doi:10.5128/ERYa14.14

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 14, 229–245

COMPILING THE DICTIONARY OF WORD ASSOCIATIONS IN ESTONIAN: FROM SCRATCH TO THE DATABASE

Ene Vainik

Abstract. The present paper describes the project titled "The Dictionary of Word Associations in Estonian" undertaken by the author at the Institute of the Estonian Language. The general aim of the Dictionary is to provide insights into Estonians' common-sense mind. It is meant to be a tool of self-reflection for Estonian native speakers and a guide for the foreigners who are eager enough to make themselves familiar with the Estonian cultural patterns of thought. The Dictionary will be published online. The number of keywords was initially limited to approximately 800. Specific emphasis is given to the stage of data collection by implementing the principles of citizen science.*

Keywords: word association, mental lexicon, lexicography, e-dictionary, citizen science, crowdsourcing, Estonian

1. Introduction

The idea of studying word associations was expressed in the Institute of the Estonian Language (Tallinn) for the first time in relation with a grant application proposal targeting the user's structure of Estonian vocabulary[1]. It was proposed that besides the indisputable theoretical interest in the so-called mental lexicon (Aitchison 2012, Pavlenko 2009), a systematic and explicated dictionary-like resource containing information about the "hidden connections" between lexical units would enjoy a great deal of practical relevance both for the general audience and for professionals in different fields.

The general aim of the "Dictionary of Estonian Word Associations" (DEWA) is to provide insights into Estonians' common-sense mind. It is meant to be a tool of self-reflection for Estonian native speakers and a guide for the foreigners who are eager enough to make themselves familiar with the Estonian cultural patterns of thought. One of its general aims is, thus, to provide a tool for facilitating cultural integration.

* This study was supported by the base funding of the Institute of the Estonian Language from the Estonian Ministry of Education and Research.

[1] https://www.etis.ee/Projects/iutapplications/Display/37e415b4-af05-466c-8e20-86ab08c5be1d (11.10.2017)

From the professional point of view, there are also several fields of application. The established interconnections between lexical units would be helpful, for example, for lexicographers, in the course of compiling thesauri (e.g. Wordnet, see Fellbaum 1999, Orav, Vider 2005). Next, the list of word associations and their relative prominence is revealing of the words' senses and their cognitive salience, which is helpful for lexicographers when it comes to arranging the sense menu for an entry in an explanatory dictionary.

Another field of application is L2 teaching. Although nothing can replace a real teacher in a real class, there will be more and more applications on the market providing helpful tools for L2 learners; a dictionary of word associations may easily be one such tool. It has been found that a smart way of improving one's mental lexicon to the level of a native speaker would be to train oneself, consciously, to restore words in one's memory together with the connections that are similar to the L1 speakers' associations (see e.g Peppard 2007). Yet another field of application is compiling L2 course books – it makes sense to follow the structure of the natural "memory chunks" existing in the shared cultural subconsciousness of the language community and organise the material in the course books in a way that mimics the framed structure of the native speakers' common sense knowledge (for *frame*, see e.g Fillmore 1985).

Applicability in the real-life linguistic tasks that the Institute of the Estonian Language is in charge with is important, as the Institute is striving to complete its tasks as the maintainer of the Estonian language and the developer of its supporting tools in an ever better and more up-to-date way. Providing support and smart tools for L2 learners is one of the prospects that the Institute is about to follow in the near future[2].

The potential relevance of DEWA reaches beyond this, though. Once the network of word associations has been explicated in an easy-to-access electronic way, the data can be used by professionals of different fields: language technology, anthropology, linguistics, psychology, sociology, literature, journalism, marketing research, advertising etc.

2. Background

2.1. Definitions of some basic notions

The term *association* refers to a generic psychological notion meaning 'mental connection between concepts, events, or mental states that usually stems from specific experiences' (Klein 2012).

The term *word association* (WA) has a narrower scope being defined as 'a person's lexical response to a lexical stimulus' (Krasnõh 2001). A trivial example of WA is such: if one says "cat" the reply might be "dog", or if the stimulus would be "bread" the response could be "butter", see also Figure 1. A word association consists of two lexical entities: a stimulus and the response(s). As such, it refers to a psycholinguistic kind of phenomenon that can be explicated by testing people. The term *word association* should not be mixed up with the *collocation patterns* of words,

[2] The program "Digipööre" 'Digital Focus': https://www.hm.ee/en/activities/digital-focus (10.10.2017)

as established by statistical analysis of text corpora (which sometimes is the case, see e.g the Word Association Network[3] or the paper by Church and Hanks (1989).

A common *word association test* (WAT) is a test of *free associations*, in which case the person being tested is provided a list of trigger words and he/she would respond as quickly as possible with word(s) that come to his/her mind (Nelson et al. 2000).

A *dictionary of word associations* (DWA) is *a lexicographic term* referring to a substantial collection of word associations (couplings of stimuli and their responses) that have been gathered from a sufficient number of testees (at least 100 per stimuli). In that respect, DWA differs from the notion of *association norms*, as developed in the field of psychology (see the details in the next section).

The terminology used for the structural components of word association varies in the literature of different fields. See Figure 1 for clarification. *A* is referred to by the term *stimulus* or *cue* in the context of psycholinguistic testing or alternatively by *headword or keyword* in the context of lexicography. *B* is referred to as a *list of responses* in the context of testing but may occur as *list of associations* in the context of lexicography and in general usage.

The present paper contains descriptions of both the keyword selection process and the testing procedures, therefore the reader will find different terms used in this paper as well.

A	B	C
UKS 'door'	*aken* 'window'	126
	link 'doorhandle'	33
	lahti 'open'	13
	maja 'house'	12
	valge 'white'	10
	lukk 'lock'	9
	ava 'doorway'	7
	kinni 'closed'	5
	tuba 'room'	5
	ukselink 'doorhandle'	5
	korter 'appartment'	4
	võti 'key'	4

D
305

Figure 1. Responses elicited in association with the Estonian stimulus *UKS* 'door'.
A – stimulus; B – responses listed in hierarchical order, C – raw frequency;
D – number of people tested

[3] https://wordassociations.net/en/ (10.10.2017).

2.2. Theoretical background

The concept of an "association of ideas" originates in the age of Enlightenment, as do the "types" of associations such as *resemblance, contiguity in time or place, and relations of cause and effect*[4]. The roots of testing, i.e studying people's verbal responses to certain stimuli, date back to the 19th century when the English explorer, amateur scientist, and psychologist Sir Francis Galton (1879) pioneered this technique in the field of criminology.

The method was further developed in the beginning of the 20th century and applied in psychiatry as one of the "projective" techniques revealing the client's subconscious ideas and connections (e.g. Jung 1910, Stevens 1994). Besides the interest in the individual differences in reaction times to different stimuli, the normative value of associations also became topical. The first list of association norms was published by Grace H. Kent and Aaron J. Rosanoff (1910). Their original list of 100 stimuli has been translated, adapted and tested in many languages, Estonian included (Toim 1980). The standard procedure included eliciting one response per stimulus by 1,000 respondents under the conditions of time pressure. The list of association norms is meant to be the baseline for detecting individual differences and was used in psychiatry and psychometric testing. The peak time of conducting association studies in psychology was during the sixties and seventies of the last century (see e.g Rosenzweig 1961, Kiss et al. 1973, Postman, Keppel 1970). An example of a relatively late collection of association norms for English is the "The University of South Florida word association, rhyme, and word fragment norms"[5] (Nelson et al. 2004).

More recently, WAT has become topical in psycholinguistics. The studies have adopted a lexical focus, and have investigated the development and organization of the mental lexicon and the influence of specific variables on lexical access (Fitzpatrick et al. 2015). In applied linguistics, the interest has most often been in the integration of L2 items into the lexicon, and the ways in which WA responses might reflect the development of L2 proficiency (many authors; for an overview, see Meara 1982, 2009). The theoretical prerequisite for the studies in both psycholinguistics and in applied linguistics is the existence of the so-called *mental lexicon* – the "store of words in one's mind" with its internal structure and connections where the associations have a central role to play.

The concept of mental lexicon (in terms of associatively interconnected lexical units) – has been extended from individuals also to group behaviour (e.g Fitzpatrick et al. 2015). It has been claimed that free association reveals commonality within a social unit that arises because of similar experience (Nelson et al. 2000). Moreover, a shift has been made to collective behaviour and understanding of the world: it has been claimed that a vocabulary of a language owes its natural semantic structure to associations between the lexical units (Morkovkin 1970). A network model as the organizing principle of lexical semantics has made its way into textbooks (e.g. Cruse 2000). According to some theorists, the associative structure of the vocabulary of a given language is believed to reflect the naïve world-view of the speakers (Apresjan 2000). This is the main theoretical prerequisite for compiling word association dictionaries (see e.g. Buk 2009).

[4] https://www.britannica.com/topic/association-psychology (2.10.2017).
[5] http://w3.usf.edu/FreeAssociation/ (10.10.2017).

In order to provide a reflection of the naïve picture of the world, the set of keywords in a DWA has to be representative of the central semantic fields of the given language, which dictates that the number of stimuli has to be much larger than the sets included in the classical association norms. Another distinctive feature of a DWA is its purely descriptive nature: it is meant to provide information about the associative habits of the given language community during a certain time span. It is not normative in any sense. There exists no such thing as "the correct response" to any of the given stimuli, although there do exist typical responses, which are supposed to reveal the relevance of the underlying concepts for the national-lingual-cultural community (Krasnyh 2001: 39).

An example of such a dictionary is *Russkij associativnyj slovar'* (RAS) 'the Russian dictionary of associations', the authors of which claim that it reflects the mental-emotional state, verbal memory and linguistic consciousness of an average Russian-speaking person. The dictionary project lasted for ten years during the 1990s, and it contains responses of 11,000 Russian speaking college students elicited in response to 7,000 stimuli. As a result, it contains 105,000 different WAs, which makes 15 per keyword, on average. This is a dictionary printed on paper, which makes the lookup of further associations and/or keywords not very convenient for the user.

Another example of a large-scale collection of WAs is a web based project titled "Dutch Word Associations". The online dictionary contains more than 10,000 stimuli together with the responses, also elicited mostly from college students[6]. This project has also lasted over a decade, and its characteristic feature is that the number of keywords keeps cumulatively growing – the responses occurring in high frequency are fed back as cues into the inquiries. Another distinctive feature is that instead of just one response, three responses are collected from the testees. The average number of different associations per cue is 11. The lookup works in both ways: from cue to associations and from associations to cues. Unfortunately, the cues occurring in the lists of responses are not directly linked; surfing in the network of associations is not as easy, yet, as it could be.

The need to add different new kinds of information to dictionaries has been accepted in lexicographic thought (e.g. Atkins, Rundell 2008). As a lexical resource in its own right, a DWA belongs rather to the periphery of the field. In comparison with the prototypical dictionaries such as bilingual, etymological or explanatory dictionaries, it provides a rather restricted range of information. Typically, the structure of a DWA consists of only the keywords (ordered alphabetically), each being accompanied by the list of its associations in the decreasing order of their elicitation frequency. The entries have, thus, a hierarchical structure but no other types of information (e.g part of speech, grammar, examples of usage etc). Figure 1 illustrates an entry of a DWA.

In the case of an e-dictionary, links can be added between the circular items, i.e a link from an item in the list of responses back to the keywords that have evoked this particular item. As such, a DWA enjoys a structure similar to a thesaurus, except that the nature of the links/associations is not explicated in terms of the semantic relations (synonymy, antonymy etc). One of the biggest databases of word associations was titled the Edinburgh Associative Thesaurus[7] (Kiss et al. 1973).

[6] They report about 9,000 cues on the web page http://www.kuleuven.be/semlab/interface/index.php (27.9.2017) and more than 12,000 in Deyne et al. (2013).

[7] EAT: http://www.eat.rl.ac.uk (retrieved 16.5.2016); unfortunately, the web link no longer works.

Summing up the theoretical background one can say that there are three main theoretical prerequisites for compiling a DWA: i) words can be characterised by the associations that they make with other words in the mental lexicon; ii) explicating the associations in a generalised way is a relevant lexicographic task; ii) associations can be detected only by psycholinguistic testing.

3. The Dictionary of Estonian Word Associations (DEWA)

3.1. The goals

DEWA is planned to be an electronic DWA which is representative of the central semantic fields in Estonian. The number of keywords was initially limited to approximately 800, because it is a small-scale low-budget project carried out by one person[8]. As one of the dictionary's general aims was to create a useful tool for cultural integration, it was suggested that the content of the dictionary (i.e the list of keywords) would reflect the most topical concepts in Estonian culture. The dictionary was planned to exist only in electronic form and not to be printed. Therefore, searchability on both the fields of the keywords and the associations was desired.

3.2. The completed stages of the project

3.2.1. Preliminary stage (3.3.2016–1.7.2016)

The preliminary stage involved studies of the theoretical background, formulating the principles and making a plan for compiling the dictionary and carrying out the pilot of gathering associations via a web questionnaire[9].

As a result, it became clear that recruiting a sufficient number of testees would be the challenge for the project. A simple calculation showed that with the minimum rate of 100 responses per stimuli, by distributing the keywords into tests of 100 stimuli in each, at least 800 respondents are needed for the whole project. The commercial poll company which was contacted[10] was uninterested in cooperating because of the low-budget nature of the project. The only affordable solution seemed to be to create a *citizen science/crowdsourcing* type of campaign in order to recruit the volunteers.

Citizen science and *crowdsourcing* are terms that are sometimes used interchangeably. Both are endeavours where ordinary people are invited to contribute in gathering data for a scientific study or forming another type of big data pool (see e.g Cohn 2008). Compiling a web-dictionary is a good example: via specific web platforms, volunteers can be invited to select keywords, formulate definitions, provide examples, cross-edit the entries and to do other lexicographic tasks (Čibej et al. 2015). The two characteristic features of *crowdsourcing* are: splitting the process into microtasks that can be completed with little effort, and 2) gamification – a further emphasis on pleasure rather than effort (Benjamin 2015).

[8] Duration: 3.3.2016–30.6.2019.
[9] The pilot was done by Anna-Liisa Männik (2016), as her Master's Thesis, co-supervised by the author.
[10] Kantar EMOR, see http://www.emor.ee (10.10.2017).

The pilot test has shown that filling in one WAT test with 100 stimuli via the web would take approximately 15–30 minutes, which did not seem like a micro-task. Therefore, the term *citizen science* seemed to match the task better. The idea was to invite people to become partners who would contribute not just once but every now and then. In that way the absolute number of people involved could be reduced to a manageable size.

3.2.2. Preparatory stage (1.7.2016–8.10.2016)

The preparatory stage involved compiling the list of keywords and distributing them between the tests, as well as creating a special web page with the background information of the project and the links to the tests.

In order to make sure that the list of keywords would represent all the necessary semantic fields, a decision was made to compile the initial list of keywords by reducing the pre-existing list of keywords that were originally selected for the Basic Dictionary of Estonian[11]. The list of its content words (N = 3015), originally selected on the basis of frequency in the corpus and the inclusion of some closed semantic categories (such as numerals, days of the week, months etc), has also been subjected to the detection of emotional valence (Vainik 2012).

The procedure of reducing was done as follows. The original list of candidate words (N = 3015) was tagged according to formal and semantic criteria. The formal criteria included the complexity of the word, its length in characters, part of speech and frequency in the corpus. The semantic criteria included emotional valency and tentative semantic category (ascribed by the author during the tagging). The proportions of the classes in the original list were calculated and applied to the desired length of the list of keywords (800). The actual selection of keywords took place keeping in mind the achieved proportional model. The result of the selection (N = 1500) was cross-checked in respect of avoiding redundancy (e.g. of two related words *sõbralik* 'friendly' (adj) and *sõbralikkus* 'friendliness' (noun) only the adjective was selected) and, at the same time, keeping the semantic categories as complete as possible (e.g. the terms for food, clothing, parts of buildings, names of occupations, colour terms etc). The result was a list of 978 words, which was 22% more than was initially planned. The proportions in terms of emotional valency, part of speech, word length etc are roughly in accordance with the parameters of the original word list.

The selected words were distributed into ten WA tests, each containing 97–98 items[12]. The order of the stimuli in the tests was randomised with respect to the semantic categories. A test consisted of four parts: a) an instruction that contained an example of the word association; b) the stimuli, one per sheet (placing something was obligatory); c) some general questions about the respondent (age, gender, occupation etc – all on the same page); d) the "Thank You!" page.

Keeping in mind the other goal mentioned above – the need to include the words that stand for the most topical concepts in Estonian culture – a special inquiry was made in order to find out what those words are, and, if they were not included yet in the list of keywords, to make sure that they would be.

[11] This is a student's dictionary http://www.eki.ee/dict/psv/index.cgi (10.10.2017), for a description of the project see Kallas and Tuulik (2011).

[12] 100 stimuli in a test is the standard, generally followed in WAT since the early times (Jung 1910, Kent, Rosanoff 1910).

The inquiry was carried out in the form of a web questionnaire[13] of free elicitation of the terms "related to Estonia and Estonians", followed by 13 more specific elicitation tasks such as "Estonian nature", "Estonian culture", "important persons", "important events", "Estonian customs", "language", "the character traits of a typical Estonian", "politics in Estonia", "values" etc (see Vainik 2017). All the responses over the 14 tasks were summarised (N = 1809) and the list of results was ordered according to decreasing frequency. The list of the responses occurring three or more times (N = 157) was compared with the existing list of keywords, and, if missing, the items were included. Two additional tests were created, ca 60 stimuli in each. In sum, the number of the keywords in DEWA is 1100, which is 38% more than was planned initially.

Another task during the preparatory stage was creating a special web page for the project[14]. This was needed in order to carry the recruiting process out as a proper Citizen Science campaign. It is a simple site (connected to the web page of the Institute of the Estonian Language) that contains very basic information about: i) Citizen Science; ii) the kind of projects one can contribute to (including a link to the DEWA project). As an example of a typical task, there was a link to the short inquiry of "vocabulary related to Estonia" (see above).

The page of DEWA[15] gives information about the basic notions (such as *word association* and *dictionary of word associations*) as well as information about the relevance of the project and about the person in charge. There were two WAT's ready for taking placed on a separate subpage (titled "The tests").

There was also a call to register oneself as a partner of citizen science. The registration form included some general questions (e.g. age, gender, education, occupation, how often one is ready to contribute and whether one would prefer to be an anonymous contributor or let his/her name be published in the list of contributors). One had a choice between whether to register oneself as a partner and fill in his/her identification number every time or to just take the test and leave no mark of identification.

3.2.3. The stage of Citizen Science partnership (8.10.2016–1.3.2017)

The campaign for registering as a partner of Citizen Science and filling in the WATs was launched in the traditional media (on a TV morning program[16]) and on social media (the Facebook page of the Institute of the Estonian language).

Table 1 shows the progression of the ten general tests, starting in October 2016. The first test was started by 603 and finished by 433 respondents, which shows that people were willing to contribute but the test proved to be too demanding (either in terms of difficulty or time consumption), so that roughly one third gave up in the first test. The average time spent on the test (01:39:23) showed that some people did not follow the instruction to give spontaneous responses. After removing the responses that were done too slowly (more than 45 minutes) or too quickly (less than 5 minutes), the average time spent on the test appeared to be approximately 17 minutes, which was reasonable. Out of 391 responses between 5 and 45 minutes,

[13] The questionaire is available online https://www.surveybuilder.com/s/uyJ47 (10.10.2017).
[14] http://portaal.eki.ee/aitakaasa.html (10.10.2017).
[15] http://www.eki.ee/~ene/kodanikuteadus/assotsiatsioonid.html (10.10.2017).
[16] See the last minutes of the clip http://etv.err.ee/v/meelelahutus/terevisioon/terevisiooni_lood/902c6f8e-3bdb-4f6a-9b68-3e9862e18520/ilmus-raamat-eesti-tunded-sonaportreed-see-on-raamat-eesti-keele-tundesonadest-nende-tahendusest-seostest-ja-sonade-paritolust (10.10.2017).

Table 1. Participation in the series of tests

No of the Test	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
a) Date of opening	11.10.2016	11.10.2016	21.10.2016	21.10.2016	9.11.2016	9.11.2016	5.12.2016	5.12.2016	9.1.2017	9.1.2017
b) Date of closing	3.1.2017	10.1.2017	4.12.2016	3.2.2017	14.12.2016	30.12.2016	25.1.2017	2.2.2017	7.2.2017	7.2.2017
c) Started by (no of persons)	607	448	343	352	403	380	482	341	377	360
d) Completed by (no of persons	433	366	304	311	331	327	381	305	339	317
e) Average time	01:39:23	2:10:57	0:32:51	0:31:19	0:54:27	2:45:38	1:56:13	0:58:27	1:19:01	7:54:15
f) No of responses between 5-45 minutes	391	338	292	296	307	300	355	284	316	294
h) Average time of those between 5–45 minutes	0:17:36	0:15:10	0:16:03	0:27:09	0:15:37	0:14:48	0:15:47	0:13:59	0:17:39	0:16:37
i) No of responses imported to the database	388	336	292	289	305	293	355	283	319	297
j) % of successful completion	64	75	85	82	76	77	74	83	85	83
k) % of responses given by registered partners	23	–	48	55	69	62	64	67	79	77
l) % of occasional contributors	77	–	52	45	31	38	36	33	21	23

Note: data about the registration numbers in Test 2 is missing due to a mistake made by the author.

388 were successfully completed, for an overall success rate of 64%. The average completion time remained the same throughout the ten tests, but the rate of successful completion consistently increased (row j in Table 1), peaking at 85% by test no 9. The two "topical" (Estonia-related) WATs were carried out last and the timing of those was set in the end of February, close to the Estonian Independence Day (the 24th of February). By the end of February 2017, the period of data gathering was over.

One can notice a correlation between successful completion and the ratio of responses given by the partners. It is clear that: a) completing a couple of tests motivated people to register themselves as partners and to continue contributing; b) registered partners behaved in a more responsible way in terms of following the instructions and completing the tasks every time they had started.

By the end of the testing period the number of registered partners had increased to 414. Their age range is 15-90 (average 41.7; StDev 12.07). There is a bias towards female (88%) and highly educated persons (80%). Occupation-wise, however, the ratio of language-related professions (such as Estonian language teacher, translator, lexicographer or editor) was only 31%. One could have suspected these jobs prevailing because of the "language-friendly" channels of social media where the information was distributed in the first place.

The partnership relation with the contributors was maintained deliberately by the project leader. Every time there were new tests launched (once per month), all the registered partners received a personalised (by their ID number) invitation to participate via e-mail. They were also informed about the progress of the project and about the publications in the media[17] and the radio broadcasts[18]. The questions and comments of the partners were replied to as quickly as possible. The partners had the feeling of working in close partnership with the project leader. Their motivation was definitely of the psychological kind (see Čibej et al. 2015: 72).

The cycle of Citizen Science ended with a feedback questionnaire. Quite a few suggestions were made with regard to the technical details of the web questionnaire format. Many of the partners mentioned that they would like to continue contributing in such projects. At the end of the cycle the project leader asked the partners for permission to publish their names on the project's web page. Quite many of those who had begun as anonymous partners were willing to see their names on the list of contributors[19].

However, the range of contributors was also intentionally kept open throughout the whole testing period in order to include more variance and reliability in the data. New contributors were invited each time a new test was launched, mostly by creating "Facebook events of testing" and sharing the invitations via social media. The percentage of random contributions kept decreasing (row l in Table 1), approaching 20% in the last two tests.

In conclusion, the data collection was performed as a combination of partnership and random contributions. The majority of the data was elicited from partners who contributed more than once in the tests.

[17] Postimees, https://tartu.postimees.ee/3917465/keeleteadlane-uurib-meie-peas-paukselt-tekkivaid-sonaseoseid (10.10.2017).

[18] https://arhiiv.err.ee/vaata/keelekorv-keelekorv-ene-vainik (10.10.2017) and http://vikerraadio.err.ee/v/toovari/saated/897bc2ac-40bb-491b-ac0d-345e2e9f3656/toovari-ene-vainik#comments (10.10.2017).

[19] http://www.eki.ee/~ene/kodanikuteadus/partnerid.html (10.10.2017).

3.2.4. The stage of controlling for the effects of sociodemographic factors (1.12.2016–1.1.2017)

It was clear from the very beginning that women were more eager to contribute than men. Therefore, a statistical control[20] for the effect of gender (and some other variables) on the stereotypicity of the responses was carried out after completing the first test. The purpose of the control was to find out whether something should be changed in the recruitment procedures in order to have a more gender-balanced pool of respondents.

The procedures went as follows. First, the responses given to a set of stimuli (N = 50) in the first test were compared to the dominating responses in each column and the verbal data were replaced with numerical values representing similarity with the prevailing responses (5 points were given for matching the most frequent response, 4 points for matching the second-most frequent etc. The cells with no matching responses were filled with zeros).

The number of responses in Test 1 was sufficient (N = 433) after the first cleaning of the results in order to select subgroups that differed in one criterion only (e.g. gender) but were comparable in other respects. Such subgroupings were made (by using the "matching technique") for gender and education (the latter normalised as "higher" vs "not higher" and profession (normalised as "language-related" vs "not language-related") and the data ranges were subjected to a T-test.

The results showed that gender had no statistically significant effect ($t = -0.15$, $p = 0.88$) nor had the age of the respondents or their professional occupation. A significant effect was found only with regard to education ($t = -2.32$, $p = 0.021$). Those with higher education tended to give more stereotypical responses.

The project leader admitted in light of this statistical analysis that the method of data gathering could bring together a data pool biased toward a higher level of stereotypicity than was possibly characteristic of the general population. As the purpose of the project was to compile a dictionary representing steady and typical connections between words and not a set of association norms in the psychological sense, this result was taken as an advantage, encouraging continued work with the citizen science partners.

3.3. The database

The results of the tests could be downloaded as tables (.csv or Excel worksheets) containing columns for each question (i.e. more than 100 columns) and rows containing the responses. The results were filtered with respect to duration (5–45 minutes) and completeness. In total, there were more than 300,000 cells in the tables containing information.

As the goal of a DEWA is to provide data about the words in a summarised manner, the perspective for the compiler was to do the summarising calculations more than a thousand times (once for each keyword) and enter the results as text in the XML database pre-existing in the Institute of the Estonian Language[21] . That would have meant losing the data about the test persons (e.g. age, gender, education

[20] The statistical analysis was technically performed by Uku Vainik https://www.researchgate.net/profile/Uku_Vainik (10.10.2017).

[21] Termeki was suggested, which is originally developed for managing terminolgy https://term.eki.ee/ (10.10.2017).

etc) for possible future research. In that stage, the DEWA project enjoyed no special IT support for building a special database for both data storage, performing the calculations and developing the web based user interface needed in the future. It seemed reasonable to use MS Office database software (Access), which was affordable as well as manageable for the person in charge. Once stored there, the data can be exported into different formats, XML included.

The procedures of importing were carried out as follows. Each of the twelve tables with the test results was imported separately and the occurring errors were corrected, which meant deleting some records where part of the data was missing. The data from the raw tables was copied to a system of related tables, see Figure 2 for illustration, so that the main data table has 338,270 records (matching the number the individual valid responses to the stimuli, see Table 2). All the data in the main table characterising the records is represented by codes, which are keys (ID numbers) of related tables where supplementary information is kept. There are three main categories of tables (and their subtables) in the relational database: i) information about the stimuli (part of speech, emotional valence, semantic category, length, frequency in the corpus etc); ii) information about the responses (the original response, its corrected form, lemma etc); iii) information about the respondents (the registration ID, age, gender, education, occupation etc). All this information can be retrieved from the database in a combined way via queries which can be compiled according to the current needs.

There is an important advantage to keeping the responses in a table of types instead of tokens: it means much less work when it comes to data correction. In the case of the present database, the token-type ratio of responses is ca 10:1 (see Table 2).

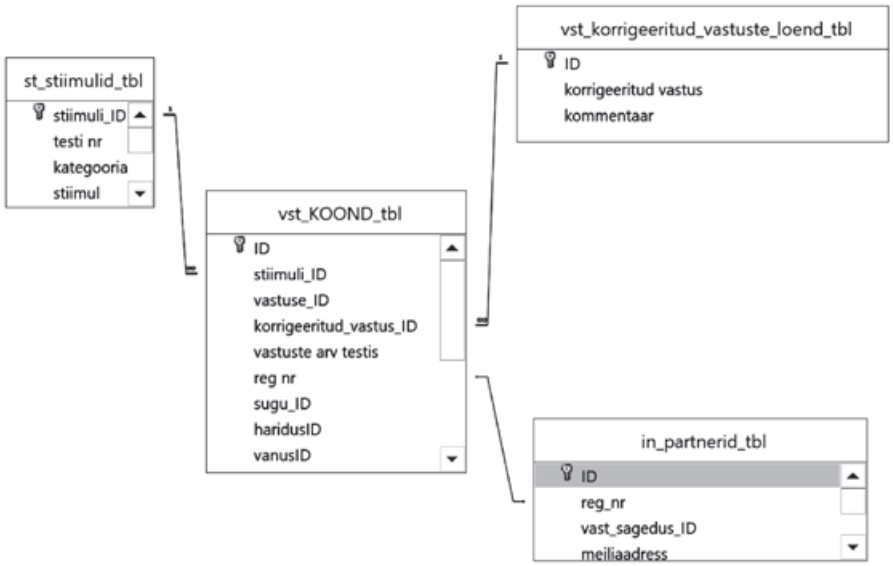

Figure 2. The simplified core structure of the relational database. The labels: st_stiimulid_tbl 'stimuli', vst_KOOND_tbl 'the main table'; vst_korrigeeritud_vastuste_loend_tbl 'the list of corrected responses (the types)'; in_partnerid_tbl 'the partners'

Table 2. The database in numbers

Parameter	Specification	No	%
General	Records in the database	338270	
	Keywords/stimuli	1100	
	Responses per keyword	307,5	
Associations (= pairs of stimuli and response)	Associations (= tokens)	338270	100,00
	Unique associations (= types of associations)	99281	29,3
	Associations occurring just once	66420	19,6
	Recurrent (F > 1) types of associations	32861	9,7
Responses	Responses (= tokens)	338035	100,00
	Unique items as responses (= types of responses)	33740	9,98
	Items occurring just once as a response	19809	5,86
	Recurrent (F > 1) types of responses	13931	4,12
Keyword-response circularity	Keywords matching a response	1084	98,37
	Keywords matching no response	18	0,01
	Recurrent types of responses matching no keyword	12860	38,11

4. Final remarks

Table 2 reflects some numerical parameters of the data stored in the database at its present stage. One can notice that the average number of responses per stimuli is rather high (307), which is due to the active participation by the volunteers. The average number of different associations per keyword is ca 90, which is many times more than in the case of the two examples introduced in the section of theoretical background. From the perspective of compiling a dictionary, recurring associations are more relevant than the individual unique word forms occurring in the pool of responses (hapax legomena). The average number of recurring responses per stimuli is 30, which means that there is plenty of variation within them as well.

The process of selecting the keywords to be used in the tests turned out to be rather successful. Table 2 shows that most of the keywords had a potential for circularity, i.e. they did show up among the responses. This indicates their status as cognitively salient items in the memory of the respondents and their belonging to the active vocabulary of the language. Table 2 also shows that there is a large number of recurring responses that had no match in the list of the keywords. This set of cognitively salient words can be taken as a reserve when it comes to a follow-up stage of the DEWA project and new stimuli are needed.

The large numbers in Table 2 reflect the effectiveness of the digital data gathering process. One cannot compare the digital era to those early days when WATs were performed either in oral or written form. No more transcriptions of the raw data nor manual calculations are needed. The electronically gathered data is easily storable and convertible. During present project, it was possible to collect the raw material for the DEWA in a very short time and at very low costs. This success would not have happened without the contribution of the volunteers recruited via the campaign of Citizen Science.

One of the values of the data is that it is collected from people of varying ages and occupations and not just from college students, as is often the case. The motivated partners of Citizen Science were diligent in compiling the tests, their completion rate was rather high and, most likely, they made fewer typos and gave fewer non-word responses than did the less motivated occasional contributors. It is a matter of further analysis to find out whether this impression turns out to be true or not.

Carrying out the Citizen Science campaign–the PR, gathering the database of partners and maintaining the partnership relation–was a challenge in itself. The task of organising and managing social processes takes time and effort, too, and does not belong to the traditional tasks of a lexicographer nor of a scientist. The ease of digital data collection was, thus, at least partly deceptive.

Digital access to people has its disadvantages, too. The channel itself dictates a preselection of the respondents: only people with an access to a computer could contribute, which meant that wealthier and more educated people had an advantage. The original idea that the partners would also collect data from their family members with no habit of computer usage did not work.

In addition, there was a further self-selection at work. Most likely, those who became partners shared a personality with a higher than average level of the traits Agreeableness and Conscientiousness[22]. This is, of course, an advantage when it comes to cooperation and completing the tasks, but one must be aware that their responses might be slightly biased, as well. For example, one can hypothesise that the level of those traits correlates positively with the stereotypicity of the associations, a tendency noticed in the statistical control of the results of the first test. The personality measures were not involved in this particular analysis, but the higher level of education that demonstrated a statistically significant effect on stereotypicity may easily be itself an effect of Conscientiousness as a personality trait at work. The possible effect of personality on the behaviour of eliciting associations is a topic for future research. Therefore, as was concluded in the section 3.2, the results of the present tests cannot be taken as the psychological association norms representative of whole population.

However, the disadvantage of self-selection might turn out to be an advantage for the purpose of compiling the dictionary. The dictionary entries are expected to reveal associations that the speakers generally agree upon and therefore the typical (and stereotypical) associations are desired; the individual and unique responses will be kept in the database for further analysis but not published in the dictionary.

Due to the lack of space we reproduce no results of the WATs in terms of particular responses given to certain stimuli, except the illustration presented in Figure 1. Analysing the gathered data from the point of view of the topical associations, their absolute and relative frequency etc will be the topic of a further publication. The database has tremendous potential as a source of further studies.

The remarks above summarised and discussed only the stages of dictionary making reported in this paper, and not the whole project, which is still in progress (correcting the data, reducing the varying linguistic forms, exporting the database to XML format, and building the web-based user interface)[23].

[22] These are two traits of the five known in the Five Factor Model of personality (Digman 1990).

[23] By the time the paper is being published, there is the first version of the dictionary available on the web: http://www.eki.ee/dict/assotsiatsioonid/index.cgi?Q=*&F=M&C06=et (16.3.2018).

References

Aitchison, Jean 2012. Words in the Mind: An Introduction to the Mental Lexicon. 4rd ed. Wiley-Blackwell.

Apresjan, Juri 2000. Systematic Lexicography. Oxford University Press.

Atkins, B. T. Sue; Rundell, Michael 2008. The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.

Benjamin, Martin 2015. Crowdsourcing microdata for cost-effective and reliable lexicography. – Lan Li, Jamie Mckeown, Liming Liu (Eds.), Proceedings of AsiaLex 2015, Hong Kong. Hong Kong Polytechnic University, 213–221.

Buk, Solomyia 2009. Lexical base as a compressed language model of the world (on material from the Ukrainian language). – Psychology of Language and Communication, 13 (2), 35–44. https://doi.org/10.2478/v10057-009-0008-3

Church, Kenneth W.; Hanks, Patrick 1989. Word association norms, mutual information, and lexicography. – Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics. Vancouver, 76–83. https://doi.org/10.3115/981623.981633

Čibej, Jaka; Fišer, Darja; Kosem, Iztok 2015. The role of crowdsourcing in lexicography. – I. Kosem, M. Jakubíček, J. Kallas, S. Krek (Eds.), Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 conference, 11–13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana, Brighton: Trojina, Institute for Applied Slovene Studies, Lexical Computing Ltd., 70–83. https://elex.link/elex2015/conference-proceedings/ (10.10.2017).

Cohn, Jeffrey P. 2008. Citizen science: Can volunteers do real research? – BioScience, 58 (3), 192–197. https://doi.org/10.1641/B580303

Cruse, David Alan 2000. Meaning in Language. An Introduction to Semantics and Pragmatics. New York: Oxford University Press.

De Deyne, Simon; Navarro, Daniel J.; Storms, Gert 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. – Behavior Research Methods, 45 (2), 480–498. https://doi.org/10.3758/s13428-012-0260-7

Digman, John M. 1990. Personality structure: Emergence of the five-factor model. – Annual Review of Psychology, 41, 417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221

Fellbaum, Christiane 1999. Wordnet. An Electronic Lexical Database. London: The MIT Press.

Fillmore, Charles 1985. Frames and the semantics of understanding. – Quaderni di Semantica, 6 (2), 222–254.

Fitzpatrick, Tess; Playfoot, David; Wray, Alison; Wright, Margareth J. 2015 (2013). Establishing the reliability of word association data for investigating individual and group differences. – Applied Linguistics, 36 (1), 23–50. https://doi.org/10.1093/applin/amt020

Galton, Francis 1879. Psychometric experiments. – Brain, 2 (2), 149–162. https://doi.org/10.1093/brain/2.2.149

Jung, C. Gustav 1910. The association method. – The American Journal of Psychology, 21 (2), 219–269. https://doi.org/10.2307/1413002

Kallas, Jelena; Tuulik, Maria 2011. Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamispõhimõtted [‘The Basic Dictionary of Estonian: The historical context and the principles of compilation’]. – Eesti Rakenduslingvistika Ühingu aastaraamat, 7, 59–76. https://doi.org/10.5128/ERYa7.04

Kent, Grace H.; Rosanoff, Aaron J. 1910. A study of association in insanity. – American Journal of Insanity, 67 (1–2), 37–96.

Kiss, George R.; Armstrong, Christine; Milroy, Robert; Piper, James 1973. An associative thesaurus of English and its computer analysis. – A. J. Aitken, R. W. Bailey (Eds.), The Computer and Literary Studies. Edinburgh: University Press, 153–165.

Klein, Stephen 2012. Learning: Principles and Applications. 6th ed. SAGE Publications.

Krasnõh, V. 2001. Osnovõ psihholingvistiki i teorii kommunikatsii. Moskva: Gnosis.
Postman, Leo; Keppel, Geoffrey (Eds.) 1970. Norms of Word Association. Elsevier.
Männik, Anna-Liisa 2016. Assotsiatsioonid eesti ja inglise keeles. Käsikirjaline magistritöö. Tallinna Ülikooli humanitaarteaduste instituut.
Meara, Paul 1982. Word associations in a foreign language: A report on the Birkbeck Vocabulary Project. – Nottingham Linguistic Circular, 11 (2), 29–37.
Meara, Paul 2009. Connected Words. John Benjamins. https://doi.org/10.1075/lllt.24
Morkovkin, Valery V. 1970. Ideographic Dictionaries. Moscow, USSR.
Nelson, Douglas L.; McEvoy, Cathy L.; Schreiber, Thomas A. 2004. The University of South Florida word association, rhyme, and word fragment norms. – Behavior Research Methods, Instruments, & Computers, 36 (3), 402–407. https://doi.org/10.3758/BF03195588
Nelson, Douglas L.; McEvoy, Cathy L.; Dennis, Simon 2000. What is free association and what does it measure? – Memory & Cognition, 28 (6), 887–899. https://doi.org/10.3758/BF03209337
Orav, Heili; Vider, Kadri 2005. Estonian wordnet and lexicography. – H. Gottlieb, J. E. Mogensen, A. Zettersten (Eds.), Symposium on Lexicography XI. Proceedings of the Eleventh International Symposium on Lexicography, May 2–4, 2002 at the University of Copenhagen. Tübingen: Max Niemeyer, 549–555.
Pavlenko, Anetta (Ed.) 2009. The Bilingual Mental Lexicon: Interdisciplinary Approaches. Bristol, UK, Buffalo, NY: Multilingual Matters.
Peppard, Jason 2007. Exploring the Relationship between Word-Association and Learners' Lexical Development. An assignment for Master of Arts in Applied Linguistics. Centre for English Language Studies, Department of English, University of Birmingham, United Kingdom. https://www.birmingham.ac.uk/Documents/college-artslaw/cels/essays/lexis/PeppardMod2.pdf (23.3.2018).
RAS = Karaulov, J. N.; Tšerkassova, G. A.; Ufimtseva, N. V.; Sorokin, J. A.; Tarasov, E. F. 2002. Russkii assotsiativnõi slovar. Moskva: Astrel.
Rosenzweig, Mark R. 1961. Comparisons among word-association responses in English, French, German, and Italian. – The American Journal of Psychology, 74 (3), 347–360. https://doi.org/10.2307/1419741
Stevens, Anthony 1994. Jung: A Very Short Introduction. Oxford: Oxford University Press.
Toim, Kalju 1980. Estonian word association norms for the Kent-Rosanoff test. – Problems of cognitive psychology. (Труды по психологии. Проблемы когнитивной психологии). Tartu Riikliku Ülikooli Toimetised 522. Tartu, 60–76.
Vainik 2017. The word associations reveal: What does it take to be an Estonian? – Liisi Laineste (Ed.), Book of abstracts of the international conference „Across Borders VII. Cultures in dialogue". Tartu: ELM Scholarly Press, 104. http://www.folklore.ee/rl/fo/konve/AcrossBorders/2017/borders2017web_abstracts.pdf (10.10.2017).
Vainik, Ene 2012. Kuidas määrata eesti keele sõnavara tundetoone? ['Detecting emotional valencies for the Estonian vocabulary'] – Eesti Rakenduslingvistika Ühingu aastaraamat, 8, 257–274. https://doi.org/10.5128/ERYa8.17

Ene Vainik (Institute of the Estonian Language) has carried out research in the field of cognitive lingusitics, semantics, folk-psychology, and the interaction between language and emotions. She has published papers about literal and figurative emotion descriptions and affective computing (detecting the cues of affect/emotion in written Estonian). In a recent book ("Eesti tunded. Sõnaportreed" 'Estonian feelings. Portraits of the words'(2016)) she developed an in-depth methodology for portraying culturally relevant concepts and applied it to the 19 most relevant Estoninan emotion terms/concepts.
Roosikrantsi 6, 10119 Tallinn, Estonia
Ene.Vainik@eki.ee

EESTI KEELE ASSOTSIATSIOONISÕNASTIKU LOOMINE: TÜHJAST KOHAST ANDMEBAASINI

Ene Vainik
Eesti Keele Instituut

Artiklis kirjeldatakse "Eesti keele assotsiatsioonisõnastiku" loomise esimesi etappe kavandamisest kuni algandameid sisaldava andmebaasini. Esmalt antakse ülevaade põhimõistetest (assotsiatsioon, sõna-assotsiatsioon, assotsiatsioonisõnastik *vs*. assotsiatsiooninormid) ja kirjanduses kasutatavast terminoloogiast. Järgneb ülevaade sõna-assotsiatsioonide uurimise ajaloost ja tuuakse välja sõnastikuprojekti teoreetilised eeldused: a) sõnu iseloomustavad nende seosed teiste sõnadega; b) nende seoste väljatoomine on oluline leksikograafiline ülesanne; c) assotsiatsioone saab tuvastada üksnes inimeste testimise teel.

Järgnevas osas kirjeldatakse tehtud töid ja põhjendatakse praktilisi valikuid. Lahti seletatakse märksõnastiku ja testide koostamise põhimõtted, kodanikuteaduse kampaania käivitamise vajadus inimeste värbamiseks ning selle kulg. Artikli viimases osas põhjendatakse valikut andmete talletamise osas (relatsiooniline baas), kirjeldatakse andmebaasi struktuuri ning andmete impordi protseduure. Tabel 2 annab arvulise ülevaate sõnastiku aluseks olevast andmebaasist.

Artikli lõpus arutletakse tehtud valikute eeliste ja nõrkuste üle. Andmete kogumist kodanikuteaduse raames loeti õnnestunud ettevõtmiseks, seda nii järjest kasvava osalemisaktiivsuse kui ka sooritamisedukuse mõttes (vt tabel 1). Kuna kodanikuteaduse partnerid kalduvad olema naissoost ja kõrgema haridusega, siis kontrolliti nende tegurite mõju statistilise analüüsiga. Tulemused näitasid, et sugu, iga ja amet vastuste stereotüüpsust ei mõjutanud, küll aga kõrgem haridustase. Seega on kogutud andmestikus tõenäoliselt üldpopulatsioonist stereotüüpsemad seosed, mida autor luges aga pigem eeliseks, kuna sõnastiku eesmärk ongi just koguda tüüpilisemaid seoseid ja ainukordsed vastused jäävad andmete suure mahu tõttu igal juhul sõnastikust välja. Kõik vastused koos andmetega vastajate soo, ea, hariduse jm kohta jäävad andmebaasi alles tulevasteks uuringuteks.

Võtmesõnad: sõna-assotsiatsioon, mentaalne leksikon, leksikograafia, e-sõnaraamat, kodanikuteadus, rahvahange