MANULEX

FIELD ABBREVIATIONS

The lemma and wordform lexicons are included in two sheets. The wordform lexicon yields all possible inflected words reduced to their lemmas in the lemma lexicon (the singular for nouns and adjectives, the infinitive for verbs). Words found in each grade are in columns. The cell is empty when the word has not been found in reading books of that grade.

G1 : Words found in 1st-grade reading books (CP in French Primary Schools)

G2 : 2nd-grade (CE1)

G3 : 3rd-to-5th grade (CE2 – CM1 – CM2)

G1-5 : 1st-to-5th = whole *Manulex* textbook corpus

NLET : **Number of Letters**

SYNT : Syntactic Category : NC-noun, NP-proper name, VER-verb, ADJ-adjective, ADV-adverb, PRO-pronoun, PRE-preposition, CON-conjunction, INT-interjection, DET-determiner, ABR-abbreviation, and UEUPH-euphonic string

F : **Frequency** : the number of times the word occurs in the grade corpus. Caution: F is not frequency per million; use U for the correct estimation of frequency

D : **Dispersion index** (.00 to 1.00 range) based on the dispersion of the frequencies across readers. D is equal to .00 when all occurrences of the word are found in a single reader, regardless of the frequency (F). It is equal to 1.00 if the frequencies are distributed in exactly equal proportions across readers. Values between .00 and 1.00 indicate degrees of dispersion between these extremes.

U : **Estimated Frequency of Usage per million words** : U is derived from F with an adjustment for D. When D equals 1, U is computed simply as the frequency per million words. But when D is less than 1, the value of U is adjusted downward. When D is 0, U has a minimum value based on the average weighted probability of the word across all readers. **It is believed that U is a better reflection** of the true frequency-per-million that would be found in a corpus of an indefinitely large size, thus permitting direct comparisons to values given by the four sub-corpora.

SFI : **Standard Frequency Index** : derived directly from U and therefore has some of U's characteristics. The user should find this index to be a simple and convenient way of indicating frequency counts, once it is understood. A wordform or a lemma with an SFI of 90 is expected to occur once in every 10 words; one with an SFI of 80 can be expected to occur once in every 100 words, etc. A convenient mental reference point is an SFI of 40, the value for a wordform or lemma that occurs once in a million words. SFI is computed from U using the formula:

$$SFI = 10 * (\log_{10}(U) + 4)$$