RuPersWordAssociation: a new dataset to study individual association behavior

Tatiana Litvinova, Viktoriya Zavarzina,

Polina Panicheva, Svetlana Lyubova

Voronezh State Pedagogical University

Ivan Mamaev

Baltic State Technical University “Voenmeh”

INTERNET AND MODERN SOCIETY – IMS-2024,

JUNE 26TH 2024, SAINT PETERSBURG, RUSSIA

Word Association (WA) Problems

- WAs are often referred to as “a language of thought” [De Deyne & Storms, 2015].
- Current WA norms do not enable the study of individual word associations.
- We present a new dataset designed for studying individual associative behavior which contains individual responses, participant data, linguistic annotations, reaction times, and semantic similarity data.

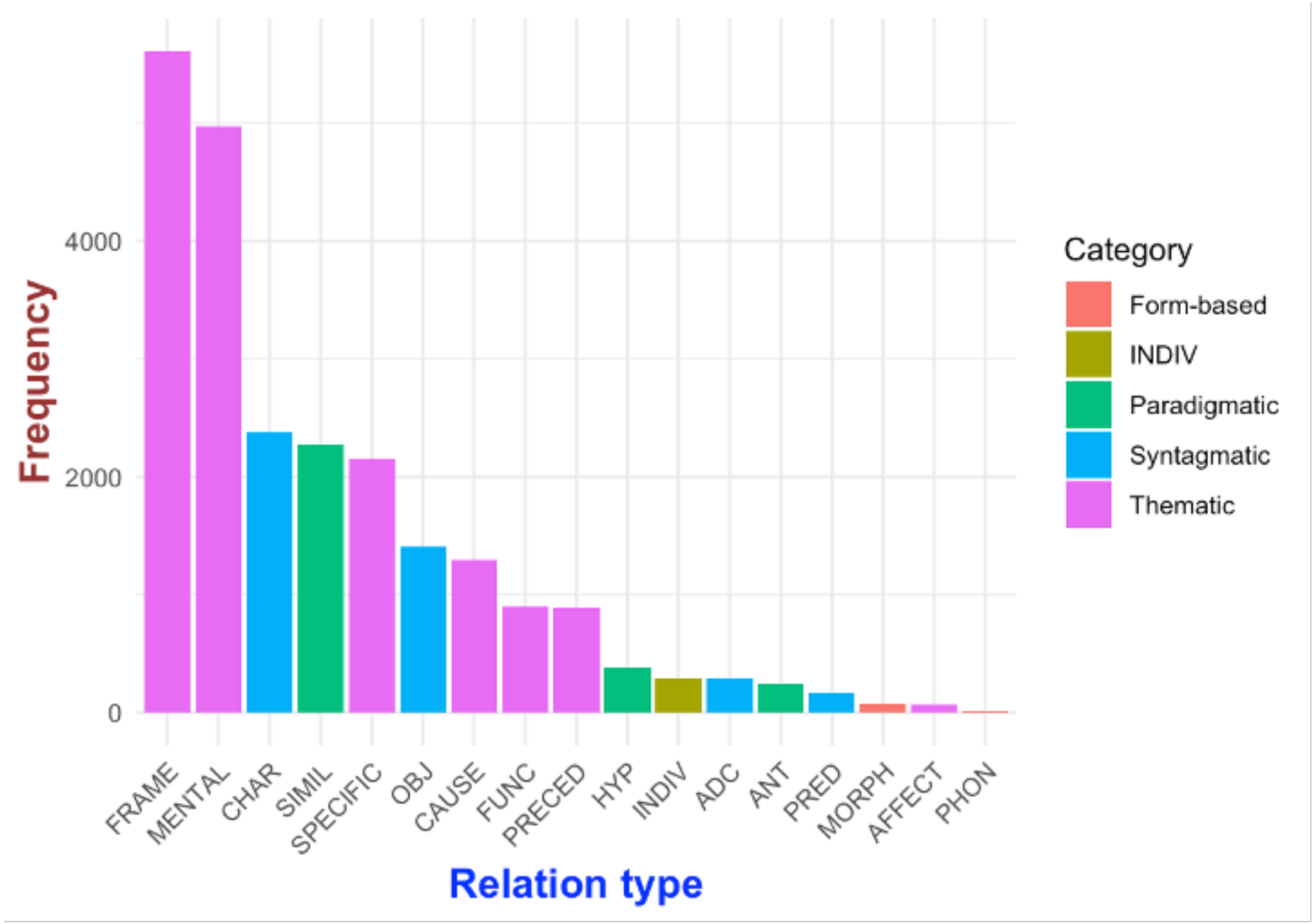
Scan to test RuPersWordAssociation!

Experimental Design

Parameter	Stage 1	Stage 2	Stage 3
Participants	49 people	10 people	10 people
Written?	+	+	-
Number of Annotated Cue – Associate Word Pairs	18143	2500	2004
Semantic Similarity Data (w2v + RoBERTa)	+	+	+
Reaction Time	+	+	-

Linguistic Annotation of Cue-Associate Word Pairs

Type	Subtypes	Examples	Meanings
Formal-based	-	MORPH / PHON	Morphological / Phonological
Meaning-based	Language-based	Syntagmatic (PRED / CHAR…)	Predicative / Attributive…
		Paradigmatic (SIMIL / HYP / ANT)	Synonyms / Hyponyms / Antonyms
	World-knowledge (Thematic)	FRAME / CAUSE…	Cue word and associates are the components of the same situation / cause-effect
Individual	-	INDIV	No explanation by annotators


Frequency
4000
2000
0
FRAME
MENTAL
CHAR
SIMIL
SPECIFIC
OBJ
CAUSE
FUNC
PRECED
HYP
INDIV
ADC
ANT
PRED
MORPH
AFFECT
PHON
Relation type
Category
Form-based
INDIV
Paradigmatic
Syntagmatic
Thematic

Data on Participants

- **The Big Five Scores**

✓ the most frequently used in modern works devoted to the psychological profiling of the author of a text and to the problem of reflecting their psychological characteristics in a text (see [Azucar et al., 2018] for review);

✓ we are unaware of any work where the above model was applied to the analysis of WA.

- **Differential Emotions Scale**

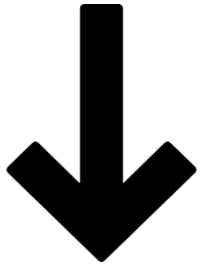
✓ studying the respondents' emotional state;

✓ each time before filling out the questionnaire.

As far as we know, our dataset is unique with respect to the range of data about WA it provides and numbers of annotated *cue* – *associate* pairs (we are aware of the only publicly available dataset which contains such annotation for English data [Chunhua et al., 2022]).

Reaction time

- There are a few studies proposing combining the analysis of associative data with other types of data, in particular, reaction time (RT), i.e., the duration of pauses between stimulus and response.
- This area of research has a long history [Wallenhorst, 1965], a number of interesting results have been obtained (e.g., that paradigmatic responses showed longer RT than syntagmatic ones, and that RT are also influenced the type of a relation between cue and associate [Minto-García et. al., 2020]).
- The absence of any sustained effort to collect WA RT makes it impossible to compare WA data against the numerous other paradigms for which large-scale RT databases already exist [Singh, 2017].



Our dataset is the first one with RTs for typed WA.

Reaction time

- **Respondent's reaction** time is the duration of the pause between the end of the cue word (or the previous associate) and the beginning of the first associate (or subsequent associates).
- **The main heuristic for calculating pause time**: the pause times of the Shift symbol after the last letter of the cue word or response, a comma (colon), a space and the first letter of the next word were summed up.
- Since there might be a variety of ways for tracking RT, it is possible to supplement the dataset with new types of the parameters related to RT for datasets.

NB! Annotation Difficulties: the information about associated words is in log-files, but not in the final doc-files, users' copying previously entered associations, etc.

Semantic Measures

Word2Vec
(ruwikiruscorpora_upos_cbow_300_10_2021)

RoBERTa
(symanto/sn-xlm-roberta-base-snli-mnli-anli-xnl)

- Lemmata for image stimuli were not assigned, and cosine distance measures were not calculated.
- If a user provided an associate in the form of a lexical construction (e.g., *ADJ_NOUN*), all the elements of the construction were subject to lemmatization.

Association between Metadata: Reaction Time and Relation Types

- Kruskal-Wallis rank sum test, $p = 0.0001765$, $Kruskal - Wallis\ \chi^2 = 32.302$, $df = 9$.
- Wilcoxon signed-rank test with Bonferroni-Holm correction revealed differences in RTs between following relation types (smaller values of RT are highlighted in bold):

- **MENTAL** and CHARDIR ($p = 0.016$),
- **MENTAL** and OBJIND ($p = 0.016$),
- **FRAME** and OBJIND ($p = 0.016$),
- **FRAME** and PRECED ($p = 0.040$),
- **FRAME** and SIMIL ($p = 0.020$),
- **MENTAL** and SIMIL ($p = 0.016$),
- **MENTAL** and PRECED ($p = 0.026$).

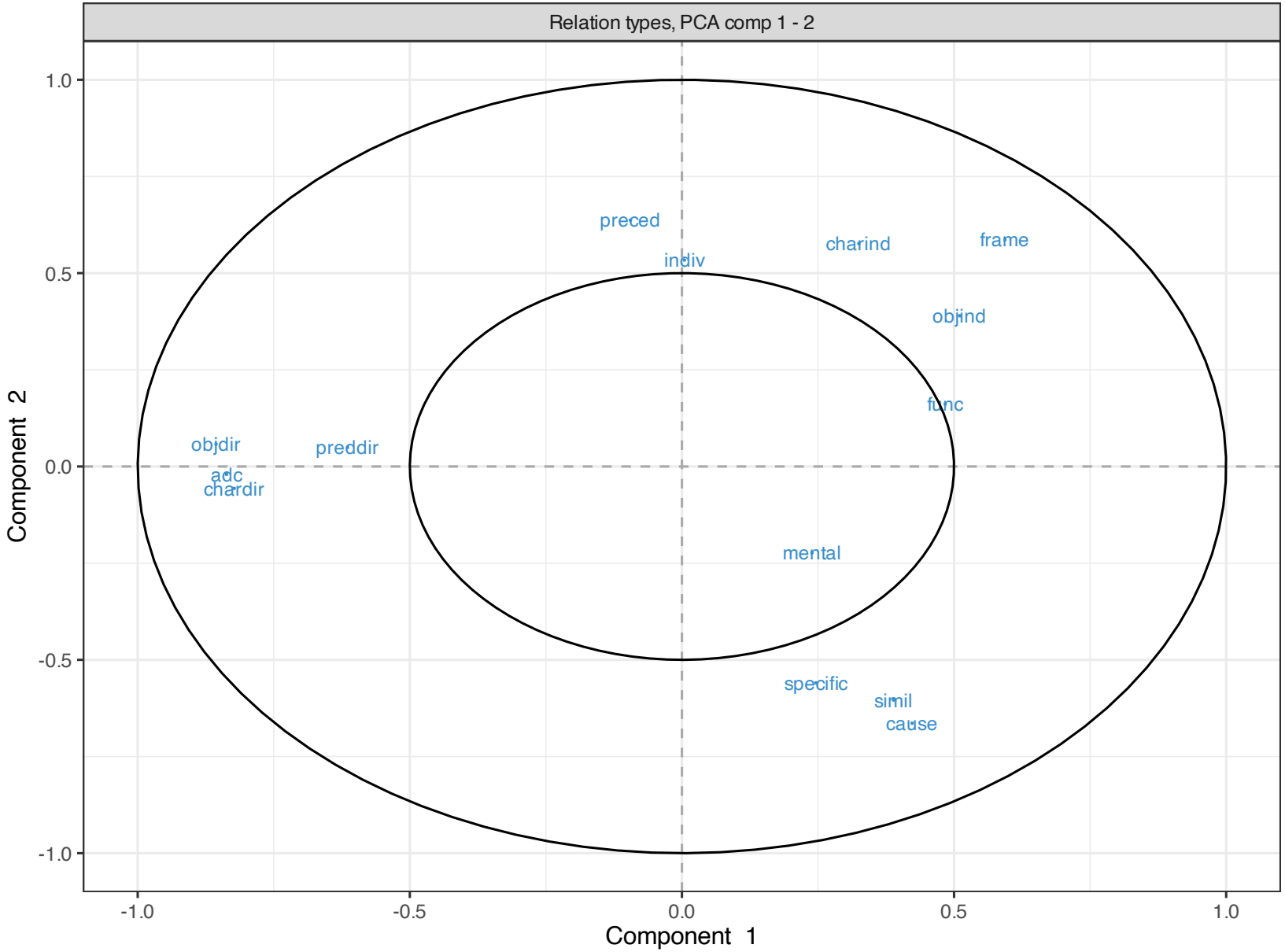
Associations between Metadata: Reaction Time and Semantic Distance

- Kruskal-Wallis rank sum test, $Kruskal - Wallis\ \chi^2 = 1328.6, df = 9, p < 2.2e-16.$
- Wilcoxon signed-rank test with Bonferroni-Holm correction revealed differences for most of the pairs except for CAUSE and FUNC, FUNC and FRAME, FRAME and CAUSE.
- The highest values of semantic similarity (over both models) were observed for the following relation types:

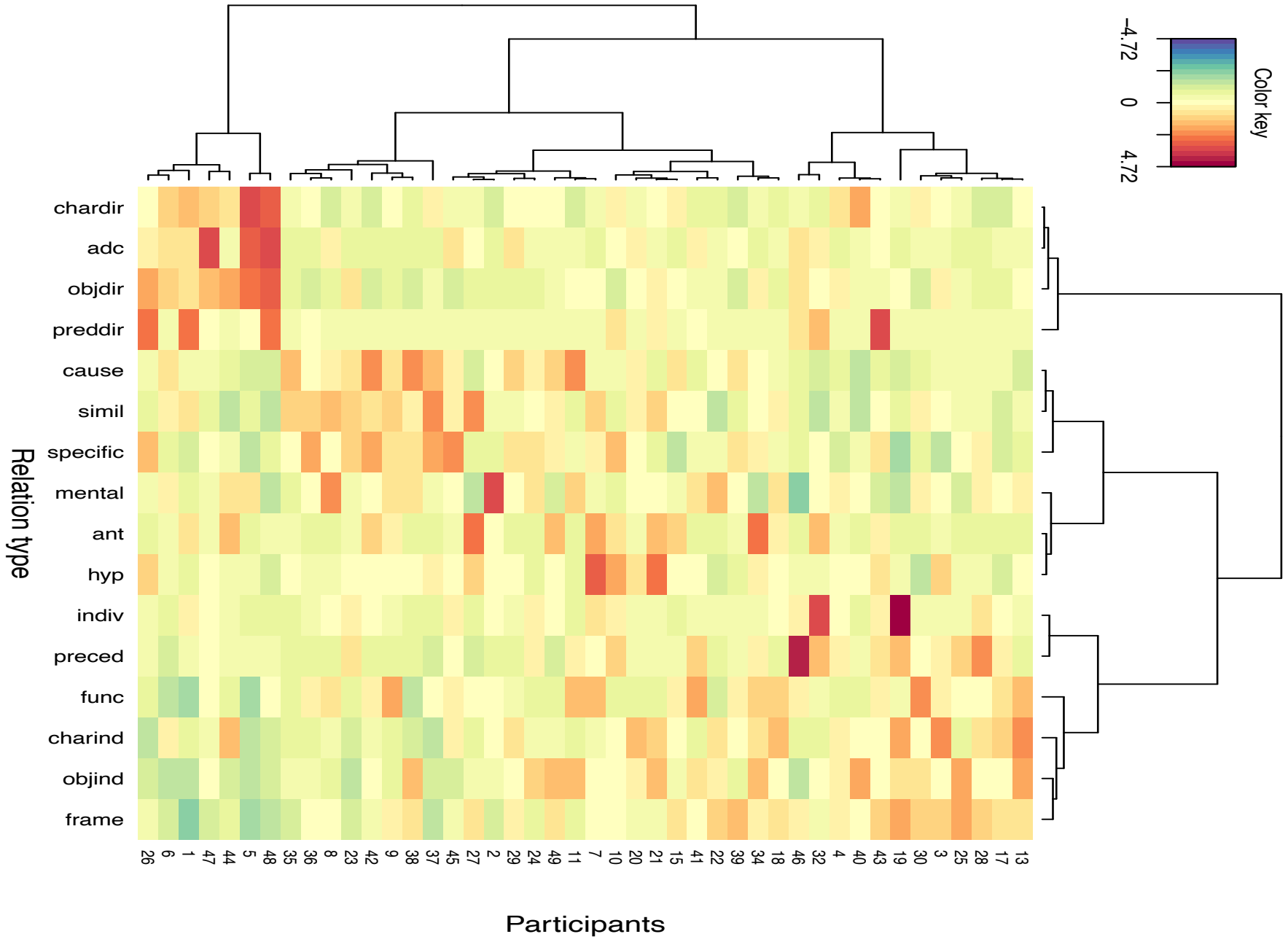
1) SIMIL (1st),
2) CAUSE, FUNC and FRAME (2nd),
3) SPECIFIC (3rd),
4) the lowest – for CHARDIR.

Individual Association Profiles

- Stage 1 only.
- 2577 association series with appropriate metadata.
- Each association series is presented as a text consisting of relation type labels, e.g. CHARDIR, CHARDIR, FRAME, CHARDIR, FRAME.
- Cue words, respondent IDs and metadata were considered as docvars for processing of the corpus with R package quanteda.
- Resulting Document-Feature Matrix: 2577 documents, 187 features.



Correlation Plot of Variables



Clustering of variables and participants

Conclusions

- Many unresolved WA issues still exist.
- More WA datasets with diverse metadata are needed
- Using the current dataset, associations were established between *cue* – *associate* word pair relation types, reaction times, and semantic distance
- Individual profiles in associative behavior were identified, confirming the heterogeneity in norms that must be considered in future research.

Acknowledgment

This study was supported by the grant
of **Russian Science Foundation**,
no **21-78-10148**

Thank you for your attention!

centr_rus_yaz@mail.ru