LINGUA:
Language-Independent Neighbourhood Generator of the University of Alberta

Chris Westbury, Geoff Hollis, & Cyrus Shaoul
Department of Psychology, University of Alberta, Edmonton, AB, T6G 2E9 Canada.

INTRODUCTION

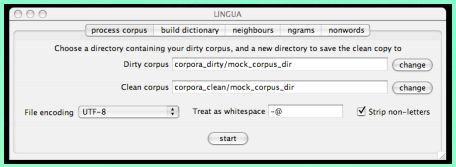
It is well-established that word frequency counts and statistical measures of word form similarity, such as orthographic neighbourhood (ON), summed N-gram frequencies, and word body counts, play an important role in lexical access. Many of the available tables of these variables are dissatisfactory for one reason or another: because they were drawn from too small a corpus, because the corpus was out of date, or because the corpus contained entries that we would prefer not to include. Moreover, these values are not available for most languages other than English. We offer a solution to these problems: the Language-Independent Neighbourhoood Generator of the University of Alberta (LINGUA). **LINGUA is a freely-available, platform-independent (Java) application that we have developed specifically to enable the calculation of various statistical measures from a corpus of (not necessarily English) text**. The calculable measures include frequency, ON, summed bigram frequencies, and Markov-chained nonwords (which respect the statistical distribution of letters in the real words in the input corpus)

As well as distributing LINGUA itself, we will freely distribute frequency dictionaries and other calculations for corpora in three languages: French, Spanish, and English.

AN OVERVIEW OF LINGUA

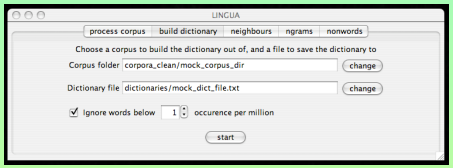
LINGUA is structured as a series of five tabbed panels that allow the user to specify and run different kinds of calculations.

1.) Corpus Processing



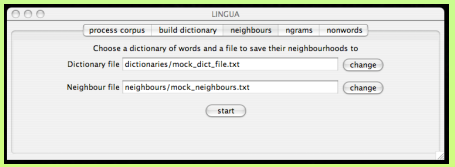
The first panel allows user to process the corpus, by removing punctuation and non-alphabetic characters, and converting all text to uppercase UTF-8 encoding. The input files can be in any encoding, allowing for the program to be used with text from most languages.

2.) Dictionary Building



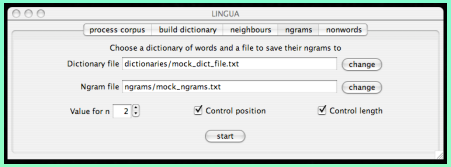
The second panel allows the user to build a frequency dictionary from the cleaned corpus text that was generated using the first panel. Users may optionally specify a lower cut-off frequency below which words will not be counted, as a rough way of cleaning nonwords and other garbage out of the dictionary.

3.) Neighbourhood Calculation



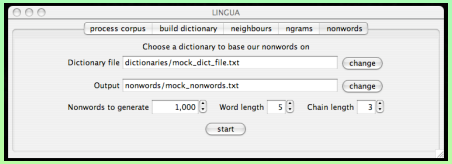
The third panel allows the user to build an orthographic neighbourhood dictionary from the dictionary text that was generated using the second panel (or from any similarly formatted dictionary). The output file lists the number of neighbours, their average frequency, and the neighbours themselves.

4.) N-gram Calculation



The fourth panel allows the user to calculate summed N-grams, where N is a user-specifiable integer. The input file is any properly formatted dictionary. The output file lists the frequency per million words of every N-gram in every word in the dictionary.

5.) Generating Plausible Nonwords



The fifth panel allows user to generate nonwords using Markov chaining. This method ensures that every N-gram in every nonword occurs in at least one real word, and that the distribution of N-grams in the nonwords is roughly the same as the distribution of N-grams in real words. In general nonwords generated this way seem like real words in the language of the dictionary. The user can specify how many words to generate, how long each nonword should be, and what length of N-gram to use. Nonwords can be added to a dictionary file (with a frequency of 0) in order to get neighbourhoods of the nonwords, using the third panel described above.

AVAILABILITY

LINGUA and sample output files in French, Spanish, and English are all available from our website at:
www.psych.ualberta.ca/~westburylab/publications.html

Acknowledgements: This work was supported by the Natural Sciences and Engineering Research Council of Canada.