MANULEX-INFRA

SHORT DESCRIPTION

You will find at http://leadserv.u-bourgogne.fr/bases/manulex/manulex_infra/index.htm (hereafter Leadserv.u-bourgogne.fr) additional information about the *Manulex-infra* database (Phonological Codes, IPA Correspondences, Syllabification, Graphemic Parsing, Field Description, Statistical Description, …). The files which list phonographic neighbors, homographs, … are also downloadable (see below).

Our quantitative descriptions of the orthographic and phonological characteristics of French words encountered by children in elementary school were based on the ***Manulex*** lexical entries (wordform lexicon) and their corresponding frequency norms (Lété et al., 2004). Our main reason for choosing *Manulex* was the fact that it provides separate lexicon and frequency norms from Grade 1 to Grade 5. Word frequency in *Manulex* was computed separately **as a function of grade level**: grade 1 (CP in French education), grade 2 (CE1) and grades 3 (CE2) to 5 (CM2). A fourth computation index provides word frequency for all grades considered together.

All entries in the ***Manulex-wordform* lexicon** (Lété et al., 2004) were used for computations except abbreviations, euphonic strings, interjections, and compound entries (entries that contain a space, an apostrophe, or a dash). Because many proper names listed in *Manulex* have ambiguous or unknown pronunciations, only those with a frequency value of at least .10 in G1-5 levels were considered in the computations. The total number of entries in G1-5 is 45080. Among these, 10861 occurred in G1, 18131 in G2, and 42422 in G3-5.

The main file **Manulex-infra.xls** at Manulex.org groups together the computations of the four grade-level lexicons.

The computations fall into two categories: **word-length characteristics** and **grade-level characteristics**. The word-length characteristics are the numbers of letters, phonemes, graphemes, and syllables in the word. Contrary to word-length characteristics, grade-level characteristics are function of the word corpus analyzed. They were computed on the four *Manulex-wordform* lexicons corresponding to the four levels, G1, G2, G3-5, and G1-5, that is, words found in first-grade readers, second-grade readers, third-to-fifth-grade readers, and all readers.

Computations are **type-based** and **token-based computations**. Type-based computations are computations made on each word occurring in a lexicon, whatever its lexical frequency. Thus, a common word like "dans" (in) has the same weight as a word rare like "rang" (rank), despite their large difference in frequency. Token-based computations are computations on each word occurring in a lexicon (word type) weighted by its lexical frequency (taken from the *Manulex* U index).

The ambiguity of phonological encoding from orthographic input, and the ambiguity of orthographic encoding from phonological input, are generally estimated by consistency index. In *Manulex-infra*, the **GP-consistency index** is equal to the frequency at which a particular grapheme-phoneme mapping occurs divided by the total frequency of the grapheme, no matter how it is pronounced. For example, the GP-consistency index of the association "ch"->/S/ (as in the word "chat" /Sa/) is obtained by dividing the frequency of occurrence of the "ch"->/S/ association by the frequency of the grapheme "ch", irrespective of its pronunciation (including /S/, but also /k/ for example, as in "choral" /koRal/). The GP-consistency index was then multiplied by 100. Its maximal value (total consistency) is 100. Similarly, the **PG-consistency index** is equal to the frequency at which a particular phoneme-grapheme mapping occurs, divided by the total frequency of the phoneme multiplied by 100, no matter how the phoneme is spelled.

Consistency can differ greatly as a function of the serial position of the units in the word. In particular, due to the derivational morphology of French, word endings are often silent, so spelling is less transparent. To better characterize the orthography-phonology mappings of French, frequency and consistency were computed **as a function of the relative serial position** (initial, middle, final) of the units in the words. The results are included in each word entry, together with the average value and the value of the word's least frequent and the least consistent associations.

The file **Manulex-infra_sublexical tables.xls** at Manulex.org furnishes two sheets (GP sheet and PG sheet) which provide statistics about grapheme-phoneme mappings and phoneme-grapheme mappings, respectively.

Bigrams, biphones, and syllable frequencies were computed for each entry at the four levels. Bigram frequency is the frequency of occurrence of each two-letter sequence in the word list. Transposed to phonology, biphone frequency is the frequency of occurrence of each two-phoneme sequence in the word list. Finally, syllable frequency was computed from the syllabic segmentations of phonological wordforms. Computations were type-based and token-based, and were performed separately for the different units (bigram, biphone, syllable) as a function of their relative serial position in the word (initial, middle, final).

The file **Manulex-infra_sublexical tables.xls** at Manulex.org furnishes supplementary sheets which provide summary statistics on bigrams, biphones, and syllable frequencies. Letter, phoneme, and trigram frequency sheets are also available.

Orthographic and phonographic neighborhood. Lexical neighborhood density was computed to assess lexical similarities between words. Orthographic neighbors are operationally defined as words that can be generated from the base letter string by a single letter substitution. For example, FACE, RICE, RATE, and RACK are orthographic neighbors of the word RACE. Because **orthographic neighborhood density** depends on the specific orthographic wordforms known by the children, values at the four levels were computed separately (first grade, second grade, third to fifth grades, all grades). Neighborhood density is type-based and token-based. The type-based measure corresponds to the raw number of orthographic neighbors. The token-based measure takes neighbor frequency into account by summing the frequency of all neighbors. Note that if two (or more) neighbors are homographs, they are counted only once, and their frequencies are summed.

Adult reading aloud (Peereman & Content, 1997) has been shown to be facilitated by a particular subtype of orthographic neighbor, namely ones that are also phonologically similar to the target word. These lexical neighbors, referred to as **phonographic neighbors,** are both orthographically and phonologically similar to the word to be pronounced. **Phonographic neighborhood density** was therefore computed in addition to orthographic neighborhood. Phonological similarity between words was estimated by applying the orthographic-neighbor operationalization to phonological forms. Hence, words were considered to be phonologically similar when they differed by a single phoneme. The density and frequency computations are incorporated in the main file **Manulex-infra.xls**. At Leadserv.u-bourgogne.fr, you will find the complete lists of orthographic neighbors and phonographic neighbors along with their frequency.

The **number of homophones** and the **number of homographs** for each entry were also computed at the four levels. Again, type-based and token-based computations were performed. Type-based and token-based values were added to the databases. Again, the words entering into the computations (heterophonic HomoGraphs and heterographic HomoPhones) are listed in separate files at Leadserv.u-bourgogne.fr, along with their frequency.

In studies on auditory word recognition, the **phonological unicity point** is traditionally defined as the serial position of the phoneme (counting from the first phoneme in the word) at which the target word diverges from other lexical candidates. Transposed to orthographic forms, uniqueness point refers to the serial position of the letter (counting from left to right) at which the target word diverges from any other lexical candidates. Orthographic uniqueness point is given for each word in each grade level.

MANULEX-INFRA

FIELD ABBREVIATIONS

WORD DESCRIPTIONS

ORTHO :orthographic code

PHON : phonological code (phonetic codes)

SYNT : syntactic class (NC: noun; NP: proper name; VER: verb; ADJ: adjective; PRO: pronoun; PRE: preposition; CON: conjunction; DET: determiner)

U : word frequency per million words = frequency of Usage; see Lété et al. (2004) for computation details

PSYLL : syllabic segmentation with '.' as separators; see Peereman et al. (2007) and Leadserv.u-bourgogne.fr for segmentation details

NBSYLL : number of syllables

GSEG : graphemic segmentation with '.' as separators; see Peereman et al. (2007) and Leadserv.u-bourgogne.fr for segmentation details

PSEG : phonemic segmentation matching the graphemic one

GPMATCH : grapheme-phoneme associations. This field allows to find words including a particular association; "-" between grapheme and corresponding phoneme, "." between grapheme-phoneme associations (e. g., (ch-S.a-a.r-R) for the word 'char' /SaR/). The leftmost character is a '(' that indicates the beginning of the word. The rightmost character is a ')' that indicates word ending. These two characters can be used to find words including grapheme-phoneme association specifically at the begining or at the end of the words (e. g., searching with '(ch-S.' or '.ch-S)' provides the list of words including the ch-S association at the beginning and at the end of the words, respectively.

nbLET : number of letters

nbPHON : number of phonemes

nbGRAPH : number of graphemes

puortho : orthographic unicity point

Homophones and Homographs

nbHPty : number of HomoPhones, Type count

nbHGty : number of HomoGraphs, Type count

nbHPNGty : number of HomoPhones Not homoGraphic, Type count

nbHGNPty : number of HomoGraphs Not homoPhonic, Type count

nbHPto : number of HomoPhones, Token count

nbHGto : number of HomoGraphs, Token count

nbHPNGto : number of HomoPhones Not homoGraphic, Token count

nbHGNPto : number of HomoGraphs not homoPhonic, Token count

Bigrams and Biphones

frBIGtty : mean BIGram frequency, Type count

frBIGtto : mean BIGram frequency, Token count

frBIGity : frequency of the Initial BIGram, Type count

frBIGito : frequency of the Initial BIGram, Token count

frBIGmty : mean frequency of the middle BIGrams, Type count

frBIGmto : mean frequency of the middle BIGrams, Token count

frBIGfty : frequency of the Final BIGram, Type count

frBIGfto : frequency of the Final BIGram, Token count

frBIPtty : mean BIPhone frequency, Type count

frBIPtto : mean BIPhone frequency, Token count

frBIPity : frequency of the Initial BIPhone, Type count

frBIPito : frequency of the Initial BIPhone, Token count

frBIPmty : mean frequency of the Middle BIPhones, Type count

frBIPmto : mean frequency of the Middle BIPhones, Token count

frBIPfty : frequency of the Final BIPhones, Type count

frBIPfto : frequency of the Final BIPhones, Token count

SYLLABLES

frSYLity : frequency of the Inital SYLlable, Type count

frSYLito : frequency of the Initial SYLlable, Token count

frSYLmty : mean frequency of the Middle SYLlables, Type count

frSYLmto : mean frequency of the Middle SYLlables, Token count

frSYLfty : frequency of the final SYLlable, Type count

frSYLfto : frequency of the final SYLlable, Token count

LEXICAL NEIGHBORHOOD

nbONty : number of Orthographic Neighbors, Type count

nbONto : number of Orthographic Neighbors, Token count

nbPGNty : number of PhonoGraphic Neighbors (phonological AND orthographic neighbors), Type count

nbPGNto : number of PhonoGraphic Neighbors (phonological AND orthographic neighbors), Token count

MEAN FREQUENCY AND CONSISTENCY OF GRAPHEME-PHONEME AND PHONEME-GRAPHEME ASSOCIATIONS

frGPtty : mean frequency of Grapheme-Phoneme associations, Type count

frGPtto : mean frequency of Grapheme-Phoneme associations, Token count

coGPtty : mean consistency of Grapheme-Phoneme associations, Type count

coGPtto : mean consistency of Grapheme-Phoneme associations, Token count

coPGtty : mean consistency of Phoneme-Grapheme associations, Type count

coPGtto : mean consistency of Phoneme-Grapheme associations, Token count

Minimal Frequency and Consistency of Grapheme-Phoneme and Phoneme-Grapheme Associations

frGPmity : frequency of the Grapheme-Phoneme associations having the Minimal value on Type count

frGPmito : frequency of the Grapheme-Phoneme associations having the Minimal value on Token count

coGPmity : consistency of the Grapheme-Phoneme associations having the Minimal value on Type count

coGPmito : consistency of the Grapheme-Phoneme associations having the Minimal value on Token count

coPGmity : consistency of the Grapheme-Phoneme associations having the Minimal value on Type count

coPGmito :consistency of the Grapheme-Phoneme associations having the Minimal value on Token count

Frequency and Consistency of Grapheme-Phoneme and Phoneme-Grapheme Associations by Position (initial, middle, final)

frGPity : frequency of the Inital Grapheme-Phoneme association, Type count

frGPito : frequency of the Initial Grapheme-Phoneme association, Token count

frGPmty : mean frequency of the middle Grapheme-Phoneme associations, Type count

frGPmto : mean frequency of the middle Grapheme-Phoneme associations, Token count

frGPfty : frequency of the Final Grapheme-Phoneme association, Type count

frGPfto : frequency of the Final Grapheme-Phoneme association, Token count

coGPity : consistency of the Inital Grapheme-Phoneme association, Type count

coGPito : consistency of the Initial Grapheme-Phoneme association, Token count

coGPmty : mean consistency of the middle Grapheme-Phoneme associations, Type count

coGPmto : mean consistency of the middle Grapheme-Phoneme associations, Token count

coGPfty : consistency of the Final Grapheme-Phoneme association, Type count

coGPfto : consistency of the Final Grapheme-Phoneme association, Token count

coPGity : consistency of the Inital Phoneme-Grapheme association, Type count

coPGito : consistency of the Initial Phoneme-Grapheme association, Token count

coPGmty : mean consistency of the middle Phoneme-Grapheme associations, Type count

coPGmto : mean consistency of the middle Phoneme-Grapheme associations, Token count

coPGfty : consistency of the Final Phoneme-Grapheme association, Type count

coPGfto : consistency of the Final Phoneme-Grapheme association, Token count

Notes

The frequencies of Phoneme-Grapheme associations are equal to the frequencies of Grapheme-Phoneme associations.

Token counts are based on *U* Frequency Index from *Manulex.*