Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora*

http://www.linguateca.pt/ACDC/
http://www.linguateca.pt/ACDC/corte-e-costura/

Diana Santos Diana.Santos@sintef.no
Cristina Mota cmota@ist.utl.pt

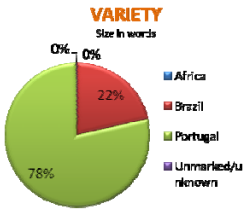
Motivation

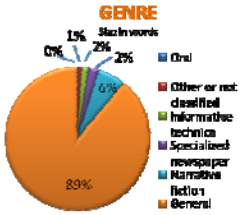
- ☹ Every corpus a different formalism and search expression
- ☹ Every corpus different undocumented assumptions and terminology
- ☹ No replication or addition to already existing corpora possible
- ☹ No revision/feedback by users possible

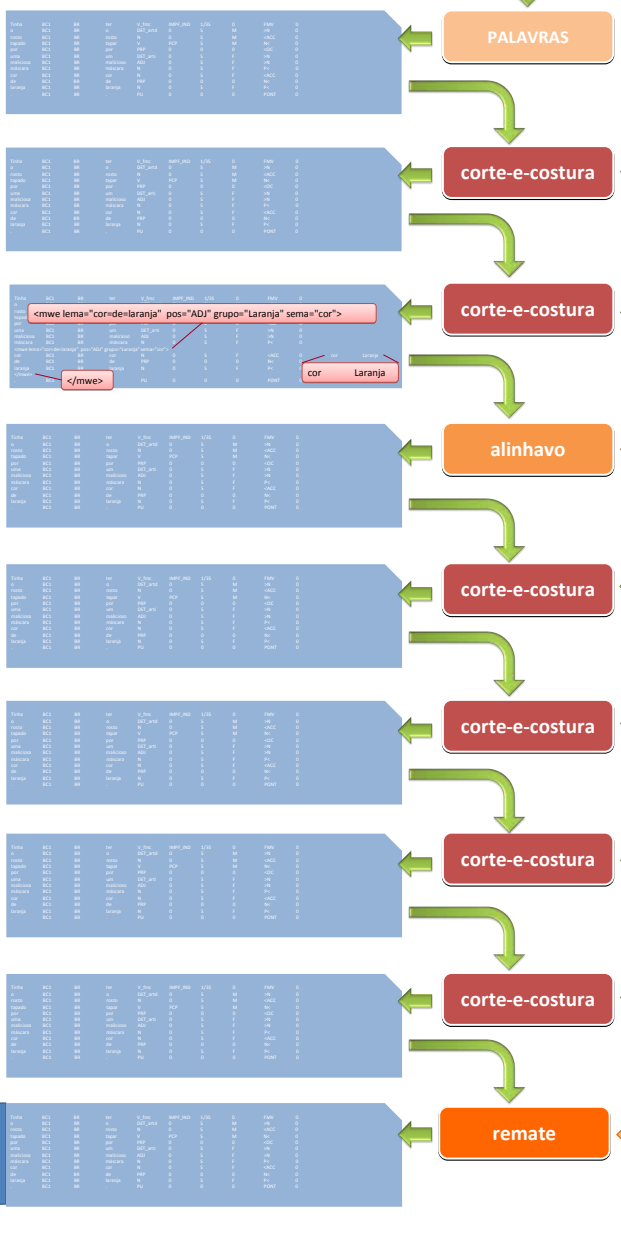
☺**AC/DC answer**: cooperative annotation of already richly-annotated corpora

AC/DC Corpora

Corpus	Units	Words	Sentences	Var.	Short description
AmostRA	127.832	98.786	4.965	B	Pos-tagged sample of NILC Corpus
ANCIB	1.690.376	1.258.764	80.992	B	(Moderated) Brazilian discussion list on librarianship
Avante!	7.766.418	6.501.257	204.414	P	Portuguese party-political newspaper Avante!, 1997-2002
CD HAREM	222.407	147.077	8.185	all	Golden collection of the First HAREM
CETEMPúblico	232.543.379	189.575.095	7.665.410	P	Major Portuguese daily newspaper, 1991-1998
CHAVE	123.868.725	99.478.954	4.740.448	P B	News from Público and Folha de São Paulo, 1994-1995
ClassLPPE	1.922.601	1.304.282	74.690	P	Portuguese fiction, drama and poetry, 16th and 19th century
CONDIVport	7.088.775	5.577.632	328.214	P B	Sports, fashion and health magazines (50s, 70s and 2000)
CoNE	925.230	685.225	31.561	any	Spam or general e-mail messages
DiaCLAV	7.758.467	6.651.549	232.152	P	Four Portuguese regional daily newspapers
ECI-EBR	917.127	724.015	44.381	B	Corpus Borba-Ramsey of Brazilian Portuguese
ECI-EE	32.034	27.140	839	P	Call for the EU ESPRIT program
ENPCPUB	92.693	72.389	4.371	P B	Translated fiction from English, subset of the ENPC corpus
FrasesPB	23.313	19.162	653	B	Individual sentences in Brazilian Portuguese
FrasesPP	20.049	16.233	594	P	Individual sentences pos-tagged
MuseuPessoa	517.747	375.158	27.288	P B	Transcriptions of oral interviews from Museu da Pessoa
Natura/Minho	2.156.187	1.749.083	68.910	P	Unedited version of regional Portuguese newspaper
Natura/Público	7.369.349	6.274.542	225.752	P	Two first paragraphs of each Público article, 1991-1994
NILC	42.608.038	32.342.456	1.963.795	B	Newspaper, commercial letters,textbooks, etc.
Vercial	18.854.273	14.315.992	596.869	P	Portuguese fiction, poetry and drama, 16th to 20th century
Total	457.707.958	368.107.085	16.342.734		







Correction

[word="Albino" & lema="albino"] >> [pos="PROP" & lema="Albino"]
[lema="azulão"] >> [lema="azul" & temcagr="AUM"]
[lema="pigmento|tafetá|veludo"] b:[lema="rubi"] >> b:[pos="ADJ"]

MWE

sema=cor pos=N
sangue azul
febre amarela
sema=cor pos=ADJ
amarelo alaranjado
cor de laranja

Simple Lexicon

anil, azul, azul-esverdeado, marinho, turquesa, incolor, transparente, anorak, canadiana

MWE

azul alentejano, cor de mar, cor de á gua, cor de turquesa, cor de casca de ovo, cor de champanhe, cor de pele, vestido de pele

Positive

[lema="tom"] [word="de"] c:[lema="aço|âmbar|areia|carne|malva|marfim|marinho|morango|ouro|pérola|pêssego|violeta"] >> c:[sema="cor"]
a:[lema="damasco"] [word="e"] b:[lema="melão"] >> a:[sema="cor"] b:[sema="cor"]

Negative

[lema="coroa|coroar|folha"] [lema="de"] a:[lema="louro"] >> a:[sema="0"]
a:[lema="leite"] b:[lema="creme"] >> b:[sema="0"]
a:[lema="madeira|móvel"] b:[lema="de"] c:[lema="castanho"] >> c:[sema="material"]

Specialization

a:[lema="caldo|chá|feijão|pimento"] b:[lema="verde"] >> b:[sema="cor:original"]
a:[lema="vinho"] b:[lema="branco|verde"] >> b:[sema="cor:vinho"]

Recursive

a:[lema="aço|areia|banana|borgonha|caramelo|celeste|cereja|champanhe|chocolate|cinza|cobre|coral|creme|havana|marfim|marinho|pérola|pêssego|royal|turquesa|uva|violeta" & pessnum="S"]
b:[lema=",|e|ou"] c:[sema="cor"] >> a:[sema="cor"]

Simple Lexicon

Azul: anil, azul, azul-esverdeado, marinho, turquesa, ...
Ausência: incolor, transparente, ...
CasacoImpermeável: anorak, canadiana, ...

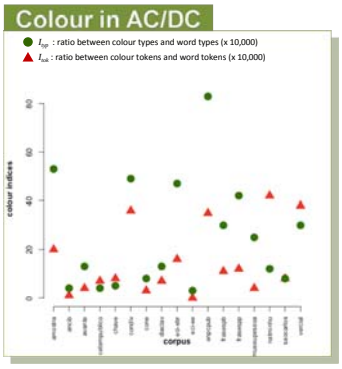
MWE

Azul: azul alentejano, cor de mar, cor de á gua, cor de turquesa, ...
Creme: cor de casca de ovo, cor de champanhe, cor de pele, ...
Vestido: vestido de noite, ...

Rule activation & Corpora changes

Corpus	Rules fired	Words changed	Word types changed	Attr. changed	New MWE	MWE changed
AmostRA	30	312	142	570	5	3
ANCIB	93	388	108	641	21	3
Avante!	707	4237	372	7641	40	34
CHAVE	19965	132244	1998	242486	2796	700
CoNE	27	393	105	642	1	0
DiaCLAV	1973	7513	481	13116	444	25
ECI-EBR	150	1969	412	3699	25	3
ENPCPUB	225	377	162	637	6	2
MuseuPessoa	71	751	137	1451	7	1
Natura/Minho	553	1388	224	2093	39	2
Vercial	2921	50246	1553	93831	546	99

Colour in AC/DC



Conclusions

- ❖ **It is possible with little effort and with a large user community behind**
- ❖ **Remains to be seen it the remaining annotation corpora are used**

Note: Reuse of **OpenCWB** and **(Open)PALAVRAS** a must! Without them we would again have started from scratch...

- OpenCWB http://cwb.sourceforge.net/
- (Open)PALAVRAS http://visl.sdu.dk/visl/pt

* All numbers dated May 5th, 2010