Research Group in Computational Linguistics

Parallel Corpora

processing, building and using in research

Maria Kunilovskaya

7LN002/UM1: Corpus Linguistics with R

Wolverhampton, April 01, 2020

Survey results (out of 14 respondents)

1. Which step in corpus research seems most challenging to you?
   - corpus design and pre-processing: 6
2. Can you create a sentence-aligned corpus from plain text files?
   - I want to know more about .. parallel corpus building: 9
3. Do you use (or have you used) all or any of the following:
   - UDpipe and/or TreeTagger: 0
   - Sketch Engine for building corpora: 1
4. Do you know how to:
   - use shell commands to ... ?: 1
   - handle compressed files without decompressing them: 0
5. What is your operating system?
   - Windows: 10
6. Which file formats have you worked with?
   - csv/tsv 6; TMX/XML 2; plain text 10; conllu 0
7. Do you plan to use multilingual data in your thesis research?
   - I am not sure,... I am exploring my options 10

Planning and Outcomes

This session: Theoretical + Practical parts

1. Understanding of available parallel corpora: Are they right for your (research) task
2. Corpus building skills: How-tos and know-hows (inc. crawling, annotation and alignment)
3. Skills in extracting descriptive statistics and frequencies for your targeted items

(based on adjustable python3 scripts)

Clone or download the GitHub repository parcorp to try out the practical tasks (included in repo).

All links in this presentation are clickable!

Outline

1. Preliminaries
2. Use available: Overview of resources
   - Institutional and domain-specific
   - Web-crawled
   - Dedicated projects
3. Build your own: Design and pre-processing
   - Parallel texts: DIY
     - scraping a multilingual website
     - annotation: parsing
   - Translation Memories
     - alignment
4. Extraction from parallel corpora
   - research design
   - their tools
   - take control

"The beginning of any corpus study is the creation of the corpus itself ...
The results are only as good as the corpus"

(Sinclair, 1991)

Major uses and typical research questions

1. Training/evaluating machine translation models since 1980s (Nagao, 1984; Brown et al., 1990)
   - official WMT19 training for EN<>RU: 38M token parallel, and 10M monolingual
2. Contrastive analysis, bilingual lexicography, grammar induction
   - reveal typological differences between languages (Song, 2017)
   - comparative discourse analysis, ex. use of DM, preferred ways of info packaging (Fabricius-Hansen et al., 2005)
   - NLP: word sense disambiguation
3. Translation studies
   - detecting patterns of (inevitable) modifications in translation (Baker, 1993)
     - Are phrasal verbs less numerous in FR > EN than in Ge > EN translations? (Cappelle and Loock, 2017)
   - studies of shifts (Čulo et al., 2017), strategies (Carl and Buch-Kromann, 2010), techniques, solutions
   - reference in translation teaching and in professional translation (Zanettin et al., 2014) + Reverso, Linguee

Issues and formats

Parallel corpora – collections of originals and their translations

'the exploitation ... remains a bottleneck'

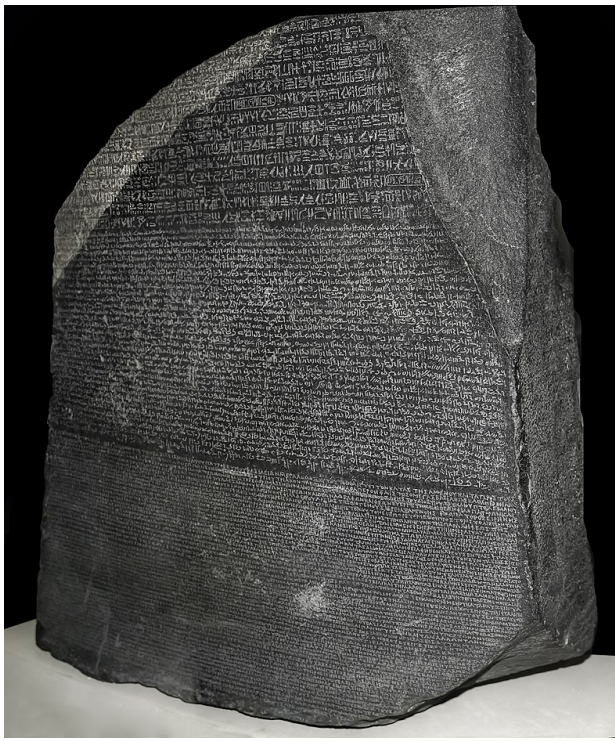
- query tools?
- best database formats?
- quality of the information extraction?
- corpus quality (metadata, balance, noise)?

(Hansen-Schirra et al., 2017)

Availability: download or search online?

if downloadable:

- alignment level: doc, sentence, phrase
- alignment quality: auto, corrected
- plain text, XML (inc. TMX, XLIFF)



Rosetta Stone: same decree in two Ancient Egyptian scripts and Ancient Greek

1. Preliminaries
2. Use available: Overview of resources
   - Institutional and domain-specific
   - Web-crawled
   - Dedicated projects
3. Build your own: Design and pre-processing
   - Parallel texts: DIY
     - scraping a multilingual website
     - annotation: parsing
   - Translation Memories
     - alignment
4. Extraction from parallel corpora
   - research design
   - their tools
   - take control

(1) Institutional, domain-specific and multilingual

1. UN corpus:
   official records and other parliamentary documents 1990-2014;
   identical trees of folders of XML for SL and TL (2.5GB);
   + ? plain-text bitexts with ids and auto-alignment info?;
   available in SketchEngine (SkE) as part of OPUS 2
2. EuroParl:
   a corpus of parliamentary debates, 21 languages with ca. 40M tokens per language, 1996-2011 (Koehn, 2005); tsv+txt
   - available in SkE
   - tailored for translationese research (Karakanta et al., 2018)
3. News Commentary corpus:
   political and economic commentary crawled from Project Syndicate website
   - current version 15 is available as tsv (en-ru: 319,242 sentence pairs) and unaligned blank-line separated documents in one plain text file (ru: 8,968 docs)

NB! where possible use data from the latest release by Workshop on Machine Translation (WMT)

(2) Web-crawled, usually for a language pair

1. Parallel Corpora for European Languages, v6:
   paracrawl, 24 EU langs + Russian
   - building tools (crawler, parallel sentences evaluator, Bicleaner, converter to tsv)
   - cleaned data: "... TMX format or plain TXT format"
   - concatenated plain text for each language, sentence-aligned
2. OPUS project:
   40 languages, "translated texts from the web"; 18 subcorpora (inc. UN, TedTalks, Europarl, Subtitles) (Tiedemann, 2012)
3. Yandex corpus:
   random 1M sentence pairs from a web-crawled corpus; txt, sentence-aligned

(3) Translation varieties and national corpora

NB! richly annotated, with metadata, higher quality, but available through web interface

1. Student and professional translations:
   - MeLLANGE Learner Translator Corpus: multilingual, error-annotated, 300 text pairs:
   - Multilingual Student Translation (MUST): work in progress since 2016
   - RusLTC: EN-RU, mostly mass media (> 500 English sources), multi-parallel, text-level quality annotation and error annotation, TMX and plain text, own search interface
   - VarTRA: EN-DE, CroCo extension, 110 text pairs in 8 registers, plain text, CQPweb
2. Parallel subcorpus of Russian National Corpus and Czech National Corpus
3. Mozilla transvision application localization corpus: TMX, very short sentences, EN-RU: 17528 sentence pairs > 2 tokens

Characteristics: What do you get?

What you want to know about the provided corpus:

1. size in text pairs, sentence pairs, total tokens ?
2. who translated? assumed translation quality?
3. what types of texts are included? any register/sample size balance?
4. sentence alignment?
5. document segmentation?
6. data format: plain text, TMX, XML
7. how is the metadata stored?
8. language pair? multilingual? directionality

Is your research question original enough to be applied to a well-known resource?

Are you looking to improve state-of-the-art (SOTA)?

1. Preliminaries
2. Use available: Overview of resources
   - Institutional and domain-specific
   - Web-crawled
   - Dedicated projects
3. Build your own: Design and pre-processing
   - Parallel texts: DIY
     - scraping a multilingual website
     - annotation: parsing
   - Translation Memories
     - alignment
4. Extraction from parallel corpora
   - research design
   - their tools
   - take control

If text-level alignment is enough and you have a (web) source

1. extract plain text from (your) TM
2. crawl a website with parallel data, using links heuristics or in-text indicators like 'Translated by ...' and 'See original text' (written in TL)

Examples of paired links patterns

- `https://www.ted.com/talks/mary_roach_10_things_you_didn_t_know_about_orgasm/transcript?language=en -- https://www.ted.com/talks/mary_roach_10_things_you_didn_t_know_about_orgasm/transcript?language=ru`
- `http://www.bbc.com/travel/story/20151026-the-sleepy-village-that-stopped-the-black-death -- https://www.bbc.com/russian/uk/2015/11/151110_vert_tra_village_that_stopped_black_death`

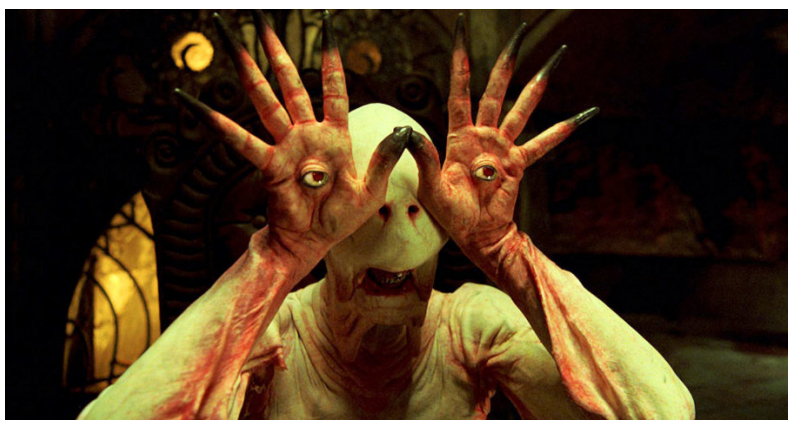
3. copy-paste, scan, OCR, convert from pdf. Make sure you can explain the how the data was harvested: how consistent and reproducible is your selection?

Designing and structuring a corpus

- folder names and structure (e.g. raw/, clean/en/, clean/ru/, parsed/, lempos/)
- version management
- filenames conventions (e.g. en_7.txt - ru_7.txt)
- corpus format: raw text files, one-word-per-line (conllu, ttagged), lemmatised, lemmatised and PoS-tagged, one-sentence-per-line, one-text-per-line
- sample-size balance, representativeness
- the same approach to filtering, spelling standardization, annotation

1. corpus stats and description (as sanity measure)
2. inspect the texts manually:

see code for scraping parallel texts from BBC



from Pan's Labyrinth by Guillermo del Toro

Annotation: "trust the text" (Sinclair, 2004)?

You cannot avoid annotation, if you want

- to go beyond string matching (even that is tricky without lemmatisation with the small data in a morphologically-rich language)
- to explore lexico-grammatic patterns like "translational correspondences for passive voice or imperfective aspect forms or "light verb + deverbal" constructions

Most existing resources are enriched with

- linguistic annotation (PoS, grammar categories, syntactic dependencies, semantic/discourse info),
- manual mark-up (translation errors, translation shifts)[1]
- meta-data that can reflect human judgments about items in the corpus (ex. genre, quality of translation).

[1] beyond the scope of this class

Automatic multilingual annotation

Typical tasks in the pipeline applied to 'clean text':

- word tokenisation
- sentence boundaries
- lemmas + PoS + MS tags
- syntactic relations

Universal Dependencies (Straka and Straková, 2017)

- install UDpipe
- select a model (ex. *english-ewt-ud-2.3-181115.udpipe*)
- decide on settings: spelling unification, filtering, handling Numerals, punctuation, function words

Tree Tagger (Schmid, 2013)

- why use it?
- avoid installation: use SkE corpus builder and download it as *.vert (see tagset); not for bitext format!
- install with Windows GUI for painful file-by-file tagging
- enjoy step-by-step handling

see *parsing* and *build_your_own* folders in parcorp repo, including to produce lemmatised or PoS versions of your corpus

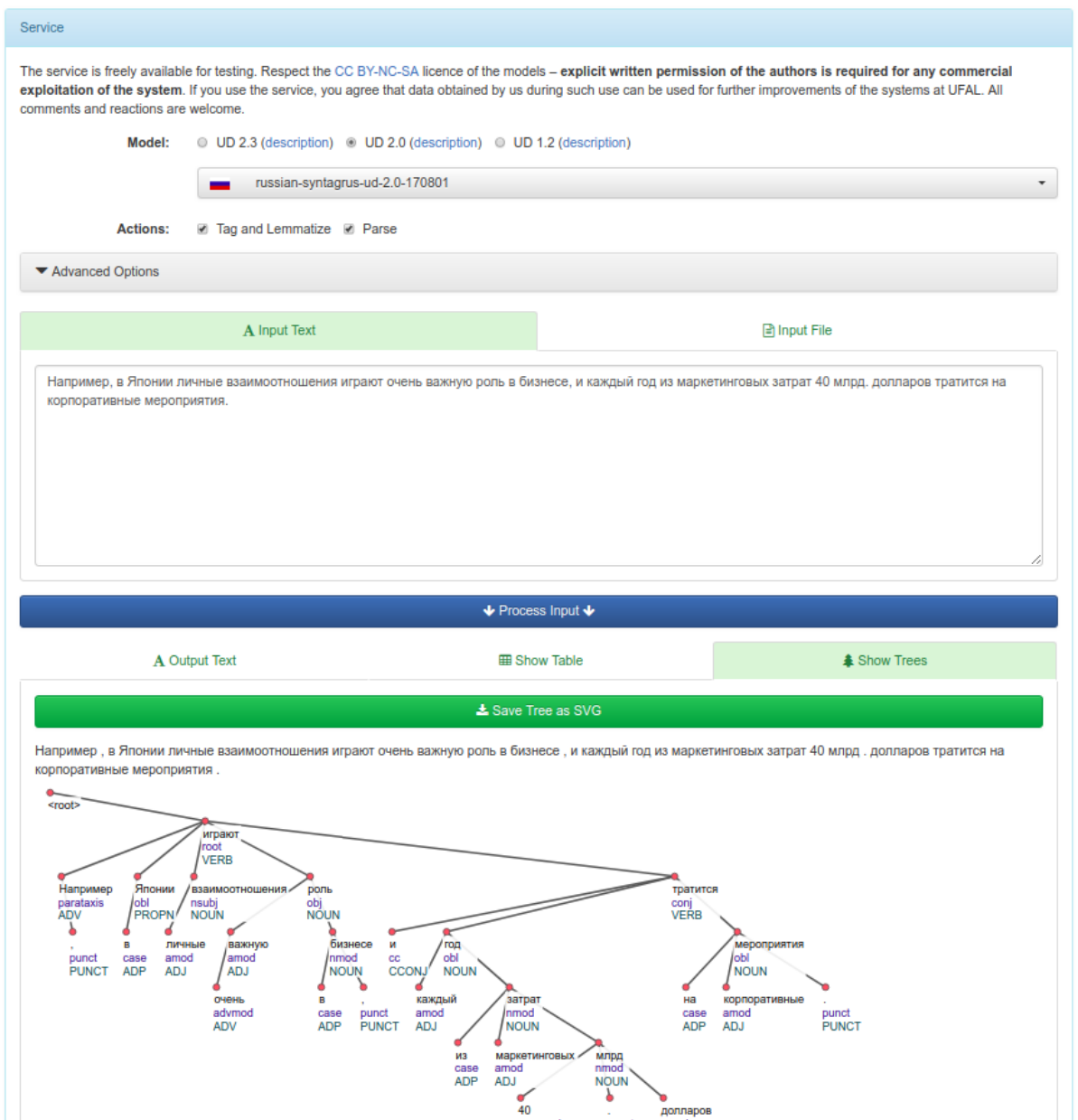
.conllu format

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
# newdoc									
# newpar									
# sent_id = 1									
# text = He was a middle-aged child that had never shed its baby fat, though some gifted tailor had almost succeeded in camouflaging his plump and spankable bottom.									
1	He	he	PRON	PRP	Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs	7	nsubj	_	_
2	was	be	AUX	VBD	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	7	cop	_	_
3	a	a	DET	DT	Definite=Ind|PronType=Art	7	det	_	_
4	middle	middle	ADJ	NN	Number=Sing	6	compound	_	SpaceAfter=No
5	-	-	PUNCT	HYPH	_	6	punct	_	SpaceAfter=No
6	aged	aged	ADJ	JJ	Degree=Pos	7	amod	_	_
7	child	child	NOUN	NN	Number=Sing	0	root	_	_
8	that	that	PRON	WDT	PronType=Rel	11	nsubj	_	_
9	had	have	AUX	VBD	Mood=Ind|Tense=Past|VerbForm=Fin	11	aux	_	_
10	never	never	ADV	RB	_	11	advmod	_	_
11	shed	sh	VERB	VBN	Tense=Past|VerbForm=Part	7	acl:relcl	_	_
12	it	it	PRON	PRP	Case=Nom|Gender=Neut|Number=Sing|Person=3|PronType=Prs	15	nmod:poss	_	SpaceAfter=No
13	s	be	PART	POS	_	12	case	_	_
14	baby	baby	NOUN	NN	Number=Sing	15	compound	_	_
15	fat	fat	NOUN	NN	Number=Sing	11	obj	_	SpaceAfter=No
16	,	,	PUNCT	,	_	7	punct	_	_
17	though	though	SCONJ	IN	_	23	mark	_	_
18	some	some	DET	DT	_	20	det	_	_
19	gifted	gift	VERB	VBN	Tense=Past|VerbForm=Part	20	amod	_	_
20	tailor	tailor	NOUN	NN	Number=Sing	23	nsubj	_	_
21	had	have	AUX	VBD	Mood=Ind|Tense=Past|VerbForm=Fin	23	aux	_	_
22	almost	almost	ADV	RB	_	23	advmod	_	_
23	succeeded	succeed	VERB	VBN	Tense=Past|VerbForm=Part	7	advcl	_	_
24	in	in	SCONJ	IN		25	mark		SpacesAfter=\n

tab-separated fields:
own id, token, lemma, PoS
morph features, head id, relation

On-line services for parsing and visualizing syntactic trees

UDpipe



Finally, the quality of automatic annotation

see quality evaluation report for the latest versions

Treebank	Sents	UPOS	XPOS	Feats	Lemma	UAS	LAS
English-EWT	79.8	93.6	92.9	94.4	96.1	80.8	77.6
German-GSD 2.3	81.8	91.4	80.4	63.3	95.4	77.0	71.7
German-GSD 2.2	**95.6**	**91.6**	79.4	71.1	**96.0**	**80.2**	**75.5**
German 2.0	79.3	90.7	**94.7**	**80.5**	95.4	74.0	68.6

excerpts from performance tables for several models (English and German)

Truly parallel: Sentence Alignment

A recent overview on translation technology confirms the choice (Bruckner, 2019)

LFAligner (batch mode)

1. download + install *perl*
2. adjust languages in *LF_aligner_setup.txt*
3. create shell script that calls *perl* for every text pair
4. concatenate all TMXs
5. delete TMX headers, except the first one
6. use Notepad++ for (3) and (5)

Hunalign statistic algorithms

- Correlation in sent length (Gale and Church, 1991)
- Dynamic vocabularies based on co-occurrence stats
- Iteration from sentence level to word level

Other approaches:

- shared numbers
- orthographic cognates
- number and order of PoS (Damerau-Levenshtein)

TMX editing

Heartsome TMX Editor 8

1. download from Developer's github
   (use *All installer* link; select the correct version!)
2. extract the archive
3. make *Heartsome TMX Editor* file executable (Properties > Permissions > Allow executing file as a program)
4. setup the correct Java (it did not run on the current Java 11.0.6, I fell back for java 1.8 (>= 1.6 is required)
5. Video tutorial

Alternative TMX Editors

- native LF Aligner GUI (manually correct each text pair)
- Olifant
- no-name TMX-editor (depends on *LF_aligner_4.05_win.zip*)

Research design and corpora

Old wisdoms:

- "corpus-linguistic analyses are always based on the evaluation of some kind of frequencies" (Gries, 2009)
  hmmm, number crunching?
- discovery potential:
  "Corpus Linguistics produces not only new facts but facts of unsuspected kind" (Stubbs, 1996) (both in corpus-based and corpus-driven approaches)
- Formulate a research question: paraphrase linguistic hypotheses into frequencies
- Is the attested language text (relevant to your RQ) 'duly recorded' (Firth 1957)?

An example

distinctive-collexeme analysis (Gries and Stefanowitsch, 2004)

What explains the speaker's choice of alternating pairs?

- the dative 'alternation':
  John sent Mary the book / John sent a book to Mary
- particle 'movement':
  John picked up the book / John picked the book up
- the s-genitive vs. the of-construction:
  the university's budget / the budget of the university

RQ: some lexemes are semantically more aligned with the meaning of the construction and tend to occur in one and not the other
Result: certain nouns/verbs do have bias to certain constructions

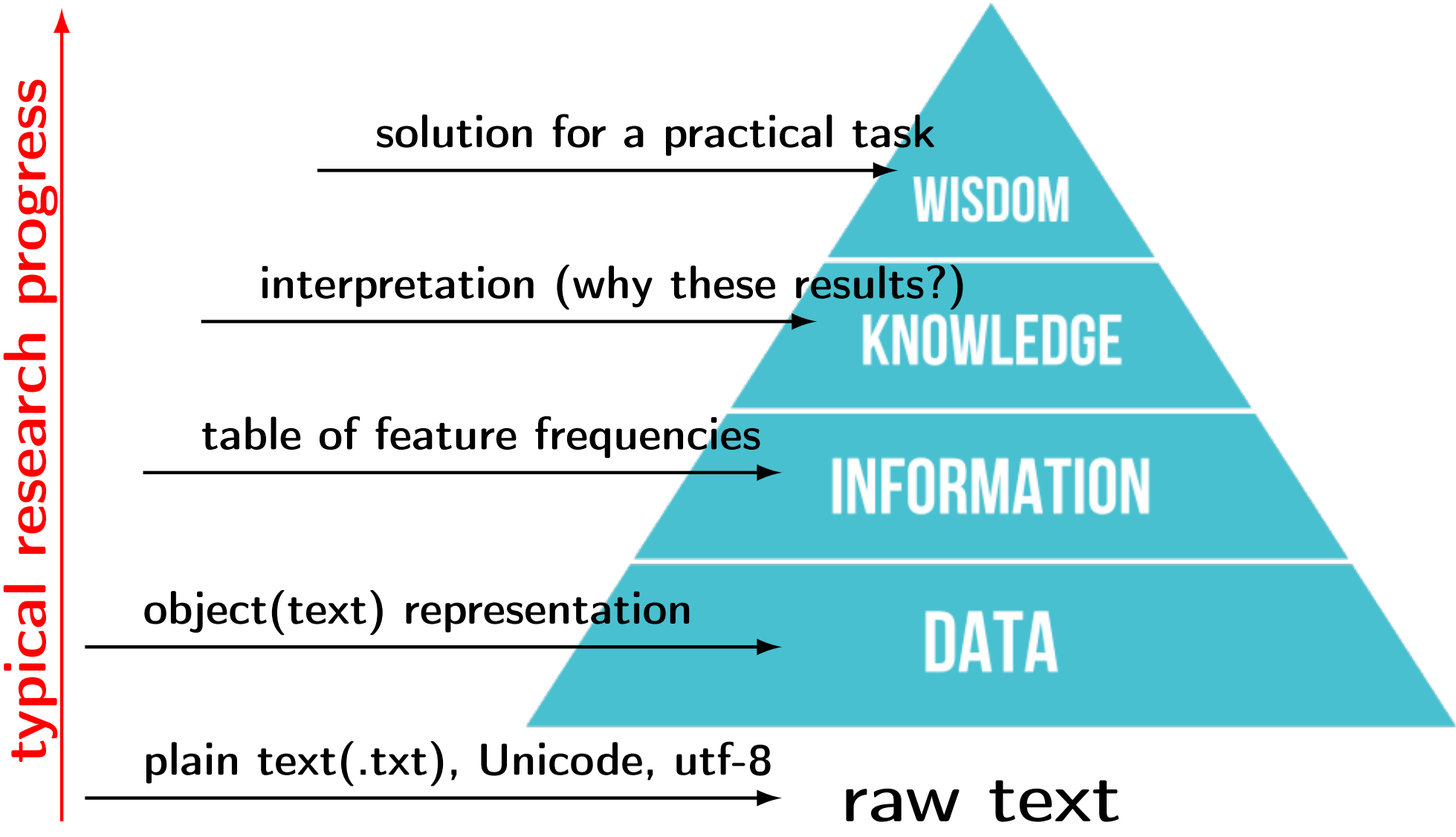
e.g. there is one significantly distinctive collexeme for each construction: make for [try to V] and get for [try and V]

Steps in a corpus-based study

1. Set the task, put forward a hypothesis (based on previous knowledge, have expectations!)
2. Plan the corpus structure
3. Collect/build the raw corpus
4. Pre-processing
5. Think of data representation and info extraction tools
6. Add mark-up and annotation if necessary
7. Extract and arrange the quantitative info
8. Interpret the results: support or reject the hypothesis

In practice the steps are usually recursive!

Information hierarchy in corpus research



Lexical translation solutions and bilingual correspondences

(Re)search ideas based on pre-defined lists:

- language-specific items (lexical lacoons)
  – inferiority, vendor lock-in, overqualified, engaging: RU ???
  – пошлость (cheap, vulgar?), лебезить, канючить: EN ???
- 'false friends'
- differences in lexicalisation of concepts (based on linguistic relativity hypothesis)
- discourse markers (differences in cohesion)
  DM overused in translation: where do they come from? Is there always a ST trigger?
- frequency calques (e.g. modal predicates)

NB! the study of translational properties per se require the comparison with a corpus of non-translations!

Getting frequencies from bitext formats: Overview

Sketch Engine (with a license)

- an easy way to tag a corpus
- intuitive (5 min video)
- use your own data
- stay in touch with your data
- learn CQL for complex searches
- manual collection/arranging of frequencies for analysis

Dedicated interfaces

- learn the functionality
- be content with the corpus behind

AntPConc

- use your own data
- work off-line
- same as ParaConc, but free
- can't use TMX directly
- manual stats collection

TMX editors

- use bilingual search...
- but why?

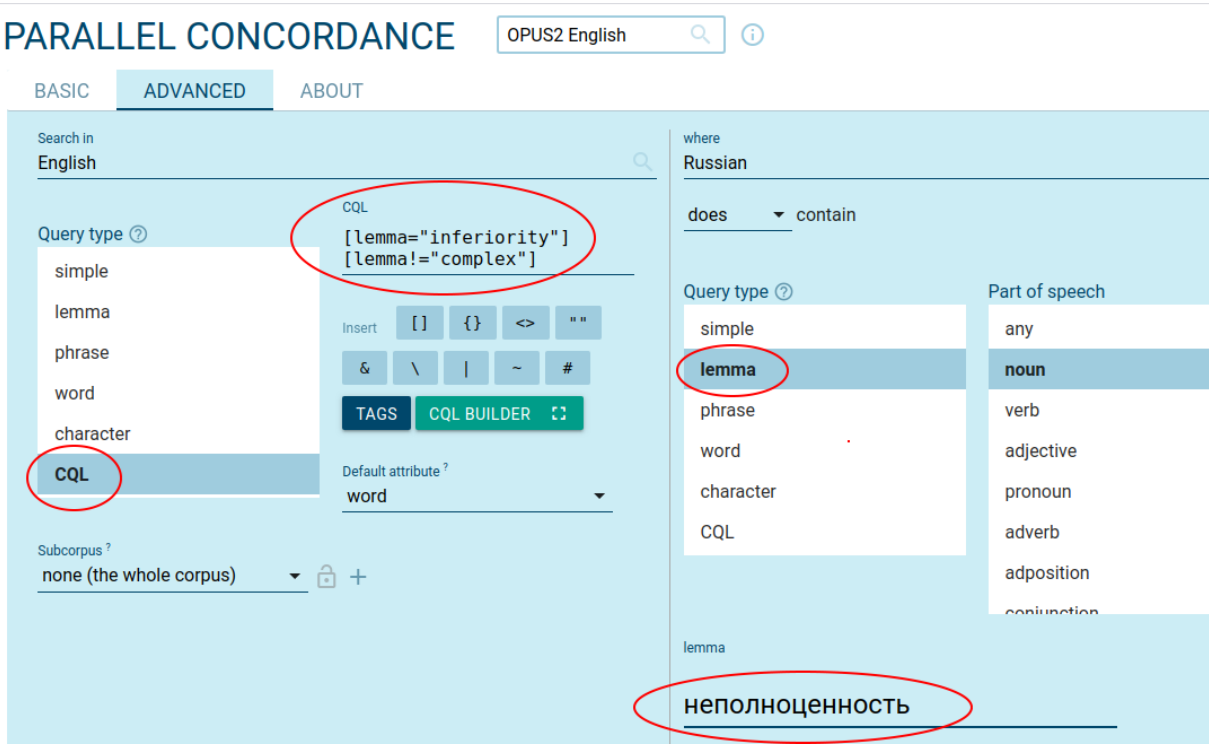
own scripts

- get stats arranged in one go
- need some scripting skills

Sketch Engine

Available parallel corpora:

- DGT
- Europarl
- EUR-lex
- OPUS2
  - UN corpus
  - TedTalks
  - Europarl
  - Subtitles
- Bible and Quran



NOT happy?

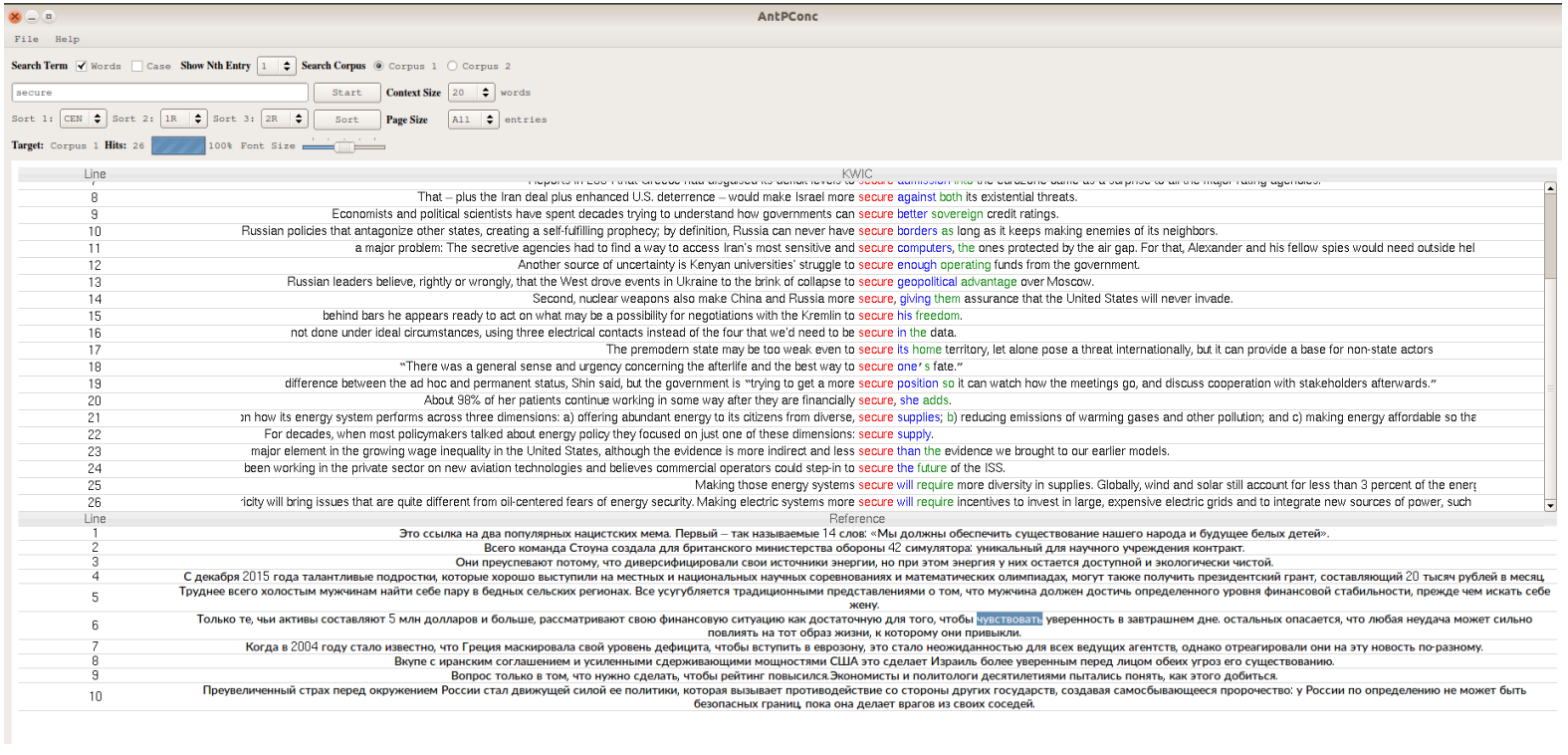
Upload your own TMX!

Make sure you have a very focused and well-motivated research question, preferably about a small number of items.

How do you calculate dispersion?

AntPConc

- extract parallel texts from your manually corrected bitext format to separate pairs of plain text files for ST and TT: my_corrected.tmx -> en_1.txt, en_2.txt, ru_1.txt, ru_2.txt
- import to AntPConc (Anthony, 2017)
- Anthony's Plans for 2018: Add more analysis tools, such as those in AntConc



own scripts for aligned plain text docs

pre-processing

- sentence tokenisation with NLTK as a stand alone task (*en_tokenise_sentences.py*)
- parsing a multilingual tree of folders with UDpipe (*raw_multi-ling-tree2txt2conllu2lempos.py*)
- converting to lemmas format (or lempos, or PoS)
- filtering (advanced: based on texts comparability)
- controlling sample size for no-docs-separation downloads
- fixing faulty linebreaks

extraction

- detect outliers, count TTR, get freqs for listed items (item stats for a corpus or corpus stats for an item), extract UD deps
- (advanced) extract frequencies of translationese indicators)

own scripts for processing TMX

```
−<tu creationdate="20180218T130337Z" creationid="n">
  <prop type="Txt::Note">en_193-ru_193</prop>
 −<tuv xml:lang="EN">
  −<seg>
     Despite its late entry and lack of major energy corporations, South Korea can harness its edge in shipbuilding and
   </seg>
  </tuv>
 −<tuv xml:lang="RU">
  −<seg>
     «Несмотря на позднее вступление в гонку и отсутствие крупных энергетических корпораций, Южная Корея
     одной из неизведанных границ мира», - говорят эксперты.
   </seg>
  </tuv>
</tu>
```

There is no way of getting around word alignment, which is theoretically problematic for automation

bitext pre-processing

- *tag TMX with UD
- *reduce TMX to a subcorpus (on a list of filename pairs)

bitext extraction

- extract sentence pairs with TT strings for listed ST strings (build a parallel concordance)

Proceed to statistical analysis and interpretation

The next step in this sort of research is statistical analysis.

Usually you want to extract frequencies to a spreadsheet (*.tsv) to perform tests in Excel or programmatically.

Data wrangling: 'long' and 'wide' tables

See some solutions for basic statistical tests applied to tabular data in *stats* folder in parcorp GitHub repository

- t-test
- Wilcoxon's matched pairs test
- Spearman rank correlation

stats folder has scripts for both extracting frequency counts and for performing stats tests

References I

Anthony, L. (2017). AntPConc (Version 1.2.1) [Computer Software]. Tokyo, Japan: Waseda University.

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In Text and Technology: In honour of John Sinclair, pages 232–250. J. Benjamins, Amsterdam.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. 16(2):79–85.

References II

Bruckner, C. (2019). Localization Engineering für Übersetzer – mit Freeware- und OpenSource-Helferlein. In Übersetzen und Dolmetschen 4.0 – Neue Wege im digitalen Zeitalter, pages 1–8.

Cappelle, B. and Loock, R. (2017). Typological differences shining through : The case of phrasal verbs in translated English. Empirical Translation Studies. New Theoretical and Methodological Traditions, pages 235–264.

Carl, M. and Buch-Kromann, M. (2010). Correlating translation product and translation process data of professional and student translators. 14 Annual Conference of the European Association for Machine Translation, Saint-Raphaël, France, (May).

References III

Čulo, O., Hansen-Schirra, S., Maksymski, K., and Neumann, S. (2017). Empty links and crossing lines: querying multi-layer annotation and alignment in parallel corpora. Annotation, exploitation and evaluation of parallel corpora, page 53.

Fabricius-Hansen, C., Ramm, W., Solfjeld, I. K., and Behrens, B. (2005). Coordination, discourse relations, and information packaging – cross-linguistic differences. Proceedings of the Symposium on the Exploration and Modelling of Meaning (SEM-05), 530(31):85–93.

Gale, W. A. and Church, K. W. (1991). Identifying Word Correspondences in Parallel Texts. In Speech and Natural Language: Proceedings of a Workshop.

References IV

Gries, S. T. (2009). What is Corpus Linguistics? Language and Linguistics Compass, 3(5):1225–1241.

Gries, S. T. and Stefanowitsch, A. (2004). Extending collostructional analysis A corpus-based perspective on 'alternations'. International journal of corpus linguistics, 1(9(1)):97–129.

Hansen-Schirra, S., Neumann, S., and Čulo, O. (2017). Annotation, Exploitation and Evaluation of Parallel Corpora. Language Science Press.

References V

Karakanta, A., Vela, M., and Teich, E. (2018). EuroParl-UdS: Preserving and Extending Metadata in Parliamentary Debates. In ParlaCLARIN: Creating and Using Parliamentary Corpora. LREC 2018 Workshop.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In MT Summit, pages 79–86.

Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. In Banerji, A. E. and R., editors, Artificial and human intelligence, chapter 11. Elsevier Science Publishers. B.V.

References VI

Schmid, H. (2013). Probabilistic Part-of-Speech Tagging Using Decision Trees. In New methods in language processing, page 154.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford University Press.

Sinclair, J. (2004). Trust the Text: Language, corpus and discourse. Routledge, London and New York.

Song, S. (2017). Modeling information structure in a cross-linguistic perspective. Language Science Press, Berlin.

References VII

Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 88–99.

Stubbs, M. (1996). Texts and text types (chapter 1). In Text and corpus analysis: computer-assisted studies of language and culture. Blackwell Publishing Inc.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. Lrec, pages 2214–2218.

Zanettin, F., Bernardini, S., and Stewart, D., editors (2014). Corpora in Translator Education. Routledge, New York, 2 edition.