#LancsBox localisation: Instructions

Thank you very much for offering to help with #LancsBox localisation. This document includes basic instructions related to the localisation.

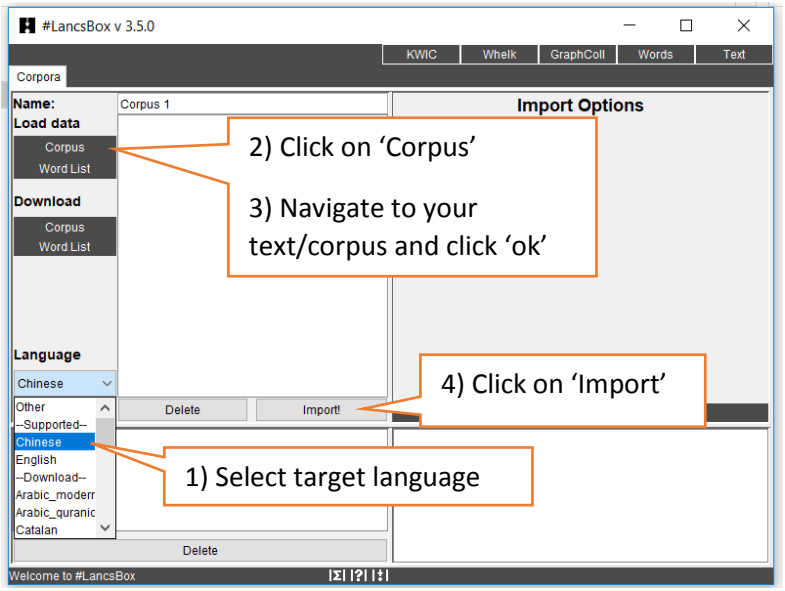
1. Find a list of part-of-speech (POS) tags for the target language on this website: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

   These are usually available under the 'Parameter files' heading, 'tagset documentation'. N.B. In some cases, the tagset might not be obvious or there might be two or more options. If unsure, email v.brezina@lancaster.ac.uk for the correct tagset.

Parameter files

- Bulgarian parameter file (gzip compressed, UTF-8, tagset documentation, trained on the Bulgarian Treebank)
- Catalan parameter file (gzip compressed, UTF8, tagset documentation)
- A Chinese parameter file and tokenizer created by Serge Sharoff are available here
- A Coptic parameter file created by Amir Zeldes is available here
- Czech parameter file (gzip compressed, UTF-8, trained on the Czech Academic Corpus)
- Danish parameter file trained on the ePAROLE corpus (gzip compressed, UTF-8, tagset documentation)
- Dutch parameter file (gzip compressed, UTF-8, tagset documentation)
- Another Dutch parameter file (gzip compressed, UTF8, trained on the Eindhoven corpus, tagset documentation (starts on page 9))
- English parameter file (PENN tagset) (gzip compressed, UTF8, tagset documentation, trained on the Penn treebank)
- English parameter file (BNC tagset) (gzip compressed, UTF8, tagset documentation, trained on the British National Corpus)

2. Load a text or corpus representing the target language into #LancsBox. If you don't have a text/corpus, copy/paste a text in the target language from the internet into a Word document and load this text. Make sure that you select the target language in #LancsBox setting before loading the text (corpus).



3. Open the 'localisation.xlsx' file in Excel. English settings are given as an example. Go through each sheet and convert the information provided for English into the information appropriate for the target language. Use the information about the tagset from step 1. You can test searches using the KWIC tool in #LancsBox and the text/corpus you loaded under step 2.

   a) First, provide info about the tagset as described online (step 1).

A	B	C	D	E	F	G	H
ID	Tag	Desctiption	Website info				
1	CC	Coordinating conjunction	http://www.cis.uni-muenchen.de/~schmid/tools/Tree				
2	CD	Cardinal number					
3	DT	Determiner					
4	EX	Existential there					
5	FW	Foreign word					
6	IN	Preposition or subordinating conjunction					
7	JJ	Adjective					
8	JJR	Adjective, comparative					

   b) Next, define POS categories; these are categories traditionally recognised in the description of the target language. These do not have to 100% map onto the English categories. N.B. If you are not well versed in regular expressions, use the 'Notes' column (highlighted) to describe these.

Category	POS_regex	Notes
adjective	J.*	All tags starting with J
adverb	W?R.*	All tags starting with R or WR
connector	(IN|CC)	All tags that are either IN or CC
noun	N.*	All tags starting with N
pronoun	(PP\$?|WP\$?)	All tags starting with PP and WP optionally followed by $
verb	[VM].*	All tags starting with V or M
other	.*	Anything else remaining

c) Next, define searches. Searches include POS categories as defined under b), punctuation searches, and complex searches. Each search can be defined at the level of a Word, Headword and POS or combination of these. N.B. If you are not well versed in regular expressions, use the 'Notes' column (highlighted) to describe these.

Search	Word	Headword	POS	Notes
NOUNS			N*	All POS tags staring with N
PROPER NOUNS			NP*	
VERBS			[VM].*	All POS tags staring with V or M
ADJECTIVES			JJ*	
ADVERBS			W?R.*	
MODALS			MD	
CONNECTORS			/(IN|CC)/	
PRONOUNS			(PP\$?|WP\$?)	
?	/.*\?/pi			
!	/.*\!/pi			
.	/.*\./pi			
,	/.*\,/pi			
PASSIVES			/VB. (R.*)	All POS tags re presenting the verb to
COMPLEX NOUN PHRASE			/(JJ.?){1,5}NN.? /	
PAST TENSE			/V.D/	
NOMINALIZATIONS	/.{3,}(tion|tions|ment|ments|ness|nesses|ity|ities)/i			
SPLIT INFINITIVE			/TO R.* V.*/	
PRESENT TENSE			/V.[PZ]/	
PAST TENSE			/V.D/	
PLACE ADVERBIALS		/aboard|above|abro		A list of following words: abroad, abc
TIME ADVERBIALS		/afterwards?|again|earlier|early|eventually|formerly|imn		

d) Optionally, Provide a list of clitics. These are words that are attached to another word and written together as one unit. For example, in English the clitic n't [not] is attached to verbs to form units such as *isn't, doesn't, didn't, hasn't couldn't* etc.
N.B. This doesn't apply to all languages. Not all languages have clitics.

e) Optionally, if available provide a list of most common abbreviations for the target language, if available. These are words that include the full stop (.) such as *Mr., Dr.*

Thank you very much for your help!