Part 4 - Language processing

SSOL 2024, České Budějovice

Jessica Nieder & Kellen Parker van Dam
Lehrstuhl für Multilinguale Computerlinguistik
Universität Passau, Germany

24 August 2024

Language Processing

One goal of psycholinguistic research is to explore the cognitive mechanisms behind (human) communication

Language Processing

One goal of psycholinguistic research is to explore the cognitive mechanisms behind (human) communication

- How does language processing look like?

Language Processing

One goal of psycholinguistic research is to explore the cognitive mechanisms behind (human) communication

- How does language processing look like?
- What mechanisms or routes are involved in processing?

Language Processing

One goal of psycholinguistic research is to explore the cognitive mechanisms behind (human) communication

- How does language processing look like?
- What mechanisms or routes are involved in processing?
- How can we explain language processing?

Language Processing

Language processing involves the two mechanisms of word comprehension and word production

Language Processing

Language processing involves the two mechanisms of word comprehension and word production

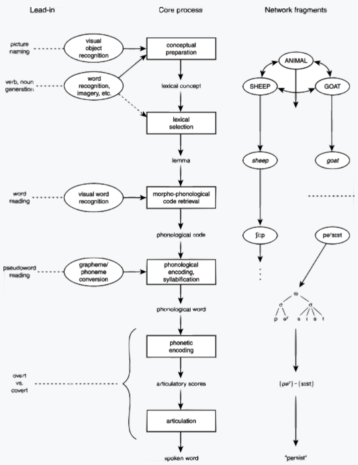
Word comprehension and production in turn involve multiple stages of lexical processing the combination of which have been described as an *extensive neural network* architecture in the pertinent literature (Indefrey & Levelt, 2004)[a]

Language Processing

Language processing involves the two mechanisms of word comprehension and word production

Word comprehension and production in turn involve multiple stages of lexical processing the combination of which have been described as an *extensive neural network* architecture in the pertinent literature (Indefrey & Levelt, 2004)[a]

A fact not everyone would agree with ;)



[a]Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. Cognition, 92(1–2), 101-144. https://doi.org/10.1016/j.cognition.2002.06.001

Language Processing

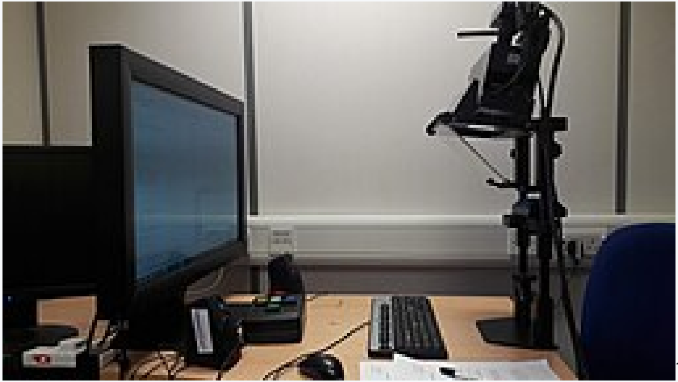
- Word production = producing a sound stream
  1. Semantic stage → conceptual preparation
  2. Lexical stage → lexical selection, grammatical encoding
  3. Phonological stage → phonological encoding
  4. Articulation

Language Processing

- Word comprehension = perceiving a sound stream
    1. Phonological stage → matching speech to phonemes
    2. Lexical stage → activation of lexical candidates, grammatical encoding
    3. Semantic stage → mapping to semantic memory

Language Processing

Typical methods used in psycholinguistics:

[1]

[1] Rai S G Bari, CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0>, viaWikimediaCommons

Language Processing

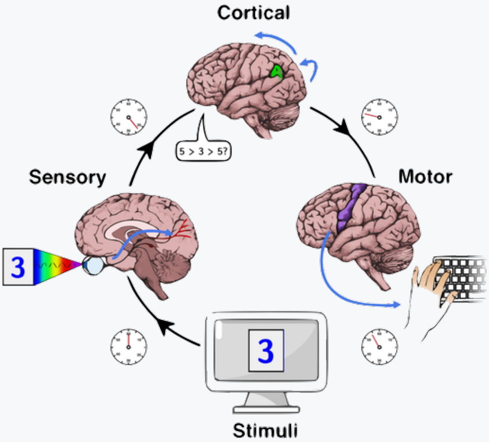
Typical methods used in psycholinguistics:


[2]

[2] Chris Hope, CC BY 2.0 https://creativecommons.org/licenses/by/2.0>, viaWikimediaCommons

Language Processing

Typical methods used in psycholinguistics:



3

[3] Emily Willoughby, CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

Language Processing

Wait, isn't this a lecture on multilingual computational linguistics?

Language Processing

Let’s zoom into the semantic stage of language processing.

Language Processing

Let's zoom into the semantic stage of language processing.

We can make use of computational models and data to help detecting semantic effects and use them for investigations of language processing e.g. word embeddings

Language Processing

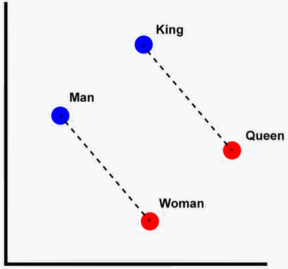
Let's zoom into the semantic stage of language processing.

We can make use of computational models and data to help detecting semantic effects and use them for investigations of language processing e.g. word embeddings

Language Processing

word embeddings = numerical meaning representations of subwords or words (in contexts) from models trained on extensive monolingual or multilingual corpora

word embeddings we can retrieve from e.g. large language models such as BERT or text classification models such as FastText



4

[4]Singerep, CC BY-SA 4.0 `https://creativecommons.org/licenses/by-sa/4.0>,viaWikimediaCommons`

Language Processing

- BERT [5]: Understands word meaning by looking at the context in which a word appears. It considers the surrounding words to better grasp the different meanings of the same word in different sentences.

[5] Devlin, J. et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics.
[6] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

Language Processing

- BERT[5]: Understands word meaning by looking at the context in which a word appears. It considers the surrounding words to better grasp the different meanings of the same word in different sentences.

  She sat by the bank of the river.
  He deposited money in the bank.

[5]Devlin, J. et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics.
[6]Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

Language Processing

- BERT [5]: Understands word meaning by looking at the context in which a word appears. It considers the surrounding words to better grasp the different meanings of the same word in different sentences.

  She sat by the bank of the river.
  He deposited money in the bank.

- FastText [6]: Captures word meaning by considering not just whole words, but also parts of words (like prefixes and suffixes). This helps it understand and represent the meaning of words when they are similar to other words.

[5] Devlin, J. et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics.
[6] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

Language Processing

- BERT [5]: Understands word meaning by looking at the context in which a word appears. It considers the surrounding words to better grasp the different meanings of the same word in different sentences.

  She sat by the bank of the river.
  He deposited money in the bank.

- FastText [6]: Captures word meaning by considering not just whole words, but also parts of words (like prefixes and suffixes). This helps it understand and represent the meaning of words when they are similar to other words.

  bank is different from banker and banking

[5] Devlin, J. et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics.
[6] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

Applying word embeddings in language processing - some studies

Word embeddings in language processing: Study 1

STUDY 1:

Nieder, J., Chuang, Y., van de Vijver, R., & Baayen, H. (2023). A discriminative lexicon approach to word comprehension, production, and processing: Maltese plurals. Language 99(2), 242-274. https://dx.doi.org/10.1353/lan.2023.a900087.

Word embeddings in language processing: Study 1

Can the semantics of a computational model (LDL, see the work of Baayen et al. in Tübingen) equipped with word embeddings from FastText predict the processing of plurals in Maltese?

Word embeddings in language processing: Study 1

Can the semantics of a computational model (LDL, see the work of Baayen et al. in Tübingen) equipped with word embeddings from FastText predict the processing of plurals in Maltese?

omm - ommijiet 'mothers' > concatenative, sound plural

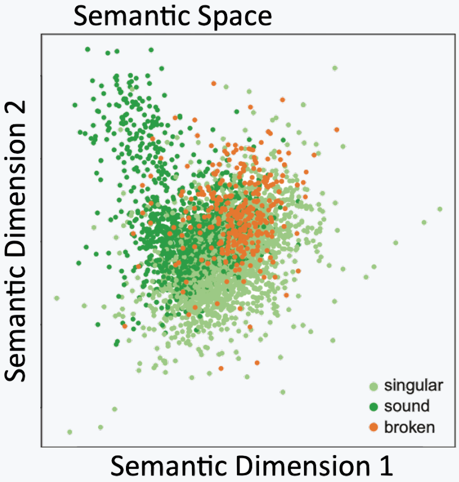
Word embeddings in language processing: Study 1

Can the semantics of a computational model (LDL, see the work of Baayen et al. in Tübingen) equipped with word embeddings from FastText predict the processing of plurals in Maltese?

omm - ommijiet 'mothers' > concatenative, sound plural

kelb - klieb 'dogs' > non-concatenative, broken plural

Word embeddings in language processing: Study 1



Word embeddings in language processing: Study 1

We calculated the semantic difference between primes and targets and the semantic support for primes from Nieder, van de Vijver & Mitterer (2021)

[7] Nieder, J., van de Vijver, R., & Mitterer, H. (2021). Knowledge of Maltese singular-plural mappings: Analogy explains it best. *Morphology, 31*, 147–170. `https://doi.org/10.1007/s11525-020-09353-7`

Word embeddings in language processing: Study 1

We calculated the semantic difference between primes and targets and the semantic support for primes from Nieder, van de Vijver & Mitterer (2021)

A model including these predictors provides a better fit for RT data from Nieder, van de Vijver & Mitterer (2021)[7] and suggests a difference in RTs for sound vs. broken plurals (but not a different priming effect!)

[7]Nieder, J., van de Vijver, R., & Mitterer, H. (2021). Knowledge of Maltese singular-plural mappings: Analogy explains it best. *Morphology, 31*, 147–170. https://doi.org/10.1007/s11525-020-09353-7

Word embeddings in language processing: Study 1

word embeddings can be used to gain insights into morphological processing

Word embeddings in language processing: Study 2

STUDY 2:

Nieder, J., & List, J.-M. (2024). A computational model for the assessment of mutual intelligibility among closely related languages. In Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (pp. 37–43). Association for Computational Linguistics. St. Julian's, Malta.

Word embeddings in language processing: Study 2

In this study we propose a computer-assisted method (again based on LDL by Baayen et al., 2019)[8] to assess mutual intelligibility in Germanic languages (German, Dutch, English cognates).

[8] Baayen, R. H., Chuang, Y. Y., Shafaei-Bajestan, E., and Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. Complexity, 2019, 1-39.

Word embeddings in language processing: Study 2

In this study we propose a computer-assisted method (again based on LDL by Baayen et al., 2019)[8] to assess mutual intelligibility in Germanic languages (German, Dutch, English cognates).

Our word embeddings are based on multilingual ConceptNet Numberbatch from Speer et al. (2017)

[8] Baayen, R. H., Chuang, Y. Y., Shafaei-Bajestan, E., and Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. Complexity, 2019, 1-39.

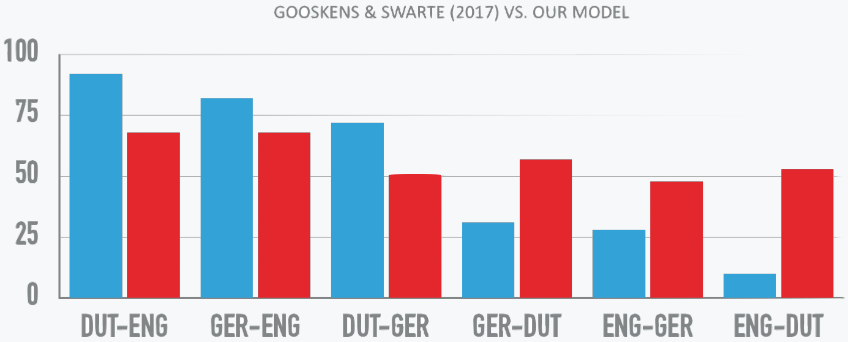
Word embeddings in language processing: Study 2

Mutual intelligibility: the ability to understand a closely related language with minor or no previous knowledge of that language

Brot > bread > brood

Highly dependent on number of shared cognates and paralinguistic parameters!

Word embeddings in language processing: Study 2



[9]

[9] Gooskens, C., & Swarte, F. (2017). Linguistic and extra-linguistic predictors of mutual intelligibility between Germanic languages. Nordic Journal of Linguistics, 40*(2), 123–147.

Word embeddings in language processing: Study 2

Our model shows *similar* results like humans and additionally allows to test mutual intelligbility without previous knowledge of languages (inherited/inherent intelligbility)

Word embeddings in language processing: Study 3

STUDY 3:

Schebesta, A. & Nieder, J. Semantic transparency affects the phonetic signal. (2024). Poster presentation accepted at 20. Jahrestreffen für Phonetik und Phonologie im deutschsprachigen Raum, Halle/Saale.

Word embeddings in language processing: Study 3

What is the influence of semantics on the phonetic signal of English NNN compounds?

Word embeddings in language processing: Study 3

What is the influence of semantics on the phonetic signal of English NNN compounds?

[health$_{\mathrm{N1}}$ care$_{\mathrm{N2}}$] law$_{\mathrm{N3}}$ = left-branching

corner$_{\mathrm{N1}}$ [drug$_{\mathrm{N2}}$ store$_{\mathrm{N3}}$] = right-branching

Word embeddings in language processing: Study 3

Using BERT embeddings we calculated the semantic transparency between the embedded constituent and the free constituent of NNN compounds adapting the procedure introduced in Buijtelaar & Pezzelle (2023)[10].

[10] Buijtelaar, L., & Pezzelle, S. (2023). A psycholinguistic analysis of BERT's representations of compounds. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 2230–2241). Association for Computational Linguistics.

Word embeddings in language processing: Study 3

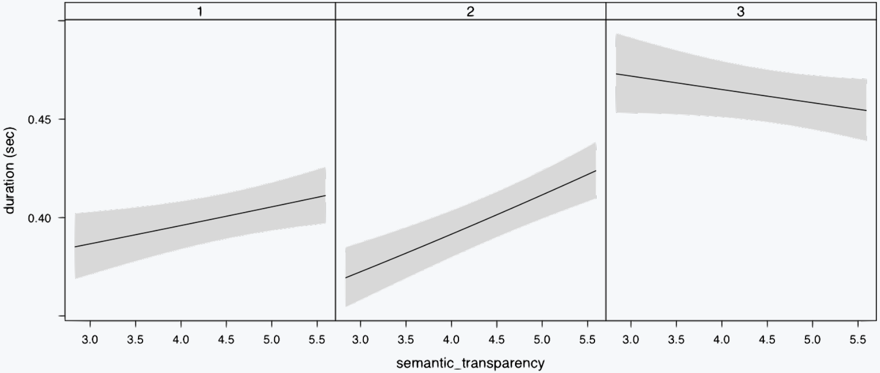
Using BERT embeddings we calculated the semantic transparency between the embedded constituent and the free constituent of NNN compounds adapting the procedure introduced in Buijtelaar & Pezzelle (2023)[10].

We analysed 10,710 constituents from 3,573 NNN compounds produced by Canadian English speakers including our semantic predictors and morphological + phonological predictors (e.g. branching direction, duration of segments etc.)

---

[10]Buijtelaar, L., & Pezzelle, S. (2023). A psycholinguistic analysis of BERT's representations of compounds. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 2230–2241). Association for Computational Linguistics.

Word embeddings in language processing: Study 3



Word embeddings in language processing: Study 3

Semantic transparency as taken from BERT embeddings as well as morphological + phonological factors affect the phonetic signal of NNN.

Hands-on example using R: Maltese insults - The pastizzi problem

Maltese insults -The pastizzi problem

The problem we are working on today: The Semitic language Maltese has an interesting concept of insults with a lot of ”harmless” words being used with a *very different* figurative meaning.

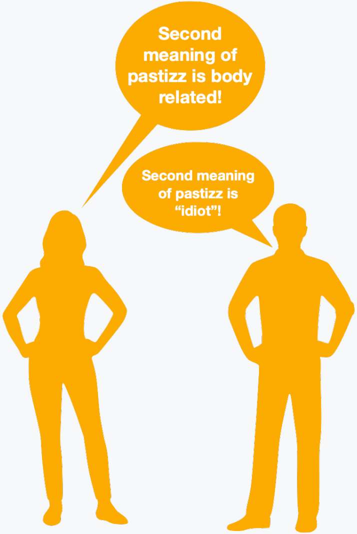
Maltese insults -The pastizzi problem

Maltese pastry: pastizzi

Figure 1: Chattacha. (2008). Malta pastizzi [Photograph]. Wikimedia Commons. CC BY 3.0. `https://upload.wikimedia.org/wikipedia/commons/f/fc/Malta_Pastizzi.JPG`

Maltese insults - The pastizzi problem

Linguists to the rescue!



Maltese insults - The pastizzi problem

Research questions to solve the pastizzi problem:

- Is there an overlap in the semantics of words for food, vulgar words/insults and words for body parts?
- If there is an overlap, are insults more likely to cluster with the meaning of body parts or with the meaning of food items?

Maltese insults - The pastizzi problem

How we will answer the RQs:

- We will retrieve FastText word embeddings for a list of Maltese word forms
- We will plot the semantic space of these word embeddings using UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) in R
- We will inspect the resulting semantic space and think about potential psycholinguistic experiments these results could be explored with

Maltese insults - The pastizzi problem

Download the data here:

Maltese insults - The pastizzi problem

The data file malti_insults.csv contains a list of 93 Maltese words belonging to the categories food, body or insult

word	category	gloss
zalzett	food	sausage
basal	food	onion
difer	body	nail
koxxa	body	thigh
vulgari	insult	gross
paxu	insult	slang term for female genitalia (vulgar)

Maltese insults - The pastizzi problem

In a next step we need to retrieve word embeddings from FastText and match them with our word list data

We use a function to read in the FastText .vec file

```
# Function to read FastText .vec file
read_fasttext_vec <- function(file_path) {
  # Read the first line to get the number of vectors and their
  ↪ dimensionality
  first_line <- readLines(file_path, n = 1) # get metadata
  first_line_split <- strsplit(first_line, " ")[[1]] # split
  ↪ first line based on spaces
  num_vectors <- as.integer(first_line_split[1])
  vector_dim <- as.integer(first_line_split[2])
  # Read the rest
  word_vectors <- data.table::fread(file_path, skip = 1,
  ↪ header = FALSE, sep = " ", quote = "")
  # We need words and vectors
  words <- word_vectors[[1]]
  vectors <- as.matrix(word_vectors[, -1])
  rownames(vectors) <- words
  list(words = words, vectors = vectors, num_vectors =
  ↪ num_vectors, vector_dim = vector_dim)
}
```

R code: Apply function to read in FastText .vec file

Maltese insults - The pastizzi problem

In this project we are working with a pre-trained model. We first need to download the .vec file for Maltese from `https://fasttext.cc/docs/en/crawl-vectors.html`.

We then apply our function to read the FastText .vec file into R file and print some information about the data.

```
# Use the function specified above
file_path <- "./cc.mt.300.vec"
fasttext_data <- read_fasttext_vec(file_path)
# Access the words and vectors and print some information
↪  about them
words <- fasttext_data$words
vectors <- fasttext_data$vectors
print(paste("Number of vectors:", fasttext_data$num_vectors))
print(paste("Vector dimension:", fasttext_data$vector_dim))
```

R code: Apply function for reading in FastText .vec file

Maltese insults - The pastizzi problem

We now read in our .csv data and match it with the vector data

```r
# Read in datafile
malti_insults <- read.csv("./malti_insults.csv")
# Initialize my matrix to store vectors
vector_dim <- fasttext_data$vector_dim
malti_insults_vectors <- matrix(NA, nrow =
↪  nrow(malti_insults), ncol = vector_dim)
# Get vectors for each word and add them to my matrix
for (i in 1:nrow(malti_insults)) {
  word <- malti_insults$words[i]
  if (word %in% words) {
    malti_insults_vectors[i, ] <- vectors[word, ]
  }
}
# Combine the original data with vectors for a full dataset
malti_insults_with_vectors <- cbind(malti_insults,
↪  malti_insults_vectors)
# Check the first few rows to ensure everything is handled
↪  correctly
head(malti_insults_with_vectors)
```

R code: Match vector data with .csv file

Maltese insults - The pastizzi problem

We clean up the data by getting rid of rows that do not have a vector representation

```r
# Select only the FastText vector columns
vector_columns <- grep("^\\d+$", names(malti_insults_with_vectors), value = TRUE)
# Create a dataframe with only these vector columns
vectors_df <- malti_insults_with_vectors %>%
  dplyr::select(dplyr::all_of(vector_columns))
# Handle missing values by removing rows with any NA in vector columns (there are quite a few)
cleaned_df <- malti_insults_with_vectors %>%
  dplyr::filter(complete.cases(vectors_df))
vectors_df <- cleaned_df %>% dplyr::select(dplyr::all_of(vector_columns))
```

R code: check for missing vectors

Maltese insults - The pastizzi problem

Finally, we perform a UMAP analysis and plot the resulting semantic space for our data

```r
# Perform UMAP analysis
set.seed(13)
umap_results <- umap::umap(as.matrix(vectors_df))

# Create a new data frame for plotting
plot_df_umap <- cleaned_df %>%
  dplyr::select(words, type_of_word) %>%
  dplyr::mutate(umap1 = umap_results$layout[, 1],
                umap2 = umap_results$layout[, 2])

# Plot semantic space using ggplot2 for UMAP results
ggplot2::ggplot(plot_df_umap, ggplot2::aes(x = umap1, y =
↪  umap2, color = type_of_word)) +
  ggplot2::geom_point(size = 10) +  # Increase point size here
  ggplot2::labs(title = "UMAP Visualization of Semantic
  ↪  Space",
                x = "UMAP Dimension 1",
                y = "UMAP Dimension 2") +
  ggplot2::theme_minimal() +
  ggplot2::scale_color_discrete(name = "Type of Word")
```

UMAP plot of the semantic space

Maltese insults - The pastizzi problem

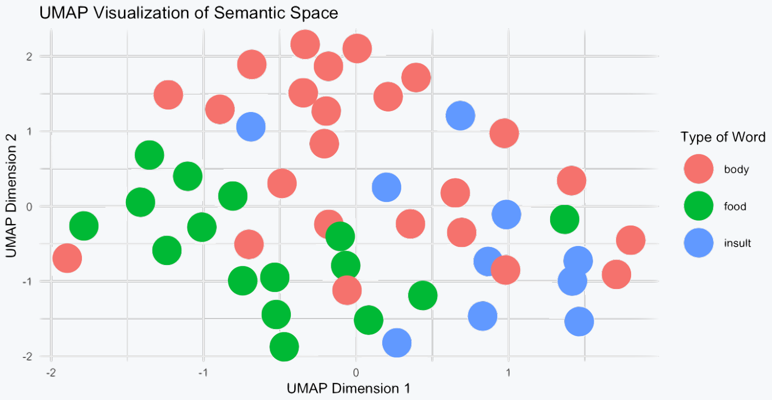


Figure 2: UMAP results for the semantic space of Maltese food, insult and body words

Maltese insults - The pastizzi problem

We find an overlap between body part words and insults. Food words build their own cloud but show a slight overlap with body part words.

Research questions:

- Is there an overlap in the semantics of words for food, vulgar words and words for body parts?

  Yes, there is an overlap.

- If there is an overlap, are insults more likely to cluster with the meaning of body parts or with the meaning of food items?

  Insults, in this toy example, cluster with body parts.

Maltese insults - The pastizzi problem

What do these results mean for language processing? A possible theoretical background:

- In language processing, theories of figurative language processing can be divided into direct access vs. indirect access theories (Gibbs, 2002; Weiland et al., 2014)[11][12]
- direct access = figurative meanings are directly available in processing
- indirect access = the figurative meanings come to play later in processing, first the literal meaning is accessed and rejected

[11] Gibbs, R. W., Jr. (2002). A new look at literal meaning in understanding what is said and implicated. Journal of Pragmatics, 34(4), 457-486. https://doi.org/10.1016/S0378-2166(01)00046-7
[12] Weiland, H., Bambini, V., & Schumacher, P. B. (2014). The role of literal meaning in figurative language comprehension: Evidence from masked priming ERP. Frontiers in Human Neuroscience, 8, Article 583. `https://doi.org/10.3389/fnhum.2014.00583`

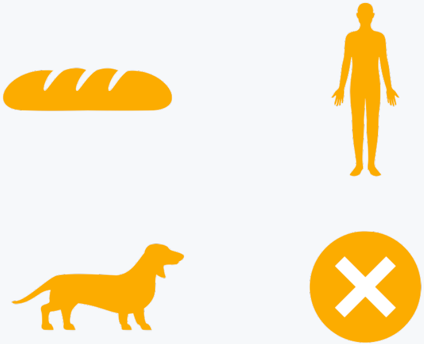
Maltese insults - The pastizzi problem

What does that mean for the pastizzi problem?

- If we follow the direct access theory, based on our computational analysis, we might want to expect an overall slowed down processing when forms are in competition but no difference between *core* meaning and other meaning
- If we follow the indirect access theory, we might want to assume a faster processing of the *core* meaning as opposed to other meaning when in competition

Maltese insults - The pastizzi problem

How can we test that? One idea: Eye-tracking with visual-world paradigm.



Discussion

Take-home message

It is possible to enrich psycholinguistic investigations with computational approaches. Word embeddings from large language models offer us the possibility to retrieve semantic representations for word forms based on an enormous amount of corpus data. These representations can then be verified with human judgements or used for predictions of already collected data from human participants.

Take-home message

However, we always need to keep limitations in mind. Computational models might not include the words we are looking for. Some models are somewhat intransparent and it is unclear how to really interpret semantics in a meaningful way. These models are often not meant to show human-like behaviour, we need to be careful in claiming that and provide checks for our claims!

Děkuju

grɐtsɪ	Maltese
tʰaimi hai	Wolam Khiamniungan
ketʒu əʒuŋ	Muishaung Tangsa
daŋkə	Central German, Bottrop dialect
θæŋk ju	Inland Northern American English