Generating vocabulary clusters using word association games

Max Baraitser Smith
Universitat der Kunste Berlin
Technische Universität Berlin
MA Design & Computation
m.baraitsersmith@gmail.com

Abstract

In foreign language vocabulary learning, it is common to learn clusters of words which are all related to a given topic (e.g. work, sport, shopping, etc). This paper looks at how these clusters can be generated using associative word games. A multiplayer online game was developed and used to compare three strategies for data collection, which was tested by 92 players. These were evaluated on the basis of statistical measures, distribution of words in a force-directed graph, and manual analysis of individual words. The results show that while each strategy generates strongly associated word clusters, there is no significant difference in their performances. Recommendations are made for further study and refinement of game-based approaches to measuring word associations.

I. Introduction

The term 'word association' describes the act of responding to a stimulus word with the first word which comes to mind. So far this has been studied by asking participants to write down their individual associations in response to a pre-set list of cue words, resulting in a set of stimulus-response word pairs. The data from the test is then used to create word association norms [2] or word association networks [1]. These methods have been used extensively to explore theories about the mental lexicon, especially in the field of second language learning [11] or in order to reveal conceptual structures around a certain topic [7].

This paper looks at an alternative game-based approach to study word associations. A word association game involves two or more players who choose a starting word and then take turns responding to each other with the first new word which comes to mind. The game-based results differ from those of the standard word association test in two main ways: 1) Games generate long chains of associatively-related words, rather than a set of word pairs; 2) Games use 1 starting word per round rather than a pre-set list of cue words.

I will evaluate word association games as a tool for generating vocabulary clusters, which are groups of connected words commonly used to learn a foreign language. Existing studies have compared different methods of clustering vocabulary words, e.g. semantic (mouse, ears, tail), thematic (mouse, cheese, jump), or random (mouse, cloud, radio). However, there is no consensus as to which method performs best [10]. A word associative approach to clustering, following the classification from Kess [6], would likely mix paradigmatic (mouse, intruder, burglar), syntagmatic (mouse, trap, door), taxonomical (mouse, rodent, animal), and rhythm/rhyme (mouse, house, louse) connections. The nature of these responses are likely to reflect the players' linguistic, social and geographical contexts. For example, the starting word 'red' may produce the word chain red-danger-fear in one part of the world, and red-luck-money in another part of the world. Note that this could not be done using the standard word association test without making assumptions about how the players might respond, and manually adding 'danger' and 'luck' as cue words. This is a potential advantage of word association games over other clustering methods, provided that players and language learners share similar backgrounds.

The underlying assumption for generating vocabulary clusters using word association games is that word associations are a good estimate of word usage and meaning. This question has been heavily debated. Fitzpatrick writes that ”L2 word association research is driven by the belief that association behaviour can reveal information about the development and organisation of the mental lexicon. However, studies have failed to produce consistent findings” [4]. More recently, De Deyne et al., have argued the opposite, claiming that ”the association data themselves provide a strong indication of the similarity (and thus the meaning) of a word”, and go on to give promising results estimating semantic similarity using word association data [2]. I outline a more extreme version of the latter claim in the following thought experiment:

Consider a variation of the philosopher Ludwig Wittgenstein’s claim that ”the meaning of a word is its use in the language” [12], but with regards to word associations: *The meaning of a word is defined by the words people associate with it*. A word’s meaning would therefore be accumulated recursively based on each associated word, association of an association, etc. This could be represented as an associative network, which is a graph format commonly used for associative data. Such a network is a sample from an infinite number of possible associations, which, taken together, represent the complete meaning of every word. The most useful samples are therefore those which are most representative of the complete network, and the best estimations of its meaning.

Word associations probably do not represent all aspects of a word’s meaning. Nevertheless, the above statements captures the core approach of this paper. While vocabulary clusters are not classically designed to convey the meaning of the words through the structure of their constellation, network-based learning methods could make this a feasible method of vocabulary learning.

Figure 1: *Screenshot from wordassociation.org which is currently the largest database of word associations. The website allows the user to play word association games with the computer [5].*

Problem definition

This paper looks at how word association games can be optimised for the purpose of generating vocabulary clusters. The aim is to generate vocabulary clusters which are as relevant as possible to a starting word without becoming overdetermined by it.

The main difficulty with this application of word association games is that words spoken by the players quickly lose their relation to the starting word (often after 2 or 3 turns). This may be the reason why such games have not yet been formally studied. The lack of control over the gameplay makes it difficult to collect detailed data around a specific set of words without collecting data from a very large number of players.

To solve this problem, I propose three alternate versions of the word association game which aim at constraining gameplay around one or multiple starting words. To be consistent with other studies, I refer to these starting words as cue words. Assuming that a word’s meaning can be defined in terms of its associations, the ideal word cluster would only include words which are most strongly associated with the cue words. Therefore, the strategy which generates the highest proportion of words which are strongly associated with the cue words can be considered the most successful.

II. Methods

Constraining the word association game

The first approach which I tried used word substitutions to constrain the word association game around a set of cue words. Every response given by a player would be substituted with a synonym which is slightly closer to one of the cue words, before being displayed to the other player. For example, if the cue word is 'scary' and a player types 'poodle', then it might be substituted with 'pit bull' before being sent to the other player, on the grounds that 'pit bull' is associatively closer to 'scary' than 'poodle' is. I prototyped this method in Python using the function most_similar from Gensim's Word2Vec model with german word embeddings [3]. However, an initial look at the results revealed that they did not relate closely enough to the response words, and as a result, the approach was abandoned (1).

Word	Cosine similarity
niedlich	0.7480849027633667
drollig	0.7219110131263733
knuffig	0.7209312319755554
Knuddeln	0.7119846343994141
entzueckend	0.7115506529808044

Table 1: *Output of Word2Vec's most_similar() function for the words 'gruselig' and 'Pudel'.*

The second strategy which I tried involved inserting pre-defined cue words into gameplay at regular intervals. This was a simpler approach compared to substituting synonyms, and therefore would likely produce more reliable results. I designed three variations, each of which had different rules for dealing with cue words 2. All three variations seemed to have potential, therefore I decided to use them all in my data collection experiment.

Strategy	Rules
1	A single cue word inserted at regular intervals
2	A single cue word, with next cue words randomly selected from words which previous players associated to the original cue word, and inserted at regular intervals
3	A set of multiple cue words. One is randomly selected and inserted at regular intervals

Table 2: *Strategies for inserting cue words.*

Data collection

I developed a web-based multiplayer game in German where players were given a stimulus word and instructed to type the first word which came to mind. A response from one player was then displayed to the other player as a stimulus word. Players were asked to type their name at the start of the game, so that returning players could be recognised. Players were shown the text "Du bist dran" with the stimulus word when it was their turn, and the text "Sie sind dran" when it was the other player's turn. There was no other information displayed on the screen. Responses were logged by the game and saved as a text file.

The data was collected over 3 days as part of the Design & Computation end of term exhibition in Febuary 2024. Members of the public were invited to sit and play the game at two neighbouring consoles. They continued playing for as many rounds as they wanted. Three game variations were tested on days 1, 2 and 3 respectively. No data was collected about the participants (age, gender, German proficiency, etc). After playing the game, participants could see a visualisation of their responses and the responses from other people in the form of a force-directed graph.

I chose the topic "klima" for the study because it is broad enough so that people of different ages and language proficiency would be easily able to respond to it. This decision was inspired by Kostova and Radoynovska's use of the word 'biodiversity' in their related word association-based experiment [7].

I chose 20 as the interval for which cue words would be inserted into gameplay. Due to a lack of existing research on this interval, my decision was based on the intuition that shorter intervals might make the gameplay feel disjointed and reduce the quality of the data.

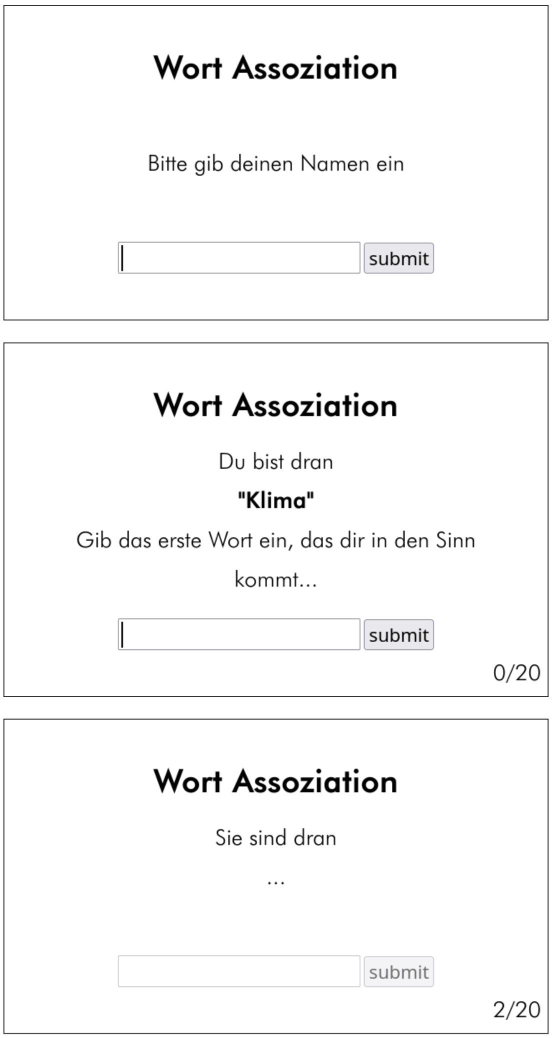

Figure 2: *Multiplayer game interface.*

For strategy 1, ”Klima” was used as the first cue word and then reinserted into the word chain every 20 words. For strategy 2, ”Klima” was used as the first cue word, with subsequent cue words randomly selected from words which previous players had associated to ”Klima” and inserted into the word chain every 20 words. For strategy 3, ”Klima” was used as the first cue word with one of four synonyms from a synonym dictionary randomly chosen and inserted every 20 words. The synonyms were ”Atmosphäre”, ”Stimmung”, ”Wetter”, ”Verhältnis” [9].

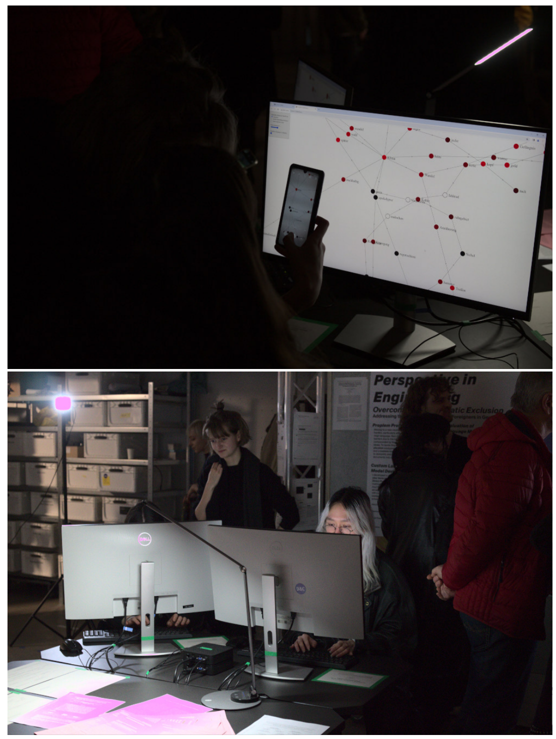

Figure 3: *Game consoles at the D&C exhibition Feb 2024.*

III. Analysis

Initial processing of the data

The data was extracted from the text file. All non-alphabetical characters were removed from the responses. Responses with different capitalisation were merged. There was no attempt made to merge multiple spellings.

Constructing an associative network

I used the data to construct an associative network, which is a graph format commonly used for associative data [1]. Each node in the graph corresponds to a word, and each edge corresponds to a stimulus-response pair made by a player during the word association game. A directed, rather than an undirected graph was chosen, because this fitted more closely with the original data.

Statistical measures of associative strength

Each strategy was to be evaluated based on how many words it contained which are strongly associated to a cue word. To evaluate this, I needed a measure of associative strength which works for both direct and indirect associations. I came up with 3 statistical measures for this purpose (3).

Manual analysis of the words

In order to get a better word-level understanding of the data, I conducted a manual analysis of individual words. All words which were mentioned more than once were included in the sample. Then, I manually separated the words which I considered to be frequently present in climate discussions taking place in Berlin. Two entries, 'schlecht' and 'bad', were merged because they were clear translations of each other.

Visual analysis using force-directed graph

To better understand the structure of the associative network, I conducted a visual analysis of each strategy using force-directed graphs, implemented in D3.js [8]. On each graph, the words were drawn as circles and associative connections were drawn as straight lines with arrows. Cue words were fixed at the centre. Words which had been selected during the manual analysis as climate-related were highlighted in green. Each associated node was given a link force with equal distance and strength. A global outward pull was applied so that less connected nodes would drift drift away from the centre. To aid visual analysis, I separated the graph into 7 regions using concentric circles, showing increasing distances from the centre. I will refer to these as regions 1-7 where 1 is at the centre and 7 is at the edge of the graph.

Statistical measure	Description
Number of unique words / Number of responses.	This broadly measures how diverse the data is, and would be low for strongly associated word clusters.
Frequency of word mentions	This gives a more detailed measure of how diverse the data is, with strongly associated clusters resulting in higher frequencies.
Mean distance of each word to the closest cue word / number of mentions.	This is the most detailed test. Strongly associated words would be expected to have lower distances and more mentions.

Table 3: *Statistical measures*

IV. Results

During the experiment, 92 individual players played the game, with a total of 141 rounds (chains of 20 words). Strategy 1 was played by 29 players for 48 rounds. Strategy 2 was played by 40 players for 49 rounds. Strategy 3 was played by 23 players for 44 rounds.

Figure 4 shows the mean rounds per player for each strategy. The graph shows that the majority of players only received 1 prompt.

Figure 5 shows that all strategies had a similar ratio between number of unique responses and total number of responses. This can be interpreted as a reliable yet small amount of convergence in the players' associations.

Figure 6 shows the frequencies of word mentions. This shows that the overwhelming majority of words were mentioned only once.

Figure 7 shows the relationship between the number of word mentions and the mean path length from a given word to the nearest cue word. All three strategies show a fairly even distribution of word mentions with respect to mean path length, with highly frequent words (mentions >= 4) having mean path lengths between 3 and 9. This contradicts the previous expectation that the most frequent words would occur nearest the centre.

Figure 8 show the force-directed graphs. For all strategies, the density of intersections between words was highest in region 1, and decreased gradually in regions 2-7. The density of intersections for strategy 1 is noticeably more contained than strategies 2 and 3, with almost all intersections occuring in regions 1-5. All strategies generated graphs where climate-related words occurred near the centre (highlighted in green). Closeups of these central clusters are shown in figure 9.

Tables 4, 5 and 6 show words from each strategy which occured multiple times. The tables shows that all three strategies produced significant numbers of climate-related words, and that the most frequent word was always climate-related.

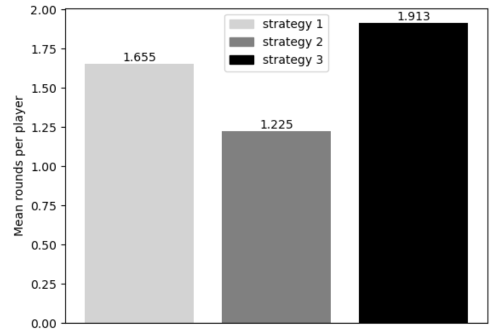

Figure 4: *Mean rounds per player.*

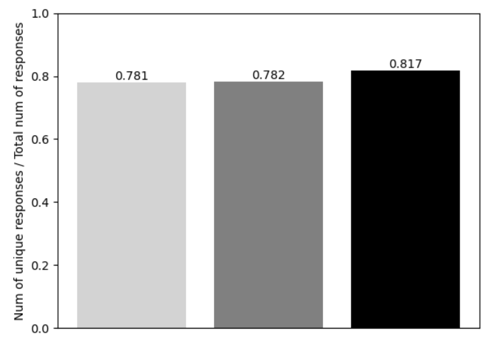

Figure 5: *Number of unique responses / Total number of responses.*

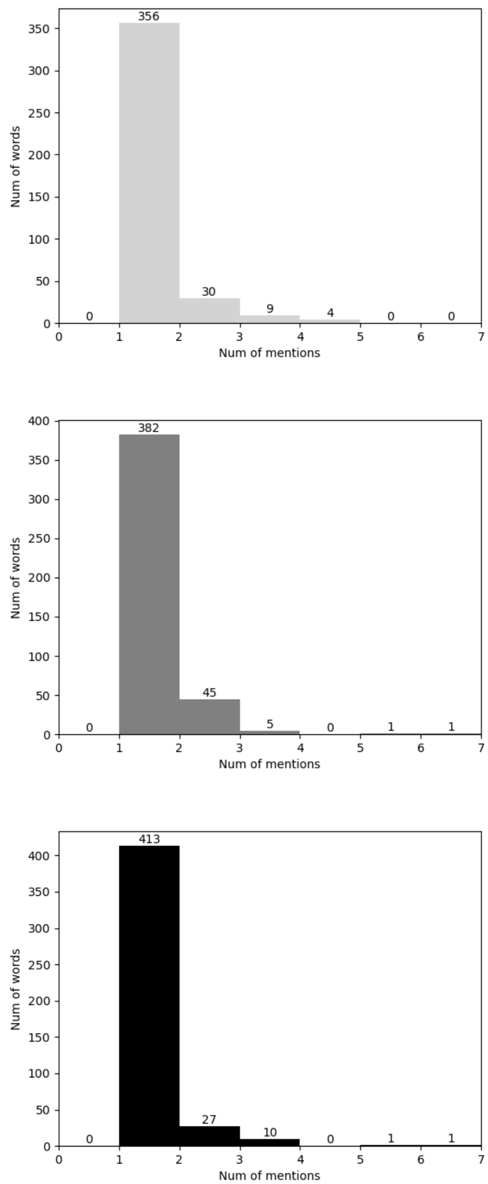

Figure 6: *Frequencies of word mentions in responses for each strategy (top=strategy 1, middle = strategy 2, bottom = strategy 3).*

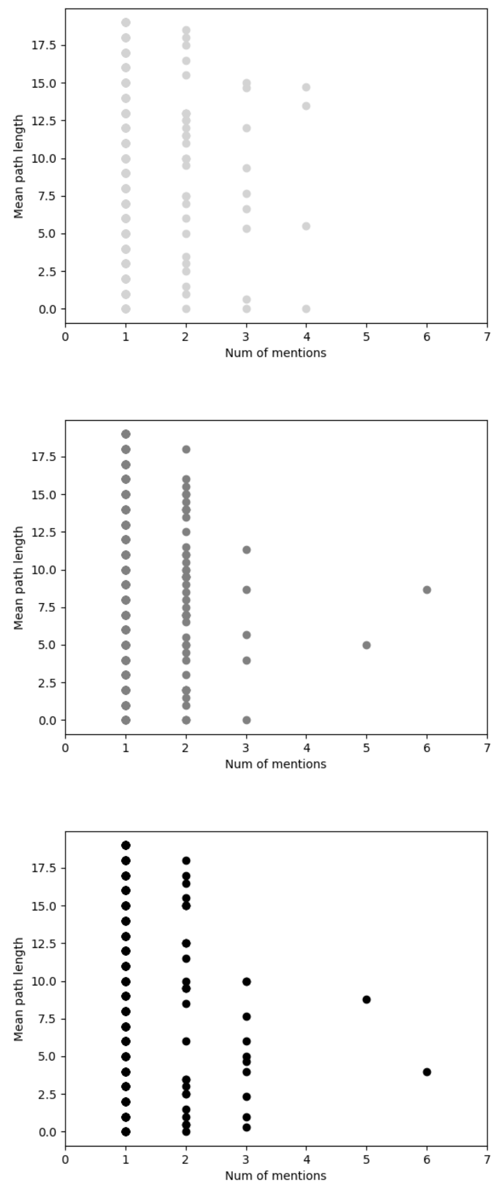

Figure 7: *Mean chain length of words to a cue word plotted against their number of mentions (top=strategy 1, middle = strategy 2, bottom = strategy 3).*

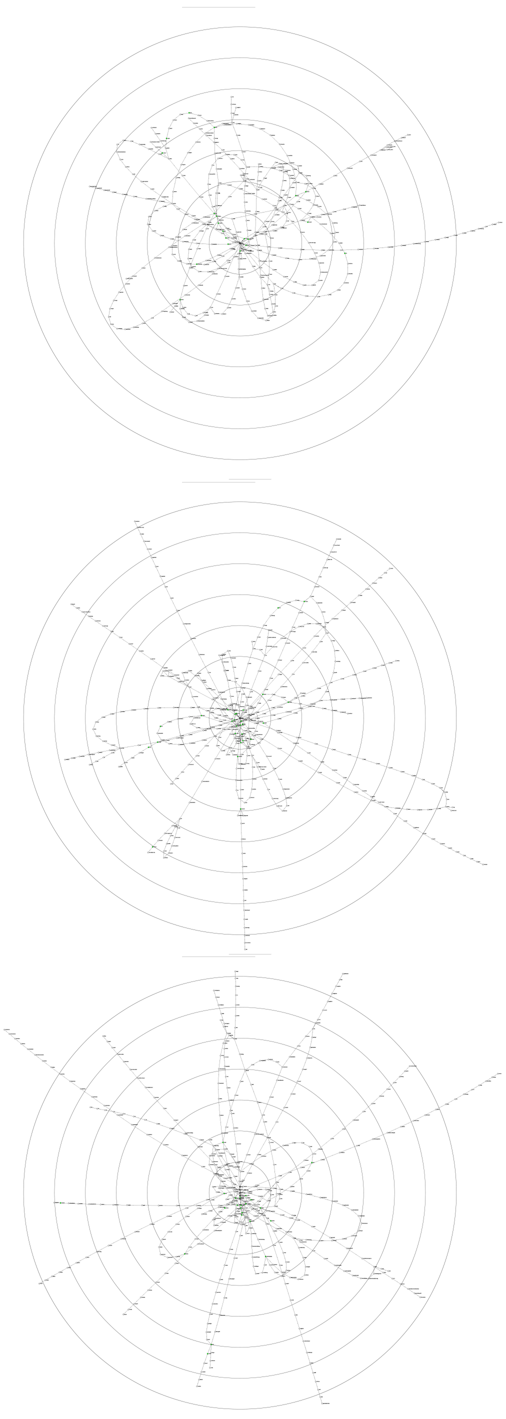

Figure 8: *Force directed graphs (top=strategy 1, middle = strategy 2, bottom = strategy 3).*

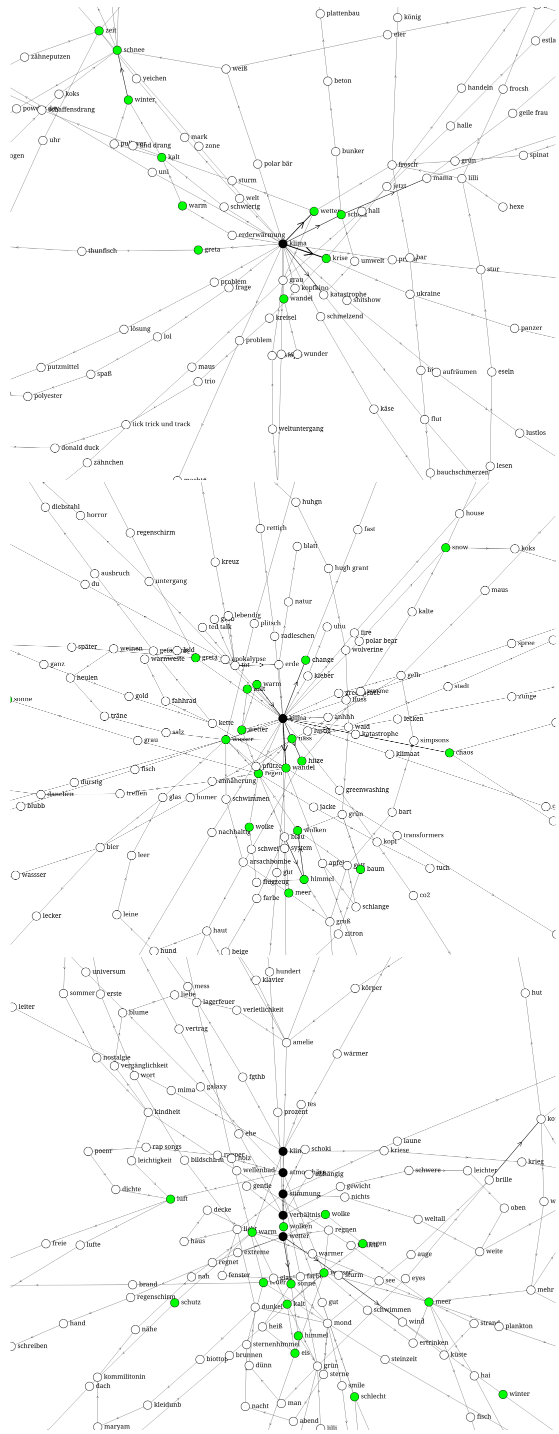

Figure 9: *Closeup of the centre of force directed graphs (top=strategy 1, middle = strategy 2, bottom = strategy 3).*

Climate related	Num of mentions
krise	4
wetter	3
schutz	3
schnee	3
schwamm	2
berlin	2
zeit	2
auto	2
winter	2
Climate unrelated	
mensa	3
essen	3
gelb	3
blau	3
sand	2
dunkel	2
grau	2
platt	2
frosch	2
halbstark	2
rot	2
kochen	2
film	2
wei	2
keller	2
nase	2
monopoly	2
sis	2
bro	2
scheie	2
max	2
gold	2
lilli	2

Table 4: *Words from strategy 1 which occured multiple times*

Climate related	Num of mentions
wasser	6
regen	5
schlecht / bad	4
control	2
change	2
wolken	2
himmel	2
warm	2
wetter	2
chaos	2
meer	2
nass	2
berlin	2
greta	2
traum	2
snow	2
hitze	2
arbeit	2
baum	2
wandel	2
Climate unrelated	
ich	3
blau	3
trippy	2
gro	2
gott	2
ok	2
weihnachten	2
jesus	2
sad	2
willie	2
bum	2
liebe	2
sex	2
nervig	2
kette	2
was	2
schlaf	2
lol	2
apfel	2
rot	2

Table 5: *Words from strategy 2 which occured multiple times*

Climate related	Num of mentions
wasser	6
meer	5
feuer	3
angst	2
arm	2
kalt	2
schlecht	2
anstrengung	2
eis	2
sonne	2
schnell	2
wolke	2
luft	2
warm	2
Climate unrelated	
amelie	3
langweilig	3
grn	3
nichts	2
schwester	2
sex	2
gut	2
mehr	2
weite	2
dunkel	2
mond	2
licht	2
fuball	2
ball	2
kopf	2

Table 6: *Words from strategy 3 which occured multiple times*

V. Discussion

The strategies presented in this paper were designed to generate clusters of words which are strongly related to the topic "Klima". The results from the individual word analysis show that climate-related words were generated by all strategies, but none of the strategies performed significantly better than the rest. This may be because most players played only one round, and therefore did not experience any difference between the strategies. This leads to the conclusion that the cue word interval of 20 was too large, and that the experiment should be repeated with a smaller interval, which could generate less diverse data and show clearer differences between the strategies.

The results show evidence against the initial assumption that frequently mentioned words tend to be direct/close neighbours of a cue word. In contrast, the most frequently mentioned words were between 3 and 9 degrees of separation from the nearest cue word. This suggests that a slightly higher number (e.g. 10) could be a better choice for the cue word interval.

The force-directed graph analysis showed central clusters of climate-related responses for all three strategies, and could be considered a potentially powerful tool for the analysis of word clusters. Further research could focus specifically on this application of force-directed graphs.

The vast majority of words were only mentioned once. While some diversity is to be expected with such free data-collection tools, this result may be partly due to the lack of grouping of responses, for instance identifying and merging synonyms, direct translations, or multiple/incorrect spellings.

This study is the first formal exploration of word association games, and suggested a number of strategies for understanding and interpreting them. It also prototyped a new experimental method where a large amount of word association data can be collected in a short period of time.

The study was limited by the fact that no data was collected on the players, which could have been used to divide them according to specific characteristics such as language proficiency. Due to the exhibition setting of the experiment, it wasn't possible to determine to what extent the different groups were similar with regard to their associative behaviours. There was no attempt to account for existing relationships between players, their effect on the data, e.g. friends might be more likely to make jokes with one another. The decaying random walks algorithm proposed by De Deyne et al.

could also be implemented and tested with this data [2].

Further studies on the use of word association games to generate word clusters could consider taking measures to ensure that players retain their anonymity, or directly studying and making use of these real-life interactions between players. They could also experiment with more restricted versions of the word association game, such as for players to only respond with synonyms. Subsequent experiments could also be carried out on the effectiveness of associative networked vocabulary clusters as practical vocabulary learning tools.

Further studies on the use of word association games in other contexts could consider extending the existing research on the affect of other factors such as age on people's word associations, looking at word association games as a tool for mapping social groups. The techniques described in this paper could also be used to assess other kinds of group knowledge, such as the contents of a conversation, a lecture, or a book. An alternative approach would be to use cue words from two different topics and investigate the associative paths which exist between them. Additional research could also be conducted on the force-directed graphs presented in this paper as a tool for starting discussions, rather than ends in themselves.

VI. Conclusion

The results did not exhibit clear differences between three strategies for the purpose of generating vocabulary clusters. However, the findings are limited by the fact that most of the players only received 1-2 prompts. It is recommended to re-run the experiment with a shorter prompt interval.

VII. Acknowledgments

I thank Prof. Daniel Hromada (Berlin University of the Arts) for support and guidance on this project. I thank the students of the class "Introduction to Optimization and Problem-Solving in Coding and Making" from the M.A. Design & Computation (Berlin University of the Arts & Technical University Berlin). I thank @devmount for providing German word embeddings.

References

Javier Borge-Holthoefer and Alex Arenas. Semantic networks: Structure and dynamics. *Entropy*, 12(5):1264–1302, 2010.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, 51:987–1006, 2019.

Devmount. Germanwordembeddings.

Teresa Fitzpatrick. Word association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics*, 17(3):319–331, 2007.

Simon Holliday. Word association, 2003.

Joseph F. Kess. *Psycholinguistics: Psychology, linguistics, and the study of Natural Language*. Benjamins, 1992.

Zdravka Kostova and Blagovesta Radoynovska. Word association test for studying conceptual structures of teachers and students. *Bulgarian Journal of Science and Education Policy (BJSEP)*, 2(2):209–231, 2008.

Observable. D3 force.

opentheasaurus.de. Klima. `https://www.openthesaurus.de/synonyme/Klima`. Accessed: 2024-03-01.

Amanda Wilcox and Almitra Medina. Effects of semantic and phonological clustering on l2 vocabulary acquisition among novice learners. *System*, 41(4):1056–1069, 2013.

Clarissa Wilks and Paul Meara. Untangling word webs: Graph theory and the notion of density in second language word association networks. *Second Language Research*, 18(4):303–324, 2002.

Ludwig Wittgenstein. *43*. Blackwell, 1953.