

Manual

CarolAnn Edie and Jeffrey Heinz

June 19, 2008

Contents

1	Welcome	5
1.1	Structure of the Database	5
1.2	Data Tables	5
1.3	Linking Tables	6
1.4	Features	6
2	Tables	7
2.1	Data Tables	7
2.1.1	Languages	7
2.1.1.1	Fields	7
2.1.1.2	The Language Page	8
2.1.2	Primary Patterns	8
2.1.2.1	Fields	9
2.1.3	Secondary Patterns	9
2.1.4	Phonotactics	10
2.1.5	Sources	10
2.1.6	Sizes	10
2.1.7	Weights	10
2.1.8	Notes	10
2.1.9	Syllables	11
2.1.10	FSA Tables	11
2.1.10.1	Learner Results nd (fsa)	11
2.1.10.2	Learner Results gram/prec (fsa)	13
2.1.10.3	Distinct Pattern Properties (fsa)	14
2.2	Linking Tables	16
2.2.1	Languages Primary	16
2.2.2	Languages Secondary	16
2.2.3	Languages Phonotactics	17

2.2.4	Language Sources	17
2.2.5	Language Ethnologue Code	17
2.2.6	Language Notes	17
3	Syllable Priority Codes	19
3.1	Bailey's Work	19
3.2	Heinz's Expansion	19

Chapter 1

Welcome

Welcome to the Heinz stress database online. This project is meant to present an atheoretical description of the documented stress patterns of the world's languages, as well as to provide a reference for those interested in stress patterns across languages. It also includes results of the Forward-Backward Learner (Heinz 2007). This database is open source (mysql) and is freely available. The stress patterns in the database are collected primarily from the typologies of Gordon (2002) and Bailey (1995). Like all databases, this is a work in progress. Please send any questions/corrections etc. to Jeff Heinz: heinz@udel.edu. There is another stress database on the web that includes information not present here. People interested in stress should also consult their database. StressTyp is available here: <http://stresstyp.unleiden.net/form2b.htm>.

1.1 Structure of the Database

The database interface is comprised of 18 tables. These tables may be Data Tables or Linking tables. These tables describe languages, their FSA representations, and the results of the learning model.

1.2 Data Tables

Data tables simply list different values for one datum. For example, the Language table lists the languages and values relevant to them. Similarly, in the Primary Stress table, each record represents a primary stress pattern. The

data tables in the database are: Languages, Primary Patterns, Secondary Patterns, Phonotactics, Sources, Sizes, Weights, Notes, Syllables, and the three views of the FSA table (Learner Results nd (fsa), Learner Results gram/prec (fsa), Distinct Pattern Properties (fsa)).

1.3 Linking Tables

Linking Tables, such as the Languages Primary table, link one record from one table with a record in another table. For example, the Languages Primary table associates each language with a primary stress pattern (see 3). These tables might also include some other field which is associated with this linkage, such as the ‘sizes’ field in the Languages Primary table. In this case, this indicates what size words to which the pattern applies in this language.

1.4 Features

The database includes atheoretical descriptions of stress patterns in English, Finite State representations of the stress patterns, and the results of the learning model. Additionally, we hope to add information about the complexity of the patterns with respect to the Subregular Hierarchy (McNaughten and Paret 1971, Pullem and Rogers 2007).

Chapter 2

Tables

While navigating the database, you will find there are 18 tables listed at the top of the page. The following section specifies what these tables mean and what they are comprised of. It is separated into two halves: Data Tables and Linking Tables.

2.1 Data Tables

2.1.1 Languages

This table contains information about each language, the domain in which the stress pattern is valid, and other information more generally about the language. A language may be listed more than once if different stress patterns occur in different domains, or if different researchers provide different descriptions.

2.1.1.1 Fields

Lg The ‘Lg’ field displays the languages that are included in the database.

LgTeX The ‘LgTeX’ field translates the characters in the ‘Lg’ field into tex-friendly characters.

per The ‘per’ field indicates a primary source which provides all of the information about the language. This account may differ from other accounts of the same language, or it may be the only account available for a rare or extinct, or understudied language.

p-domain The ‘p-domain’ field refers to the phonological domain of the stress pattern: root, stem, or phrase.

ms-domain The ‘ms-domain’ field refers to the morpho-syntactic domain (ex. Verbs, nouns, roots, etc) of the stress pattern.

weight The ‘weight’ field indicates the number of syllable types between which the language distinguishes in the placement of stress.

syl_id The ‘syl_id’ field links to a record in the Syllables table which describes the syllable types in a Quantity Sensitive pattern. (See Language Page: Syllable Weight Hierarchy, Syllables table)

fsa_id The ‘fsa_id’ field links to a record in the FSA table which contains information about the finite state acceptor which recognizes the stress pattern of this language. (See Distinct Patterns Properties (fsa) table: 2.1.10.3)

notes The ‘notes’ field is an internal development table that is used to keep track of things.

status The ‘status’ field is an internal development table that has been used to keep track of things.

2.1.1.2 The Language Page

On the language table, clicking on the link which is the language ID field will bring you to the Language Page. This page brings together much of the salient information in the database about one particular language. This includes the relevant ethnologue code(s), if any, the type of stress pattern ((un)bounded, quantity (in)sensitive), prose statements of primary and secondary stress patterns, a Syllable Weight Hierarchy chart, statement(s) of relevant phonotactic information, sources, and finite state acceptor informational charts and diagrams.

2.1.2 Primary Patterns

This table lists each primary stress pattern included in the database by a code. (See information on SPC codes: section 3)

2.1.2.1 Fields

SPC The ‘SPC’ field lists the Syllable Priority Code patterns of primary stress that are present in the database. More information about Syllable Priority Codes may be found in 3.

English The code is explained in the ‘English’ field in prose. This field should in most cases be prefaced with “Primary stress falls on”.

Notes Any relevant notes referring to the primary stress pattern are located in the ‘Notes’ field.

Weight Quantity (in)sensitivity is specified in the ‘Weight’ field with the values QI or QS.

Bounded Boundedness of the pattern is specified in the ‘Bounded’ field (yes or no).

Unbounded Unboundedness of the pattern is specified in the ‘Unbounded’ field (yes or no).

fsa The finite state acceptor for the primary stress pattern is listed in the ‘fsa’ field. This fsa uses features.

2.1.3 Secondary Patterns

This table lists each secondary stress pattern included in the database by a code. (See information on SPC codes: 3)

Code The ‘Code’ field lists the Syllable Priority Code patterns of secondary stress that are present in the database. More information about Syllable Priority Codes may be found in 3.

English The code is explained in the ‘English’ field in prose. This field should in most cases be prefaced with “Secondary stress falls”.

Notes Any relevant notes referring to the secondary stress pattern are located in the ‘Notes’ field.

Weight Quantity (in)sensitivity is specified in the ‘Weight’ field with the values QI or QS.

Howmany The number of secondary stresses placed by the pattern is indicated in the ‘Howmany’ field.

Bounded Boundedness of the pattern is specified in the ‘Bounded’ field (yes or no).

Unbounded Unboundedness of the pattern is specified in the ‘Unbounded’ field (yes or no).

2.1.4 Phonotactics

This table lists some phonotactic pattern that researchers have described, along with the finite state acceptor that can account for it.

2.1.5 Sources

This table lists the bibliographical information of sources referenced in the database.

2.1.6 Sizes

This table explicates the size-code, as seen in the Languages Primary and Languages Secondary tables. These codes, found in the ‘size’ field, for example “5+”, are elaborated in prose in the ‘English’ field. 5+ would be ‘five or more syllables’, 5- would be ‘less than five syllables’, and just 5 would be ‘exactly five syllables’.

2.1.7 Weights

This table lists the possibilities for the numbers of weights that are distinct in a language. W0 means there are no weight distinctions, W1 means there is a light/heavy distinction, and so on. For each possibility, there is a finite state acceptor in the ‘fsa’ field that generates it.

2.1.8 Notes

This table is a simple list of notes that were made on various languages. These notes are associated with certain languages in the Language Notes table. See 2.2.6.

2.1.9 Syllables

This table indicates types of syllable patterns. It is organized from lightest to heaviest, from W0 (weight zero) to W4 (weight four). The information within the cells refers to syllable structure and consonant and vowel features.

2.1.10 FSA Tables

The following three tables are in actuality three views of one single table. This table contains the distinct stress patterns that exist in the database.

2.1.10.1 Learner Results nd (fsa)

This table displays a list of representative languages (that is, there is one language listed here for each combination of primary and secondary stress patterns in the database), summaries of their primary and secondary stress patterns in the expanded SPCs (see 3) and phonotactic information, then present information about the learners and FSA.

name The ‘name’ field contains the name of the representative language.

main The ‘main’ field contains the SPC(s) for the primary stress pattern(s) and in parentheses, the word size in which the pattern is valid. For example, Bhojpuri’s entry under ‘main’ is “3R (4+), 2R (3-)”, meaning that in words of four or more syllables, primary stress is placed on the antepenultimate syllable, and in words of three or fewer syllables, primary stress is placed on the penultimate syllable.

secondary The ‘secondary’ field contains the SPC(s) for the secondary stress pattern(s) and in parentheses, the word size in which the pattern is valid. For example, Cayuvava’s entry under ‘secondary’ is “None (2-), i3@3R (3+)”, meaning that in words of two or fewer syllables, there is no secondary stress, and in words of three or more syllables, secondary stress is placed iteratively on every third syllable counting from the antepenultimate syllable.

phonotactics Any phonotactic information that is relevant to the representative language is listed in the ‘phonotactics’ field.

The number in the next three fields dictates that we needed to present words with up to this number of syllables in order for the learner to converge to the pattern.

- fl** The number in the ‘fl’ field dictates that the Forward Learner converged to the target pattern when given a sample consisting of all word types with 1 to n syllables.
- bl** The ‘bl’ field dictates that the Backward Learner converged to the target pattern when given a sample consisting of all word types with 1 to n syllables.
- fbl** The ‘fbl’ field dictates that the Forward Backward Learner converged to the target pattern when given a sample consisting of all word types with 1 to n syllables.

The next nine fields have values true or false.

- tail_nd** This field shows a value of true if the tail canonical acceptor is 1-1 neighborhood distinct.
- head_nd** This field shows a value of true if the head canonical acceptor is 1-1 neighborhood distinct.
- tc_mpt_nd** This field shows a value of true if the tail canonical acceptor of the merged prefix tree, given the sample as indicated in the fbl field, is 1-1 neighborhood distinct.
- hc_mpt_nd** This field shows a value of true if the head canonical acceptor of the merged prefix tree, given the sample as indicated in the fbl field, is 1-1 neighborhood distinct.
- tc_mst_nd** This field shows a value of true if the tail canonical acceptor of the merged suffix tree, given the sample as indicated in the fbl field, is 1-1 neighborhood distinct.
- hc_mst_nd** This field shows a value of true if the head canonical acceptor of the merged suffix tree, given the sample as indicated in the fbl field, is 1-1 neighborhood distinct.
- mpt_x_mst_nd** This field shows a value of true if the intersection of the merged prefix tree and merged suffix tree is 1-1 neighborhood distinct.

tc_mpt_x_mst_nd This field shows a value of true if the tail canonical acceptor of the intersection of the merged suffix tree is 1-1 neighborhood distinct.

hc_mpt_x_mst_nd This field shows a value of true if the tail canonical acceptor of the intersection of the merged prefix tree is 1-1 neighborhood distinct.

2.1.10.2 Learner Results gram/prec (fsa)

This table displays a list of representative languages, summaries of their primary and secondary stress patterns in the expanded SPCs (see 3) and phonotactic information, then present information about the learner results. This information is included in the following fields:

name The ‘name’ field contains the name of the representative language.

main The ‘main’ field contains the SPC(s) for the primary stress pattern(s) and in parentheses, the word size in which the pattern is valid. For example, Bhojpuri’s entry under ‘main’ is “3R (4+), 2R (3-)”, meaning that in words of four or more syllables, primary stress is placed on the antepenultimate syllable, and in words of three or fewer syllables, primary stress is placed on the penultimate syllable.

secondary The ‘secondary’ field contains the SPC(s) for the secondary stress pattern(s) and in parentheses, the word size in which the pattern is valid. For example, Cayuvava’s entry under ‘secondary’ is “None (2-), i3@3R (3+)”, meaning that in words of two or fewer syllables, there is no secondary stress, and in words of three or more syllables, secondary stress is placed iteratively on every third syllable counting from the antepenultimate syllable.

phonotactics Any phonotactic information that is relevant to the representative language is listed in the ‘phonotactics’ field.

2gram The number in the ‘2gram’ field indicates the smallest sample size a bigram learner needed in order to converge to the attested pattern. If the entry is 0, this means that no sample size will allow the learner to converge; the learner has failed. If the entry is greater than 0, then the pattern is Strictly 2-Local.

3gram The number in the ‘3gram’ field indicates the smallest sample size a trigram learner needed in order to converge to the attested pattern. If the entry is 0, this means that no sample size will allow the learner to converge; the learner has failed. If the entry is greater than 0, then the pattern is Strictly 3-Local.

4gram The number in the ‘4gram’ field indicates the smallest sample size a 4-gram learner needed in order to converge to the attested pattern. If the entry is 0, this means that no sample size will allow the learner to converge; the learner has failed. If the entry is greater than 0, then the pattern is Strictly 4-Local.

prec The number in the ‘prec’ field indicates the smallest sample size a precedence learner needed in order to converge to the attested pattern. If the entry is greater than 0, then this pattern belongs to the class of Precedence languages.

pr2g If the number in the ‘pr2g’ field is greater than 0, then the language obtained through the intersection of the language obtained with the bigram learner and the language obtained with the precedence learner is the target language.

pr3g If the number in the ‘pr3g’ field is greater than 0, then the language obtained through the intersection of the language obtained with the trigram learner and the language obtained with the precedence learner is the target language.

pr4g If the number in the ‘pr4g’ field is greater than 0, then the language obtained through the intersection of the language obtained with the 4-gram learner and the language obtained with the precedence learner is the target language.

Since the precedence learner returned languages which were strict supersets of the languages returned by the n-gram learners, the prNg fields track the n-gram fields exactly.

2.1.10.3 Distinct Pattern Properties (fsa)

This table displays a list of representative languages, the order in which they are presented in Heinz 2007 (disseration) which is based upon the type of

pattern, summaries of their primary and secondary stress patterns in the expanded SPCs (see 3) and phonotactic information, specifies the type of pattern, (Quantity (in)sensitive, (un)bounded), and presents information about the FSA. This information is included in the following fields:

name The ‘name’ field contains the name of the representative language.

ord The ‘ord’ field indicates the order in which these languages are presented in Heinz 2007 (dissertation).

main The ‘main’ field contains the SPC(s) for the primary stress pattern(s) and in parentheses, the word size in which the pattern is valid. For example, Bhojpuri’s entry under ‘main’ is “3R (4+), 2R (3-)”, meaning that in words of four or more syllables, primary stress is placed on the antepenultimate syllable, and in words of three or fewer syllables, primary stress is placed on the penultimate syllable.

secondary The ‘secondary’ field contains the SPC(s) for the secondary stress pattern(s) and in parentheses, the word size in which the pattern is valid. For example, Cayuvava’s entry under ‘secondary’ is “None (2-), i3@3R (3+)”, meaning that in words of two or fewer syllables, there is no secondary stress, and in words of three or more syllables, secondary stress is placed iteratively on every third syllable counting from the antepenultimate syllable.

phonotactics Any phonotactic information that is relevant to the representative language is listed in the ‘phonotactics’ field.

type The ‘type’ field indicates quantity (in)sensitivity: qi or qs and (un)boundedness of pattern: b or ub.

subtype The ‘subtype’ field indicates which type of pattern the fsa is converging to: single, dual, binary, ternary or multiple. A single stress pattern is one in which there is exactly one stressed syllable per word. Similarly, a dual stress pattern is one in which there are exactly two stressed syllables per word. In binary stress patterns, roughly every other syllable or mora is stressed. And similarly, in ternary stress patterns, roughly every third syllable or mora is stressed. A multiple stress system is one in which multiple stresses are placed, usually on quantitatively more prominent syllables.

tail_nd This field shows a value of true if the tail canonical acceptor is 1-1 neighborhood distinct.

tail_st This field indicates the number of states in the tail canonical acceptor.

tail_tr This field indicates the number of transitions in the tail canonical acceptor.

tail_fi This field indicates the number of final states in the tail canonical acceptor.

head_nd This field shows a value of true if the head canonical acceptor is 1-1 neighborhood distinct.

head_st This field indicates the number of states in the head canonical acceptor.

head_tr This field indicates the number of transitions in the head canonical acceptor.

head_in This field indicates the number of initial states in the head canonical acceptor.

2.2 Linking Tables

2.2.1 Languages Primary

This table links languages in the database with the syllable priority code (see 3) for its primary stress pattern (if any). The size (in number of syllables) of words in which this pattern is valid is indicated in the ‘1sizes’ field (see 2.1.6). The ‘notes’ field is an internal development table that is used to keep track of things.

2.2.2 Languages Secondary

This table links languages in the database with the syllable priority code (see 3) for its secondary stress pattern (if any). The size (in number of syllables) of words in which this pattern is valid is indicated in the ‘2sizes’ field (see

2.1.6). The ‘notes’ field is an internal developemnt table that is used to keep track of things.

2.2.3 Languages Phonotactics

This table associates languages with their relevant phonotactic information. The ‘notes’ field is an internal developemnt table that is used to keep track of things.

2.2.4 Language Sources

This table associates each language with the relevant source id, and in some cases, a source’s page number from which the language information was taken. Full reference text may be found in the Sources table or on the Language Page. The ‘notes’ field is an internal developemnt table that is used to keep track of things.

2.2.5 Language Ethnologue Code

This table associates each language ideally with one ethnologue code in order to aid language identification. Some languages are listed without an associated ethnologue code. In these cases, there was no clear indication of which ELC referred to a language which corresponded the best with the one listed in the database. There were other instances in which there were several ELCs which were possible codes. In those cases, multiple entries were created for the language and one possible ELC is associated per entry. If you have more information about the appropriate ELC for a language, please e-mail Jeff Heinz: heinz@udel.edu.

lg_id The ‘lg_id’ field lists the languages that are in the database.

1id The ‘1id’ field lists the ethnologue code that could be associated with the language in the ‘lg_id’ field.

2.2.6 Language Notes

This table relates languages to the notes that are relevant to them. Topics range from notes on stress patterns, to syllable structure, and some general language notes.

Chapter 3

Syllable Priority Codes

3.1 Bailey's Work

The Syllable Priority Code system was developed by Bailey (1995) as a shorthand for indicating primary stress assignment rules (Heinz, 2007; Bailey, 1995). A brief description of Bailey's system may be found here ([/url-http://www.cf.ac.uk/psych/subsites/ssdb/syllableprioritycode/index.html](http://www.cf.ac.uk/psych/subsites/ssdb/syllableprioritycode/index.html)).

3.2 Heinz's Expansion

Heinz has expanded Bailey's system to include secondary stress systems. Included in the expansion are the symbols *i*, meaning that stress is applied iteratively, *@mL* and *@mR*, which mean that stress is applied to the left counting from the main stress, and to the right, counting from the main stress, respectively, *H*, indicating that stress falls on heavy syllables only, and a system of *Hs* and *Ls* in parentheses to describe foot-based stress patterns. A full review of Heinz's system may be found on pages 192-194 of Heinz 2007 (<http://phonology.cogsci.udel.edu/~heinz/diss/heinz-2007-UCLA-diss.pdf>). These pages have been copied below.

The 'Main' column contains the Syllable Priority Code (SPC), which was developed by Bailey (1995) as a shorthand for indicating primary stress assignment rules. The last character of the SPC (*L* or *R*) indicates from which edge of the word to begin counting. Thus the initial syllable is designated *1L*, the peninitial *2L*, the penultimate *2R*, and the final syllable *1R*. Thus the simplest SPC codes, such as *1L* (Afrikaans), simply mean main stress

falls on the initial syllable.

Generally, more complex SPCs can be read as a series of if-then-else statements. Slashes indicate a quantity-sensitive rule with rules governing heavier syllables occurring left of the slash. Thus the SPC *12/2L* (Maidu) unpacks to the following: If the initial syllable is heavy, it gets stress, else if the peninitial syllable is heavy, it gets stress, else stress falls on the peninitial syllable. If the numbers are suffixed with @s, it means primary stress is assigned if the syllable position carries secondary stress.

Unbounded patterns, where the stress can fall any distance from the word edge, use the *12..89* construct. For example, the SPC for Amele *12..89/1L* unpacks to the following: If the first syllable counting from the left is heavy then it receives primary stress, else if the second syllable counting from the left is heavy then it receives primary stress . . . otherwise (if there are no heavy syllables) the first syllable counting from the left receives primary stress. Since words are unbounded in length, Bailey (1995) uses *..89* to indicate “and so on” in the increasing order for any length. Thus *89* do not literally mean the 8th or 9th syllable. Rather *9* means the farthest syllable from the relevant edge and *8* means the next-to-farthest syllable from the relevant edge and so on. See Bailey (1995) for more details.

SPCs that are followed by $(n+)$ means the code only applies to words that have at least n syllables. Likewise SPCs that are followed by $(n-)$ means the code only applies to words that have at most n syllables.

The ‘Secondary’ column contains extensions I made to the SPC in order to describe secondary stress patterns. ‘None’ of course means that no secondary stress is present. ‘Not included’ means that source material reports secondary stresses, but that either 1) the source material did not describe it, usually because it was deemed too complex, or 2) the source material did describe it, but the pattern was either unclear or too complicated for me to incorporate into the study due to the usual suspect: time.

Since secondary stress patterns are often iterative (that is can be described recursively once the position of one stress is known), I indicate secondary stress patterns that can be described iteratively with the prefix *i-*. The prefix *i2* means the second syllable from a stress receives a stress (in both directions). The first stress is indicated with a SPC suffixed with a @ symbol. Thus *i2@1L* (Bagandji) indicates secondary stresses fall on odd syllables from the left, whereas *i2@2R* (Anejom) indicates secondary stresses fall on even syllables from the right. *@m* means that the first stress upon which the iterative procedure is based is the position of main stress. *@mL*

means the iterations proceeds only leftwards of main stress. Likewise, $@mR$ means the iterations proceeds only rightwards of main stress.

When the secondary stress rules are quantity-insensitive, I use H,L,X to designate heavy, light, and either heavy or light syllables, respectively. Thus a typical trochaic pattern is designated $i('H, 'LL)$ and a typical iambic pattern $i(H', LX')$. If the iterative procedure begins from the word edge (as opposed to from a particular position), I forgoe the connective @ and just suffix L or R to indicate whether the pattern proceeds from the left or right edge, respectively. Thus $i('H, 'LL)R$ (Inga) means trochees are iteratively constructed from the right word edge.

Whenever only heavy syllables bear secondary stress, I indicate this with H . Sometimes it is necessary to explicitly mention that secondary stress only precedes main stress (as in cases describable with foot extrametricality), in which case I use the symbol \downarrow .