Manulex_Morpho : user manual

The *Manulex_morph o*database (Peereman, Sprenger-Charolles & Massaoud-Galusi, 2013) was generated to examine the contribution of morphological information (mainly inflectional morphology) to the consistency of grapheme-phoneme (GP) and phoneme-grapheme (PG) correspondences in French elementary-school readers. The corpus corresponds to the 9,949 lexical entries of *Manulex* (Lété, Sprenger-Charolles, &Colé, 2004)whose frequency of textual occurrence estimated by the U index was higher than 2.99. For each word, the frequency and the consistency of GP and PG mappings were determined as a function of their relative positions in the word (mappings occurring in initial position, in final position, or in internal positions). Mean consistency and frequency are also estimated without considering relative positions. Several variants of *Manulex_morpho* have been generated; details are provided below.

A list of the different GP and PG associations in the word corpus is given in a separate file, together with their corresponding frequency and consistency values.This list can be useful to estimate consistency of GP or PGmappings in pseudowords or in words not included in the *Manulex_morpho* corpus.This file is described in Annex 1.

The phonetic codes (Annex 2) are similar to those used in Manulex_infra (Peereman, Lété & Sprenger-Charolles, 2007). Four categories of morphological cues were used to mark graphemes and phonemes (Annex 3): gender and number inflections, verbal inflections, final grapheme *–ent* of the adverbial derivation in *–ment*, and graphemes allowing inflection or derivation (e. g., the final « d » in « grand » is silent, but heard in inflected or derived words (grande, grandeur). Details are provided in Peereman et al. (2013).

Manulex_morpho is available in the Excel format.

References.

Peereman, R., Sprenger-Charolles, L., & Massaoud-Galusi, S. (2013).The contribution of morphology to the consistency of spelling-to-sound relations: A quatitative analysis based on French elementary schoolreaders. *L'Année Psychologique.*

Previous developments :

Lété, B., Sprenger-Charolles, L., &Colé, P. (2004). MANULEX : A grade-level lexical database from French elementary-school readers . *Behavior Research Methods, Instruments, & Computers, 36*, 156-166.

Peereman, R., Lété, B., &Sprenger-Charolles, L. (2007). Manulex-Infra: Distributional characteristics of grapheme-phoneme mappings, infra-lexical and lexical units in child-directed written material.*BehaviorResearchMethods,* 39, 579-589

Manulex_morpho

Manulex_morpho provides frequency and consistency values of GP and PG mappings occurring in a set of 9,949 words found in French primary school readers (Lété et al., 2004). The word corpus corresponds to approximately 20% of the lexical entries occurring in Lété et al. (2004) but to 98% of the words encountered by children in their schoolbooks (textual occurrences).Frequency and consistency are determined as a function of the relative position of the GP or PG mappings:initial GP and PG associations, internal ones, and final ones (the grapheme or the phoneme in the initial or final positions is respectively the first and the last in the word, e.g., in the words "ami", "amis", "amie", "amies", the last graphemes are respectively "i", "s", "e", and "s").Frequency and consistency computations were realized taking into account either lexical frequency (i.e. by-type count) or textual frequency (i.e., by-token count) of the GP and PGmappings. In the first case, the frequency of each association is determined by the number of words in the lexical database that include the association.In the second case, the frequency of each mapping is weighted by the textual frequency of the words that include the mapping.

Four different variants of *Manulex_morpho* are available. Differences between variants are related to:

1. the nature of the word frequency index (from Lété et al., 2004) used in the by-token counts
    - U index(frequency per million words weighted by the dispersion of the words across the different books)
    - F index (frequency per million words)
2. how vowelsweredistinguished. Computations were realized either after removal of several distinctions between vowels (/o/and /ɔ/, /e/ and /ɛ/, and /ø/ and /œ/ for PG mappings, and removal of orthographic distinctions resulting from diacritics for GP mappings; e. g. a-â-à, i-ï,…)[1]or preserving these distinctions[2].

[1]See Delattre (1965) for a discussion of the distinctions /o/- /ɔ/, /e/ - /ɛ/, /ø/ - /œ/, and /a/ - /ɑ/, and Walker (1984) for comparison with the French Canadian vocalic system.

[2]The difference between these two approaches can be illustrated with the distinction between /O/ and /ɔ/. According to the first approach ("without distinction"), the consistency of each graphemic association (e.g., o, ô, …) is determined as a function of all graphemes associated with /O/ or /ɔ/. According to the second approach ("with distinction"), consistency values are estimated independently for /O/ and /ɔ/, as it is the case for different phonemes (e.g., /i/ and /O/). In this case, the sum of the consistency values of each grapheme associated to /O/ is 100%, and the sum of the consistency values of each grapheme associated to /ɔ/is 100%. Similarly, for GP consistency, the « without distinction » computations consider that graphemes differing only by a diacritic are identical (e. g., à-â ;é-è-ë-ê). Conversely, the « with distinction » computations preserve the differences and the orthographic variants are considered as distinct graphemes.

The different variants of *Manulex_morpho* are provided in distinct excel files:

Prefered frequency index	**Without or with the V distinctions**	**Filename of *Manulex_morpho***
U	without	Manulex_morpho-U-without
	with	Manulex_morpho-U-with
F	without	Manulex_morpho-F-without
	with	Manulex_morpho-F-with

Information provided in *Manulex_morpho*

- orthographic code of the word
- phonetic code of the word

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

- word frequency (according to *Manulex* ; Lété et al., 2004). Depending on the variant of the *Manulex_morpho* database, the textual frequency index used in the by-token estimations is either U (frequency per million words weighted by the dispersion of the words across the different books) or F (frequency per million words)
- graphemic segmentation of the word ('.' denotes a graphemic boundary)
- phoneme segmentation
- grapheme-phoneme mappings of the word. This column 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 beginning 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.

Word length
- number of letters
- number of phonemes
- number of graphemes
- number of syllables

Frequency of grapheme-phoneme mappings
(*notes*. Depending on the variant of the *Manulex_morpho* database, the textual frequency index used in the by-token estimations is either U (frequency per million words weighted by the dispersion of the words across the different books) or F (frequency per million words); the frequency of Phoneme-Grapheme associations is identical to the frequency of Grapheme-Phoneme associations)

- mean frequency of GP mappings, Type count
- id., Token count

- frequency of the Initial Grapheme-Phoneme association, Type count
- id., Token count

- mean frequency of the Internal Grapheme-Phoneme associations (non-initial and non-final mappings), Type count
- id., Token count

- frequency of the Final Grapheme-Phoneme association, Type count
- id., Token count

Consistency of grapheme-phoneme mappings
(*notes*. Depending on the variant of the *Manulex_morpho* database, the textual frequency index used in the by-token estimations is either U (frequency per million words weighted by the dispersion of the words across the different books) or F (frequency per million words)

- mean consistency of GP mappings, Type count
- id., Token count

- consistency of the Initial Grapheme-Phoneme association, Type count
- id., Token count

- meanconsistency of the Internal Grapheme-Phoneme associations (non-initial and non-final mappings), Type count
- id., Token count

- consistency of the Final Grapheme-Phoneme association, Type count
- id., Token count

Consistency of phoneme-grapheme mappings
(*notes*. Depending on the variant of the *Manulex_morpho* database, the textual frequency index used in the by-token estimations is either U (frequency per million words weighted by the dispersion of the words across the different books) or F (frequency per million words)

- meanconsistency of PG mappings, Type count
- id., Token count

- consistency of the Initial Phoneme-Grapheme association, Type count
- id., Token count

- meanconsistency of the Internal Phoneme-Grapheme associations (non-initial and non-final mappings), Type count
- id., Token count

- consistency of the Final Phoneme-Grapheme association, Type count
- id., Token count

References

Delattre P. (1965). *Comparing the phonetic features of English, French, German and Spanisch*.Heidelberg :Jumius Gross Verlag.
Walker, S. (1984). *The pronunciation of Canadian French*. Ottawa : University of Ottawa Press.

Annex 1. Consistency of GP and PG mappings

Consistency and frequency values of the GP and PG mappings, estimated as a function of the morphological cues, are provided in the file « mappings ». Frequency and consistency are determined as a function of the relative positions of the mapping in the word (mappings occurring in initial position, in final position, or in internal positions). Consistency and frequency values are also determined without considering relative positions

The file « mappings » includes 4 different sheets:

- Sheets 1 and 2 :GP and PG values as a function of the textual frequency index U (frequency per million words weighted by the dispersion of the words across the different books) of Manulex (Lété et al., 2004).
- Sheets 3 and 4 :GP and PG values as a function of the textual frequency index F (frequency per million words) of Manulex (Lété et al., 2004).

In each case, consistency values are computed in two different ways :

1) After removal of the distinction between the vowels après /o/and /ɔ/, /e/ and /ɛ/, and /ø/ and /œ/ for PG consistency, and after removal of the distinctions beween letters differing only by diacritics(a-â-à, i-ï,…) for GP consistency. These values appear in the **columns labeled *« without distinction of V »***

2) Keeping the distinctions between the V. **Columns *« with distinction of V »***

Summary

Preferedfrequency index	**Frequency/consistency**	**Without or with the V distinctions**	**Where can I find the information ?**
U	GP	without	sheet 1, « without distinction of V »
		with	sheet 1, « with distinction of V »
	PG	without	sheet 2, « without distinction of V »
		with	sheet 2, « with distinction of V »
F	GP	without	sheet 3, « without distinction of V »
		with	sheet 3, « with distinction of V »
	PG	without	sheet 4, « without distinction of V »
		with	sheet 4, « with distinction of V »

Annex2. Phonetic codes and corresponding computer characters

Phonetic codes corresponding to standard computer characters were used to ensure compatibility across platforms. The characters are the same as those used in *Manulex_infra*, except that the distinction between the obligatory schwa (coded with the "%" character; e.g., "e" in "table" because of resyllabification) and the optional one (coded with the "°" character; e.g., "gare") was introduced. The IPA values corresponding to the computer characters are given in the table. Note that a hatch mark (#) is also used in grapheme-phoneme mappings to indicate a silent grapheme (e.g., the grapheme "t" at the end of a word "fort"; the mute "e" in "fée" or "année").

IPA	Codes	Examples
Vowels		
i	i	lire, vie
u	u	joue, ours
y	y	bulle, sud
e	e	fée, nez
ɛ	E	jouet, aile
a	a	date, plat
ɑ	A	tâche, bois
ø	2	deux, peu
œ	9	neuf, fleuve
ə	%	renard, chèvre
	°	adulte, cerise
ɔ	O	roche, sol
o	o	jaune, mot
ɔ̃	§	nom, pont
ɛ̃	5	cinq, plein
ɑ̃	@	vent, blanc
œ̃	1	un, brun
Glides		
j	j	feuille, lieu
w	w	soie, watt
ɥ	8	huit, fruit

IPA	Codes	Examples
Consonants		
p	p	loupe, pain
t	t	terre, vite
k	k	qui, bec
b	b	cube, brosse
d	d	danse, aide
g	g	gare, bague
f	f	foule, phare
s	s	tasse, cerf
ʃ	S	chat, vache
v	v	vent, rêve
z	z	zéro, rose
ʒ	Z	gel, juge
m	m	main, femme
n	n	nage, laine
ɲ	N	ligne, peigne
l	l	lune, pull
ʀ	R	rue, air
ŋ	G	viking, ring

Annex3. Morphological cues

Computer codes used in Manulex_morpho are described below. See the manuscript for details on the different morpho-graphemes(Peereman, Sprenger-Charolles & Massaoud-Galusi, 2013).

The characters 3, 4, 6 and 7 are used to code morphemic units (note that the other number-characters are used to code phonemes; see Annex 2).

The code "3" is used to code gender and number inflections. For example, the word "photos" is coded using the final morpho-grapheme "3s" which is associated to the phonological code "3#" (the "#" character indicates a silent grapheme; see Annex 2). Conversely, in the word "parfois", the final "s" does not correspond to a number inflection and it is coded "s", associated with the phonological code "#".

The code « 4 » is used to code verbal inflections. For example, the word "aiderai" is coded using the final morpho-grapheme "4ai" which is associated to the phonological code "4E". In the word "vrai", the "ai" grapheme does not correspond to a verbal inflection and it is therefore coded "ai", associated to "E".

The code «6 » is used to code inflectional and/or derivational markings of word finals. For example, the final "s" of the word "gris" is silent but heard when the word is inflected ("grise") and in derivatives ("griser"). The "s" grapheme is thus coded as "6s", associated to the phonological code "6#". Similarly, the silent "d" of the word "grand" is silent but heard in "grande" or "grandeur". Consequently it is coded as "6d". The full list of morpho-graphemes is provided in the manuscript (Peereman, Sprenger-Charolles&Massaoud-Galusi, 2013).

The code « 7 » is used to cue the « ent » grapheme occurring in adverbs ending in « -ment ». For example, the grapheme "7ent" is associated to the phonological code « 7@ » in the adverb « simplement ».