Onset Cluster Database – Description and Instructions

Katherine Bare, Jonathan Bauman, Jack Campbell, Katja Jablonski, Stephen Jay, David Leonardi, Timothy Palmer, Steve Parker, William Payne, Moriah Rose, Jenna Sawyers, Emily Scheie, Lydia Stebbins, Matthew Woods, and Shaiann Wyatt-Murphy

October 31, 2023

corresponding author: steve_parker@diu.edu

Background

This document summarizes the purpose and contents of an accompanying Excel worksheet. That spreadsheet ("Onset Cluster Database") is available on this same webpage. This project began in 2014 and has been overseen since then by Steve Parker. The contributors (co-authors) to this project consist of students, interns, and faculty members who have been associated with Dallas International University at one time or another. This school was formerly called the Graduate Institute of Applied Linguistics.

Motivation

The goal of this project is twofold: (1) to compile a robust yet non-exhaustive sample of languages which are reported to permit consonant clusters in onset position, i.e., at the beginning of syllables; and (2) to document which specific combinations of complex onsets are attested in each of these languages. The onset combinations are reported here in terms of phonological natural classes based on the sonority hierarchy. The sonority scale we use assumes the standard four-way split among consonants: G > L > N > O (glides > liquids > nasals > obstruents, from higher sonority to lower sonority).

Disclaimer

This corpus of 1257 languages, and the data contained therein, represents a work in progress. As such, we make no guarantee that any of the details reported here are necessarily accurate in every respect. In many cases we have not had the opportunity to directly corroborate the facts by carefully reading the primary sources ourselves. Furthermore, some of the entries are incomplete because we lacked the time to fill in every cell. With this caveat in mind, this open access project is being made freely available in order to facilitate research by others. We urge serious users to follow up and confirm the data themselves before publishing their results. You are welcome to use and cite this project for non-commercial purposes in academic and educational contexts. However, we request that in doing so you please note the preliminary and tentative nature of this work. As time permits, we may add new entries and data to this corpus in the future. Consequently, we welcome feedback, questions, and corrections (additional or more current information).

Design

Our primary aim in undertaking this project is to shed light on which combinations of onset clusters are permitted vs. prohibited in natural human languages worldwide. Accordingly, we have drawn data from a number of extant resources which provide information about phonotactic inventories. The list of previous databases we consulted is discussed below (Column E). Given our focus on sonority as a theoretical force which universally constrains syllable structure, we summarize and report onset cluster types in each language in terms of their membership in one of four categories: obstruent, nasal, liquid, or glide. In any given language this can give rise to a maximum of 4 × 4 or 16 potential combinations of biconsonantal onsets. In addition, some languages are reported to permit syllable-initial strings of three (or even more) consonants. However, the prosodic affiliation (representation) of such sequences is a matter of debate. Nevertheless, although the focus of our list is on inventories of onset clusters consisting of exactly two segments, we do make note of such cases in a separate column.

Column by Column Explanation

In the remainder of this document we describe in more detail the contents of the accompanying spreadsheet (Onset Cluster Database.xlsx). The discussion here follows the order of the columns in that file, from left to right. Their names are displayed in bold font in the top row (the column headers). In the corner of each title cell (row 1) is a drop-down text filter box to facilitate searching based on the values in that column.

Column A: language

The leftmost column in the database lists the name of each language or lect. These 1257 names follow the online version of the Ethnologue (Eberhard et al. 2023) whenever possible. In cases of different (competing) names, we give main consideration to the primary sources referenced at the end of each row (entry) in the spreadsheet. These begin in column K. Since we have compiled this database over the course of many years, the language names listed here may not always coincide with the most recent version of the Ethnologue. Nevertheless, since language names generally tend to be stable over time, we hope that the number of such discrepancies is relatively small.

Column B: stock

This is the top-most genetic phylum or family to which each language belongs. In the default case these also follow the classification system of the Ethnologue. We are aware that in many situations such details are disputable, but we do not have the time or expertise to resolve these.

Column C: country

These cells display the primary country where the language is spoken, as reported by the Ethnologue and/or the primary sources. In several cases up to three or four countries are listed here. The country listed first in each row is by assumption the one where the greatest number of speakers have historically resided.

Column D: ISO

This is the three letter ISO code conventionally assigned to each language. In most cases these are taken from the corresponding entries in the Ethnologue.

Column E: database

This cell lists the previously compiled (extant) resources which we consulted in order to gather data and populate each entry (language) while constructing our database. These include the following sources (in alphabetical order), some of which are not publicly available at the moment:

Name	Resource	Reference(s)
Greenberg	Greenberg	Greenberg 1978
Kreitman	Kreitman	Kreitman 2006, 2008
LAPSyD	Lyon-Albuquerque Phonological Systems Database	Maddieson et al. 2014-2016
Parker	Parker	an unpublished corpus compiled by Steve Parker (see Parker 2012:110)
Phonotacticon[1]	Phonotacticon	Joo and Hsu 2023
SylTyp	Syllable Typology Database	Hulst 2012
WALS	The World Atlas of Language Structures Online	Dryer and Haspelmath 2013
WPD	World Phonotactics Database	McElvenny et al. 2013

The order in which these sources are listed in each cell in column E in our database is somewhat random, and no significance (priority) should be ascribed to it. Rather, in most cases the order in the spreadsheet cells simply corresponds to the stage in the project in which each entry was filled in.

Column F: CG/CL

This cell contains a summary indication of the types of bisegmental onset clusters present in each language. The abbreviation *CG* stands for combinations in which C2 (the offset) is always a glide, that is, OG, NG, LG, and GG. *CL* indicates languages in which C2 is restricted to a liquid of some kind. These abbreviations follow the conventions in Parker (2012, 2017) and Jay and Parker (2020). Cells annotated as *both* indicate languages which permit both glides and liquids in offset position (C2). An entry such as *CG+* stands for a language in which C2 can consist of nasals and/or obstruents (in addition to glides), but normally not a liquid. Similarly, *CL+* means the language permits liquid offsets as well as other segment types, but not glides. In some cases when the facts are unclear, we add a question mark at the end of this cell to indicate ambiguity or doubt.

[1]Thanks to Ian Joo for making the Phonotacticon database available to us.

Column G: CC rising/plateau

This column contains an exhaustive list of the specific types (combinations) of two-member onset clusters present in each language, minus any reversed sonority clusters. Thus, the set of CC clusters indicated in this cell includes those in which sonority rises from C1 to C2, as well as those in which C1 and C2 exhibit the same sonority rank (i.e., plateaus). This yields a maximum of ten possible types. Any of these which are attested in the language in question (according to our sources) are entered in this cell in the same fixed order, separated by a comma, with no space in between: OG,NG,LG,GG,OL,NL,LL,ON,NN,OO. However, certain complications inevitably arise when categorizing phonological segments into one of these four natural classes. A particularly difficult issue is whether to classify glottal consonants (/h/ and /ʔ/) as obstruents or something else, such as glides. In the default case we assign glottal consonants to the class of obstruents. Nevertheless, when sources posit a different analysis (of details like this) and provide compelling language-specific arguments to back it up, we sometimes follow those conclusions instead.

Column H: CC reversal

There are six potential types of reversed clusters, i.e., those in which the sonority index falls (descends) from C1 to C2. These are indicated in the cells in this column in the order GL,LN,GN,NO,LO,GO. In reporting these cluster types we follow the descriptions in the primary sources of data. Many phonological analyses assume, but do not prove, that such clusters are tautosyllabic, usually by virtue of appearing in word-initial position. In our opinion this is a huge theoretical and methodological oversight. Sequences of consonants which occur at the beginning of words but are never attested medially in that same language are dubious candidates for canonical syllable onsets. Such cases are confounded by the fact that instrumental measurements often demonstrate the extra (first) word-initial consonant in many languages to be a prosodic appendix — a segment which does not reside in any syllable (Hermes et al. 2013, Gafos et al. 2020, Durvasula et al. 2021). Nevertheless, in the interest of being as exhaustive as possible, we report here any reversed sonority clusters which are posited in the references listed.

Column I: CCC

In this column we give a brief indication of whether the sources report the existence of triconsonantal clusters at the beginning of syllables and/or words. In some cases this cell is simply annotated as *yes*, with no further details. For other languages we specify the quality of one or more of these three consonants, or an overall summary of the main types of sequences which occur.

Column J: notes

This cell mentions any complications we encountered in the description and analysis of particular languages. One of the most frequent problems involves the interpretation of potential clusters in which C2 is a glide, such as [pj]. In many cases the primary sources may disagree among themselves. Nevertheless, if the majority of the facts indicate that the [j] portion is truly a separate consonant rather than (say) palatalization, and that it pertains to the syllable onset rather

than being part of a nuclear diphthong, then we consider this to be a true CC cluster and list it as type OG in column G. Consequently, the cluster types we report for each language in columns F and G are the ones which, in our impression, are most strongly and consistently supported among the consulted references. Other issues noted in this column include dialect differences, cluster types limited to loanwords and/or ideophones, incomplete data, etc.

Column K (and following): References

These are the sources of data provided by the databases listed in column E for each language. The order in which multiple references are displayed in each row is random — they are not necessarily alphabetical, nor are they prioritized in terms of quality, reliability, accessibility, nor any other factor.

Acknowledgements

We are indebted to Ian Joo for kindly sharing with us the latest version of the Phonotacticon database.

References

Dryer, Matthew S. and Martin Haspelmath (eds.). 2013. WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at https://wals.info, Accessed on 2023-10-23.)

Durvasula, Karthik, Mohammed Qasem Ruthan, Sarah Heidenreich, and Yen-Hwei Lin. 2021. Probing syllable structure through acoustic measurements: case studies on American English and Jazani Arabic. Phonology 38:173-202.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2023. Ethnologue: Languages of the World. Twenty-sixth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Gafos, Adamantios I., Jens Roeser, Stavroula Sotiropoulou, Philip Hoole, and Chakir Zeroul. 2020. Structure in mind, structure in vocal tract. Natural Language & Linguistic Theory 38:43-75.

Greenberg, Joseph H. 1978. Some generalizations concerning initial and final consonant clusters. In Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik (eds.), Universals of Human Language, volume 2: Phonology, 243-279. Stanford: Stanford University Press.

Hermes, Anne, Doris Mücke, and Martine Grice. 2013. Gestural coordination of Italian word-initial clusters: the case of “impure s.” Phonology 30:1-25.

Hulst, Harry van der. 2012. Syllable Typology Database. DANS-KNAW.

Jay, Stephen, and Steve Parker. 2020. Onset cluster typologies. Occasional Papers in Applied Linguistics (OPAL). Dallas: Dallas International University. https://www.diu.edu/academics/opal/

Joo, Ian, and Yu-Yin Hsu. 2023. Phonotacticon: A cross-linguistic phonotactic database, 16 August 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-3269302/v1].

Kreitman, Rina. 2006. Cluster buster: A typology of onset clusters. In Jacqueline Bunting, Sapna Desai, Robert Peachey, Christopher Straughn, and Zuzana Tomková (eds.), Proceedings from the Annual Meeting of the Chicago Linguistic Society, vol. 42, 163-179. Chicago: Chicago Linguistic Society.

Kreitman, Rina. 2008. The phonetics and phonology of onset clusters: the case of Modern Hebrew. Ph.D. dissertation. Cornell University.

Maddieson I., S. Flavier, E. Marsico, and F. Pellegrino, 2014-2016. LAPSyD: Lyon-Albuquerque Phonological Systems Database, Version 1.0. https://lapsyd.huma-num.fr/lapsyd/

McElvenny, J., Mark Donohue, and R. Hetherington. 2013. World Phonotactics Database. Software.

Parker, Steve. 2012. “Sonority distance vs. sonority dispersion – a typological survey.” In Steve Parker (ed.), The Sonority Controversy, 101-165. Berlin: De Gruyter.

Parker, Steve. 2017. Reconsidering sonority dispersion and liquid vs. glide offsets: what do the typological facts indicate? WINAK: Revista de Estudios Interculturales 26:11-42.