

Description of the ASJP database format

This is a general description of the ASJP database of wordlists. Some aspects of the database that are specifically intended for ASJP software are described more thoroughly by Holman (2011a, 2011b).

The first line is specific to ASJP software. For users of that software, the 2 in col. 6 is the maximum number of synonyms read for each item, the 24 in cols. 11-12 is the minimum number of items that must be attested in a wordlist for that wordlist to be analyzed by the software, and the 1700 in cols. 15-18 is a date that can specify which lists are read.

The next line gives the format for reading the immediately following list. The list itself consists of the 40 most stable items, as determined by Holman et al. (2008), in the 100-item list of Swadesh (1955). Most of the wordlists in the database contain these 40 items, or as many of them as are attested in the sources. About 314 wordlists contain as many items as are attested from the full 100-item Swadesh list. The English names of the items are less important than the preceding numbers, which are used to identify the items in the wordlists.

The next list consists of the ASJPCode symbols that are used to transcribe the wordlists. These are described by Brown et al. (2008).

Then there is a wordlist for each language, on consecutive lines. Lists are ordered according to the classification in WALS (Haspelmath et al. 2005). Families ordered geographically, genera are ordered alphabetically within families, and languages are ordered alphabetically within genera.

The first line for each list gives the name of the language followed within curly brackets by its position in three classifications, without any blank spaces. The name is taken from the source of the list; it never starts with a number or a blank. Between { and | is the classification of the language in WALS. It's of the form Fam.GENUS, with the family name abbreviated and the genus name spelled out. Between | and @ is the classification of the language in *Ethnologue* (Lewis 2009). Names of taxonomic groups and subgroups are separated by commas and ordered from most inclusive to least inclusive. Languages not in WALS or *Ethnologue* are classified from information in the source for the list. If this information is insufficient for WALS, the family and genus are called Unknown. If it's insufficient for *Ethnologue*, the sequence of subgroups is continued only as far as the information permits, including in some cases no groups at all. Between @ and } is a (not yet published) classification by Harald Hammarström. This is a new feature of Version 14 of the database. The classification is similar to that of *Ethnologue* except as far as the definitions of the highest taxonomic units (families) are concerned. Hammarström's classification is more conservative in this regard than those of WALS and *Ethnologue*.

The second line gives properties of the languages, in fixed format separated by blanks (not tabs), so the columns are important.

Col. 2: 3 if the language is the first one in a new WALS family, 2 if it's the first language in a new WALS genus, 1 otherwise.

Col. 4-10, right justified: latitude in degrees and hundredths of a degree; minus means South.

Col. 12-18, right justified: longitude in degrees and hundredths of a degree; minus means West.

Latitudes and longitudes were ascertained from WALS, or from the maps in *Ethnologue* or Moseley and Asher (1994), or from information in the source for the list.

Col. 19-30, right justified: number of speakers, from *Ethnologue*. This number always refers to the entire language, as defined in *Ethnologue*, even if the list itself refers to a dialect. The number is 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. In the ASJP software, 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 read.

Col. 34-36: three-letter WALS code, if any.

Col. 40-42: three-letter ISO639-3 code from *Ethnologue*, if any. This code is included for languages in previous editions of *Ethnologue* even if they aren't in the 16th edition. Languages that lack an ISO639-3 code but can be placed in the *Ethnologue* classification are given a code consisting of two letters followed by the number 0 (for use in ASJP software).

Each of the next lines refers to an item in the list, until the next language begins. Items may be in any order. The line always begins with the item number, starting in Col. 1, left justified. This number always identifies the same item throughout the database. The number is followed after some spaces by the name of the item, usually in English but sometimes in Spanish. This is followed by a tab (the only tab in the line), which always signals the beginning of the transcription. Immediately after the tab is the transcribed word or phrase; words in a phrase are separated by a space (which is ignored by ASJP software), and synonyms are separated by a comma. A % sign before a word or phrase identifies it as a loan. XXX means that the item isn't attested for the language; alternatively, unattested items may be omitted from the list. The end of the transcription is indicated by a space and then, normally, //. Two consecutive spaces also signal the end of the transcription. The // may be followed by additional information, such as the donor language for loans.

References

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, and Bernard Comrie (eds.). 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press. (<http://wals.info/>)

Holman, Eric W. 2011a. Programs for calculating ASJP distance matrices (version 2.1). <http://email.eva.mpg.de/~wichmann/software.htm>

Holman, Eric W. 2011b. Program for calculating ASJP dates (version 1.1). <http://email.eva.mpg.de/~wichmann/software.htm>

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, Pamela Brown, and Dik Bakker. 2008. Explorations in automated language comparison. *Folia Linguistica* 42:331-354.

Lewis, Paul M. (ed.). 2009. *Ethnologue*. 16th Edition. Dallas: SIL International. (<http://www.ethnologue.com>)

Moseley, Christopher, and R. E. Asher (eds.). 1994. *Atlas of the World's Languages*. London: Routledge.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21:121-137.