WordGen: A Tool for Word Selection and Non-Word Generation in Dutch, English, German and French

Wouter Duyck[1], Timothy Desmet[1], Lieven Verbeke[1], & Marc Brysbaert[2]
[1] Department of Experimental Psychology, Ghent University, Belgium
[2] Royal Holloway University of London, UK

E-mail: wouter.duyck@UGent.be

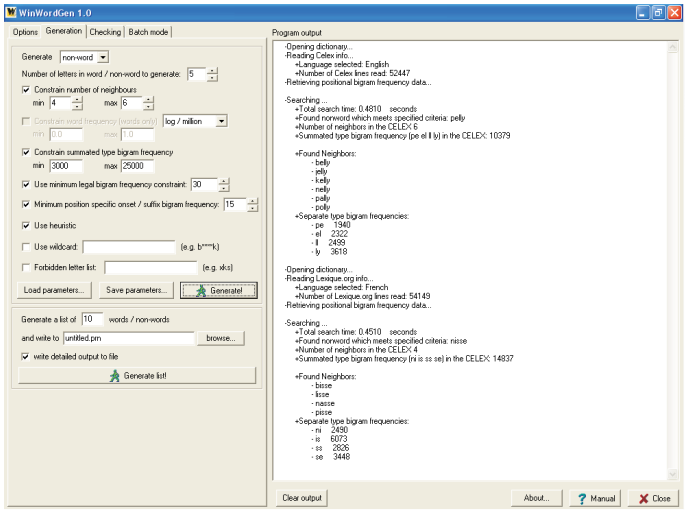
WordGen program and poster available at:
http://expsy.ugent.be/wordgen.htm
Manual and theoretical framework:
Behavior Research Methods, Instruments & Computers, 36(3)

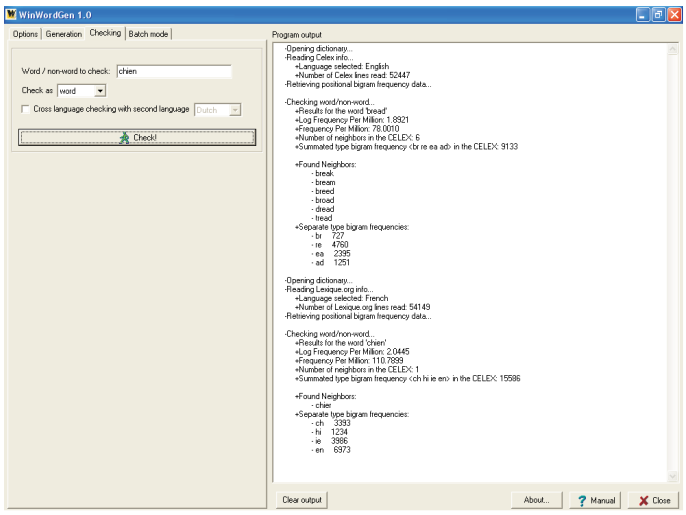
Abstract

WordGen is an easy-to-use program that uses the CELEX and Lexique lexical databases for word selection and non-word generation in Dutch, English, German and French. Items can be generated in these four languages, specifying any combination of seven linguistic constraints: The program also has a module to calculate the respective values of these variables for items that have already been constructed (either with the program or taken from earlier studies). Stimulus queries can be entered through WordGen's graphical user interface, or by means of batch files. WordGen is especially useful for (1) Dutch and German item generation, because no such stimulus selection tool exists for these languages, (2) the generation of non-words for all four languages, because our program has some important advantages over previous non-word generation approaches and (3) psycholinguistic experiments on bilingualism, because the possibility of using the same tool for different languages increases the cross-linguistic comparability of the generated item lists. WordGen can be downloaded freely from the following URL: http://expsy.ugent.be/wordgen.htm.

Features

- Word selection:
  - Dutch, English and German: word selection with a simple graphical user interface (GUI) from the CELEX lemma database (Baayen et al., 1993, 1995)
  - French: GUI word selection from the Lexique.org lemma database, which is compiled from Frantext (New et al., in press)
- Non-word generation: random or heuristic non-word generation. Wordlikeness and pronounceability can be manipulated through several parameters. With optimized parameters, WordGen's random letter generation algorithm produces pronounceable non-words at a rate of 80% (up to 8 letters).
- (Psycho)linguistic parameters: word selection and non-word generation satisfying any combination of the following constraints:
  - word length
  - neighborhood size
  - frequency per million words (plain/logarithmic; words only)
  - summated type bigram frequency
  - mimimum (onset/suffix) type bigram frequency
  - use heuristic non-word generation (change 1 random letter from existing words)
  - position-specific letter inclusion and exclusion
- Checking feature: within- and cross-language computation of the above parameters for existing sets of words and non-words for Dutch, English, German and French
- Batch mode: Although WordGen is designed to provide an easy-to-use 'click and retrieve' GUI for word selection and nonword generation, repetitive queries can be highly automated using the batch mode feature. Complete stimulus sets can be generated without human intervention, using a simple syntax which is entered through batch files or the command line.





Contributions to the field

- Dutch and German psycholinguistic studies:
  - No published counts of neighborhood size and bigram frequency
  - No published non-word generation tools or databases
- French psycholinguistic studies:
  - WordGen is complementary to the Lexique database, which does not contain bigram frequency measures
  - No published non-word generation tools or databases
- English psycholinguistic studies:
  - Word selection: WordGen is complementary to the MRC psycholinguistic database (Coltheart, 1981), which does not contain bigram frequency measures or neighborhood size counts
  - Non-word generation: WordGen is complementary to the ARC nonword database (Rastle et al., 2002), that only contains multisyllabic non-words, which are also always pseudohomophones or very wordlike non-words.
- Studies on bilingualism: using the same norms and selection tool for different languages increases cross-language comparibility
- WordGen allows manipulating and measuring the 'wordlikeness' of non-words for all four languages
- Easily extendable to other languages for which a reliable lexical database is available (e.g. the Spanish LexEsp corpus is currently being integrated into WordGen)

References

- Baayen, R.H., Piepenbrock, R., & van Rijn, H. (1993, 1995). *The CELEX lexical data base (CD-ROM)*: Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.
- Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33(A)*, 497-505.
- Duyck, W., Desmet, T., Verbeke, L.P.C., & Brysbaert, M. (in press). Wordgen: A Tool for Word Selection and Non-Word Generation in Dutch, English, German and French. *Behavior Research Methods, Instruments, & Computer, 36(3)*.
- New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (in press). Lexique 2.5: A New French Lexical Database. *Behavior Research Methods, Instruments, & Computer, 36(3)*.
- Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology, 55(A)*, 1339-1362.