

Advantages and Challenges of Using Corpora in Translation Practice

Ana Frankenberg-Garcia

Tuesday, 07 July 2015

Translators use language that is not their own



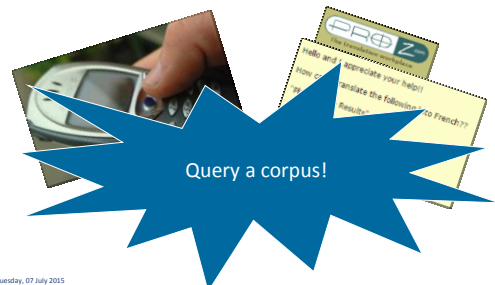
Tuesday, 07 July 2015

Translators use language that is not their own



Tuesday, 07 July 2015

Translators use language that is not their own



Tuesday, 07 July 2015

Why should translators use corpora?

- Instant access to combined intuitions of dozens (or hundreds) of language users
- Without informants having to think
- Without translators having to bother informants
- Answers to questions that have not been documented in dictionaries, termbases and other edited resources
 - Equivalence
 - Terminology
 - Phraseology
 - Translation decisions

e.g. Aston 1999, Bowker & Pearson 2002, Zanettin et al. 2003, Beeby et al. 2009, Rodríguez-Inés 2010, Kubler 2011, Zanettin 2012, Gallego-Hernández 2015, etc.

Tuesday, 07 July 2015

Why should translators use corpora?




Tuesday, 07 July 2015

Why should translators use corpora?

Business Letter Corpus
A collection of around 1 million words of U.S. and U.K. business letters compiled into a corpus by Professor Yasumasa Someya for his MA at the University of Tokyo

Free online access
<http://www.someya-net.com/concordancer>



Tuesday, 07 July 2015

Business Letter Corpus
Online KWIC Concordancer
(How to Use Me -- A Quick Reference)
INSTRUCTIONS for more details). Hope you will find it useful in writing business messages o

Search String:
Search Type:
Line Width:
Search Corpus:
Sort Type:

Click or

Online BLC KWIC Concordancer Search Result
Search String: contain "looking forward to hearing"
Search Corpus: 01. Business Letter Corpus (BLC2000)
No. of Hits: 19

1 programme sounds interesting indeed and I shall be looking forward to hearing all about it from
2 [BLC2:04:01087] I'm looking forward to hearing from you in the ne
3 [BLC2:04:03240] I am certainly looking forward to hearing from you again so
4 [BLC2:31:01360] We are looking forward to hearing from you in the ne
5 [BLC2:04:03537] I am looking forward to hearing from you in the ne
6 [BLC2:29:00058] I'm looking forward to hearing from you soon.
7 assistance in making our final decision, and are looking forward to hearing from you soon.
8 [31:03180] Bee appreciates your interest and I am looking forward to hearing from you soon.
9 [BLC2:32:01976] I am looking forward to hearing from you soon.
10 [BLC2:32:02042] I am looking forward to hearing from you soon.
11 [BLC2:32:03159] We are looking forward to hearing from you soon.
12 [BLC2:01:00221] I am looking forward to hearing from you.
13 [BLC2:31:00769] We are looking forward to hearing from you.
14 [BLC2:32:00772] We are looking forward to hearing from you.

Business Letter Corpus
Online KWIC Concordancer
(How to Use Me -- A Quick Reference)
INSTRUCTIONS for more details). Hope you will find it useful in writing business messages o

Search String:
Search Type:
Line Width:
Search Corpus:
Sort Type:

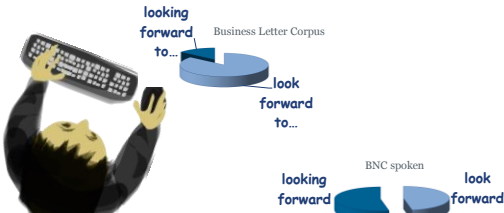
Click or

Online BLC KWIC Concordancer Search Result
Search String: contain "look forward to hearing"
Search Corpus: 01. Business Letter Corpus (BLC2000)
No. of Hits: 212

1 r your prompt attention to this matter, and shall look forward to hearing a favorable reply from
2 BLC2:15:03672] We hope you enjoy your course, and look forward to hearing about it.
3 [BLC2:36:01390] We look forward to hearing favorably from you so
4 [BLC2:36:02613] We look forward to hearing favorably from you.
5 [BLC2:25:07827] I look forward to hearing favourably from you.
6 [BLC2:32:02614] We look forward to hearing from teachers in some
7 [BLC2:32:02519] We look forward to hearing from the prospective
8 service we need in the Lebanon, and we shall now look forward to hearing from them.
9 BLC2:32:02749] We thank you for your interest and look forward to hearing from Whatawhata High S
10 [BLC2:15:01073] I look forward to hearing from you about this ic
11 [BLC2:15:05975] I look forward to hearing from you after you ha
12 you for placing your order with our company and I look forward to hearing from you again in the
13 [BLC2:22:01273] We look forward to hearing from you again in the
14 [BLC2:05:00064] We look forward to hearing from you again soon.

Interpreting corpus data


- Empirical evidence
- But no clear explanations and rules
- It's up to users to interpret the data
- And results are only as good as the corpus



Tuesday, 07 July 2015

Interpreting corpus data

- Interpretation is a matter of common sense
- Corpora are made of natural language
- They contain
 - Mistakes
 - Non-standard usage
- What is interesting is what is **conventional**



Tuesday, 07 July 2015

What queries can be useful to translators?



Concordances?
Frequencies?
Collocations?
Word lists?

Tuesday, 07 July 2015

Concordances

Portuguese > Casada com um francês
Literal English translation > *Married with a Frenchman



Is this right?

Tuesday, 07 July 2015

Concordances

BNC concordance query: *married* sort right

health and complementary medicine. **Married** with a son, she is the author of The Well
health and complementary medicine. **Married** with a son, she is the author of The Well
erpool Sunday Referees Society. He is **married** with a son and daughter and lives in Mossley
d at weekends. </p><p> Mike - 43 and **married** with a son of ten - has already worked
> was from Colchester, Essex, and was **married** with a three-year-old son. </p> Dairy jobs
aduate of Forrest High School in 1975, **married** with a three-year-old daughter and a one-year
Services. </p><p> Mr. Benson, who is **married** with a two-year-old son, lives near Havant
ano teacher in London. Now 43, she is **married** with a two-year-old daughter and says it
way. The Arbutnots prospered. They **married** with a view to inheritance, attended the
will manage , although you are a year **married** with a wee child learning to walk, and
ry </p><p> Roger Morrison was 26 and **married** with a young child. He joined the Gibraltar
questioning two men. Mark Perrins was **married** with a young daughter and two children
a , they know for six months, they get **married** with all the gear, the next year they got
his career in England. </p><p> Jones, **married** with an 18 month-old daughter, was adamant
<p><p> Police say the attack on Ashiq, **married** with an 18-month-old daughter, appeared
er to escape'. </p><p> Kathryn, who is **married** with an 18-month-old daughter, said she
put 'single' on her CV, although she is **married** with an 18-month-old baby. 'It's vital
teacher. All his classmates were either **married** with brats or disgustingly overweight.
be of slob sitcoms such as Roseanne or **Married** with Children , and it's hard not to wonder

Tuesday, 07 July 2015

Concordances

BNC concordance query: *married* sort right

Julies, an ex-computer programmer **married** to a radio producer, described her teeling
<p> He and his wife have a daughter **married** to a retained firefighter at Masham. </p>
amps Elysées, while Juliette Adam, **married** to a rich banker and later to become the
o Roman Catholics and a Protestant **married** to a Roman Catholic - again in three separate
et named Nina Runich (an Austrian **married** to a Russian); she used the stage name
ry Byron, a middle-class housewife **married** to a sales representative, talks about
he next door flat, a Dutch Javanese **married** to a Scotsman, came rushing in to see if
our waters, the Lady Yolande was **married** to a Scottish King. We bear warrants from
in the colonial culture of India and **married** to a senior military figure. In her fifties
ever been kissed, and he had been **married** to a sensational young actress. The contras
driver Mirsad Nevsetovic, a Muslim **married** to a Serb, expressed the apparently forlorn
entieth centuries. An Englishwoman **married** to a Serb (Mrs Lawton-Mijatovićacut):
was an East London waitress, first **married** to a shrimp-seller and later living with
are delicious, Simeti (an American **married** to a Sicilian) admits that some are of
William's brother Thomas was one. **Married** to a sickly wife, then having to cope as
il awareness that comes from being **married** to a singer - his compassion - his steadfast
with 'il Signor Conte', and a period **married** to a sisal farmer in Portuguese East Africa
pharnais and the Murats and he was **married** to a sister of the King of Portugal. All

Tuesday, 07 July 2015

Concordances

English > They are very flexible nowadays.
Portuguese literal translation > Eles são muito flexíveis **hoje em dia**



Does this sound good?

Tuesday, 07 July 2015

Concordances

BNC concordance query (sentence view): *nowadays*

Few contested actions are **nowadays** decided in favour of media defendants. </p>
I think **nowadays** we all know, that the flagships of the conservative party, crime is no longer a thing they can fly high.
It isn't merely a matter of training and qualifications, it's a matter of religious commitment as well and most teachers **nowadays** take the view that to try and teach this subject without that religious commitment is sheer hypocrisy.
'The things you girls do **nowadays** .
'It's what happens **nowadays** .
erm, and er, sometimes even pleased to say well I'll try to get you in purple, but er, **nowadays** erm, we we don't go in for that because there's, there's so many er shops now that selling lighting fittings erm and selling nothing else, there isn't one actually in Harlow, but there's one in Epping and there's one in Stortford, erm that er, it's, it's riding a new car people want to see er a good variety of them, have a look at a lot or a washing machine or anything like that, you want to see many before you buy decorated type of shades, erm usually **nowadays** in, in plastic but erm
Not much in the way of floor standards now, erm I don't think there's many floor standards sold **nowadays**
and really always has been, erm, the idea of a centre is, is simply because you consider how much it costs to have a point wired when the house is, is built, er several people **nowadays** is ha well they buy an old house, they erm, say well we'll take out the centre light and put us in some wall lights instead you see.
they were, they were all, all going, people had them, but er you don't see them **nowadays**
<p> Just a little point of interest, how many people are on this thing **nowadays** ?

Tuesday, 07 July 2015

Concordances

p(TenTen concordance query (sentence view): *hoje em dia*

Eu acho que as possibilidades ambivalentes do conhecimento e do conhecimento científico são possibilidades que **hoje em dia** explodem cada vez mais .
Hoje em dia nós ligamos a televisão e podemos assistir à morte com grandes diferenças.
<p> Compra seu Maveric pq é um puta carro, tem muito macaco **hoje em dia** , todos querem se achar especiais, mas acabam sendo iguais seguem a opinião geral.
<p> Estas imagens podem não parecer tão extraordinárias à primeira vista , comparadas com as feitas **hoje em dia** , mas quando se descobre que foram tiradas entre 1909 e 1912, ou seja , há mais ou menos um século atrás é difícil não ficar surpreso.
<p> **Hoje em dia** se encontra muito generalizada a ideia do poder e influência do pensamento.
<p> **Hoje em dia** só vive preso a um grupo restrito de pessoas
Pois bem , **hoje em dia** é não é muito difícil encontrar pessoas inteligentes e cultas.
<p> É impossível pensar em telecomunicações **hoje em dia** sem pensar em programação.
<p> **Hoje em dia** , essa falta de educação do público é um fator que afasta ainda mais as pessoas do cinema.
Hoje em dia , em cada dez animações lançadas em cinema, oito são bem sucedidas, e poderia dizer que as oito têm ótima qualidade.
<p> Por outro lado , a ideologia nazista teve (e tem **hoje em dia**)
<p> **Hoje em dia** , todos querem ser originais, especialmente os gurus.
Hoje em dia nem sempre existe mais aquele Oceano, nem Albertino.
Em cada linha a ver com os senadores de **hoje em dia** .
em julho de 2004, recebi algumas mensagens da amiga-virtual TERESA Rios Van DUSEN, brasileira mas que mora **hoje em dia** em Dallas, Texas, nos Estados Unidos . </p>

Tuesday, 07 July 2015

Concordances

English > In my opinion
Literal French translation > Dans mon opinion

Is this right?

Tuesday, 07 July 2015

Parallel concordances

EuroParl corpus concordance query: *in my opinion*, align with French

In my opinion , we should look further into this point , and I also think that the Maastricht Treaty , which comes into effect in 2001 , should be looked at very closely . There are certain weaknesses, the greatest of which , **in my opinion** , is that not enough consideration is being given to the situation of small businesses : after all , the car industry is more than just large companies I welcome these even though , **in my opinion** , they do not go far enough . The charter should at least deal with the citizens ' fundamental rights , the political rights , the social rights and , **in my opinion** , the rights of minorities too , and it should form a supplement to what we already have . With reference to the argument of subsidiarity , I should simply like to tell you that , **in my opinion** , it is also important to put forward the argument of citizenship in fiscal matters . **In my opinion** , the most urgent need is to create , as soon as possible , in all Member States and in the candidate countries , the legal and technological capability to search the contents of the Internet for child pornography . **In my opinion** , this is indeed a huge challenge which must be accepted . I would like to list three elements which are lacking **in my opinion** .

Je crois que nous devons avancer sur ce point et que la convention Polmar qui prend cours en 2001 doit être examinée attentivement . Elle témoigne de certaines faiblesses . Selon moi , la faiblesse principale est que la situation des PME n'est pas suffisamment prise en considération , alors que l'industrie automobile n'est pas seulement constituée de grands groupes . Je m'en félicite , même si , à mon avis , ce n'est pas suffisant . La Charte doit , en tout cas , traiter des droits fondamentaux des citoyens , des droits politiques et des droits sociaux , ainsi qu'à mon sens , des droits des minorités , et elle doit compléter les dispositions qui existent déjà à l'heure actuelle . Me référant à l'argument de la subsidiarité , je voudrais simplement vous dire qu'à mon sens , il est également important de mettre en avant l'argument de la citoyenneté . Selon moi , le plus urgent est que , dans tous les États membres , mais aussi dans les pays candidats à l'adhésion , les capacités juridiques et techniques d'inspecter les contenus d'Internet à la recherche de matériel pédopornographique soient automatisées sans tarder . Il s'agit effectivement , à mes yeux , d'un immense défi à relever . Je souhaite évoquer trois questions qui , selon moi , font défaut

Tuesday, 07 July 2015

Frequencies

enTenTen frequency query: *baked|roast|roasted fish*

Portuguese> peixe assado
English translation > baked fish?
roast fish?
roasted fish?

Which is better?

Portuguese	English translation	Frequency
baked fish	baked fish	423
roasted fish	roasted fish	101
roast fish	roast fish	20

Tuesday, 07 July 2015

Frequencies

BNC frequency query, distribution per text type: *preoccupied, worried*

Portuguese> preocupado
English translation > preoccupied
worried

Text Type	preoccupied	worried
SPOKEN	38.11	38.11
FICTION	23.34	23.34
MAGAZINE	40.83	40.83
NEWSPAPER	71.37	71.37
NON-ACAD	24.55	24.55
ACADEMIC	18.7	18.7

Tuesday, 07 July 2015

Collocations

enTenTen collocation query: *opinion*

What verbs can I use before *opinion*?

What adjectives can I use before *opinion*?



voice
express
form
differ
respect
share
sway
value
concur
bias
issue
solicit
state
divide
influence
render
reflect

unanimous
worthless
mine
contrary
irrelevant
valid
divergent
invalid
subjective
uninformed
dissenting
unpublished
welcome
admissible
non-binding
inadmissible
unfounded

Tuesday, 07 July 2015

Collocations

pTenTen collocation query: *opinião*

How about *opinião* in Portuguese?



expressar
emitir
exprimir
respeitar
partilhar
manifestar
compartilhar
manipular
expor
ouvir
externar
dividir
sensibilizar
formar
corroborar
auscultar
contrariar
mobilizar
dar
pedir
influenciar
recoher


divergente
público
alheio
formado
pessoal
unânime
sincero
contrário
favorável
diferente
desfavorável
discordante
contraditório
negativo
dominante
semelhante
abalizado
subjetivo
parecido
consensual
político
majoritário

Tuesday, 07 July 2015

Collocations: subtle differences

Portuguese > elétrico
English translation > electric or electrical

Which one?



Tuesday, 07 July 2015

Collocation: subtle differences

Electric or electrical?

electric/electrical

enTenTen [2012] freqs = 546,415 | 414,714

	electric	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	electrical
vehicle	24,381	206	7.7	2.2					
guitar	16,965	357	8.8	3.3					
battery	7,601	344	7.3	2.9					
motor	21,873	1,321	9.1	5.2					
cigarette	7,731	524	7.6	3.9					
razor	3,077	205	7.5	3.9					
shaver	3,942	310	8.0	4.7					
toothbrush	2,970	252	7.5	4.2					
scooter	4,237	401	7.8	4.7					
heater	8,125	894	8.3	5.3					
utility	5,147	648	7.2	4.4					
bill	13,952	2,355	7.2	4.6					
appliance	1,748	7,818	5.7	8.0					
impulse	442	2,738	4.5	7.8					
stimulation	411	3,815	4.4	7.8					
wiring	539	5,424	4.9	8.5					
conductivity	163	1,971	3.4	7.3					
engineer	156	6,846	1.9	7.5					
contractor	91	5,041	1.2	7.2					
engineering	127	7,802	1.6	7.7					

Tuesday, 07 July 2015

Collocation: subtle differences

Safety or security?

safety/security

enTenTen [2012] freqs = 1,358,002 | 2,014,001

	safety	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	security
occupational	2,286	0	7.4	--					
pedestrian	1,218	0	6.6	--					
road	7,149	0	6.3	--					
patient	9,358	95	8.8	1.3					
workplace	2,323	88	6.6	1.5					
fire	5,025	171	6.3	1.2					
basic	10,742	890	7.5	3.7					
public	24,897	3,485	7.8	4.9					
relative	2,223	430	6.5	3.7					
cyber	168	2,627	3.7	7.0					
economic	253	5,231	1.7	6.0					
financial	668	13,995	2.7	7.0					
airport	163	4,201	2.3	6.6					
social	683	20,382	2.6	7.4					
national	383	36,878	2.4	8.9					
homeland	27	2,263	1.3	8.4					
marketable	0	1,508	--	6.0					
border	0	2,756	--	6.4					
Treasury	0	2,076	--	6.5					
mortgage-backed	0	5,396	--	8.2					

Tuesday, 07 July 2015

Collocation: looking for equivalence I

PT - dente

EN - tooth

object_of	280,702	0.4
brush	24,312	10.74
whiten	11,589	9.88
grit	6,624	9.51
clench	3,297	8.42
grind	4,358	8.28
stain	3,742	8.14
sink	4,121	8.05
discolor	1,997	7.76
straighten	2,163	7.61
gnash	1,663	7.56
chip	1,677	7.38
bare	1,512	7.37

V obj dente_N	23,939	0.1
escovar	4,056	11.56
palitar	204	8.05
arreganhar	143	7.52
clarear	199	7.26
cerrar	171	7.09
lavar	912	6.75
branquear	94	6.7
ranger	129	6.62
trincar	87	6.6
afiar	121	6.51
descascar	128	6.46
cravar	140	6.04

Tuesday, 07 July 2015

Collocation: looking for equivalence II

Kilgariff et al. (2013)

Personagem [character]		Character	
fictício	fictitious? 8.05	main	43241 8.79
principal	principal? 8.02	cartoon	11580 8.32
jogável	511 7.89	fictional	7293 7.95
bíblico	640 7.46	favorite	11268 7.58
secundário	1109 7.39	female	8897 7.5
feminino	feminine? 7.38	lead	4154 6.89
favorito	1055 7.37	moral	4687 6.84
carismático	383 7.14	Chinese	6127 6.83
preferido	378 7.13	Disney	3783 6.7
marcante	752 6.98	central	5587 6.61
		comic	3070 6.54

Collocation: looking for equivalence III

Navigating through bilingual word sketches (Kilgariff et al. 2013)

brown				marrom				arroz_N mod ADJ				parboilizado				cozido				carolino				integral				
modifies 223516 0.3				N mod marrom-ADJ 7998 0.4				acúcar_N mod ADJ 9946 0.0				cabelo_N mod ADJ 47199 0.1				liso 2512 9.88				comprido 2375 9.85				castanho 1861 9.68				
rice	13109	9.38		linhaca	48	6.9		borra	42	6.55		mascao	1564	12.08		refinado	1347	11.34		recuperavel	139	8.4						
trout	3425	8.26		aranha	100	6.35		lapis	87	5.68																		
sugar	11161	8.21		couro	53	5.59		borrinha	12	5.58																		
bear	2754	7.11		percevejo	14	5.41		terno	51	5.28																		
hair	14175	7.02						camurça	13	5.25																		
eye	15461	6.78		casaco	55	5.23																						
dwarf	830	6.68																										
ale	842	6.47																										
pelican	545	6.23																										
envelope	918	6.08																										
leather	1733	6.04																										
spot	4512	5.96																										
bread	1383	5.66																										

Collocation: specialized vs non-specialized language

Goal

Verb collocates of *goal* in general English corpus (enTenTen)

- Accomplish
- Attain
- Reach
- Set
- Pursue
- Achieve
- Score
- Fulfill

Verb collocates of *goal* in DIY football corpus

- Score
- Disallow
- Award
- Concede



Word lists

Becoming acquainted with new terminology

mg	autorisation
vial	affections
ribavirin	medicament
olanzapine	comprimé
excipients	insuline
ritonavir	hypoglycémie
expiry	olanzapine
pharmacist	ribavirine
administration	hémoglobine
ml	excipients
insulin	ritonavir
dose	administration
authorisation	authorisation
syringe	posologie
warning	precautions
dosage	mg

EMEA English corpus compared with BNC via Sketch Engine

EMEA French corpus compared with BNC via Sketch Engine

Word lists

Becoming acquainted with new terminology

oral use

marketing authorisation for injection

medicinal products

authorisation holder

film-coated tablets

special warning

batch number

expiry date

injection site

active substance

package leaflet

pre-filled syringe

special precautions

adverse reactions

hepatic impairment

EMEA English corpus compared with BNC via Sketch Engine

From terminology to phraseology

Phraseology?

mg
vial
ribavirin
olanzapine
excipients
ritonavir
expiry
pharmacist
administration
ml
insulin
dose
authorisation
syringe
warning
dosage

EMEA English corpus compared with BNC via Sketch Engine

From terminology to phraseology

UNIVERSITY OF SURREY

Less than 2 % of the **dose** is **excreted** in the urine

no problem to **administer** this **dose** at a different time of the same day.

The maximum **tolerated dose** (MTD) of Beromun for ILP is 4 mg,
to **adjust doses** as appropriate

Ammonaps is taken in **divided doses** at meal times

You will usually **inject** two **doses** of 0.5 ml in a day.

The recommended **starting dose** is 15 mg or 30 mg once daily.
after Day 4, **reduce** next **dose** by 25 %

If you accidentally **miss** a daily **dose**, just take the next dose as normal.
take the full daily **dose prescribed** by your doctor.

You can **dial** your **dose** one unit at a time.
increase the **dose** gradually

EME English corpus

Tuesday, 07 July 2015

Consistency

UNIVERSITY OF SURREY

ipcc
INTERGOVERNMENTAL PANEL ON climate change

CLIMATE CHANGE 2013
The Physical Science Basis

- 375 MB, 1552 pages
- Two-week deadline

Tuesday, 07 July 2015

Consistency

UNIVERSITY OF SURREY

Huge source text, impossible deadline

Split translation

Tuesday, 07 July 2015

Consistency

UNIVERSITY OF SURREY

IPCC report 2013 – volume 1 – The Physical Science Basis

- Extract ST terminology
 - Aerosol
 - Tropospheric
 - Stratospheric
 - Cryosphere
 - Paleoclimate
 - Interannual variability
 - North Atlantic Oscillation (NAO)
 - Global mean surface temperature
 - Coupled Model Intercomparison Project Phase 5 (CMIP5)
- Agree on TT terminology beforehand to ensure consistency among different translators
 - Feed it to a terminology management system like SDL MultiTerm

Tuesday, 07 July 2015

Interpreting

UNIVERSITY OF SURREY

Preparing for IPCC interpreting job

- Extract ST terminology
 - Aerosol
 - Tropospheric
 - Stratospheric
 - Cryosphere
 - Paleoclimate
 - Interannual variability
 - North Atlantic Oscillation (NAO)
 - Global mean surface temperature
 - Coupled Model Intercomparison Project Phase 5 (CMIP5)
- Research TT equivalent terms
- Sight translate concordances with terms to prepare and practice (Xu 2015)

Tuesday, 07 July 2015

More

UNIVERSITY OF SURREY

Populating translation memories with existing parallel corpora

OPUS ... the open parallel corpus

- Books - A collection of translated literature (DOGC2014-07-17.tar.gz - 236 MB)
- DGT - A collection of EU Translation Memories provided by the JRC
- DOGC - Documents from the Catalan Government (DOGC2014-07-17.tar.gz - 702 MB)
- ECB - European Central Bank
- EMEA - European Medicines / GB)
- The EU bookshop corpus (EUbookshop)
- EUconst - The European constitution
- EUROPART v7 - European Parliament (8.4 GB)
- EUROPART - European Parliament (8.4 GB)
- GNOME - GNOME localization
- The Croatian - English WAC
- JRC-Acquis- legislative EU texts


Tuesday, 07 July 2015

Do translators use corpora?



Tuesday, 07 July 2015

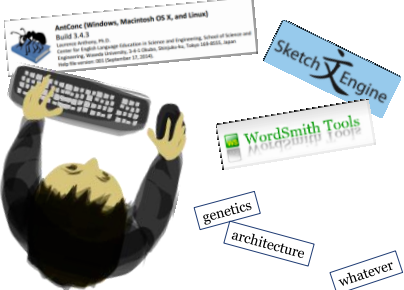
Access to corpora today



Tuesday, 07 July 2015


Corpus compilation today

Thousands of digital texts available



Tuesday, 07 July 2015

Removing barriers is not enough



Tuesday, 07 July 2015

Awareness of corpora

- Survey by Bernardini (2006) – 623 responses, mostly UK and Europe
 - 42% never heard of corpora
- Survey by Gough (2013) – 540 responses, mostly EU
 - 62% Least used technology (compared with term bases, translation memories, glossaries, web searches)
- Survey by Gallego-Hernández (2015) – 526 responses, Spain
 - 50% never or almost never used corpora

Tuesday, 07 July 2015

Do translators want training?

- No pressure from the industry
 - Many translation jobs demand use of TM systems
 - Software developers/agencies/corporate clients profit
 - Corpora not mentioned
 - No immediate gains for big stakeholders
- Why bother?
 - Translations don't get cheaper ☹
 - No obvious productivity gains ☹
 - Quality ☹
 - Reassurance ☹
 - Flexibility and autonomy ☹

Tuesday, 07 July 2015

Implementing corpora in translator education

- Corpus linguistics modules often available, but...
 - Focus on linguistics (not specific to translation)
 - Corpora & translation mostly translation studies research (especially after Baker 1993)
- Not simple to teach about corpora *for translation practice*
 - Especially in UK
 - Students translate into and out of many languages
- Corpus instructors vs translation instructors
- General imbalance regarding corpora of different languages
 - Even when available (e.g. BNC, CREA, DeReKo), interfaces and query languages differ

Tuesday, 07 July 2015 50

Corpora in translator education

- Is it worth the trouble?
- Optional corpus module *for translation practice* at Surrey
- Focus not on research, but on everyday translation
- 11 weeks, 22 hours
- 13 students
- Complicated multilingual setting

French > English	English > German	Spanish > English
Russian > English	Portuguese > English	English > Chinese
German > English	English > Greek	English > Italian
- Task-based, consciousness-raising activities (Frankenberg-Garcia 2012)

Tuesday, 07 July 2015 51

Corpora in translator education

- Guided tasks with English corpora
- Students try out similar queries using corpora of other languages


- Main corpora used
 - BYU: COCA & BNC (Davies 2004, 2008)
 - OPUS collection (Tiedeman 2012)
 - EMEA, ECB, EuroParl, OpenSubtitles
 - Sketch Engine (Kilgarriff et al. 2004, 2014)
 - Via the same interface and using same query language, access to
 - BNC, OPUS collection
 - Huge webcrawled corpora in 60 languages (Jakubíček et al. 2013)
 - DIY corpora
 - Pre-selected texts
 - Crawl the web via WebBootCaT, (Baroni et al. 2006)
 - TMX (parallel)

Tuesday, 07 July 2015 52

Reactions

Two types of data

- End-of-semester anonymous questionnaires
- End-of semester written assignments



Tuesday, 07 July 2015 53

Questionnaire

Before the MA Translation

I had never heard of corpora before my MA.	12 true 1 false
I had already used a corpus hands-on before I started my MA.	1 true 12 false


Tuesday, 07 July 2015 54

Questionnaire

Self-assessment after completion of the module

I understand the strengths and limitations of different types of corpora.	3 strongly agree 6 agree 4 neither agree nor disagree 0 disagree 0 strongly disagree
---	--


Tuesday, 07 July 2015 55

Questionnaire 

Self-assessment after completion of the module

I can carry out simple word queries to retrieve KWIC concordances.	4 strongly agree 9 agree 0 neither agree nor disagree 0 disagree 0 strongly disagree
--	---


Tuesday, 07 July 2015 56

Questionnaire 

Self-assessment after completion of the module

I can carry out queries involving more than one word.	2 strongly agree 11 agree 0 neither agree nor disagree 0 disagree 0 strongly disagree
---	--


Tuesday, 07 July 2015 57

Questionnaire 

Self-assessment after completion of the module

I understand the difference between looking up lemmas and looking up plain words.	4 strongly agree 8 agree 1 neither agree nor disagree 0 disagree 0 strongly disagree
---	---


Tuesday, 07 July 2015 58

Questionnaire 

Self-assessment after completion of the module

I can use part-of-speech tags in my queries.	2 strongly agree 8 agree 1 neither agree nor disagree 2 disagree 0 strongly disagree
--	---


Tuesday, 07 July 2015 59

Questionnaire 

Self-assessment after completion of the module

I can use corpora to retrieve information about collocation.	5 strongly agree 8 agree 0 neither agree nor disagree 0 disagree 0 strongly disagree
--	---

Tuesday, 07 July 2015 60

Questionnaire 

Self-assessment after completion of the module

I am able to compare the frequencies of different words or combinations of words within a corpus.	4 strongly agree 8 agree 1 neither agree nor disagree 0 disagree 0 strongly disagree
---	---

Tuesday, 07 July 2015 61

Questionnaire

Self-assessment after completion of the module

I am able to use normalized frequencies to compare words across different corpora or sub-corpora.	1 strongly agree 6 agree 0 neither agree nor disagree 3 disagree 1 strongly disagree
---	---

Tuesday, 07 July 2015

Questionnaire

Self-assessment after completion of the module

I am able to build a simple corpus on my own.	5 strongly agree 7 agree 0 neither agree nor disagree 0 disagree 1 strongly disagree
---	---

Tuesday, 07 July 2015

Questionnaire

Opinion about corpus output

I find concordances helpful.	8 strongly agree 5 agree 0 neither agree nor disagree 0 disagree 0 strongly disagree
I find word lists & frequencies helpful.	7 strongly agree 4 agree 2 neither agree nor disagree 0 disagree 0 strongly disagree
I find collocation queries helpful.	10 strongly agree 3 agree 0 neither agree nor disagree 0 disagree 0 strongly disagree

Tuesday, 07 July 2015

Questionnaire

Present uses of corpora outside module

I use corpora to help me when I am writing in my native language.	0 very often 2 often 6 sometimes 3 rarely 2 never
I use corpora to help me when I am writing in a language that is not my native language.	3 very often 5 often 3 sometimes 2 rarely 0 never

Tuesday, 07 July 2015

Questionnaire

Present uses of corpora outside module

I use corpora to help me with my translation assignments.	1 very often 5 often 6 sometimes 0 rarely 0 never
I use corpora for other purposes.	0 very often 2 often 3 sometimes 4 rarely 4 never

Tuesday, 07 July 2015

Questionnaire

Future uses of corpora

I am likely to look things up in corpora during my translation exams or for writing my MA dissertation.	3 strongly agree 7 agree 2 neither agree nor disagree 1 disagree 0 strongly disagree
I am likely to carry on using corpora in the future in my work as a translator.	6 strongly agree 6 agree 1 neither agree nor disagree 0 disagree 0 strongly disagree

Tuesday, 07 July 2015

Student assignments

- Graded piece of assessment about student uses of corpora in translation with examples
- 3000 word limit (excluding references)
- Corpus of 47,123 running words
 - Overall picture
- Essays read from beginning to end
 - Detailed analysis

Tuesday, 07 July 2015 68

Student assignments: corpus analysis

Search terms (lemmas)	Concept mentioned	Concept used
frequency, hit, occurrence, token	13	13
concordance	12	13
lemma	5	13
collocation, collocata, word sketch	12	12
word/frequency list	7	7
keyword list, keyness	6	5
part-of-speech, part of speech, pos, tag	8	3
relative/normali(zs)ed frequency	2	2

Tuesday, 07 July 2015 69

Student assignments: more details

Translation decisions

- English > Chinese technical translation
- Student not sure whether translation of *mRNA* should keep the English form *mRNA* or use the Chinese form 信使RNA
- Looked up frequencies zhTenTen corpus
 - mRNA - 1674
 - 信使RNA – 106
- Corpus helped her decide to use English form

Tuesday, 07 July 2015 70

Student assignments: more details

Collocation

- French > English translation of *victoire total* [total victory]
- Looked up collocates of *victory* in the BNC
- Unhelpful results
 - Labour/Conservative/great victory
- Reformulated query so as to look up synonyms of *total* in the context of *victory*
 - final/outright/complete/conclusive victory
- Immediately spotted what she considered to be best option in context of translation: *outright victory*

Tuesday, 07 July 2015 72

Student assignments: more details

Collocation

- Spanish > English business translation
- Student not sure how to translate *cuadro macroeconómico* into English
- Tried out several concordance queries in enTenTen
 - Macroeconomic picture – 67
 - Macroeconomic projection – 35
 - Macroeconomic prediction – 9
- Chose to use literal translation but not too happy

Tuesday, 07 July 2015 73

Student assignments: more details

Collocation

Had student tried a collocation query to identify which nouns follow *macroeconomic*...

Source: enTenTen corpus

policy
fluctuation
equilibrium
forecast
framework
slowdown
disequilibrium
dynamicsyou
equilibria
conditionality
cauldron
coordination
policy-making
environment
indices
shock
backdrop

Tuesday, 07 July 2015 74

Student assignments: more details

Terminology

- German > English business translation
- Compiled a small corpus of different types of companies to become more familiar with terminology in this area
- Used her DIY corpus to look up terms with *liability* and came up with
 - *joint liability*
 - *non-current liabilities*
 - *interest-bearing liabilities, etc.*
- Added terms to her glossary of business terminology

Tuesday, 07 July 2015

Student assignments: more details

Translation decisions

- Russian > English translation of short story
 - *Ира - это не девушка - а мальчик*
 - [*Ira - not a girl - but a boy*]
- ruTenTen: *Ira* very common and always a female
- BNC: *Ira* used mostly for *Irish Republican Army*, occasionally for male name
- COCA: *Ira* usually a man's name
- Decided to make the translation more explicit for English readers
 - *Ira was not, as the name seemed to imply, a girl, but a boy.*

Tuesday, 07 July 2015

Student assignments: more details

Opinions about parallel corpora

“Of all the types of corpora available, parallel are undoubtedly the easiest for translators to draw conclusions from because the necessary information can be accessed immediately and terms can be directly compared to their equivalents in another language”

“The parallel corpus often produced few results.”

Only a very small part of what people in general say or write ever gets to be translated, which seriously limits the number and types of texts available for the compilation of parallel corpora. Indeed, this is one of the main reasons why parallel corpora are usually much smaller in scale than monolingual corpora. Frankenberg-Garcia (2009: 60)

Tuesday, 07 July 2015

Student assignments: more details

Opinions about DIY corpora

“Although my corpus was put together in only a matter of minutes, it still allowed me to study terminology and phraseology related to astronomy in a reasonable amount of depth”

“I find that compiling corpora is more suitable for researchers, linguists and teachers, rather than translators and interpreters.”

Tuesday, 07 July 2015

Student assignments: more details

Opinions about learning to use corpora

“The translator spend a huge amount of time familiarise him or her with the tool and then spend extra effort on mastering the code and tag language these things, but he or she may never use some of the functionalities in a corpus”

“Overall, it has been a useful resource but has been limited by my relative inexperience of applying the available functions and occasional searches taking too long”

“the use of corpora [...] takes some time to get used to but has proved to be a good resource for translation practice”.

Tuesday, 07 July 2015

Student assignments: more details

Opinions about using corpora

“One thing that can make using corpora time-consuming is that once concordances are begun, in my experience, I can find myself looking further and often find interesting things out that I wasn't looking for in the first place, which isn't necessarily a negative observation.”

“Compared to dictionaries, they [corpora] offer translators with extensive genuine examples in various contexts, thus can be a powerful complementary tool for understanding the usage of language. However, it is also noted that translators should be careful with their own interpretations for data presented by corpora and examine the reliability of some examples in corpora before making further analysis.”

“Although corpus is highly informative, it is no substitute for other authoritative resources like dictionaries. A better solution would be to combine them both and utilise the advantages of both.”

Tuesday, 07 July 2015

Student assignments: more details



Opinions about the usefulness of corpora

“my comparable [DIY] corpora saved me time and effort.”

“unexpected insights on the native language [...] a precious resource especially in regards with working into a non-native language, in this case English, during the writing of essays.”

“Producing an authentic-sounding TT is, however, especially difficult when you are working out of your native language and I therefore found corpora to be especially useful when translating a text about an Aztec artefact from English into my non-native language, German”

Tuesday, 07 July 2015

83

Conclusion



- Students not power users of corpora
 - Some much better than others
- Could do with a lot more training
 - POS queries
 - Normalized frequencies
 - Extracting terminology
 - Researching phraseology
- But were able to successfully carry out many corpus queries for which online dictionaries, glossaries and web searches and other more conventional resources wouldn't have provided satisfactory answers
- As with any new technology, it is likely that the more they use corpora the better they will be able to use them

Tuesday, 07 July 2015

83

More details



Frankenberg-Garcia, A. (forthcoming, 2015) Training translators to use corpora hands-on: challenges and reactions by a group of 13 students at a UK university. *Corpora*, 10/2.

Tuesday, 07 July 2015

83

References



- Aston, G. (1999) 'Corpus use and learning to translate'. *Textus*, 12, pp. 289-314.
- Aston, G. (2009) Foreword. In Beeby, A., Rodríguez, P. & Sánchez-Gijón, P. (eds.), ix-x.
- Baker, M. (1993), 'Corpus linguistics and translation studies. Implications and applications', in M. Baker, G. Francis and E. Tognini-Bonelli (eds) *Text and Technology: In Honour of John Sinclair*. Amsterdam and Philadelphia: John Benjamins, 233-250.
- Baroni, M., Kilgariff, A., Pomikalek, J. & Rychly, P. (2006) 'WebBootCaT: Instant domain-specific corpora to support human translators'. *Proceedings of EAMT-2006*, pp. 247-252.
- Beeby, A., Rodríguez, P. & Sánchez-Gijón, P. (2009) (eds.) *Corpus use and translating*. Amsterdam and Philadelphia: John Benjamins.
- Bernardini, S. (2006) *Corpora for Translation Education and Translation Practice: Achievements and Challenges*. *Proceedings of the Third International Workshop on Language Resources for Translation Work, Research & Training (LR4Trans-III)* Available online at: <http://www.ifi.unizh.ch/cl/yuste/LR4Trans-III/materials/silvia.pdf>
- Bowker, L. and Pearson, J. (2002) *Working with Specialized Language: a practical guide to using corpora*. London: Routledge.

84

References



- Davies, M. 2004. *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). Available online at <http://corpus.byu.edu/bnc/>.
- Davies, M. 2008. *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at <http://corpus.byu.edu/coca/>.
- Frankenberg-Garcia, A. 2009. 'Compiling and Using a Parallel Corpus for Research in Translation'. *International Journal of Translation*, XXI, 1, pp 57-71.
- Frankenberg-Garcia, A. (2012) *Raising Teacher's awareness of corpora*. *Language Teaching*, vol 45, 4, pp 475-489.
- Gallego-Hernández, D. (2015) 'The use of Corpora as translation resources: a study based on a survey of Spanish professional translators'. *Perspectives: Studies in Translatology*, DOI 10.1080/0907676X.2014.964269.
- Gough, J. (2013) *Survey of professional translators' use of on-line resources for terminology research*. Unpublished interim PhD report, September 2013, University of Surrey.
- Jakubiček, M., Kilgariff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013) *The TenTen corpus family*. Paper presented at 7th International Corpus Linguistics Conference, Lancaster, July 2013.
- Kilgariff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014) *The Sketch Engine: ten years on*. *Lexicography*, 1/1, 7-36.

83

References



- Kilgariff, A., P. Rychlý, P. Smrz and D. Tugwell (2004) *The Sketch Engine*. *Proceedings of Euralex*. Lorient, France. Kubler, N. (2011) *Working with Corpora for Translation Teaching in a French-speaking setting*. In Frankenberg-Garcia, A., Flowerdew, L. and Aston G. (eds) *New Trends in Corpora and Language Learning*. London: Continuum, 62-80.
- Rodríguez-Inés, P. (2010) 'Electronic Corpora and Other ICT (Information and communication technologies) tools: an integrated approach to translation teaching'. *The Interpreter and Translator Trainer*, 4, 2, pp. 251-282.
- Tiedemann, J. (2012) *Parallel Data, Tools and Interfaces in OPUS*. [pdf] In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012)*, 2214-2218.
- Xu, Ran (2015) *Using comparable corpora for simultaneous interpreting preparation*. Paper presented at the IV International Conference on Corpus Use and Learning to Translate (CULT), Alicante, 27-29 May 2015.
- Zanettin, F., Bernardini, S. & Stewart, D. (2003) (eds.) *Corpora in Translator Education*. Manchester: St. Jerome.
- Zanettin, F. (2012) *Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*. Manchester: St. Jerome.

84

Advantages and Challenges of Using Corpora in Translation Practice

UNIVERSITY OF SURREY

ernal and external audits. </p><p> Thank you for your attention and cooperation with these audits. KPMG
</p><p> Job seekers (of all ages). </p><p> Thank you for your attention and for being part of the GoliathJobs fam
</p><p> 00% More Jackoff Magic blog, and I thank you for your attention . Sylvia, I have the grave misfortune of
</p><p> he is valued by her clients. </p><p> Thank you for your attention ! </p><p> Gail S. Carmel, IN </p><p> Dee
</p><p> eared for our global responsibility. Thank you for your attention to this urgent matter." </p><p> Since ope
</p><p> and medicine that works. </p><p> Thank you for your attention . </p><p> Cassie Travaini, RSHom, CCH V
</p><p> her standard for trapping. </p><p> Thank you for your attention in this matter and it really needs attentio
</p><p> ndon its decades old blackout rules. Thank you for your attention to this matter. </p><p> Congresswoman
</p><p> e would be most welcome. </p><p> Thank you for your attention to this vital issue, Joan, Nathalie and
</p><p> umstances. </p><p> I would like to thank you for your attention to this critical issue and I welcome your
</p><p> ffect the length of a year? </p><p> Thank you for your attention and any information will be greatly appre
</p><p> comes. Itâ€™s that simple. </p><p> Thank you for your attention to detail and great writing style. Your
</p><p> s in magazines.... [groan] </p><p> Thank you for your attention . [stepping off the (unscented) soap box
</p><p> listens to the customers Innovative Thank you for your attention Customer's Interview Customers' Intervie
</p><p> h, like his candidacy, is refreshing. Thank you for your attention to this matter. </p><p> Why does lextig
</p><p> for believing in the army and navy. Thank you for your attention , respectfully, Laredo Girl ... ZZZZ." The
</p><p> WASHINGTON I feel myself going. I thank you for your attention . You had better not take any more trout
</p><p> next year's budget process. </p><p> Thank you for your attention to these issues. I look forward to your
</p><p> s. </p><p> It's been a great ride, I thank you for your attention and interest. The next few years should
</p><p> e in the meantime, </p><p> While thanking you for your attention , I remain, </p><p> Sincerely yours, </p></p>

To Adam Kilgarriff

UNIVERSITY OF SURREY

12 Feb 1960—16 May 2015



Tuesday, 07 July 2015