bilingual manual word sketch:
Language: None
Corpus: None
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)

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