Eric W. Holman

Programs for calculating ASJP distance matrices (version 2.2) [2014]

New in version 2.2

Those already familiar with version 1.0 will perhaps not immediately recognize any changes. The only substantial difference between the two versions is that version 1.0 treated the modifiers (”, *, ~, and $) as separate elements in the string comparisons, whereas version 2.2 treats these symbols as originally described by Brown et al (2008): ” and * as indicating that the preceding symbol represents a different phoneme from the unmodified version of the same symbol, ~ as indicating that two preceding symbols make up a composite symbol representing one phoneme, and $ as indicating that three preceding symbols make up a composite symbol representing one phoneme. This change slightly increases the correlations between ASJP distances and the classifications in WALS (Haspelmath et al. 2005) and Ethnologue (Lewis et al. 2014).

Two other differences accommodate recent expansion in the ASJP database. First, the programs in version 2.2 can deal with up to 8000 languages. Second, asjp62x outputs distances in percentages with two decimal points but with the dot removed (alternatively the numbers can be interpreted as multiplied by 100). This is a way of saving space in the matrix which, with the current number of languages in the database, would otherwise be too large for MEGA6 (Tamura et al. 2013).

As another new feature the source codes are now also provided along with the compiled .exe files.

One other feature, described in the instructions below, allows the user to screen out lists with less than a specified number of attested items. Otherwise, the instructions are the same as those for version 1.0, repeated here for convenience.

Instructions

The programs described here calculate LDND, as defined by Bakker et al. (2009), between pairs of languages. The programs all use the same input and produce slightly different output. asjp62 produces a matrix of LDND with rows and columns labeled by the language names. asjp62x produces the same matrix in a format appropriate for use as input to the MEGA6 phylogeny package (Tamura et al. 2013). asjp62e produces 1-LDND for pairs of languages within taxonomic groups, with each pair on a separate line in a format appropriate for pasting into Excel.

To run a program, get the MS-DOS command prompt, type a command of the form

```
program < input > output
```

and then press Enter. For example, the command asjp62 < input.txt > output62.txt will run asjp62 on input.txt to produce output62.txt; the command asjp62x < input.txt > output62x.txt will run asjp62x on input.txt to produce output62x.txt; and the command asjp62e < input.txt > output62e.txt will run asjp62e on input.txt to produce output62e.txt. The input and output can be either .txt files or plain unmodified files. The computer may add a line like

```
Stop - Program terminated.
```

to the end of the output, but this can be deleted before the output is used further.

The input file must obey the following general rules.

The first line is in fixed format so the columns are important.
Col. 6: maximum number of synonyms read for each item (1 or 2).
Col. 11-12: minimum number of attested items in lists, up to 100; lists with fewer attested items are ignored.
Col. 15-18: if this number is 0, all lists are read; if it’s a positive number, it’s interpreted as a date and lists from languages extinct before that date are ignored.
Col. 24: if this is a number other than 0, transcribed words and phrases preceded by % are ignored, which allows loans to be excluded if they are identified by %.
Col. 30: Taxonomic rank of groups within which similarities are calculated by asjp62e: 3 = families, 2 = genera. Only asjp62e uses this information; asjp62 and asjp62x ignore it. ASJP uses the families and genera defined in WALS (Haspelmath et al. 2005) but the computer will accept whatever definition is specified for the languages as described below.

The next line gives the format for reading the item numbers below it. The programs described here ignore the item names so the format in the example could just as well be I4.

The next set of lines give the item numbers that will be used. There is one line for each item in the list. The item numbers must be between 1 and 100 inclusive, but they don't have to be consecutive or listed in numerical order. For I4 format, the numbers must be in Cols. 1-4, right justified. Items with numbers other than those listed here aren’t used in calculating LDND. The item names in the example are just for convenience.

There must be a blank line after the item list. Press the space bar a few times to give the computer something to read. This line tells the computer that the item list is finished.

The next set of lines give the ASJPcode symbols, one per line in Col. 1, in any order. As an alternative to ASJPcode, any ascii symbols can be used; there can be up to 100 different symbols. Symbols not on this list aren’t used in calculating LDND.

There must be two blank lines after the symbol list.

Then there is a wordlist for each language, on consecutive lines.

The first line for a language is the language name and classification, which can be anything as far as the computer is concerned, as long as it doesn't start with a number or a blank.

The second line gives properties of the languages, again in fixed format so the columns are important.
Col. 2: 3 if the language is the first one in a new family, 2 if it's the first language in a new genus, 1 otherwise.
Col. 4-10: latitude in degrees and hundredths of a degree; minus means South. The programs described here don't use this information.
Col. 12-18: longitude in degrees and hundredths of a degree; minus means West. The programs described here don't use this information.
Col. 19-30: number of speakers, from Ethnologue (Lewis et al. 2014); 0 if the number of speakers is unknown; -1 if the language is recently extinct; -2 if the language is long extinct; or if the approximate date of extinction is known, the date is preceded by a minus sign. If there is a date in the first line of the entire file, lists with earlier extinction dates here are ignored, as are lists with -2; otherwise, all lists are used.
Col. 34-36: three-letter WALS code, if any. The programs described here don't use this information.
Col. 40-42: three-letter ISO code from Ethnologue, if any. The programs described here don't use this information.

Each of the next lines refers to an item in the list, until the next language begins. Items can be in any order. The line must begin with the item number, starting in Col. 1, left justified. The next column after the number can be anything except a tab. The program then ignores everything until it reaches a tab; this part of the line can be used for the name of the item. After the tab is the transcribed word or phrase; words in a phrase are separated by a space, which is ignored in the calculations; synonyms are separated by a comma. XXX here means that the item isn't attested for the language; alternatively, unattested items can be omitted from the list. The end of the transcription is indicated by a space and then //. Two consecutive spaces also signal the end of the transcription.

There must be a blank line after the last list.

References

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman. 2009. Adding Typology to Lexicostatistics: A Combined Approach to Language Classification. *Linguistic Typology* 13.167-179.

Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Vilupillai. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *STUF – Language Typology and Universals* 61:285-308.

Haspelmath, Martin, Matthew Dryer, David Gil, & Bernard Comrie (eds.). 2005. The World Atlas of Language Structures. Oxford: Oxford University Press. (http://wals.info/)

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2014. *Ethnologue: Languages of the World, Seventeenth edition.* Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Tamura, K., G. Stecher, D. Peterson, A. Filipski, and S. Kumar. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0**.** Molecular Biology and Evolution 30:2725-2729.