Setting up for Corpus Lexicography

Adam Kilgarriff, Jan Pomikalek, Pete Whitelock

LCL & OUP

Premise

- Corpus technology can support lexicography

  making it

  - ***more accurate***
  - ***more consistent***
  - ***faster***

  Rundell and Kilgarriff 2011

  Automating the creation of dictionaries

  in Sylviane Granger's Festschrift

- This paper
  - A case study

A new Portuguese dictionary

- OUP
- Pt-En and En-Pt
- 40,000 headwords on each side
- Pt-En starts from
  - Dictionary
    - Medium-sized Pt-Dutch
  - Corpus
    - blank sheet

Agenda

1. Collect corpus
2. Process with best tools
3. From parser output to corpus system input
4. Finding good examples
5. Regional variants

Status

1. Collect corpus
2. Process with best tools
3. From parser output to corpus system input
4. Finding good examples
5. Regional variants

Corpus collection

- Big and diverse
- 100m not big enough
  - 40,000 headwords
  - 40,000th word in BNC: 27 hits

dogfish *(noun)* British National Corpus freq = 27 (0.2 per million)

modifier	7	1.0
spotted	1	6.19
odd	3	4.55
spur	1	4.41
lesser	1	4.09
breeding	1	3.75

modifies	2	0.3
skin	1	1.38

and/or	18	5.5
gurnard	1	10.25
thornback	1	10.19
dogfish	2	10.14
pollock	1	10.09
Scuba	1	9.42
pout	3	9.33
codling	1	7.83
whiting	1	7.48
dab	1	6.86
cod	2	6.8
ling	1	6.64
crab	1	5.21

pp_obj_for-p	3	10.1
Mwnt	1	12.68
mark	1	1.72

pp_obj_with-p	2	9.2
daylight	1	4.73
poor	1	1.17

pp_to-p	1	2.9
lb	1	5.16

pp_obj_o
Catch
Nerve

dogfish *(noun)* ukWaC freq = 462 (0.3 per million)

object of	77	1.3
white	2	8.59
dab	6	8.12
pout	3	7.63
catch	8	1.24

modifier	202	1.5
lesser-spotted	9	10.22
deep-sea	14	8.53
spiny	7	7.88
huss	2	7.81
wrasse	11	7.74

modifies	88	0.6
coalfish	2	7.61
pollack	3	7.15
conger	3	6.92
wrasse	3	5.97
ling	2	5.9

and/or
Huss
bullhuss
nursehound
wrasse
gulper

Where from?

- Web
- Quantity
  - Yeah
- Quality
  - As good or better
    - Keller and Lapata 2003, Sharoff 2006, Baroni et al 2009

How?

- New linguistics-specialist crawler
  - Was Heritrix, next time: Spiderling
    - Other talk
- Cleaning *including language-identification*
  - jusText
    - Pomikalek thesis
- Deduplication
  - Onion
    - Pomikalek thesis

Processing tools

- Reviewed options
  - Best: ***Palavras***
    - Bick 2000
    - + ongoing development since
  - Contacted author, negotiated licence
  - Installed
  - Applied to 2b words

Vast process

- Parsing is usually slow - would it take years?
- Parallelised in 12 processes
- Many bugs encountered, resolved with developers
- Crashed on many input files – leave them out
- Final run: 15 days

Corpus creation stats

	European	Brazilian
HTML data downloaded	1.10 TB	1.37 TB
Unique URLs	31.5 million	39.1 million
Crawling time	8 days	10 days
Final corpus size (words)	0.7 billion	1.0 billion

From dependency parse to word sketch

- Palavras: *dependency parser*
- Output for each word
  - Lemma, pos tag
  - “my governor is word N”
  - “relation is …”
- Like CONLL output
- SKEW-2, Siva Reddy
  - Word sketches from CONLL format data

To get better word sketches

- Parser output and lexicographic word sketches
  - *Not quite the same*
- Anomalies in parser output
- Large project

Preposition-article contractions

satisfação [satisfação] <cjt> <act> <percep-f> N F S @<ACC #19->17
de [de] <sam-> <np-close> PRP @N< #20->19
os [o] <-sam> <artd> DET M P @>N #21->23
nossos [nosso] <poss 1P> DET M P @>N #22->23
clientes [cliente] <Hattr> N M P @P< #23->20

19 satisfação satisfação N F:S 14,V obj %w_N/%w_V obj
20 dos de PRP 19,NIL/%w_N dep PRP
21 nossos nosso DET M:P 22,NIL/DET spec_of %w_N
22 clientes cliente N M:P 19,N _de_ %w_N/%w_N _de_ N

Verb form reconstruction

Deveria- [dever] <*> <hyfen> <fmc> <aux> V COND 3S VFIN @FS-STA #1->0
se- [se] <hyfen> PERS M/F 3S/P ACC @<SUBJ #2->1
começar [começar] <vH> <mv> V INF @ICL-AUX< #3->1

1 Dever-se-ia dever V COND:3S:VFIN 1,REFL-SUBJ
2 começar começar V INF 1,V comp %w_V/%w_V comp V

Multi-word unpacking

A=Comunidade=de=Direitos=Humanos
[A=Comunidade=de=Direitos=Humanos] <tit> <*> PROP F P @NPHR #2->0

<mwe parsed="yes" pos="PROP">
2 A o DET F:S 3,NIL/DET spec_of %w_N
3 Comunidade comunidade N F:S
4 de de PRP 3,NIL/%w_N dep PRP
5 Direitos direito N M:P 3,N _de_ %w_N/%w_N _de_ N
6 Humanos humano ADJ M:P 5,N mod %w_ADJ/%w_N mod ADJ
</mwe>

Trinary Relations/Coordination

```
um      [um] <arti> DET M S @>N  #17->18
simulador       [simulador] <H> N M S @<SC  #18->0
de      [de] <np-close> PRP @N<  #19->18
inclinação      [inclinação] <cjt-head> <percep-f> <am> N F S @P<  #20->19
e       [e] KC @CO  #21->20
direção         [direção] <cjt> <HH> <dir> <Ltop> N F S @P<  #22->20
```

```
17      um      um      DET     M:S     18,NIL/DET spec_of %w_N
18      simulador       simulador       N       M:S
19      de      de      PRP             18,NIL/%w_N dep PRP
20      inclinação      inclinação      N       F:S     18,N _de_ %w_N/%w_N _de_ N
21      e       e       KC
22      direção direção N       F:S     18,N _de_ %w_N/%w_N _de_ N;
                                        20,N e|ou %w_N/%w_N e|ou N
```

Control relations

não [não] ADV @ADVL> #3->4
é [ser] <vK> <fmc> <mv> V PR 3S IND VFIN @FS-STA #4->0
viável [viável] <nh> ADJ F S @<SC #5->4
sua [seu] <poss 3S> DET F S @>N #6->7
aplicação [aplicação] <act> <sem-r> N F S @<SUBJ #7->4

3 não não ADV 4,%w_ADV mod_of V/ADV mod_of %w_V
4 é ser V PR:3S:IND:VFIN 7,N subj_of %w_V/%w_N subj_of V
5 viável viável ADJ F:S 4,V dep %w_ADJ/%w_V dep ADJ;
7,N subj_of %w_ADJ/%w_N subj_of ADJ
6 sua seu DET F:S 7,NIL/DET spec_of %w_N
7 aplicação aplicação N F:S

Reanalysis

variados [variar] V PCP M P @>N #21->22
aspectos [aspecto] <cjt-head> <ac-cat> N M P @<SUBJ #22->17
de [de] <sam-> <np-close> PRP @N< #23->9
a [o] <-sam> <artd> DET F S @>N #24->25
tecnologia [tecnologia] <domain> N F S @P< #25->23

20 variados variar V PCP:M:P
21 aspectos aspecto N M:P 35,N subj_of %w_N/%w_N subj_of N
22 da de PRP 21,NIL/%w_N dep PRP
23 tecnologia tecnologia N F:S 21,N _de_ %w_N/%w_N _de_ N

Lemmatization

Old spelling	New spelling
acto	ato
carbónico	carbônico
cabeça-de-burro	cabeça de burro
concetual	conceptual
auto-sugestão	autossugestão

Female form	Male form
amiga	amigo

GDEX

- Good dictionary example finder
- Customise for Portuguese
  - Follow Slovene lead

Regional variation

- European *vs* Brazilian
- Method
  - Keyword list of each vs other
  - If in top 1%: add note to word sketch