Software paper for submission to the Journal of Open Research Software

To complete this template, please replace the blue text with your own. The paper has three main sections: (1) Overview; (2) Availability; (3) Reuse potential.

Please submit the completed paper to: editor.jors@ubiquitypress.com

---

Title

A Word Association Data Processor to facilitate robust and consistent categorisation and analysis of word associations.

Paper Authors

Andreas Buerki

Paper Author Roles and Affiliations

corresponding author, *Centre for Language and Communication Research*, *Cardiff University*, *Wales*

Abstract

The Word Association Data Processor (WADP) is an open-source, free software package which automates key aspects of the categorisation and analysis of word association data gathered from respondents in word association tasks. The user base of this software is expected to be linguists and others working with word association data and employing a methodology similar to that presented in Fitzpatrick et. al. (2015). The WADP offers three modules. The categoriser module provides an interface for the manual categorisation of word association responses and the automatic categorisation of responses in cases where categorisations for the relevant cue-response pairs are found in a database of past categorisations. It also facilitates the automatic storage of all new ratings in the database and the tracking of respondent IDs (and categoriser/rater IDs if provided) in all input and output files. The reporter module allows the automatic creation of individual response profiles, cue profiles and primary response profiles. Finally, the administrator module performs a number of housekeeping functions such as the merging of two database files (including the resolution of conflicting entries) and producing database files from categorised output data.

Keywords

word association, *mental lexicon*, *psycholinguistics*, *lexical relations*

(1) Overview

Introduction

The Word Association Data Processor (WADP) is a unique piece of open-source software that automates key aspects of the processing and analysis of word association data gathered from respondents in word association tasks. It does so in a way that is designed to bring significant advantages to working with and analysing word association data, compared to current approaches, while also benefiting from the stability and robustness that comes from having been applied, tested and improved over the course of over five years in successive research projects.

The background to what word association is and how it is used in research is covered in the first section below, followed by a section that sets out how the WADP approaches the task of supporting the analysis of word association data, the reasons for its approach and the significance of its benefits. The details of implementation and architecture are then presented in their own section, followed by a section on quality control. A section on availability lists the technical specifications before the final section looks at ways the software can adapted for use in a range of scenarios.

Word Association

A typical word association task involves subjects being given a series of cue words to which they are asked to name (or write down) the first word (or words) that come to mind. For example, given the cue ‘boy’, one participant might respond with ‘girl’, another might respond with ‘band’ or any other response. Such word association data have been used in psychology, psycholinguistics and related fields to study a wide range of topics including the minds and inner lives of individuals [1, 2], the organisation of words and concepts in the mind [3, 4, 5], the linguistic properties of words and classes of words [6, 7], cultural differences [8, 9, 10], first language learning in children [11, 12], differences in linguistic knowledge between first and second language speakers [13, 14] or consumer perceptions [15]. [16] provides an overview.

Despite the diversity of applications, the basic format of word association data (consisting of a number of cues, and participants’ responses to each cue) is widely shared, as are the basic ways in which word association data are analysed. One way to analyse associations is to use norms lists (i.e. a corpus of normative responses) and calculate stereotypy scores that indicate the degree of similarity of an individual's or a group's responses to a given norms list. There are published norms lists (e.g. [17, 18]) and alternatively, any comparison group’s responses can be used

as a norm or benchmark. There have been warnings, however, about biases if comparison groups are not carefully matched for a range of factors such as age, educational background or gender [19]. Another way to analyse responses is to categorise them and derive profiles for individuals or for cues on the basis of relative frequencies of response categories, rather than individual response words. For example, building on a distinction going back to Saussure [20], the *boy* → *girl* association instantiates a paradigmatic relationship (*boy* and *girl* are alternatives that can take up the same slot in a sentence such as 'the ___ fell over'). On the other hand, *boy* → *band* instantiates a syntagmatic relationship in that these words typically occur next to each other in a sentence rather than being alternatives. Adult L1 users appear generally to show a preference for paradigmatic responses [21], but the strength of this preference may differ and certain cues favour one or the other response category. Typically, however, more elaborate categorisation schemas are used in word association research, such as the fourteen basic categories proposed by Fitzpatrick et al. [19] – a categorisation schema shown to result in stable individual response profiles across time. Yet other analyses focus on primary responses (i.e. the most frequent responses given to a cue) in a certain data set or compare primary responses of one data set with those of another, or are interested in mapping out networks of association among words and concepts, via visualisation and/or probability tables [13, 22] or association chains, where responses become cues for further responses [13].

Word Association Software

There are a number of software implementations that aim to automatically generate word associations (i.e. word association responses to input cues) available from open repositories [23], as paid-for APIs [23], or commercial services [25, 26]. Others provide word-association games or tests of various descriptions, typically used to gather word association data (e.g. [27, 28]). There is also a range of APIs and libraries that offer lexicographical information which could be accessed to retrieve, for example, lexical relations of a cue word, such as synonyms, antonyms or collocations or to check if a response to a cue falls into one of the mentioned categories as shown by Gaume et al [29]. Similarly, there are a range of implementations following Church and Hanks [30] that derive relative associative strength based on textual co-occurrence. This is useful, for example, to an OCR software algorithm in deciding which of two or more possible forms is more likely correct in a certain context. An alternative approach is employed by De Deyne et al. [22], who provide a collection of R scripts that derive associative strengths based on large data sets of word-associations provided by humans rather than derived from textual data. The scripts also facilitate visualisations, cue statistics, response chaining and other analysis tools at https://github.com/SimonDeDeyne/SWOWEN-2018. Meara [13] mentions a web-based service, 'WA_Sorter', which sorts data into a format that lists responses in

descending order of frequency under each cue word thus facilitating a primary response analysis, but this tool seems no longer available. At present, word association analyses involving categorisations of cue-response pairs are still largely processed and analysed manually or with the help of general office applications like spreadsheets [19]. Especially when large amounts of data are processed in this way, this is not only highly work-intensive but also prone to errors, accidental data corruption and inconsistencies ('error-prone and tedious' according to Meara [13, p 159]). Inconsistencies can occur in the categorisation of cue-response pairs when identical pairs in the same data set receive different categorisations during data processing. Even if done consistently, keeping track of how identical cue-response pairs have been categorised manually adds significantly to the time required to score responses and derive results.

The Word Association Data Processor (WADP) was developed to address these difficulties, initially encountered in the course of a research project that required the processing of thousands of word association responses (https://www.alzheimers-brace.org/cardiff-university-prof-alison-wray/). The WADP facilitates the efficient and consistent categorisation of cue-response pairs by providing a convenient interface for manual categorisation that presents cue-response pairs to raters though a text-based interface. Raters' categorisations are automatically stored in a database and if an identical cue-response pair is encountered, the programme pulls the appropriate categorisation from the database, presenting only novel cue-response pairs to the rater. Pre-existing database files can also be used and included in the WADP is a large database of previous categorisations by researchers at Cardiff University's Centre for Language and Communication Research which can be used to substantially speed up the categorisation of cue-response pair data sets that employ the categorisation scheme proposed by Fitzpatrick et al. [19] and (some of) the same cues as those contained in the database (see the software repository on full details). Because the correct categorisation of a given cue-response pair is not always clear, rigour and consistency is typically ensured by obtaining independent categorisations by at least two different raters and comparing the results [7, 19]. The WADP facilitates this by recording the ID of each rater against their categorisation in output and database files and by providing an administrator module which facilitates the comparison (and where required the resolution) of differences between two sets of categorisations. It is also possible to automatically calculate inter-rater agreement measures.

The approach taken in the WADP is to facilitate a manual categorisation (only repeats are scored automatically) rather than attempt to construct automatic categorisations of unseen cue-response pairs. Although such automation is feasible at least in part, each specific categorisation scheme employed in the field of word association research (and tweaks to existing schemes) would require a separate

implementation whereas the WADP is entirely flexible with regard to categorisation schemes, thus making it easily adaptable for specific purposes (see the manual for how to adjust the categorisation scheme via two simple plain text edits in the code). Another important reason why automated categorisation is not pursued is that there is some evidence that different data sets may require different categorisation rules due to the sensitivity of word associations to demographic and other factors. For example, a response of ‘Tintern’ to the cue ‘Abbey’ would require a categorisation as erratic in most contexts, whereas in participants familiar with the South Wales locality, it should be categorised as a reversed syntagmatic combination (‘Tintern Abbey’ is a mediaeval ruin in the South Walian village of Tintern).

The featured automatic categorisation of repeats, however, vastly increases the efficiency of manual classification: in a sample of 300 responses to each of 100 cues (30,000 cue-response pairs), on average less than a third of responses to each cue (31%, range: 13–53%) had to be manually categorised, the remainder being repeats of already categorised pairs. This figure did vary between individual cues, as the indicated range shows. Figure 1 plots the progressive percentage of response tokens to the cue 'snap' that require manual input as the categorisation progresses in successive blocks of 25 cue-response tokens, based on data from 300 responses. Overall, 31% of response tokens to this cue required manual categorisation. As can be seen, among the first 25 categorised cue-response pairs the rate of required manual categorisation already dropped to below 50%. Among cue-response tokens 25 to 50, the rate decreased further to well below 40% and this was followed by a broadly downward trajectory in each subsequent batch of 25 cue-response token as fewer and fewer previously unseen cue-response pairs were encountered. Consequently, after the initial build-up of categorisations from zero, only rare and unusual responses require manual classification.

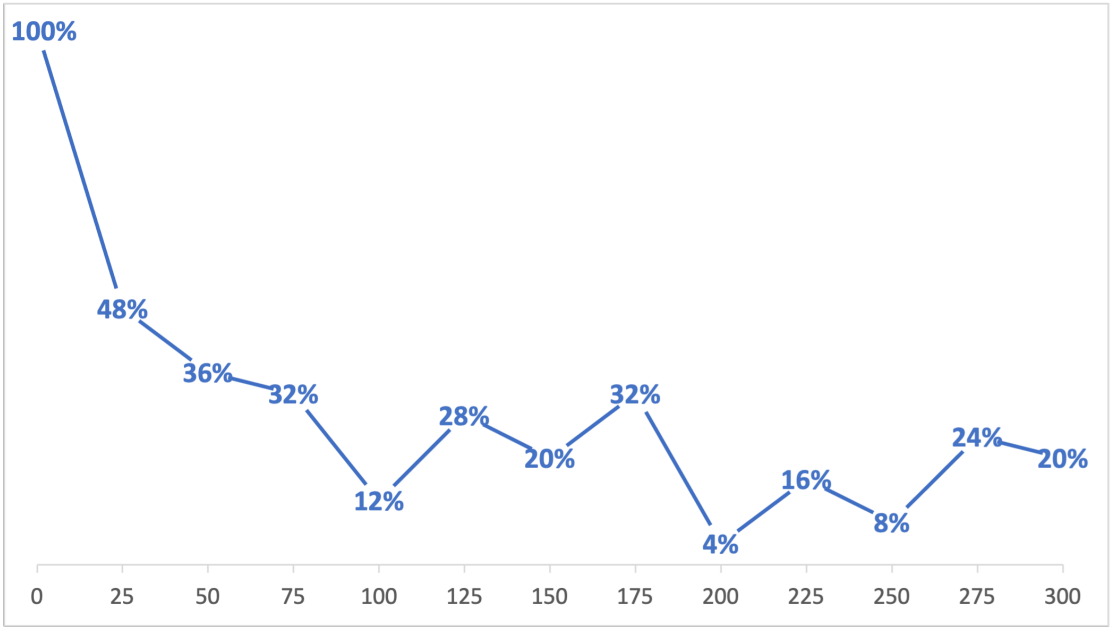

Figure 1 Progressive rate of response tokens to cue ‘snap’ that required manual categorisation in a sample of 300, in successive windows of 25 responses.

In addition to categorisation-related functions, the WADP automates the creation of three types of response profiles as detailed in the next section: category-based response profiles for individual participants, category-based response profiles for individual cues, and primary response profiles by cue. The latter does not require previous categorisation as it is based on the most frequent response words themselves. Automatically created profiles can then be compared to other profiles, including reference profiles, or used as variables in other analyses.

Overall, the WADP offers functionality that significantly enhances word association data analysis in terms of efficiency, accuracy and consistency in categorisation and category-based profile creation, a use case that has up to now had to rely on error-prone manual or semi-manual processing. The next section details the modular architecture of the WADP.

Implementation and architecture

The WADP is implemented via a shell script written in Bash v. 3.2 and makes use of a terminal window to present a text-based, interactive user interface, allowing a clean and efficient user experience while making minimal demands in terms of compatible operating systems and hardware. The WADP's architecture comprises three main modules that are accessible from a central menu.

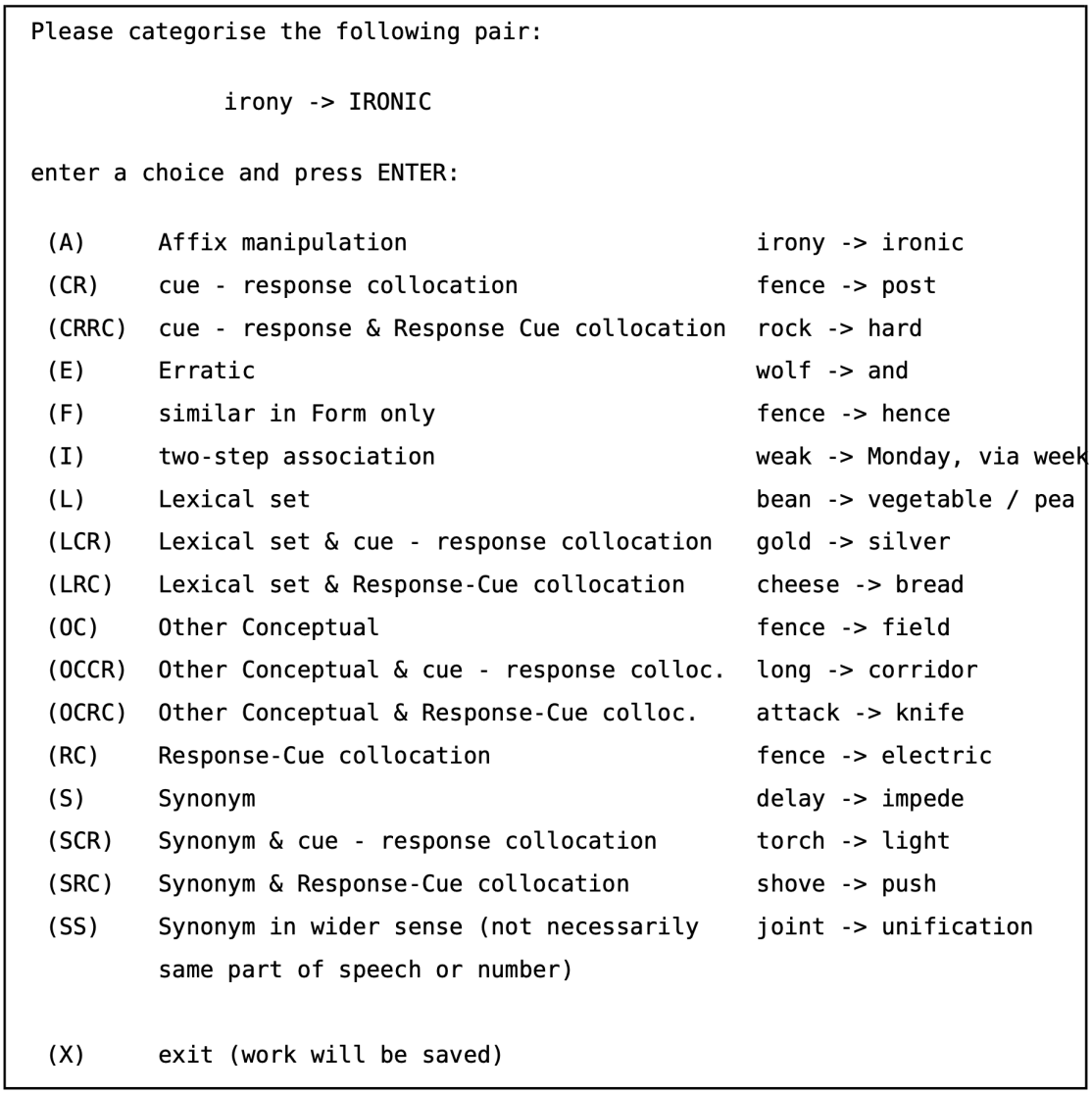

Figure 2. Categoriser interface

Categoriser

The categoriser module manages the manual categorisation of cue-response pairs as well as the automatic categorisation of repeats or pairs that are already contained in an optionally supplied database file. The main categorisation interface is shown in Figure 2. The specific categorisation scheme is editable to fit any application. The categoriser takes as input a csv file which contains cue words in the header row, followed by responses to the cues in the remaining rows. Optionally, the first column can contain

participant IDs. No internal line breaks are allowed in csv-fields, but otherwise the standard comma-separated values (csv) format [31] is accepted. As shown in Figure 3, a database of existing categorisations can optionally be supplied as well. The plaintext format of the database file is documented in the manual and can be manually edited, though this is not usually necessary. The output of the categoriser module consists of a database file (either new or an updated version of the input database file) and an output csv file which is of the same format as the input csv file, with a column inserted after each cue column that lists the category of cue-response association and a further column that lists the rater ID of the researcher who categorised each cue-response pair.

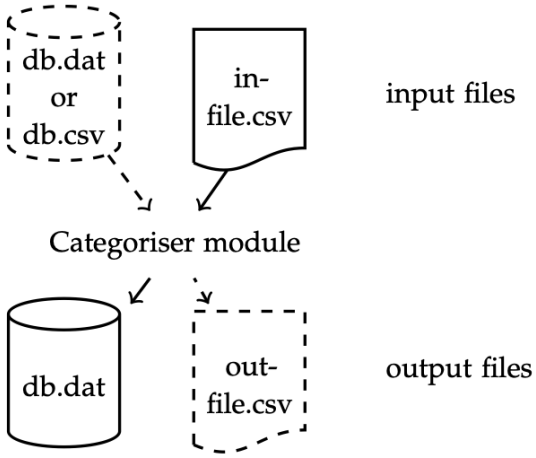

Figure 3. Input and output files of the categoriser module. Note: optional files in dashed outline

Reporter

The reporter module takes as input a categorised csv file (the output file of the categoriser module) and produces one or more of four different reports or profiles: Individual response profiles list the frequencies of different cue-response categories across all cues per input line (each input line typically corresponds to one participant) as a csv file. Cue profiles list the category frequencies by cue rather than by participant. Primary response profiles (which do not require categorised input files and can also operate with the same input files as the categoriser module) produce a csv file showing each response type and its frequency below each of the cue words. Finally, if an inter-rater agreement procedure (as outlined in the manual) was used to

categorise input data, then an inter-rater agreement report can be produced which shows the degree of agreement between raters. The input and output files of the reporter module are shows in Figure 4.

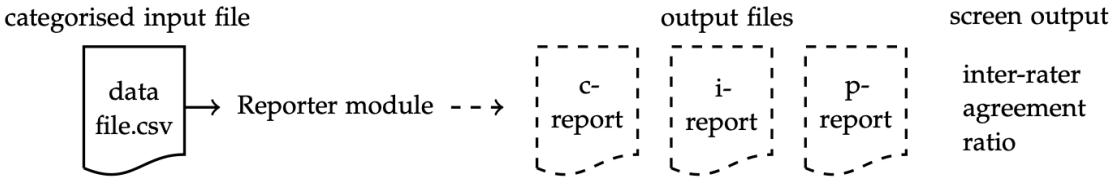

Figure 4. Input and output files of the reporter module.

Administrator

The third module provides a range of administrative functions related to database files:

- turning a rated csv file (produced as output of the categoriser module) into a database file
- producing a list of differences between two database files showing differences in cue words, in responses to cues (where cues are shared) and in categorisations of cue-response pairs (where cue-response pairs are shared across databases).
- combining two different database files into a single database file (if there are conflicts, the values of the first database file are used).
- combining two different database files, but where the two files contain conflicting information (such as alternative categorisations of the same cue-response pair), the user is given the option to prioritise the first or the second database file, to provide a new categorisation or to delete the entry from the combined database file.

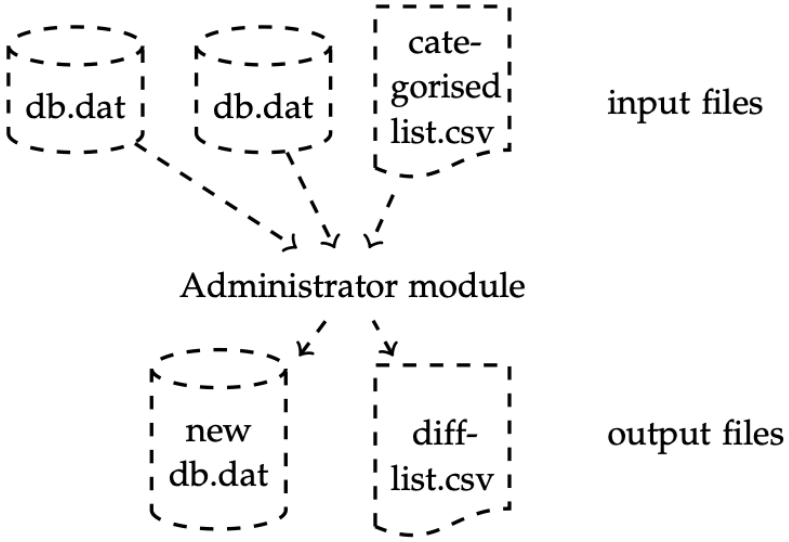

Figure 5. Input and output files of the administrator module.

Quality control
The WADP has been developed, tested, improved and used over successive research projects at the Centre for Language and Communication Research at Cardiff University since 2014. Well over 60,000 cue-response pair tokens have been processed over this time and the resulting categorisation database file is distributed together with the WADP for use by other researchers. Consequently, the WADP is now a mature tool that has benefited from the rigours of use and testing over many years and large amounts of processed data. The README file included with the software describes a series of tests to confirm the correct working of a particular installation of the software using test data.

(2) Availability

Operating system
Linux (tested: Ubuntu 20.04 LTS), macOS (tested: macOS 11), Windows via WSL (tested: Windows 10 via WSL running Ubuntu 20.04 LTS).

Programming language
Bash v3.2

Additional system requirements

none.

Dependencies
Under Microsoft Windows: Windows Subsystem for Linux (WSL) with Ubuntu

List of contributors
none.

Software location:

Archive (e.g. institutional repository, general repository) (required – please see instructions on journal website for depositing archive copy of software in a suitable repository)

Name: Zenodo
Persistent identifier: http://doi.org/10.5281/zenodo.593662
Licence: EUPL v1.2 or later
Publisher: Andreas Buerki
Version published: *1.0*
Date published: 12/09/21

Code repository (e.g. SourceForge, GitHub etc.) (required)

Name: GitHub
Identifier: https://github.com/buerki/WADP
Licence: EUPL v1.2 or later
Date published: 12/09/21

Language
English

(3) Reuse potential
The software can be used in any situation where word-association data require categorisation and analysis to produce response profiles, and indeed in any task that requires the manual classification or categorisation of word pairs. Off the shelf, the WADP displays the categorisation scheme of [19] as shown in Figure 2 during the categorisation task. Other categorisation schemes can easily be substituted by editing a plain-text section of the programme as detailed in the included manual. Different schemes include e.g. [7]: Entity Features (e.g. giraffe → long neck), Situation Features (sofa → cat), Taxonomic Categories (including superordinates, subordinates, coordinates, etc.), Introspective Features (e.g. wasp → annoying) and Lexical Features (including orthographic similarity, e.g. wine → whine), all of which can easily

be accommodated via a simple edit of the code.

As the software is licensed under an open-source licence, further adaptations are also possible. Feature requests and problems or issues can be submitted by raising an issue at the WADP's GitHub repository. The distribution includes a detailed manual and links to video tutorials covering installation and use of the software.

Acknowledgements
I wish to acknowledge members of successive research projects who have facilitated the development of this software. They include, in particular, Alison Wray and Tess Fitzpatrick who have been influential in defining the original feature specifications of the software, Rosie Dymond, Sam Collins, Michael Willett, Lauren Tate, Bilyana Zdravkova and David Schönthal who have applied and tested the software and whose input has led to many improvements. Thanks are also due to the Prevent Dementia project (preventdementia.co.uk) in collaboration with whom much of the data processed by the software in the course of its development were collected. Acknowledgements do not necessarily signal endorsement.

Funding statement
Projects during the course of which the WADP was developed and improved were funded in part by BRACE Dementia Research (https://www.alzheimers-brace.org/cardiff-university-prof-alison-wray/), The British Academy/Leverhulme (SRG / 171398) and Wellcome Trust ISSF Humanities Collaborative Award.

Competing interests
The author declares that they have no competing interests.

References
1. Jung, CG 1910 The association method. *The American journal of psychology* 21(2):219-69.
2. Merten, T 1995 Factors influencing word-association responses: A reanalysis. *Creativity Research Journal* 8(3):249-63.
3. De Deyne S, Verheyen S, Storms G 2015 The role of corpus size and syntax in deriving lexico-semantic representations for a wide range of concepts. *Quarterly Journal of Experimental Psychology* 68(8):1643-64. DOI: http://dx.doi.org/10.1080/17470218.2014.994098
4. Prior A, Bentin S. 2008 Word associations are formed incidentally during

sentential semantic integration. *Acta Psychologica* 127(1):57-71. DOI: http://doi.org/10.1016/j.actpsy.2007.01.002
5. Kostova Z, Radoynovska B 2008 Word association test for studying conceptual structures of teachers and students. *Bulgarian Journal of Science and Education Policy (BJSEP)* 2(2):209-31.
6. Thwaites P 2020 Does verb transitivity influence word association responses? *The Mental Lexicon* 15(3):464-84. DOI: https://doi.org/10.1075/ml.20019.thw
7. De Deyne S, Storms G 2008 Word associations: Network and semantic properties. *Behavior Research Methods* 40(1):213-31. DOI: https://doi.org/0.3758/BRM. 40.1213
8. Szalay LB, Deese J 1978 *Subjective meaning and culture: An assessment through word associations*. New York: Lawrence Erlbaum Associates.
9. Son JS, Do VB, Kim KO, Cho MS, Suwonsichon T, Valentin D 2014 Understanding the effect of culture on food representations using word associations: The case of “rice” and “good rice”. *Food Quality and Preference* 31:38-48. DOI: https://doi.org/10.1016/j.foodqual.2013.07.001
10. Shin JE, Suh EM, Eom K, Kim HS 2018 What does “happiness” prompt in your mind? Culture, word choice, and experienced happiness. *Journal of Happiness Studies* 19(3):649-62. DOI: https://doi.org/10.1007/s10902-016-9836-8
11. Brown R, Berko J 1960 Word association and the acquisition of grammar. *Child development* 31(1): 1-4.
12. Entwisle DR 1966 *Word Associations of Young Children*. Baltimore, MD: Johns Hopkins University Press.
13. Meara P 2009 *Connected Words: Word Associations and Second Language Vocabulary Acquisition.* Amsterdam: John Benjamins.
14. Fitzpatrick T 2009 "Word Association Profiles in a First and Second Language: Puzzles and Problems" In: *Lexical processing in second language learners*. Bristol: Multilingual Matters. pp. 38-52. DOI: https://doi.org/10.21832/9781847691538-006
15. Roininen K, Arvola A, Lähteenmäki L 2006 Exploring consumers' perceptions of local food with two different qualitative techniques: Laddering and word association *Food Quality and Preference* 17(1-2):20-30. DOI: https://doi.org/10.1016/j.foodqual.2005.04.012
16. Fitzpatrick T, Thwaites P 2020 Word association research and the L2 lexicon. *Language Teaching* 53(3):237-74. DOI: https://doi.org/10.1017/S0261444820000105

17. De Deyne S, Storms G 2008 Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods* 40(1):198-205. DOI: https://doi.org/10.3758/BRM.40.1.198
18. Nelson DL, McEvoy CL, Schreiber TA 2004 The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* 36(3):402-7. DOI: https://doi.org/10.3758/BF03195588
19. Fitzpatrick T, Playfoot D, Wray A, Wright MJ 2015 Establishing the reliability of word association data for investigating individual and group differences. *Applied Linguistics* 36(1):23-50. DOI: https://doi.org/10.1093/applin/amt020
20. de Saussure, F 1974 [1916] *Course in General Linguistics*. London: Peter Owen.
21. Cramer P 1968 *Word Association*. New York: Academic Press.
22. De Deyne S, Navarro DJ, Perfors A, Brysbaert M, Storms G 2018 The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior research methods* 51(3):987-1006. DOI: https://doi.org/10.3758/s13428-018-1115-7
23. Kim S. Auto Word Association [software] available at https://github.com/soyeonkimgithub/AWA
24. Rotmistrov YA Word Associations Network API. [online] https://wordassociations.net/en/api
25. Visuwords Visual Dictionary, Visual Thesaurus, Interactive Lexicon. [online] https://visuwords.com
26. Thinkmap Visual Thesaurus [online] https://www.visualthesaurus.com
27. De Deyne S, Storms, G. Word Association Study [online] https://smallworldofwords.org/en
28. Holliday S. Word Association [online] https://wordassociation.org
29. Gaume B, Tanguy L, Fabre C, Ho-Dac L, Pierrejean B, et al. 2018 Automatic analysis of word association data from the Evolex psycholinguistic tasks using computational lexical semantic similarity measures. *13th International Workshop on Natural Language Processing and Cognitive Science (NLPCS)*, Krakow, Poland. Available at https://hal.archives-ouvertes.fr/hal-01881336
30. Hanks P, Church KW 1990 Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22-9.
31. Shafranovich Y 2005 Common Format and MIME Type for Comma-Separated Values (CSV) Files [online] available at https://datatracker.ietf.org/doc/html/rfc4180
