How Human-Like are Word Associations in Generative Models? An Experiment in Slovene

Mojca Brglez*, **Špela Vintar***, **Aleš Žagar**†
*Faculty of Arts, University of Ljubljana
Aškerčeva 2, Ljubljana
mojca.brglez, spela.vintar@ff.uni-lj.si
†Faculty of Computer and Information Science, University of Ljubljana
Večna pot 109, Ljubljana
ales.zagar@fri.uni-lj.si

Abstract

Large language models (LLMs) show extraordinary performance in a broad range of cognitive tasks, yet their capability to reproduce human semantic similarity judgements remains disputed. We report an experiment in which we fine-tune two LLMs for Slovene, a monolingual SloT5 and a multilingual mT5, as well as an mT5 for English, to generate word associations. The models are fine-tuned on human word association norms created within the Small World of Words project, which recently started to collect data for Slovene. Since our aim was to explore differences between human and model-generated outputs, the model parameters were minimally adjusted to fit the association task. We perform automatic evaluation using a set of methods to measure the overlap and ranking, and in addition a subset of human and model-generated responses were manually classified into four categories (meaning-, position- and form-based, and erratic). Results show that human-machine overlap is very small, but that the models produce a similar distribution of association categories as humans.

Keywords: word associations, generative models, T5, multilingual, Slovene

1. Introduction

Free word associations are a widely known technique of researching the human mental lexicon and have been used well before the emergence of psycholinguistics as a discipline (Galton, 1879). Having participants give oral or written associations to a cue word with as little reflection as possible may sound like a simple task, but as it turns out the responses given by different people show great variation both in the type of semantic relation governing the cue-response pair (cat -> dog vs. cat -> black vs. cat -> rat) and the individual association style (Fitzpatrick, 2007).

Since human associations to a large extent adhere to some patterns of semantic, syntactic or orthographic proximity, the emergence of vector-space meaning representations and early language models soon motivated a number of studies comparing different notions of relatedness in the human mental lexicon and that of a language model (see Section 2).

In this work, we describe an experiment which follows a similar aim, but for the first time such a comparison can be performed for Slovene, mostly because the human association dataset (SWOW-SL) for this language has been created only recently (see Section 3). We fine-tune two generative models to perform the task of responding to the prompt *"Which words do you associate with the word [WORD]?"* and select the parameters best suited to the association generation task. Since our aim is not to achieve maximum overlap between human and machine output but to better understand the workings of the artificial semantic space, we perform a series of evaluations. These include five different metrics to measure the concordance between the human and neural responses, and a qualitative evaluation through manual annotation in order to analyse the types of associations produced by humans and the language models.

In short, our key contributions in this work are:

- We construct and describe the first general dataset for Slovene word associations;
- We are the first to use generative models to explore word associations in language models;
- We are the first to test and evaluate the associations as output by generative models, using quantitative and qualitative methods in both English and Slovene.

2. Related Work

In recent years, several studies have attempted to evaluate the ability of vector space models to represent conceptual organization. Mandera et al. (2017) perform a detailed evaluation of correlations between human semantic spaces and corpus-based vector representations, whereby for the former they use semantic priming, semantic related-

The Workshop on Cognitive Aspects of the Lexicon, pages 42–48

ness judgements and word associations, while the vector-space models tested include count (LSA and HAL) and static neural models, referred to as predict models (skipgram and CBOW). They find that predict models, especially CBOW, consistently outperform traditional count models, and that the window size used in training plays a significant role in the performance of the models. They also report that a larger training corpus did not necessarily improve the results as in several experiments models trained on a smaller subtitle corpus outperformed those trained on UKWaC.

In an experiment by Nematzadeh et al. (2017) human word associations were compared to nearest neighbours suggested by word2Vec and GloVe, and they show that overall correlation is low and that static word embeddings fail to capture certain critical aspects of human associations.

The debate about common misconceptions about what word embeddings do or do not represent from a cognitive linguistics viewpoint was continued by Günther et al. (2019). One important emphasis, relevant also for our own experiments, is that while neural models are extremely powerful in producing quantitative representations of word meaning from (almost exclusively) textual data, the original idea behind Latent Semantic Analysis (LSA) was of it being not merely a computational but an explanatory tool, shedding light to how "word meanings are acquired through experience". For this reason, models trained directly on introspective data generally outperform corpus-trained ones (De Deyne et al., 2016).

Along similar lines, Jones et al. (2018) point out that association retrieval in humans is not symmetric, hence cosine distance may not be the best way to predict association strength. A more recent detailed discussion of the complexity of human associative behaviour and neural modelling is provided by Richie et al. (2022) who also train their GloVe model on English SWOW (De Deyne et al., 2018) and achieve good prediction results using a variety of asymmetric measures.

Although many studies have explored how human associations are represented in vector space models, they have all done so through indirect intrinsic measures (cosine similarity, probability distributions). To the best of our knowledge, our study is the first to directly explore the generation of associations with large language models. Secondly, the study is the first to examine the automatic generation of word associations for Slovene.

3. Datasets

3.1. English SWOW

For the training of the English model we use word association norms created in the English Small World of Words project (SWOW-EN, (De Deyne et al., 2018)). The data set consists of over 12,000 cue words and responses from over 90,000 participants, which makes it the largest resource of its kind for English. Responses were (and still are) collected through a web interface which presents association collection as a game, and participants were recruited via social media, e-mail and university websites. Over the years the SWOW project, originally developed for Dutch, grew into an ongoing world-wide study which currently covers 19 languages, including Slovene.

As a pre-processing step on SWOW-EN, we discarded the cue words which were labelled with a meaning gloss, as in *bat*, *bat(animal)*. The data was split into a training and testing set, whereby both subsets were sampled proportionally with regard to the PoS frequency.

3.2. Constructing SWOW-SL

The data collection for SWOW-SL[1] was supported by the generous help of Simon De Deyne (University of Melbourne) and Gert Storms (University of Leuven), the authors of the original SWOW project, who kindly added Slovene as another language on the SWOW platform, imported the data for the experiment and set up the localized web pages describing the task. The experiment for Slovene is the same as for other languages; a participant is consecutively shown 18 cue words and is asked to contribute up to 3 associations to each. At the end of the experiment the participant is shown some preliminary results, such as overlap with other participants and basic project statistics.

The selection of cue words for Slovene was based on the frequency lexicon from the Gigafida 2.0 corpus (Krek et al., 2020), whereby we limited the part-of-speech for cues to nouns, adjectives, verbs and adverbs, and then selected lemmas from the top 500-1500 frequency-ranked items, but removing proper names, acronyms and adjective-adverb duplicates (e.g. *dober* - *dobro*). The data collection for Slovene started in November 2023 and has reached 671 participants and the time of writing this article. Our dataset was constructed with responses up to January 10, 2024. To that point, each cue word has received responses from 8–10 different respondents, with approximately 17 unique responses per cue word on average.

[1] https://smallworldofwords.org/sl

Before dividing the Slovene dataset into testing and training splits, we apply lowercasing in order to normalize the responses. Because previous works (De Deyne and Storms, 2008) have shown that response categories can vary by the PoS of the cue, we sample the test (evaluation) set proportionately from each PoS according to its relative frequency.

Dataset	train	test
SWOW-en	10.946	611
SWOW-sl	949	51

Table 1: Datasets and data splits. The number refers to the number of cues.

Table 2 presents the structure of a training example for Slovene and English. Both consist of an immutable input prompt integrating the **cue** and the expected output with all the human *responses*.

4. Experimental Setup

4.1. Models

We employed two state-of-the-art models, namely SloT5 (Ulčar and Robnik-Šikonja, 2023) and mT5 (Xue et al., 2021). Both models are rooted in the transformer architecture, characterized by an encoder-decoder framework, and have been pretrained to generate text effectively.

For our experimentation, we utilized preprocessed datasets described in the previous section. The SloT5 model was deployed in a monolingual setting, focusing solely on the Slovene language version of the dataset. In contrast, the mT5 model, known for its multilingual capabilities, was trained on a concatenated dataset comprising both Slovene and English versions, thereby facilitating a multilingual experiment.

4.2. Evaluation

To assess the trained generative models on how well they align with human associative networks, we evaluate 1) the overlap between human and model responses, 2) the ranking of responses, and 3) the categories of the responses. We employ five distinct automatic metrics for the first and second aspect, and perform manual annotation on a sample of data for the third aspect.

Automatic Metrics To assess the performance of our trained generative models, we employ five distinct automated metrics, including 4 similarity and 1 distance metric:

1. **Jaccard** similarity compares the size of the intersection of two sets to their union, i.e. it provides a measure of how much overlap exists between two sets of items with regard to the whole number of distinct items in both sets.

2. Rank-based Overlap (**RBO**, Webber et al., 2010), used especially in information retrieval, evaluates the similarity between two sets by considering both the overlap and the ranking of items. This metric assigns a larger weight to items appearing higher in the list of the gold standard (here, our the human responses).

3. **Precision** and **Recall**: Conventional metrics used to evaluate the accuracy of the responses.

4. Word Mover's Distance (**WMD**, Kusner et al., 2015) assesses the minimal "transportation cost" needed to move from one set of word embeddings to another. We use this metric to measure semantic similarity beyond direct word matches, where a lower score means a shorter distance travelled and thus a more similar set of words.

To enhance the clarity and interpretability of the automatic metrics, we also assess each metric on the overlap of human-generated associations. We simulate the human subsets by first concatenating all responses for a specific cue according to their associative strength. Then, we distribute these responses into two groups, alternating between each group for every response in the ranked list.

Manual Annotation of Response Categories In addition to evaluating the performance of the trained models with the methods described above, our aim was to better understand the behaviour of neural models with respect to the types of responses they produce. As proposed by (Fitzpatrick, 2007), human associations can be classified into four broad categories, namely:

- **Meaning-based association**: *x* means the same as *y*, *x* and *y* come from the same lexical set, *x* and *y* have some other conceptual link
- **Position-based association**: *y* follows or precedes *x* directly or with words between them
- **Form-based association**: *y* is *x* plus or minus affix, *y* looks or sounds similar to *x*
- **Erratic association**: *y* has no decipherable link to *x*

We first performed a manual annotation of categories on a random selection of 50 Slovene cues with human responses. A total of 996 responses were classified into the above categories by at least two annotators, followed by a curation phase

Input	Na kaj pomisliš ob besedi **umetnik**? [What comes to mind when thinking about the word **artist**?]
Output	Najprej pomislim na besede: *jaz, galerija, flamenko, slika, slikar, poet, slika, kri, beretka, lev, molitev, oblaki, dolina šentflorijanska, kreator, barve, pavarotti, len, ljubezen, radost, poslanstvo, čustva, picasso, lepo* [I first think of the words: *me, gallery, flamenco, painting, painter, poet, painting, blood, beret, lion, prayer, clouds, St. Florian valley, creator, colours, pavarotti, lazy, love, joyfulness, calling, emotions, picasso, beautiful*]
Input	Which words do you associate with the word **arctic**?
Output	I associate this word with: *cold, ice, snow, polar bear, circle, penguin, polar, north, white, tundra, Antarctic, North Pole, freezing, icy, ocean, penguins, Antarctica, bear, fox, freeze, monkeys, polar bears, pole, frozen, glacier, iceberg, igloo, roll, air, Arctic circle, Arctic ocean, Aurora, bears, blue, char, chilly, clear, collapse, conditions, continent, cruise ship, dappled, enema, Eskimo, expedition, explorer, extreme, far, far away, flexible, free, frigid, frost, hare, icebreaker, lights, monkey, orca, owl, p, pudding, region, resonance, Russia, sadness, Santa, seals, ship, slicer, snot, snowy, software, spare, spontaneous, temperature, war, wind, winter, wolf, zone*

Table 2: A Slovene and an English training example, consisting of an input prompt integrating the cue (in bold) and the expected output with all the human responses (in italic)

.

to resolve inter-annotator disagreements. The inter-annotator agreement between pairs of annotators was on average moderate with Cohen's Kappa score of 0.507, but the values varied greatly amongst pairs ranging from meager 0.147 to 0.80. Since the annotators were students who were given only a brief training before the actual annotation, many inconsistencies were resolved later through discussion and curation. On the other hand the task itself is somewhat ambiguous as many responses could legitimately be assigned several categories.

The category frequencies of human responses by cue part-of-speech are given in Table 3. Over one half of responses fall into the meaning-based category, with verbal cues deviating from the typical distribution of categories by favouring position-based associations. It would appear that verbs as cues are stronger triggers for collocational patterns than other part-of-speech.

A similar classification of responses was then performed for a randomly selected set of 10 cue words for all three evaluated models: SloT5, mT5-SL and mT5-EN.

PoS	Erratic	Form	Meaning	Position
Adj	12	13	72	60
N	52	27	311	114
Adv	19	2	68	27
V	16	15	71	117
Total	99	57	522	318

Table 3: Categories in human annotated associations by PoS of the cue word

5. Results

To fine-tune the SloT5 and mT5 models for our specific research task, we employed a consistent set of hyperparameters across both SloT5 and mT5 models. These included a learning rate of 5×10^{-5}, a training span of 10 epochs, a batch size of 8, and the $AdamW$ (Loshchilov and Hutter, 2019) optimizer.

For the inference phase, careful consideration was given to the selection of parameters with the aim of preserving the model parameters (i.e. the existing network and representations) at their default values and adjusting them only slightly to obtain structurally sound outputs and to reduce repetitive behaviour from the models. The parameters configured were as follows: sampling was $enabled$ to introduce variation in the outputs, the maximum sequence length was set to 128 tokens, the top-k sampling was $disabled$ to prevent constraining the sampling space, a repetition penalty of 1.2 was applied to diminish redundancy in the text generation, and the nucleus sampling threshold was established at 0.8 to manage the diversity of the generated content.

5.1. Evaluation

Automatic Metrics The automatic evaluation of results shows extremely low overlap between the human word associations and the model-generated word associations. As shown in Table 4, the overlap, ranking and semantic similarity of responses is much higher for human subsets than for any of the trained generative models. Note that the deviations are much higher for the Slovene human-human subsets due to the small size of the dataset. Overall, the multilingual model performs marginally better on the English than on the Slovene dataset according to Jaccard, Precision, and WMD metrics. On the other hand, between the monolingual and multilingual T5 model for Slovene, the monolingual performs much better, achieving a higher score on all the five metrics.

	model	RBO	Jaccard	Precision	Recall	WMD
SL	human	0.22 ±0.17	0.15 ±0.1	0.24 ±0.15	0.24 ±0.15	0.76 ±0.15
	sloT5	**0.05** ±0.06	0.03 ±0.03	0.05 ±0.04	**0.09** ±0.07	**0.95** ±0.06
	mT5	0.02 ±0.08	0.01 ±0.02	0.03 ±0.06	0.02 ±0.03	1.05 ±0.05
EN	human	0.43 ±0.08	0.3 ±0.05	0.46 ±0.06	0.46 ±0.06	0.3 ±0.07
	mT5	0.03 ±0.04	**0.04** ±0.02	**0.13** ±0.08	0.05 ±0.03	**0.95** ±0.05

Table 4: Results of automatic metrics for word associations overlap and their standard deviations. In bold: overall best score by trained models, underlined: best score on the Slovene dataset. Note that for WMD, which is a distance measure, a lower score is better.

Manual Evaluation Manual annotation of responses (see Table 5) produced by the models first revealed a much higher ratio of erratic responses (human 0.11 vs. 0.47-0.61 in models). Erratic responses are those where no meaningful relation or connection between the cue and response can be found, and such responses are typically rare in humans. Conversely, around a half of the models' associations are relatively far-fetched and thus labelled as erratic, and we can associate the ratio of erratic responses with the overall performance of the model rendering SloT5 as the winner amongst the three.

- SloT5: napisati [write] -> pes [dog], še [still], pisk [whistle], stranica [edge], informatika [informatics], ...
- mT5-SL: napisati [write] -> pravijo [they say], slikovati [*unknown word*], dogovoriti [arrange], požičiti [*unknown word*], obrazi [faces], ...
- mT5-EN: ecological -> white, creole, furry, cheerful, lively, instinctive, dark, ...

For the other three categories, Form-based, Meaning-based and Position-based respectively, the distribution of the models' responses is surprisingly similar to human categories, with very few form-based responses and a good measure of meaning-based ones (see Figure 1).

Another observation concerning the output of the SloT5 and mT5-SL models is the occurrence of unknown (invented) words as responses to cues. While the SloT5 model hardly ever forms an unexisting word (0.01%), in the mT5 model every 5th response is a newly created word (e.g. *ključak, obkreče, kuzko, nadvid, slikovati, požičiti, snemeti, avtores, pohnost, ...*

6. Discussion

Since the purpose of our experiment was to compare the associative behaviour of fine-tuned mono- and multilingual models with human association norms, low overlap between them does not necessarily mean failure. Thus, we did not use the now-popular instruction-based large language models

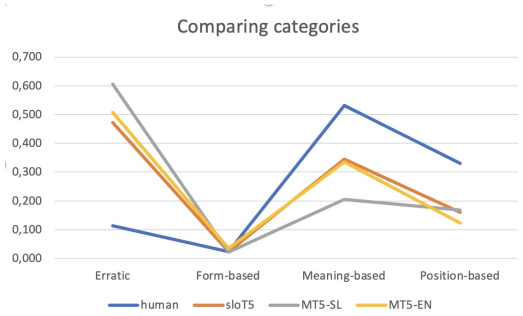

Figure 1: Distribution of categories in human responses and the three models

model	Erratic	Form	Meaning	Position
human	0.11	0.02	0.53	0.33
SloT5	0.47	0.02	0.34	0.16
mT5-SL	0.61	0.02	0.20	0.17
mT5-EN	0.51	0.03	0.21	0.12

Table 5: Distribution of categories in human responses vs. trained models

because they do not align with our experiment goal. In our experiment, we deliberately preserved model parameters at default values and see the results obtained as a kind of a baseline for a new language, for which no such study had been performed before.

The results for word associations overlap are generally low, but consistently best with the monolingual SloT5 model. Results also show that a much larger set of training data, which was the case for English, does not improve the alignment with human responses - in either the English or Slovene inference task.

The manual classification of human responses and predictions into categories shows that the models behave in a manner rather similar to humans in that they generate - if not erratic - mostly meaning-based associations which entail synonyms, hypo- and hypernyms and other more loosely related words (e.g. *natančno* [exact] -> *točno* [precise], *umetnost* [art] -> *igra* [play], *partner* [partner] -> *odnos* [relationship]). Similar to human norms, predicted words for verbal cues contain a slightly

higher number of position-based associations (e.g. *napisati* [write] -> *odgovor* [reply], *knjigo* [book], *besedilo* [text]). We speculate that the fact that the multilingual mT5-SL generated a high number of non-existing words in the erratic category, compared to both SloT5 and mT5-EN, is due to a lower quality and quantity of pre-training data.

Our research is limited in that it uses a rather small dataset for Slovene, where the number of human responses collected for each cue is considerably lower than for English. Later versions of the dataset may prove better in this respect. Another limitation is that tha manual annotation comprised only a relatively small random sample of responses, so that the overall distribution might be different for a more representative sample. We also assume that results would be different when employing newer and larger language models.

7. Conclusion

The first contribution of our work is the creation of a new resource, the first version of SWOW-SL containing human associations to 1000 Slovene cues contributed by over 600 participants, and created under the auspices of the "parent" Small World of Words project (De Deyne et al., 2018). We then use this dataset to fine-tune a monolingual T5 and a multilingual mT5 model (as well as an English one for comparison) for the word association task, but without attempting to optimize the parameters. The predictions of the models are evaluated using 4 automatic metrics, namely Jaccard, rank-biased overlap, precision and recall and Word Mover's Distance. Results show that the overlap between human and model-generated responses is very low, and that the better model for Slovene is the monolingual one. A manual classification of responses into categories is performed in order to better understand the behaviour of the models. While all models generate a high number of erratic responses (between 47 and 61 percent), the distribution of meaningful responses amongst the meaning-based, position-based and form-based categories closely resembles human norms.

8. Acknowledgements

The authors thank Simon De Deyne of Melbourne University for his kind help in setting up the SWOW experiment for Slovene, which makes possible the ongoing collection of association data for Slovene. This research was partly supported by the ARIS research programmes P6-0215 Slovene Language - Basic, Contrastive, and Applied Studies and P6-0411 Language Resources and Technologies for Slovene.

9. Bibliographical References

Álvaro Cabana, Camila Zugarramurdi, Juan Carlos Valle-Lisboa, and Simon De Deyne. 2023. The "small world of words" free association norms for rioplatense spanish. *Behavior Research Methods*, pages 1 – 18.

Simon De Deyne, Danielle Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2018. The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 1861–1870.

Simon De Deyne and Gert Storms. 2008. Word associations: Network and semantic properties. *Behavior research methods*, 40(1):213–231.

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

Francis Galton. 1879. Psychometric experiments. *Brain*, 2(2):149–162.

Fritz Günther, Luca Rinaldi, and Marco Marelli. 2019. Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, 14:1006 – 1033.

Michael N. Jones, Thomas M. Gruenenfelder, and Gabriel Recchia. 2018. In defense of spatial models of semantic representation. *New Ideas in Psychology*, 50:54–60.

Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraž Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc. 2020. Gigafida 2.0: the reference corpus of written standard slovene. In *Proceedings of the twelfth language resources and evaluation conference*, pages 3340–3345.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 957–966. JMLR.org.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78.

Ken McRae, George Cree, Mark Seidenberg, and Chris Mcnorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37:547–59.

Douglas Nelson, Cathy Mcevoy, and Simon Dennis. 2012. What is free association and what does it measure? *Memory and Cognition*, 28:887–899.

Aida Nematzadeh, Stephan C Meylan, and Thomas L Griffiths. 2017. Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *CogSci*.

Russell Richie, Ada Aka, and Sudeep Bhatia. 2022. Free association in a neural network. *Psychological Review*.

Paul Rozin, Nicole Kurzer, and Adam B. Cohen. 2002. Free associations to "food:" the effects of gender, generation, and culture. *Journal of Research in Personality*, 36(5):419–441.

Avijit Thawani, Biplav Srivastava, and Anil Singh. 2019. SWOW-8500: Word association task for intrinsic evaluation of word embeddings. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 43–51, Minneapolis, USA. Association for Computational Linguistics.

M Ulčar and M Robnik-Šikonja. 2023. Sequence-to-sequence pretraining for a less-resourced slovenian language. *Frontiers in Artificial Intelligence*, 6:932519–932519.

Ivan Vulić, Douwe Kiela, and Anna Korhonen. 2017. Evaluation by association: A systematic study of quantitative word association evaluation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 163–175, Valencia, Spain. Association for Computational Linguistics.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28:20.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Peiran Yao, Tobias Renwick, and Denilson Barbosa. 2022. WordTies: Measuring word associations in language models via constrained sampling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5959–5970, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Špela Vintar and Amanda Saksida. 2023. The anatomy of specialized knowledge: Comparing experts and non-experts through associations, frames and language models. *Lexicographica*, 39(1):165–190.

10. Language Resource References

Small World of Words - downloadable resources, https://smallworldofwords.org/en/project/research