Corpora and Language Awareness – Hands-on exercises
http://corp.hum.sdu.dk

Spelling variation

Living languages are subject to constant evolution, and at a given point in time alternate spelling variants may exist for the same word, and there may be regional differences or genre-dependent differences. Individual usage may deviate consistently from the official norm, as when old people end up spelling incorrectly simply because they insist on the spelling they learned in school. Especially interesting are loanwords from other languages, where a "democratic" process can be observed over several years before a new word settles for its final spelling and a full integration into the host language's grammatical system (e.g. gender and number).

Check the frequency of the following word forms, in a corpus or on the Internet. Are the dictionaries right?

German:
- Spontaneität - Spontanität (Google < Wikipedia, why?)

Danish
- tjek – check (betydningsforskel?), linie – linje (RO)
- viruset – virusset

Portuguese
- European-Brazilian differences: Use the Folha and Europarl corpora at the same time, searching form (a) 'registr?a.*' or (b) 'ac?tos?', then sort (by left or right edge) for "freq"(uency).
- Aids, aids, Sida, sida - what is “normal” - and where? Frequency?

Loan words

Every language uses loan words from other languages, and often the words have been in the language so long that only an etymological dictionary will tell where they really came from.

Compile lists of loanwords by using certain prefixes or suffixes. Can you find words with mixed "international" elements? Use regular expressions, e.g. '.*isation'!

	English	**Portuguese**	**Spanish**	**German**
Latin	-ise/-ize, -tion	-izar/-ização	-izar, -ización	-isieren, -ation
Greek	meta-, geo-, syn- -log, -logy	meta-, geo-, syn- -logo, -logia	meta-, geo-, syn- -logo, -logia	meta-, geo-, syn- -log, -logie
English		'th.*', 'sh.*', -ing, -man (how do you avoid false positives?)		
French	[a-z].*[áàé].* (Europarl, case-sensitive)			[a-z].*[áàé].* (Europarl, case-sensitive)
German	über.+ (Wikipedia)			

Grammatical variation

Portuguese: Existing words may acquire new meanings, triggering gender change:

- How does the language decide the gender of loan words, e.g. *e-mail, homepage, website* (*Público 98* or *Folha*)
- Can gender change? And who's winning, and where? Compare the frequencies of *a personagem - o personagem* (compare *Público 91, Público 98* and *Folha)*

German needs to inflect English loan verbs:

- find participles of English loan words: e.g. gesaved - gesavet, downgeloaded - downgeloadet, outgesourced - outgesourcet - outgesourct, eingelogt - eingelogd - eingelogged - eingelogget

Danish needs to assign gender and number to English loan nouns:

- en (web)site – et (web)site (gender)
- hooliganer (RO) – hooligans (number)

Synonyms

There may be more than one word for the same object - even in one language. Check the "who" & "where" of the following:

English: Mobile phone, Cellphone, Cellular phone, Handy, Pocket phone
Portuguese: AIDS, SIDA, Aids, Sida, aids, sida (compare: Floha corpus vs. Público)

Loan words may face competition from newly created expressions or loan translations, and it may take years before one or the other has "won". Check the following:

German

- downloaden - herunterladen (Wikipedia-corpus/Google)
- uploaden - hinaufladen (Google)

Danish

- homepage - hjemmeside
- website – websted - webside
- e-mail – e-post
- laptop - bærbar

Portuguese

- How do you say “acetatos” in Brazilian Portuguese? Is the Brazilian word used in Portugal (same meaning? frequency?)
- correio elctrónico - e-mail - email

New words: abbreviations

Not all new words are loan words or compounds. A special way to create a new word is to turn an abbreviation into a regular word - itself then subject to variation, inflexion and compounding.

Find out whether upper or lower case is preferred, and if derivation is possible:

- SMS/sms, DVD/dvd/Dvd, CD-ROM/cd-rom/CD-rom/CD-Rom
- Check verbal forms (Wikipedia/Google): [A-Z][A-Z]+'?+ing (English), [A-Z][A-Z]+'?+en (German), [A-Z][A-Z]+'?+ar (Spanish/Portuguese), [A-Z][A-Z]+'?+ede (Danish/Swedish)

Swear words

Abusive language (4-letter words etc.) are a sociolinguistic treasure grove. Find out about the unwritten rules of this rather under-researched area of language teaching.

English

- Use the Corpuseye chat corpus to find compounds with 'shit-' (shit[a-z]+) and 'fuck' (fuck[a-z]+). Is there a semantic difference as to which nouns these two prefixes can attach to?
- 'fucking' is also used on its own, as either an adjective or an adverb. Explore the syntactic restrictions for the use of 'fucking' as an adverb!

Danish

- Is there a grammatical difference between compound starting in 'skide-', 'møg-' og 'pisse-' on the one hand, and 'lorte-' on the other hand? Compare with 'super-' and 'kæmpe-'! Find other augmentative prefixes!
- Find syntactic rules for the use of 'sgu' in Danish sentences! Which wordclasses are allowed directly left and right of 'sgu'?

Language and sexual gender

In corpus-linguistic terms, gender/sex-dependent differences of context and usage are among the easier topics to check:

- English and Danish: Determine which nouns most frequently occur to the right of 'his' and 'her' (English) or 'hans' and 'hendes' (Danish). Use the statistics buttons 'freq' and 'rel' with "right context". Use the Shakespeare corpus KEMPE and the ENRON email corpus (or Korpus 90 and 2000 for Danish). If 2 corpora are chosen at the same time, frequency sorting will yield parallel lists for comparison.)
- Brazilian Portuguese: Find out which nouns typically occur left of *dele* and *dela,* respectively (Folha corpus). Note that the parser has split *dele* into *de <sam->* and *ele <-sam>.*
- Spanish, Portuguese and other Romance languages: Compile a list of typical adjectives used next to 'hombre' and 'mujer'. Use the Wikipedia corpus and "refine search" to ask for an adjective right of 'hombre'/'mujer', or use cqp-speak in the normal interface: [word="mujer"] [pos="ADJ"]. Sort the resulting concordance by "right edge" and relative frequency ("rel" button).
- Danish: Find compounds with 'mand.+' og 'kvinde.+', maybe 'dreng.+' and 'pige.+', and compare the results statistically! Use [a-zæøå]{4,} instead of .+ to rule out simple inflected forms.

Derivation

The German, Danish and Swedish lexicon is - morphologically speaking, of course - more dynamical than the Portuguese, French or English vocabulary, because the former group of languages allows productive and multiple root compounding.

- Find the longest Portuguese/German/Danish/English/French/Spanish word! Dots can be used to indicate the number of letters ('....'), but a smarter solution is an expression like [a-zæøåäöüé]{20,30}, meaning a word between 20 and 30 letters. Don't (!) use .+ after many dots - it's very hard on the server. Can you beat the Danish *'Menneskerettighedskonventionen' (Human rights convention)* or *'Tændstiketiketsamlersammenslutningen' (Matchbox label collector society),* from Korpus 2000?
- Portuguese: Find words ending in *-ódromo*! (any Brazilian/Portuguese differences?)
- Find as many verbs as possible ending in *-geben* (German) or *-sætte* (Danish). Use "left-edge" sorting with the frequency button ("freq"). For English, find phrasal constructions with *'give', 'stand' etc.,* i.e. 'give' followed by an adverb or a preposition.

Word order

Many rules restrict the order of words in a sentence, and word class is - apart from meaning - a key factor in formulating these rules.

Use the statistical tools in Corpuseye: “left edge” vs. “right edge”, and frequency listings (“freq”).

- Portuguese: Find out if object pronouns are more common right or left of (a) finite verbs, (b) infinitives, using the Brazilian Folha corpus and the "freq" statistics. What is the typical left context for the VFIN-PRON and PRON-VFIN cases, respectively? Are the results different on the European Público98 corpus?
- Various languages: Find subjects to the right of their verbs! Are certain verbs more likely to occur in these constructions than others?
- English: Find out which word class is allowed between an infinitve marker ('to') and its infinitive! Use "refine search" with 3 fields, and search by "left context, offset=1".
- English and Danish: Compile a list of adverbs that can be placed inside a prepositional phrase (i.e. between a preposition and its argument)
- Choose a language and find out where in a chain of several adjectives that language places nationality adjectives (Italian, Russian etc.).

Syntactic structure

Syntactic structure is not only governed by languages-specific rules (such as whether adjectives are placed left or right of their head noun), but also by more universal principles, such as the uniqueness principle. One such constraint is simply complexity - the human brain sets limits to the length of a determiner chain, or the number of chained subclauses, and these limits are usually stricter for spoken than for written language.

- All languages: Find a noun phrase with as many dependents as possible (CQP or treebank)
- All languages: Find a relative clause within a relative clause!
  E.g For a depth-3 search in CorpusEye: DN+fcl << (DN+fcl << DN+fcl)
  Portuguese Floresta with Milhafre search tool (linguateca.pt):
  /^N<:fcl\b.*$/=e1 << /^N<:fcl\b.*$/=e2 :

Semantic issues

- All languages: Find profession nouns, using different methods (suffix, context, special tags: <Hprof>)
- High-level tak: Find time adverbs and other time expressions and classify them!

- Portuguese: Find metaphors in the Folha_sem corpus (using cqp speak): [pos="N" & func="SUBJ>" & extra!="H[a-z]*"] [morph=".*(PR|IMPF).*" & extra="vH"], or in the Público corpus, for likely candidates like 'cara': [word="cara"] [word="de"] [pos="N"] or, with a more selective search: [word="cara"] [word="de"] [pos="N" & extra=”A”]

The exercises were originally designed for the Danish Korpus 2000, but will work with most languages and most of the Corpuseye corpora. When working on their own, first time users are advised to watch the flash-film "Guided Tour". Additional information, including search examples and an introduction to regular expressions, can be found in the "info" and "help" files.

Further course material (in Danish), for teachers of "Almen Sprogforståelse" (Language Awareness), can be found at VISL's teaching server at: http://beta.visl.sdu.dk/urkas_pearls.html. Exercises are ordered according to topic, with corpus exercises to be found in the stack of blue pearls (E), “Test & data”.

Some documentation can be found in the following publications, available in pdf at the same site:

Bick, Eckhard (2005-5), CorpusEye:Et brugervenligt web-interface for grammatisk opmærkede korpora, In: Peter Widell & Mette Kunøe (eds.), 10. Møde om Udforskningen af Dansk Sprog 7.-8.okt.2004, Proceedings. pp.46-57, Århus University

Bick, Eckhard (2005-4), Live use of Corpus data and Corpus annotation tools in CALL: Some new developments in VISL, In: Henrik Holmboe (red.), Nordic Language Technology, Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004 (Yearbook 2004). pp.171-186. Copenhaguen: Museum Tusculanum