A new tool for semantic network analysis: *fan_xplorr*

Albert Maršík (Charles University) —
Eva Maria Luef (Charles University, Hamburg University)

ABSTRACT
Lexical association data is a valuable resource in psycholinguistics, providing researchers with empirical insights into which words are perceived as 'belonging together'. One of the largest such datasets for English, the *University of South Florida Free Association Norms database*, is available freely online. The raw data format however requires some technical steps to tap into the potential it offers for network science, which can present a barrier for some researchers. This paper introduces a user-friendly command-line tool, fan_xplorr, that addresses this issue by providing linguists with access to the data. With the proposed tool, researchers can interactively display portions of the semantic network dataset in the form of interactive network graphs. The design of the tool enables linguists to access the dataset without advanced technical skills and promotes exploration within the dataset for psycholinguistic studies.

KEYWORDS
lexical network science, priming, psycholinguistics, Python, semantic associations

DOI
https://doi.org/10.14712/18059635.2024.2.4

1 INTRODUCTION

As of recently, network science has become an increasingly popular method to investigate lexical relationships in psycholinguistics (e.g., Chan & Vitevitch, 2010; Fourtassi et al., 2020; Hills et al., 2009b; Siew & Vitevitch, 2016; Vitevitch, 2021). Networks are based on dyadic relationships between entities and consist of nodes — which are words in lexical networks — and links placed between them in case of a relationship. Semantic networks are a special type of lexical network in which words are linked when they share meaning (De Deyne et al., 2017). The notion of shared meaning can be quantified in various ways (e.g., Stella et al., 2018), for instance according to taxonomic relations between words (e.g., hypernyms and their hyponyms: "flower" and "rose"), the number of shared features (e.g., synonymy: "couch" and "sofa"), syntactic co-occurrences (e.g., "sleep" and "bed") or mental associations (e.g., similar notions that people hold: "grandma" and "cookies"). Figure 1 schematizes the semantic neighborhoods of the words *keeper*, *holder*, *napkin*, *handkerchief*, *scarf* and *bandanna* and their mental association. The target words are linked to words which are commonly associated with the target words, such as "gloves" in the case of the target word "scarf".

Constructing a network based on semantic relationships in a lexicon results in an intuitive and readable structure, which the human eye and mind are apt of

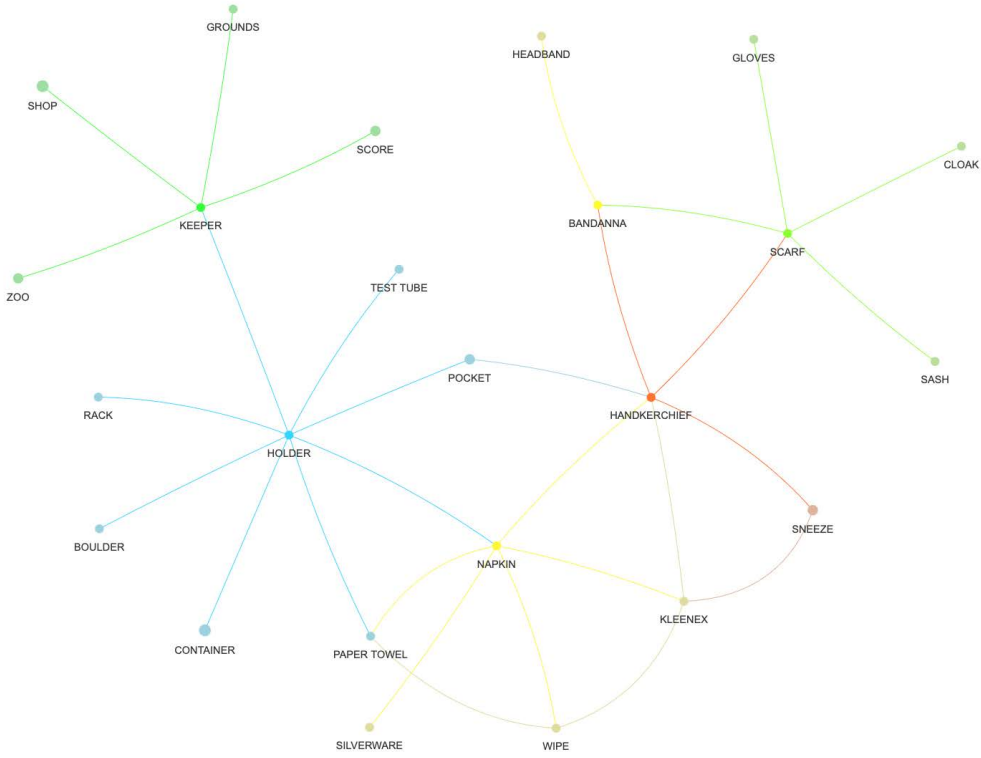

FIGURE 1. Common word associations of *keeper*, *holder*, *napkin*, *handkerchief*, *scarf*, and *bandanna*

interpreting. It allows us to draw inferences as to facilitated or delayed lexical retrieval based on the number of semantic neighbors of a target word (Reilly & Desai, 2017) and the semantic closeness of the neighbors (Mirman et al., 2010; Mirman & Magnuson, 2008). In addition, such a network can yield insights into the structural organization of the semantic lexicon in relation to semantic fluency (Siew & Guru, 2023), inform about lexical learning probabilities (Beckage et al., 2011; Steyvers & Tenenbaum, 2005), and allows us to track its changes during lexical development in young children and adults in order to gain insights into the structure of the mental lexicon (Beckage & Colunga, 2019; Beckage et al., 2011; Fourtassi et al., 2020; Hills et al., 2009b). Moreover, semantic network analysis has been useful for investigations of psycho-emotional experiences of language users in certain contexts (e.g., Oh et al., 2023; Che et al. 2023).

Semantic networks are inherently useful to psycholinguistics as they enable researchers to study semantic relationships by considering a more holistic view of the semantic lexicon and its interrelationships (De Deyne et al., 2017; De Deyne et al., 2018; Fourtassi et al., 2020; Hills et al., 2010; Hills et al., 2009a, 2009b; Kenett et al., 2014; Stella, 2020; Stella et al., 2018). Rather than focusing on target words and their semantic neighbors (i.e., 'semantic neighborhood analysis'), a lexical network incorporating relationships between all words in an entire lexicon can be much more

informative of semantic memory and retrieval, lexical competition, and acquisition. Even though semantic network analysis has been identified as a superior methodology to study cognitive processes of the human lexicon (Siew et al., 2019), the research field has been hampered by the fact that data collection and network construction are tedious and laborious, in addition to a lack of resources for researchers (Christensen & Kenett, 2023).

Graphs of links between words are the core of semantic network science. One must keep in mind, however, that what the graph tells is inherently vague. Is every entity represented by the nodes of the same status? Is every connection represented by the links of the same status? There are several visual cues that can be inserted in the graph to achieve a more detailed depiction of the information present in the dataset, such as the thickness of lines ('edge strength' in network terms), as well as varied colors and shapes of nodes according to certain criteria. In the case of semantic network graphs, what the nodes represent is clear — it is word meanings. Even though a debate can be had about what a *word* and its meaning is (e.g., Haspelmath, 2011), there are various ways in which links between words can reliably be quantified. Especially word associations are psycholinguistically clear-cut notions, as they solely represent the lexical relationships that individual speakers have in their minds. These associations are frequently collected (and used) in the context of semantic priming studies (McNamara, 2012). Priming is a cognitive process where exposure to a stimulus affects perception; in other words, hearing a target word activates other (neighboring) words that are perceived to be semantically close to the target word (Schreuder et al., 1984). Priming studies typically ask participants about their lexical associations — "*What comes to your mind when you hear/ see the word xx?*" — and harvest the complex real-world compilation of language use in relation to object and concept perception. Most often, priming is asymmetric in the sense that some target words have few primes but are asymmetrically primes to many other target words. Semantic networks can offer a new method to investigate topics such as asymmetrical priming, which has been frequently discussed in recent years (Yu, L., Zhang, Q., Ke, M., et al., 2022, Hilpert, M., 2021). In general, semantic priming has been shown to be robust across different studies and populations (Heyman et al., 2018), with the priming paradigm constituting a central tenet of psychological and psycholinguistic research. Primes can be seen as simple causal relationship between lexical items that can be represented by a link in a network. If enough participants are surveyed, we can say our data has some empirical validity and that it represents something larger than an individual, perhaps a synchronous state of a given language. Once that is achieved, we gain a reliable reference for what semantic relationships there are in a language.

The goal of the present paper is to describe and make easily accessible to a wider research audience a tool for constructing semantic networks based on primes called fan_xplorr. First, the general aim of the work will be described, followed by an outline of a practical application of the semantic network analysis as provided by fan_xplorr. For readers interested in the technical aspects of the tool, a description of the technical details and an installation guide has been compiled in the Appendix.

2 THE DATABASE

One of the largest databases of lexical associations (i.e., primes) for English as a first language (L1) is the *University of South Florida Free Association Norms* database (see Nelson, D. L., McEvoy, C. L., & Schreiber, T. A., 2004, accessible at: http://www.usf.edu/FreeAssociation/). The database contains over 5,000 words and their associations and provides a great resource for the construction of English lexical networks. Data is available online in text format. The dataset requires some manipulation to become useful for visualization of nodes and edges in a network. Since large-scale networks are too dense to be usefully visualized, it is also important to be able to zoom-in on particular areas of the network to view relevant pieces of information.

The research uses for the dataset and networks are manifold. One can explore the dataset itself, identify asymmetric primes, and cycles (e.g., as seen in Figure 1, the words *holder* and *handkerchief* are not directly associated, but there are two paths between them, each secured by a different word and semantic relationship), or examine polysemy (e.g., the word *mug* has primes like *attack* and *alley* but also primes like *beer* and *glass*). Furthermore, the dataset is prepared in a way that it facilitates the calculations of a number of network statistics, including degree centrality (i.e., the number of neighbors or semantic neighborhood density), clustering coefficient (i.e., the interlinking of neighboring nodes in a neighborhood), and node relevance (i.e., hubs).

3 DESCRIPTION OF FUNCTIONALITY

As a start, the lexical prime data in text format (Figure 2) was downloaded from the website and converted into JSON format that can capture the connections between the nodes in a way that can be processed by a computer.

```
BOULDER      FSG   BSG  MSG  OSG  QSS TSS  QFR TFR   QMC   TMC   QUC TUC

 ROCKS       0.040.0000.0000.0023  20   8   23   10  1.56 0.43    0   1
 ROCK        0.030.6560.0000.0005  13   8   75   10  0.46 0.43    1   1

 NO. OF CUES:  2
```

FIGURE 2. Original format of the data

This allows us to query the dataset for selected parts (words). Let us look at an example, the node *lemon.*

We are interested in viewing the primes of *lemon*. First, we need to find the node *lemon* in the dataset.

Next, we can have a look at which words have been recorded as its primes. This gives us a star-shaped structure involving a number of nodes where every node is connected to *lemon* as a prime. This is interesting but does not provide any extra information other than what we can already see in the textual representation of the

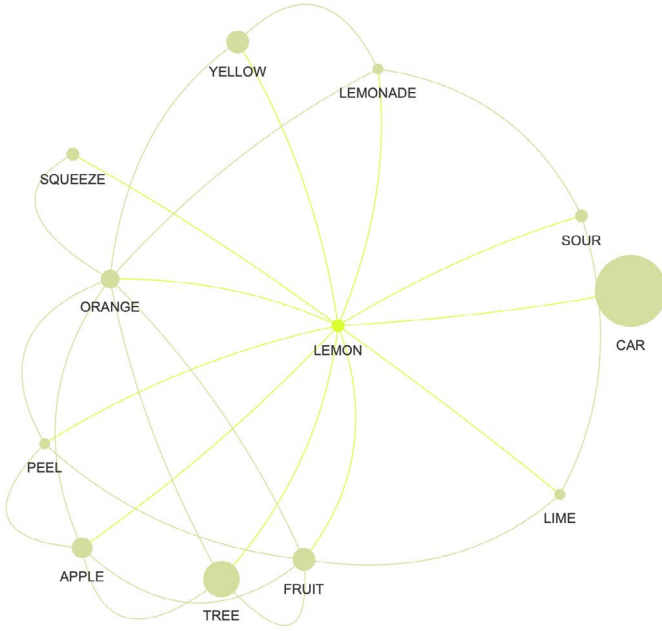

FIGURE 3. Primes of *lemon*

data. We also want to see if there are any connections between the primes themselves (i.e., clustering of the network). We might be missing a well-connected node, a hub, in the *lemon* neighborhood, which can inform us about influential primes. So how do we go about that? We take every word one by one from the current pool (in this case all the primes of *lemon*) and check the following in the dataset:

1) Has the word been a researched node?
2) If so, are any of the *words in the current pool primes of this word?*
3) Semantically similar words often share primes, so there usually will be more connections than just the ones forming the star. After adding the necessary edges to the star-shaped structure we get Figure 3.

We see that *fruit* is a prime of lemon. If we are interested in identifying the primes of *fruit* (to see how many primes it shares with *lemon*), we now focus on two nodes and all their primes. The process is the same as before, only involving two target words. Let us now turn to Figure 4, which is a representation of the two-word query. We can query fan_xplorr for as many words we like. One might expect *fruit* to have more primes. This depends on the methodology of the research and which lexical items have been researched. However, there is a way we can display the word *fruit* so that it will be a large hub in the graph. It is now that asymmetrical priming comes into play. If we adopt the hypothesis that *fruit* is subject to asymmetrical priming, we will expect the number of primes of *fruit* will be very different from the number of words

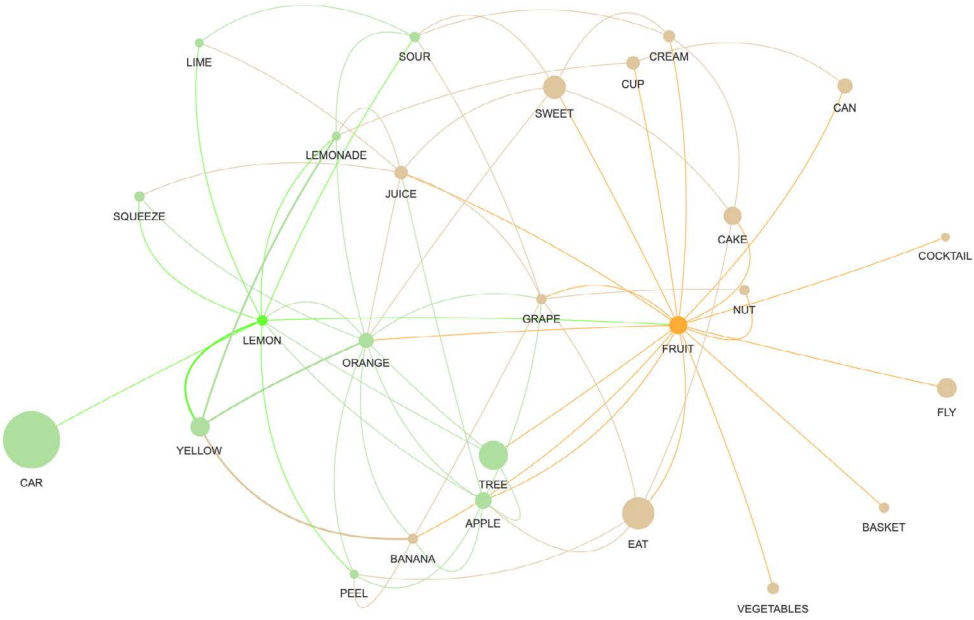

FIGURE 4. Primes of *lemon* and *fruit*

that *it is a prime to*. How do we research this? Every node in the dataset is checked to see whether *fruit* is to be found there as a prime. The visualization will be the same — *fruit* in the middle of a star-shaped network graph (again enriched by the edges between the other nodes). But this time it means something else — namely that *fruit* is the prime and the nodes around are the words that have been surveyed while gathering this data. Figure 5 shows the results. *Lemon* was excluded from the query this time for the sake of simplicity, but it can still be seen, since *fruit*, as we know, is a prime of *lemon*.

The difference in node sizes reflects the number of primes the nodes have (i.e., node size according to degree centrality of the network or the number of overall neighbors).

4 A PRACTICAL USE OF FAN_XPLORR

Apart from it being a reliable source on semantic neighborhood statistics, there are many network-related psycholinguistic uses for the fan_xplorr functions in the realm of semantic network analysis.

We suggest a study focused on the network statistic referred to as "clustering coefficient". It describes the degree to which semantic neighbors of a target word are also neighbors of themselves (Barabási, 2016). Studies on word form similarities in the mental lexicon of language users have demonstrated an effect of neighborhood clustering on the efficiency and accuracy of word retrieval. Namely, words with highly

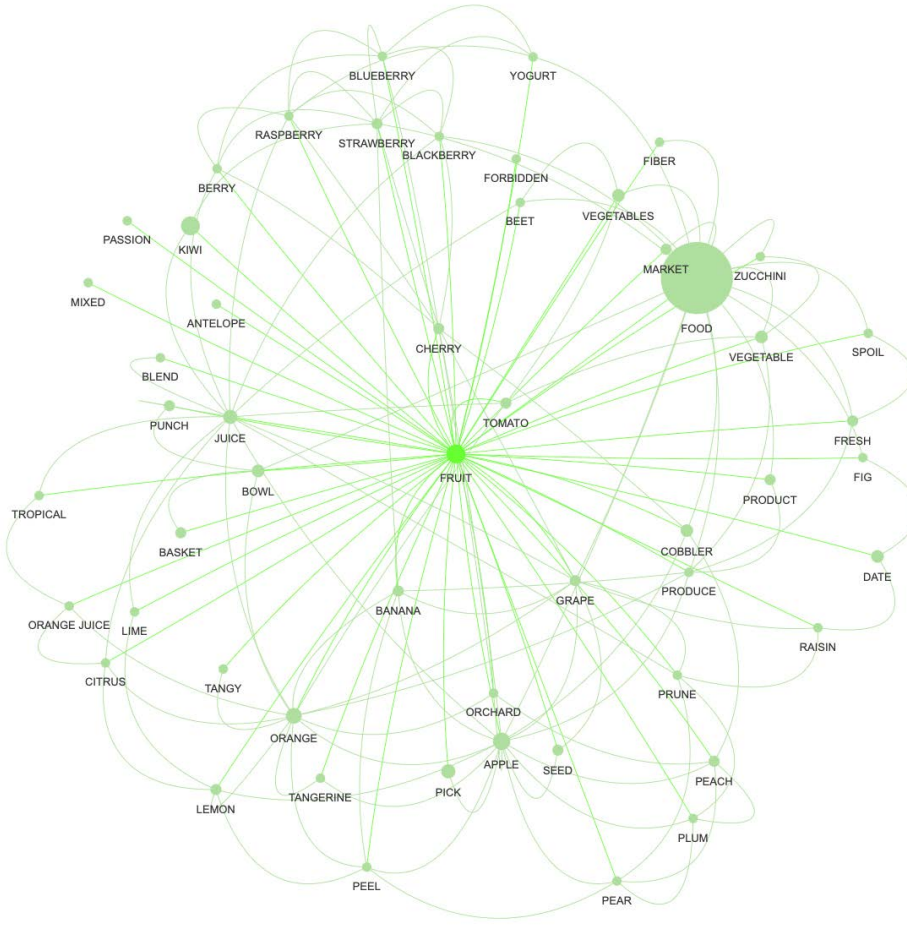

FIGURE 5. Words to which *fruit* is a prime

clustered neighborhoods — in which most neighbors are neighbors of one another in addition to the target word — pose processing challenges to speakers and lead to longer reaction times to words in lexical decision experiments (Vitevitch, 2021). Given the similarities between phonological and semantic neighborhood characteristics in relation to speech processing, it is safe to assume that a similar neighborhood dynamic might rule semantic neighborhoods.

The clustering coefficient of a semantic neighborhood is a statistic difficult to obtain from available databases but it requires a network-based approach to semantic neighborhoods, such as provided by fan_xplorr. For instance, the target word "scarf" (as illustrated in Figure 1) has a semantic neighborhood metric of 5, that is the number words English speakers associate with "scarf". This information is then added to the following equation to obtain the clustering coefficient ("CC") of the "scarf" neighborhood:

$$CC_{scarf} = 2e/(k(k-1))$$

Here, k corresponds to the number of neighbors of "scarf" (=5) and *e* is the number of linked word pairs (i.e., edges) between all neighbors of "scarf" (=1), obtained by counting all links among the neighbors (see Goldstein & Vitevitch, 2014; Schank & Wagner, 2005). CC yields a number between 0 and 1, with 0 indicating no clustering in the neighborhood (i.e., neighbors are not neighbors of one another), and 1 indicating high clustering, with all neighbors being neighbors of one another. The "scarf" neighborhood has a clustering coefficient of 0.1 and thus ranges among the lowly clustered neighborhoods. When one compares that to the "napkin" neighborhood (see Figure 1), which is characterized by CC=0.33, the difference in CC between the neighborhoods could mean a difference in lexical processing speed of the two target words. A lexical decision task can reveal whether low-CC neighborhoods have a processing advantage over high-CC neighborhoods.

The network-based representation of the mental lexicon of English speakers that is provided by fan_xplorr adds an important facet to the study of word associations. It allows a closer inspection of semantic neighborhoods by going beyond the simple target-neighbor links and computing the wider neighborhood and its interlinking. Through this, certain neighborhood metrics, such as CC, can be obtained, and introduced to a wider audience in psycholinguistics.

5 CONCLUSION

Lexical networks offer an intriguing approach to studying language, offering an empirical perspective on the intricate semantic relationships between words. The tool presented here attempts to make the exploration and analysis of semantic networks of English as a first language more accessible and user-friendly. By simplifying the process, researchers and language enthusiasts can now delve deeper into the complexities of semantic links, revealing hidden patterns and gaining insights into the ways words and their meanings shape our understanding of the world.

REFERENCES

Barabási, A.-L. (2016). *Network science*. Cambridge: Cambridge University Press.

Beckage, N., Smith, L. B., & Hills, T. T. (2011). Small worlds and semantic network growth in typical and late talkers. *PLOS ONE*, *6*(5), e19348. https://doi.org/10.1371/journal.pone.0019348

Beckage, N., & Colunga, E. (2019). Network growth modeling to capture individual lexical learning. *Complexity*, *2019*, 1–17. https://doi.org/10.1155/2019/7690869

Chan, K. Y., & Vitevitch, M. S. (2010). Network structure influences speech production. *Cognitive Science*, *34*(4), 685–697. https://doi.org/10.1111/j.1551-6709.2010.01100.x

Che, S. Wang, X. Zhang, S. et al. (2023). Effects of daily new cases of COVID-19 on public sentiment and concern: Deep learning-based sentiment classification and semantic network analysis. *Journal of Public Health (Berl.)*.

https://doi.org/10.1007/s10389-023-01833-4

Christensen, A. P., & Kenett, Y. N. (2023). Semantic network analysis (SemNA): A tutorial on preprocessing, estimating, and analyzing semantic networks. *Psychological Methods*, *20*(4), 860–879. https://doi: 10.1037/met0000463

De Deyne, S., Kenett, Y. N., Anaki, D., Faust, M., & Navarro, D. J. (2017). Large-scale network representations of semantics in the mental lexicon. In Jones M. N.: *Psychology Press eBooks* (pp. 183–189). https://doi.org/10.4324/9781315413570-18

De Deyne, S., Navarro, D. J., 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. https://doi.org/10.3758/s13428-018-1115-7

Fourtassi, A., Bian, Y., & Frank, M. C. (2020). The growth of children's semantic and phonological networks: insight from 10 languages. *Cognitive Science*. https://doi.org/10.31234/osf.io/37npj

Goldstein, R., & Vitevitch, M. S. (2014). The influence of clustering coefficient on word-learning: How groups of similar sounding words facilitate acquisition. *Frontiers in Psychology*, *5*, 1307.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, *45*(1), 31–80. https://doi.org/10.1515/flin.2011.002

Heyman, T., Bruninx, A., Hutchison, K. A., & Storms, G. (2018). The (un)reliability of item-level semantic priming effects. *Behavior Research Methods*, *50*(6), 2173–2183. https://doi.org/10.3758/s13428-018-1040-9

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. B. (2009a). Longitudinal analysis of early semantic networks. *Psychological Science*, *20*(6), 729–739. https://doi.org/10.1111/j.1467-9280.2009.02365.x

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. B. (2009b). Categorical structure among shared features in networks of early-learned nouns. *Cognition*, *112*(3), 381–396. https://doi.org/10.1016/j.cognition.2009.06.002

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, *63*(3), 259–273. https://doi.org/10.1016/j.jml.2010.06.002

Hilpert, M. (2021). *Ten lectures on Diachronic Construction Grammar*. Leiden: Brill. https://doi.org/10.1163/9789004446793

Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00407

McNamara, T. P. (2012). *Semantic Priming: Perspectives from Memory and Word Recognition*. New York: Psychology Press. http://ci.nii.ac.jp/ncid/BB13889015?l=en

Mirman, D., Kittredge, A. K., & Dell, G. S. (2010). Effects of near and distant phonological neighbors on picture naming. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*(32). 1447–1452. https://escholarship.org/content/qt5620c08n/qt5620c08n.pdf?t=op2ksz

Mirman, D., & Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *34*(1), 65–79. https://doi.org/10.1037/0278-7393.34.1.65

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods Instruments & Computers*, *36*(3), 402–407. https://doi.org/10.3758/bf03195588

Oh, M., Badu Baiden, F., Kim, S., & Lema, J. (2023). Identification of delighters and frustrators in vegan-friendly restaurant experiences via semantic network analysis: Evidence from online reviews. *International Journal of Hospitality & Tourism*, *24*(2), 260–287.

Reilly, M., & Desai, R. H. (2017). Effects of semantic neighborhood density in abstract and concrete words. *Cognition*, *169*, 4653. https://doi.org/10.1016/j.cognition.2017.08.004

Schank, T. & Wagner, D. (2005). Approximating clustering coefficient and transitivity. *Journal of Graph Algorithms and Applications*, *9*, 265–275.

Schreuder, R., D'Arcais, G. B. F., & Glazenborg, G. (1984). Effects of perceptual and conceptual similarity in semantic priming. *Psychological Research-psychologische Forschung*, *45*(4), 339–354. https://doi.org/10.1007/bf00309710

Siew, C. S. Q., & Vitevitch, M. S. (2016). Spoken word recognition and serial recall of words from components in the phonological network. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *42*(3), 394–410. https://doi.org/10.1037/xlm0000139

Siew, C. S. Q., Wulff, D. U., Beckage, N. M., Kenett, Y. N., & Meštrovic, A. (2019). Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, 2019, 1–24. https://doi.org/10.1155/2019/2108423

Siew, C. S. Q., & Guru, A. (2023). Investigating the network structure of domain-specific knowledge using the semantic fluency task. *Memory & Cognition*, *51*, 623–646.

Stella, M. (2020). Multiplex networks quantify robustness of the mental lexicon to catastrophic concept failures, aphasic degradation and ageing. *Physica D: Nonlinear Phenomena*, *554*, 124382. https://doi.org/10.1016/j.physa.2020.124382

Stella, M., Beckage, N., Brede, M., & De Domenico, M. (2018). Multiplex model of mental lexicon reveals explosive learning in humans. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-20730-5

Steyvers, M., & Tenenbaum, J. B. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, *29*(1), 41–78. https://doi.org/10.1207/s15516709cog2901_3

Vitevitch, M. S. (2021). What can network science tell us about phonology and language processing? *Topics in Cognitive Science*, *14*(1), 127–142. https://doi.org/10.1111/tops.12532

Yu, L., Zhang, Q., Ke, M., Han, Y., & Kinoshita, S. (2022). Some neighbors are more interfering: Asymmetric priming by stroke neighbors in Chinese character recognition. *Psychonomic Bulletin & Review*, *30*(3), 1065–1073. https://doi.org/10.3758/s13423-022-02207-9.

Albert Maršík
Charles University, Faculty of Arts
Institue of Czech language and theory of communication
nám. J. Palacha 2, 110 00 Prague 1, CZ

Eva Maria Luef
Charles University, Faculty of Arts
Dept. of English Language and ELT Methodology
nám. J. Palacha 2, 110 00 Prague 1, CZ
University of Hamburg, Institute of English and American Studies
Von-Melle-Park 6, 20146 Hamburg, Germany
ORCID-ID: 0000-0002-2362-2422
evamaria.luef@ff.cuni.cz

APPENDIX

A

Installation and behavior

The following section will guide you through the installation of the tool. The only thing necessary is to have Python3 installed. It can be downloaded from python.org. It is also very useful to have *git* as it will simplify installation and updating in the future.

git

If you have git installed, all you need to do is follow the steps on https://github.com/almarsk/fan_xplorr. The installation guide is within README.md. After having installed the prerequisite, you can run "wizard.py". The *wizard.py* in step 3) does most of the work. If you want to review the installation steps, they are in the *setup/setup_steps.json* for macOS/Linux and *setup/setup_steps_w.json* for Windows. If you are using Windows, it is best to use the *git cmd* application that comes with the *git* for Windows download. The setup is not suitable for *PowerShell*. If you want to use *PowerShell*, follow the manual installation steps but use the PowerShell command equivalents.[1]

If you are not yet a *git* user, it is recommended to become one. Any future updates of the *fan_xplorr* project will be easily accessible via *git*.

download ZIP

In case you are not familiar with *git*, click the green code button on https://github.com/almarsk/fan_xplorr and choose the "download ZIP" option. Then extract the downloaded folder to your chosen location. Open the command line on Windows or a terminal on macOS/Linux and navigate to the extracted folder. From then on you can run the python3 wizard.py (or python wizard.py) command.

The command line

A command line or terminal is another way of interacting with the computer other than the graphic user interface. The command line has two specificities, which need to be considered — 'path' and 'history'. Your command line is always in one place in your computer's file system which is why you need to navigate to the downloaded folder with the *cd* command. Otherwise, the wizard.py script cannot be seen. The *history* will be discussed in the *behavior* section, where it will help achieve a smooth workflow, a central tenet of *fan_xplorr*.

Manual installation

If your installation went well, you can skip this section. If you run into issues, feel free to start a thread in the issues section of the GitHub page of the project. Please try the manual installation first. It may not solve the problem but it can help isolate it.

1 There is no specific reason for this, and a contributor can create an installation wizard for PowerShell as well.

To uninstall
If you want to remove *fan_xplorr* from your computer, you can simply delete the folder. Everything will be removed then, including the processed dataset (unless you saved it elsewhere before). This action will not remove Python from your computer.

Data
During the installation, *wizard.py* downloaded and parsed the data from the USFFAN website. They can be found in the *data* folder. There is the raw text concatenated from all the alphabet sectors on the website in the *raw_florida_free_association_norms.txt* and there is *the USFFAN_just_words.json* file, which contains the readable data in the *.json* format. You will notice that it contains only words. As a part of the research output of the USFFAN project, concrete numbers have been published, too. If you want an equivalent *.json* file with these numbers included, run *python3 setup/process_data.py*. This extra information can be used for further improvements of *fan_xplorr* but has been neglected for now, as it takes up slightly more memory and time. For *fan_xplorr* to work, the *.json* file should stay in the data folder, but it can be copied elsewhere to perform other operations.

Behavior
Some thought has been put into the workflow of *fan_xplorr*. Assuming all setup steps have been successful, and the command line's location is in the root of the *fan_xplorr* folder, we can now use two commands *xplor* and *rnm*![2] If we simply run the *xplor* command, it will tell us what to do:

> *Xplor creates a network graph in the ‚graphs' directory.*
> *Follow the command up by words you want to be a part of the graph.*
> *They just need to be separated by space.*
> *To change the name of the destination file, run ./rnm*

Xplor takes the words we feed it and creates graphs as shown in the previous section. You will find them in .html format in the graphs folder. It creates an HTML file for both options from the previous section and one which is combined. A note of caution: if you just move the combined one out of the file it will not work, since it uses the two previous files as a source (i.e., it reads them).

Once you have generated a graph, you can open it in your web browser with

start graphs/nodes_both.html (Windows)
open graphs/nodes_both.html (macOS/Linux)

This is when the *history* feature of the command line becomes crucial. If you do not yet know what graph you want to create you can try out different words, and you may decide that you want to change one word but keep the rest of the currently used

2 ./xplor and ./rnm on macOS/Linux; xplor.bat and rnm.bat on Windows

words. If you hit the *arrow up* key in the command line, you will find your last command ready to be run again. This means you can modify it easily and wander around the dataset. The next graph will overwrite the old one, so it will not overburden your graphs folder with nearly identical graphs. With each new graph of the same title, the browser can simply be refreshed to see the change.

Once you are done with your graph and you want to move on to the next one, the *rnm* command comes into play. Running *rnm* with a word after it changes the name of the file into which the future graphs will be saved. This workflow was designed so that you can tune your graph or just surf the dataset.

Thanks to the Python *Pyvis* library the graphs are interactive. If there is an inconvenient overlap of nodes in the graph, you can manually drag the nodes and rearrange them.

B

What else can be done?

This section will describe some possible enhancements to the tool. It will serve as a list of suggestions for anyone who might decide to contribute.

Add --open flag

The *xplor* command could have a *--open* flag that would open the combined graph in the browser directly. This would mean checking for the flag in the *xplor* and the *src/xplor.bat* file and adding the extra command with a condition.

Incorporate numerical data

This suggestion slightly defeats the purpose of *fan_xplorr*, but the data in the .json files can be used for all sorts of other endeavors. If you have an idea about how to implement it in the workflow and visual of *fan_xplorr*, please get in touch.

Large hubs

As of right now, large hubs are an issue. The generation of the HTML file works fine but the browser takes time to load all the connections, the number of which rises exponentially. Running *./xplor food* creates an HTML file that takes up to a couple of minutes to load for example. Any suggestions or improvements on how to sort and filter the connections to keep the functionality and add performance are welcome.

Add thickness to the edges

The numerical data could be used to change the strength of the edges (i.e., weighted degree centrality in network terms). The more people have stated a certain lexical link, the thicker the links could be. This would mean parsing the full dataset and looking into which column contains this information. The Pyvis library offers customization of the edge strengths, so it is possible. The ultimate edge strength value should probably be run through a function to avoid extreme values.

Random colors are too close to each other
Sometimes the randomly generated colors are too close to each other. A checker for the RGB values could be implemented. In addition, if someone wanted to generate several graphs with the same node, the color would be different every time. Implementing the possibility to customize the color for some of the nodes might produce better results in these contexts.

Add metadata to the resulting graph
The combined graph output could be perfected to be standalone (independent of the other two graphs) and to show metadata about the nodes by default. All of this can be done but it is a question of preferred workflow. Feel free to start a thread on GitHub on this.

Check the time of writing to prevent accidental overwriting
A check could be added to the HTML files and if they have not been edited in a certain period of time (such as one hour), a warning and a confirmation request could be displayed, for instance, *„You haven't edited this destination graph in more than an hour. Are you sure you're not overwriting something that's already been finalized? Y/N*".

More target formats
More target formats (e.g. gexf) could be added to make the data available in other network visualisation/analysis software like Gephi.

GUI
A graphical user interface would make this tool much more approachable to linguists.