Tampereen yliopisto
Tampere University

From Ulyanov to Lenin: a corpus-based discourse analysis of Vladimir Lenin's works

Mikhail Mikhailov, Tampere University
mikhail.mikhailov@tuni.fi

CILC2024

Las Palmas de Gran Canaria,

22-24 May 2024

Vladimir Lenin (1870-1924)

Worship...

… and hatred

Why study Lenin's discourse now?

- Because there is a chance to do at last a non-politicized study
- Because otherwise we will not understand what is happening in Russia and former republics of USSR
- Because Lenin's works is a large body of political texts produced by a single person and translated into 80 languages that were in wide circulation for a long period of time and had substantial influence on the political discourse worldwide

Tampereen yliopisto
Tampere University

Research group in Tampere

- Data
  - Corpus of Lenin's works (Lencor)
  - Parallel corpus of Lenin's works (ParVLen)
- Research
  - Corpus-based discourse studies
  - Corpus-based translation studies
  - Data science
  - History of translations

Studying Lenin's discourse

- Focus corpus: Lenin's works
  - Complete works freely available in Internet
  - leninism.su
  - **Lencor**: complete works by Lenin (5 M running words)
- Reference corpus: Political discourse of 1890s to 1920s
  - Many works are digitized and freely available in Internet
  - **Rudire**: politicians, journalists, writers, artists, scientists, currently 1.5 M running words

Processing texts

- Downloading: web-scraping
- Parsing: Turku Universal Dependencies parser
- Querying: NoSketch Engine

Tampereen yliopisto
Tampere University

Lencor: subcorpora

Subcorpus	Tokens	Words	%
1893_1901_early_works	684,213	~ 530,358	12.06
1902_4_party_split	333,004	~ 258,123	5.87
1905_7_first_revolution	679,636	~ 526,811	11.98
1908_10_reaction	472,317	~ 366,110	8.32
1911_13_before_revolutions	497,049	~ 385,281	8.76
1914_17_WW_I	535,369	~ 414,984	9.44
1917_revolutions	539,587	~ 418,253	9.51
1919_21_civil_war	689,499	~ 534,456	12.15
1921_23_building_socialism	249,213	~ 193,174	4.39

N.B. Letters are not included!

Getting a general picture

- Keyword analysis: Kilgarriff's simple maths index
- Collocations: logDice index
- Network Coincidence Analysis (netCoin)

Kilgarriff's simple maths index

$$\frac{fpm_{rmfocus} + N}{fpm_{rmref} + N}$$

where

is the normalized (per million) frequency of the word in the focus corpus,

fpm_{rmref} is the normalized (per million) frequency of the word in the reference corpus,

N is the smoothing parameter ($N = 1$ is the default value).

- N=1000, emphasis on high-frequency words
- Cleaning up the results:
  - Removing proper names, pronouns, prepositions, conjunctions, etc.
  - Removing wrong lemmatizations

Visualizing

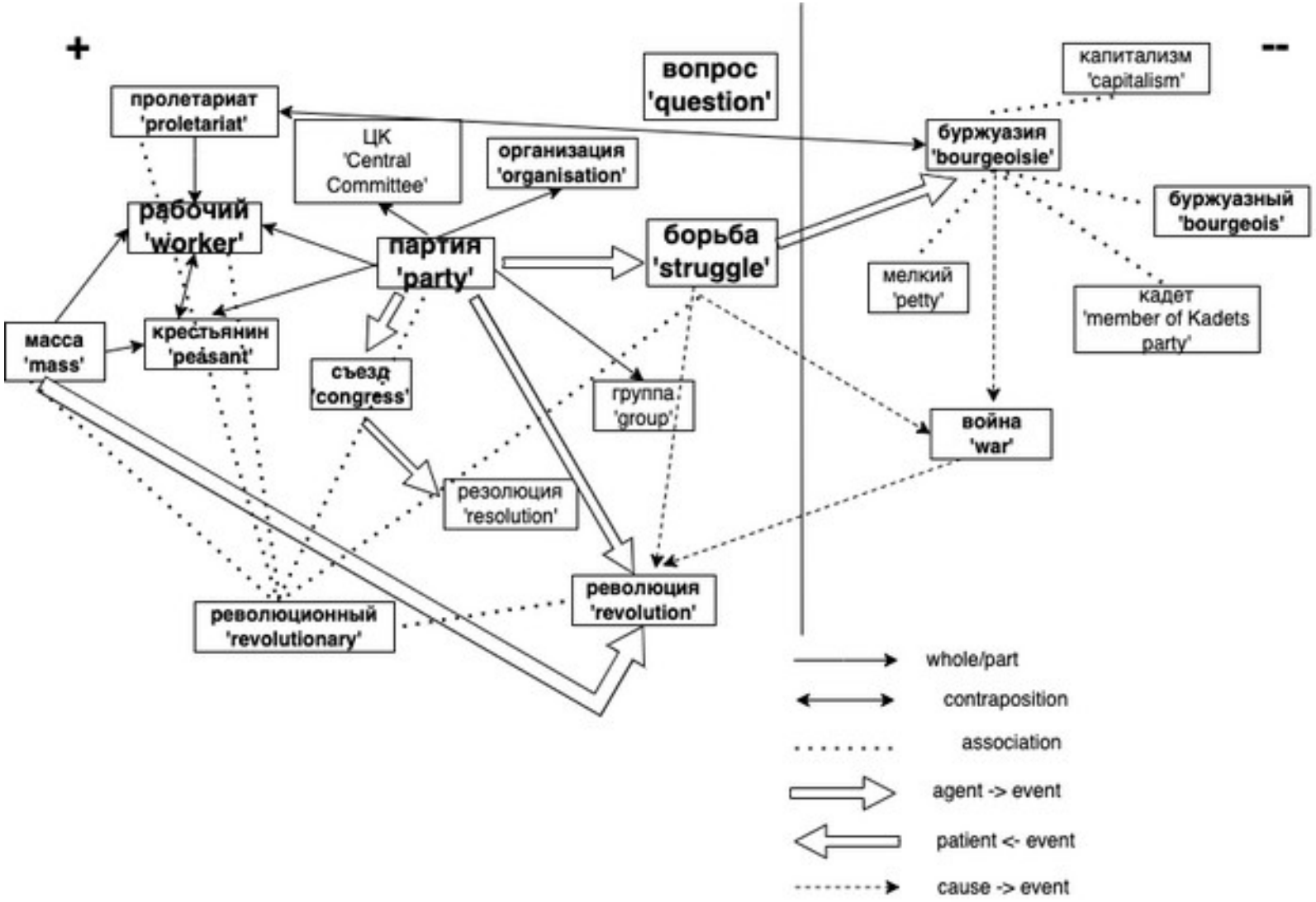
- Top 20 keywords (after cleaning) → Semantic networks
- Top 10 keywords + Top 10 collocates per keyword → netCoin → Collocation networks

Tampereen yliopisto
Tampere University

QR code for the diagrams

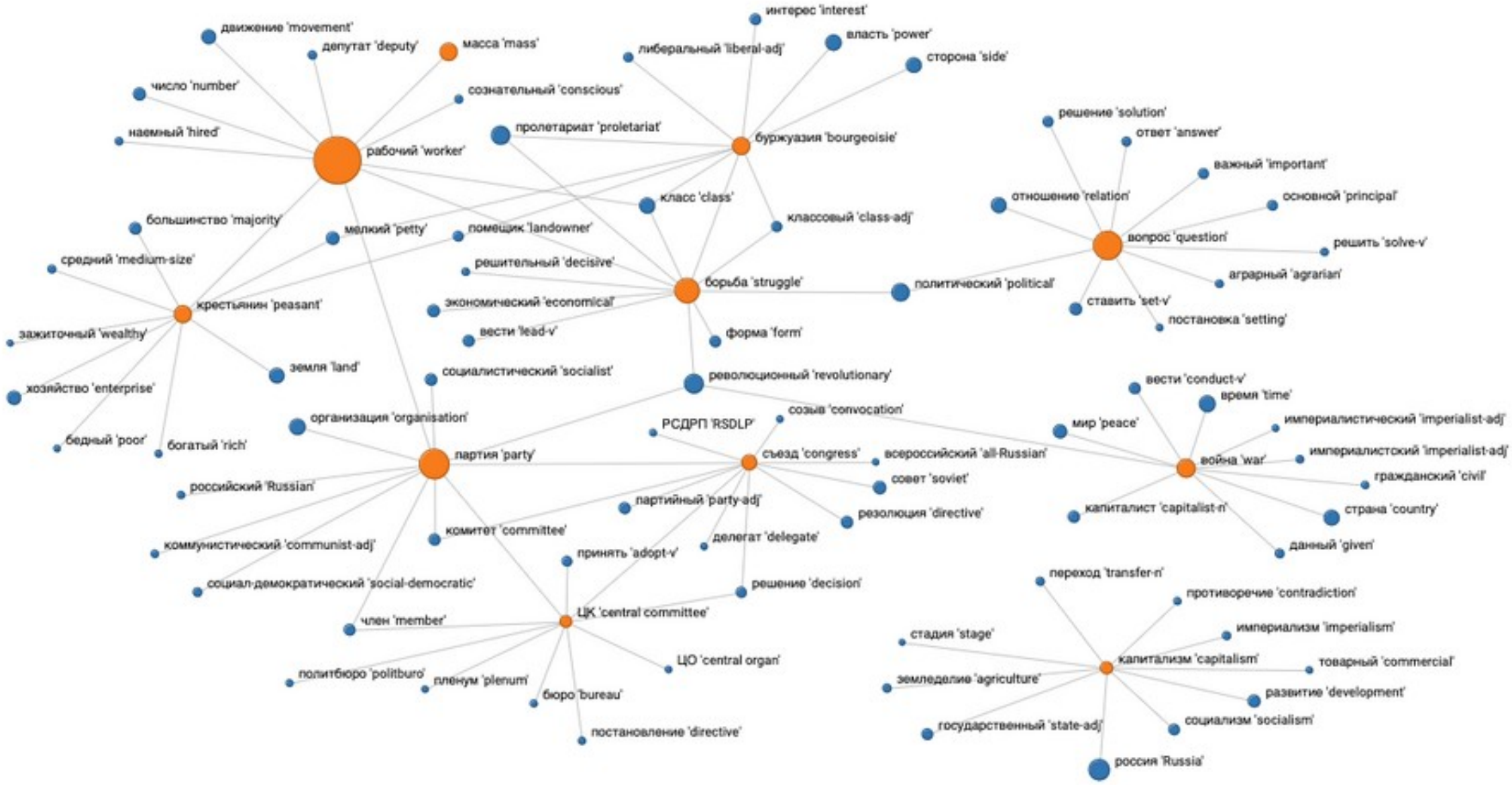
Tampereen yliopisto
Tampere University

Semantic network: whole corpus



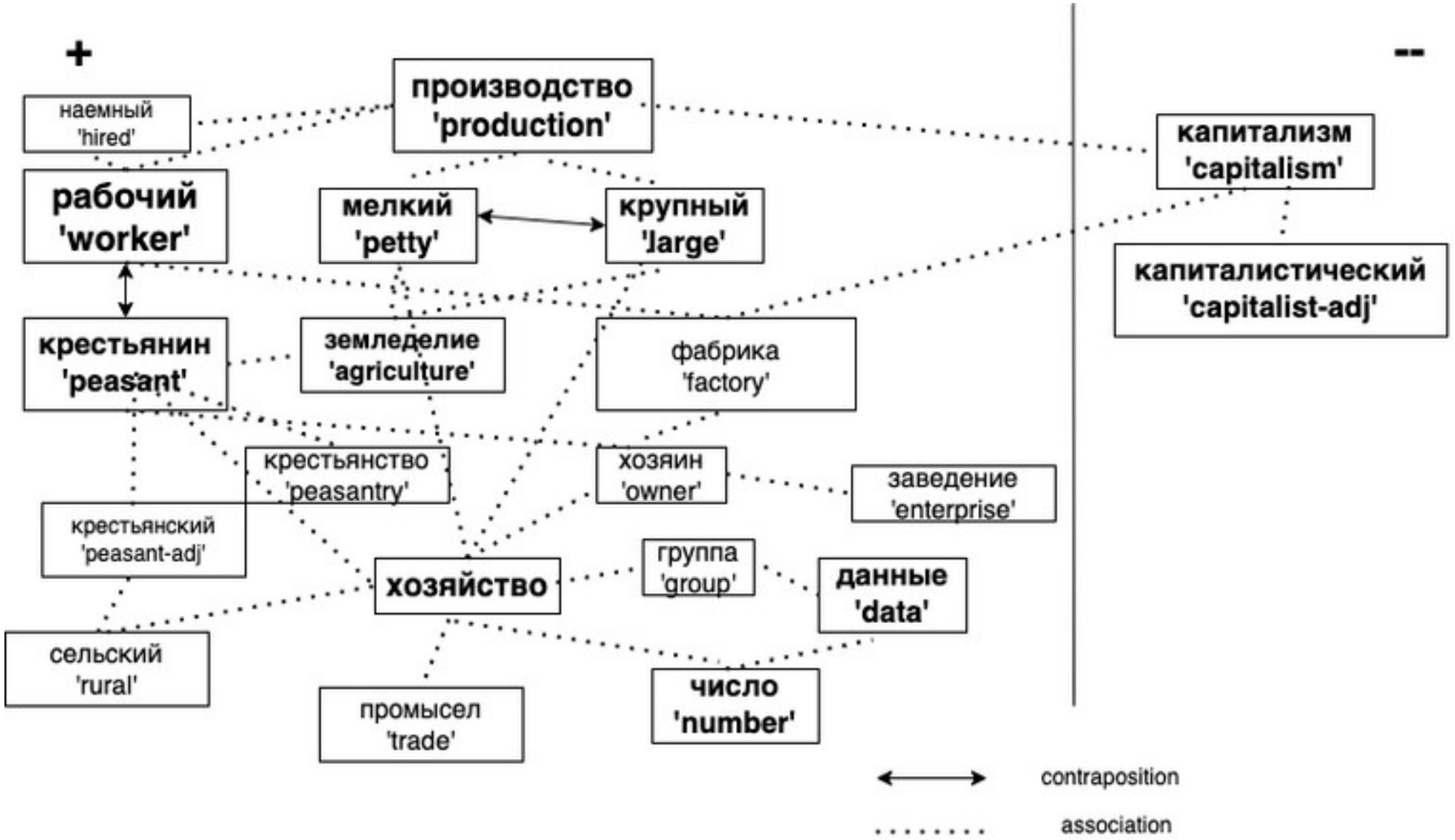
Tampereen yliopisto
Tampere University

Collocation Network: whole corpus



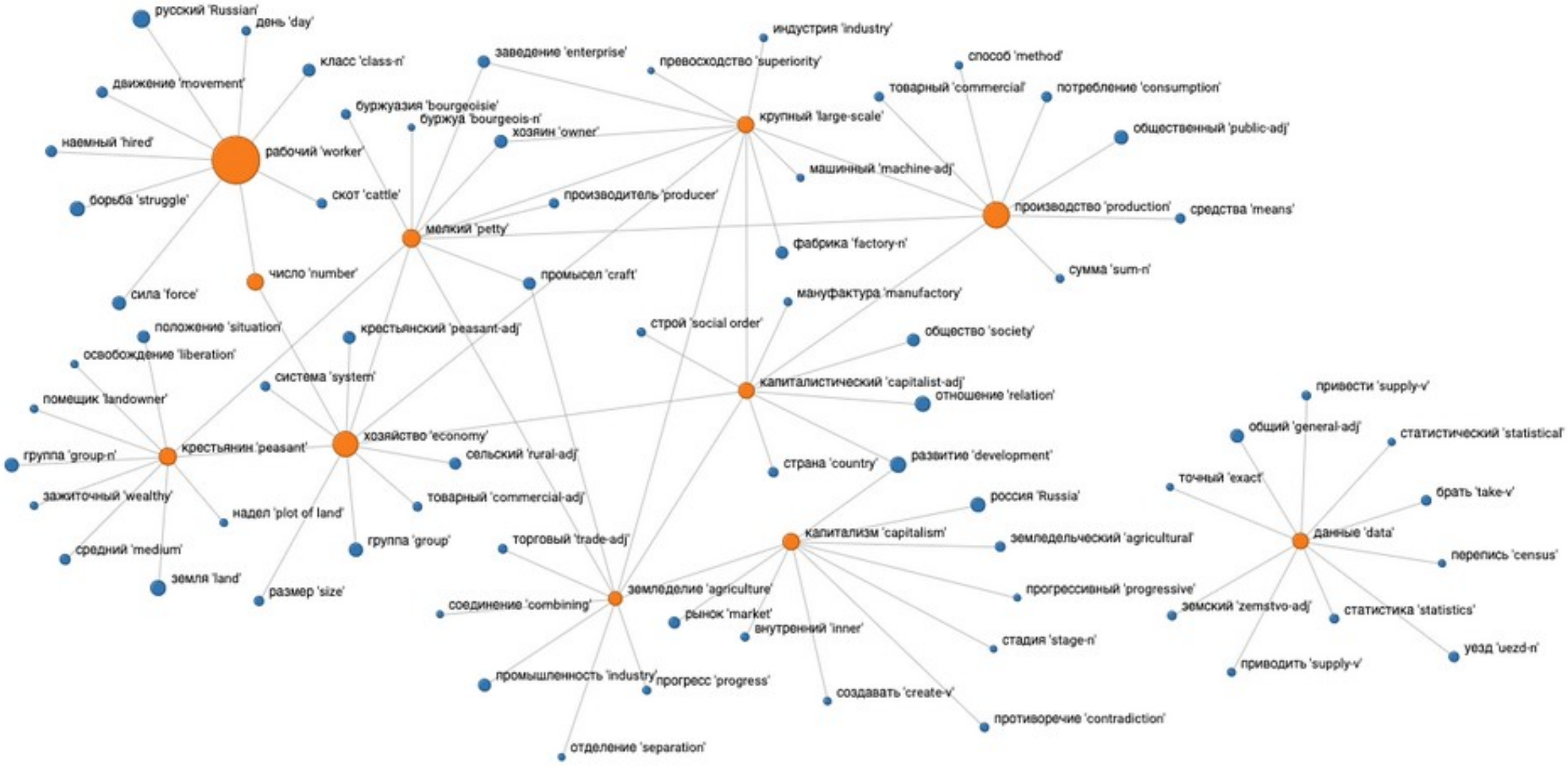
Tampereen yliopisto
Tampere University

Semantic Network: early works



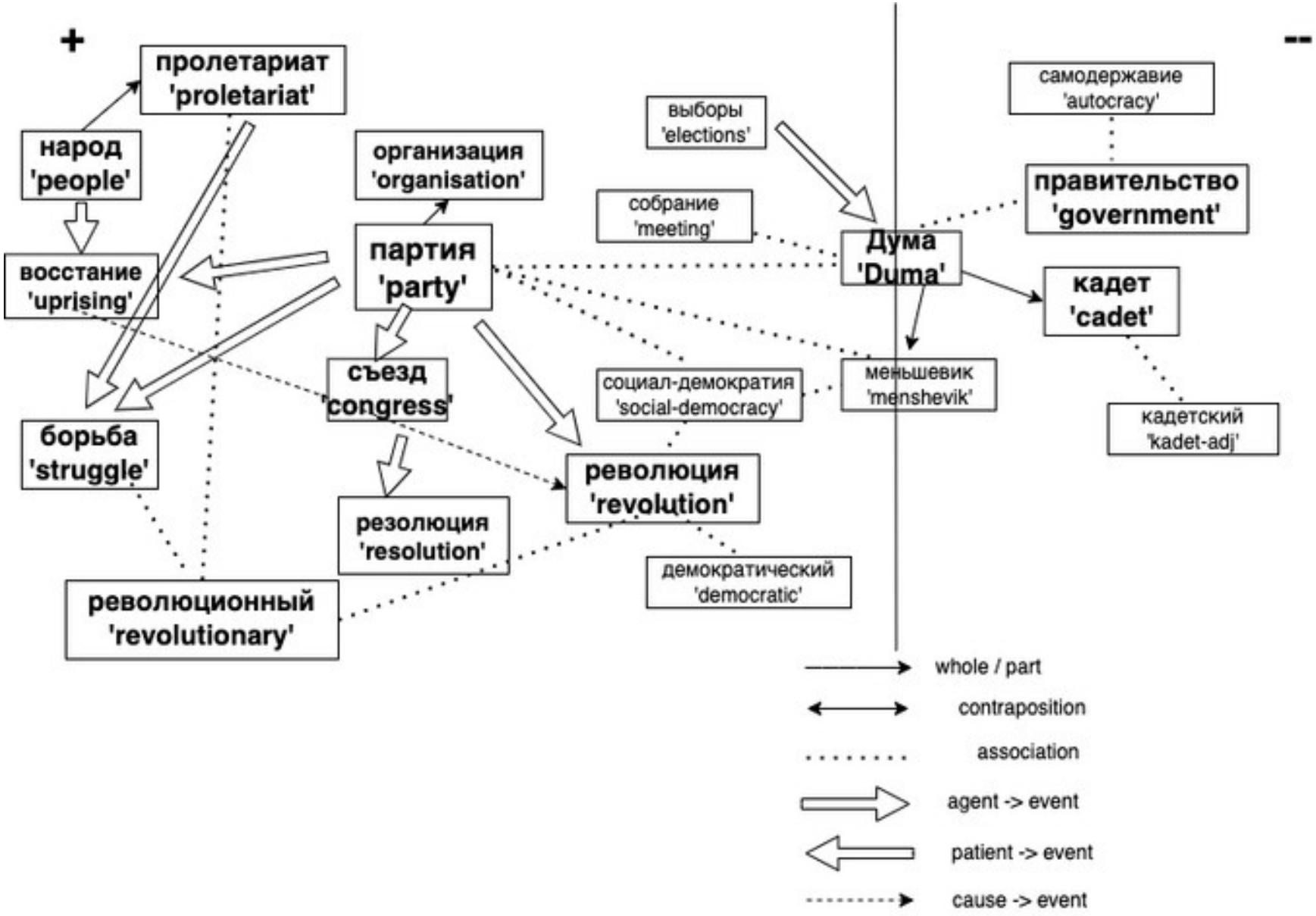
Tampereen yliopisto
Tampere University

Collocation Network: early works



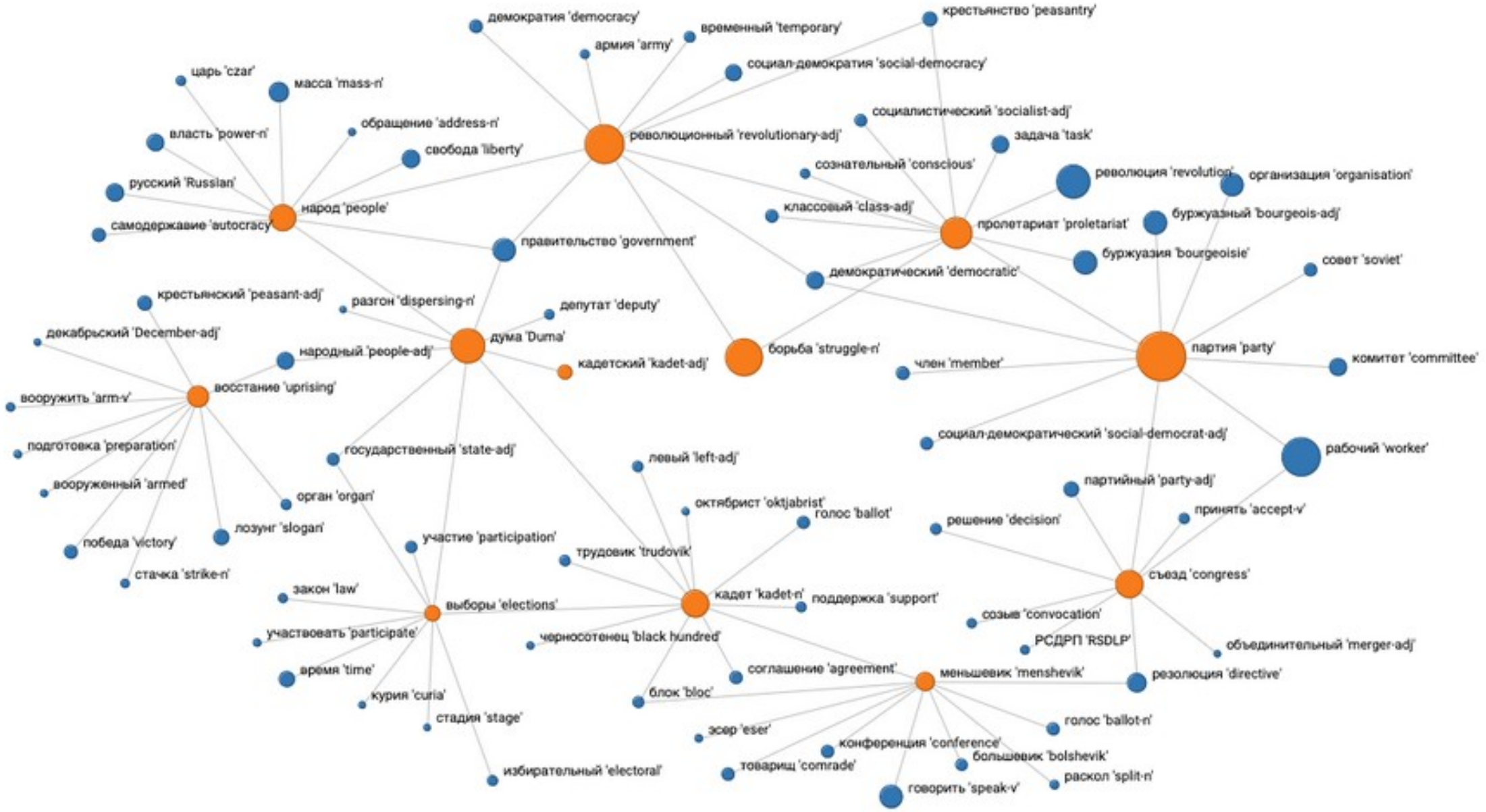
Tampereen yliopisto
Tampere University

Semantic network: Revolution 1905-1907



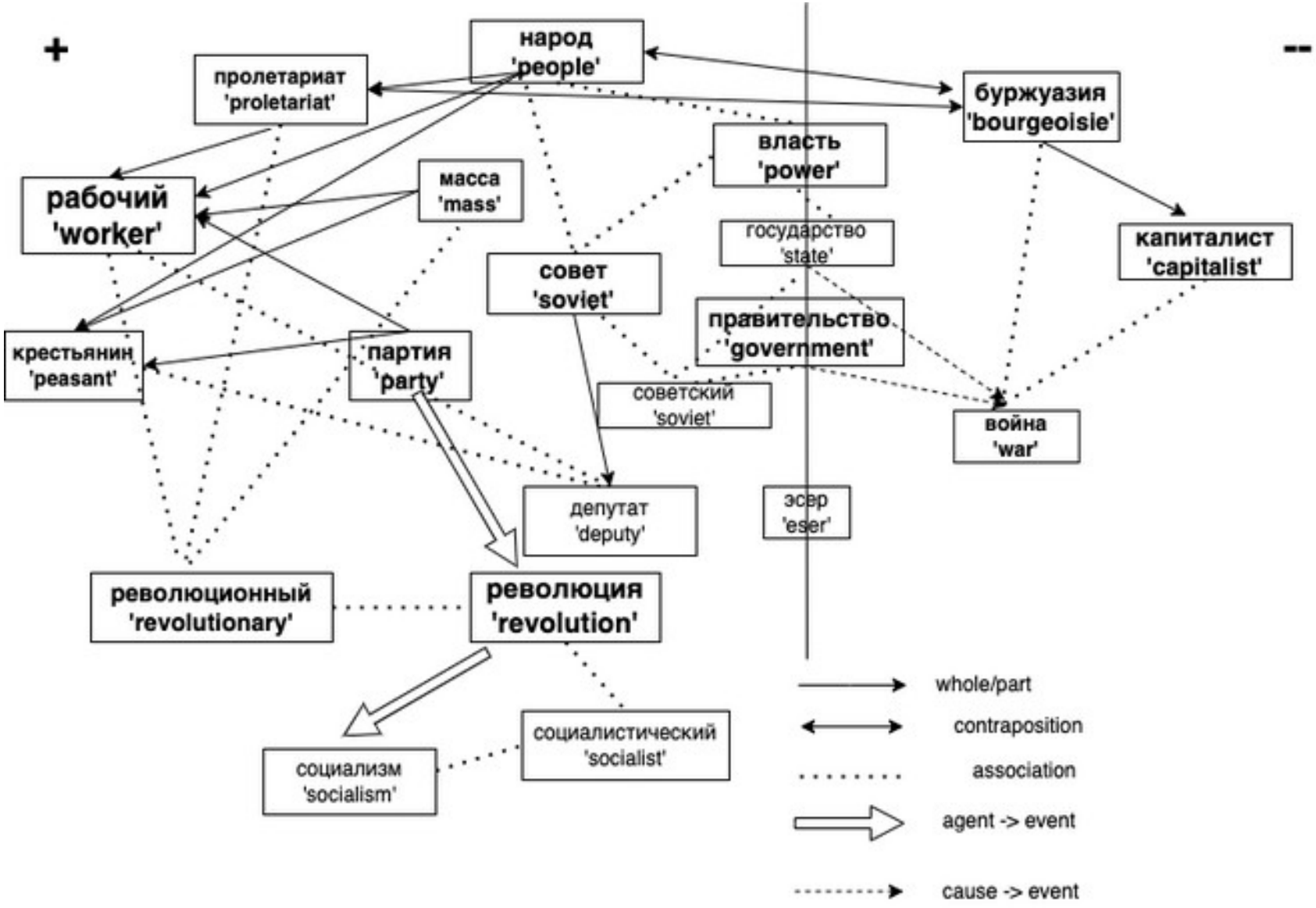
Tampereen yliopisto
Tampere University

Collocation network: Revolution 1905-1907



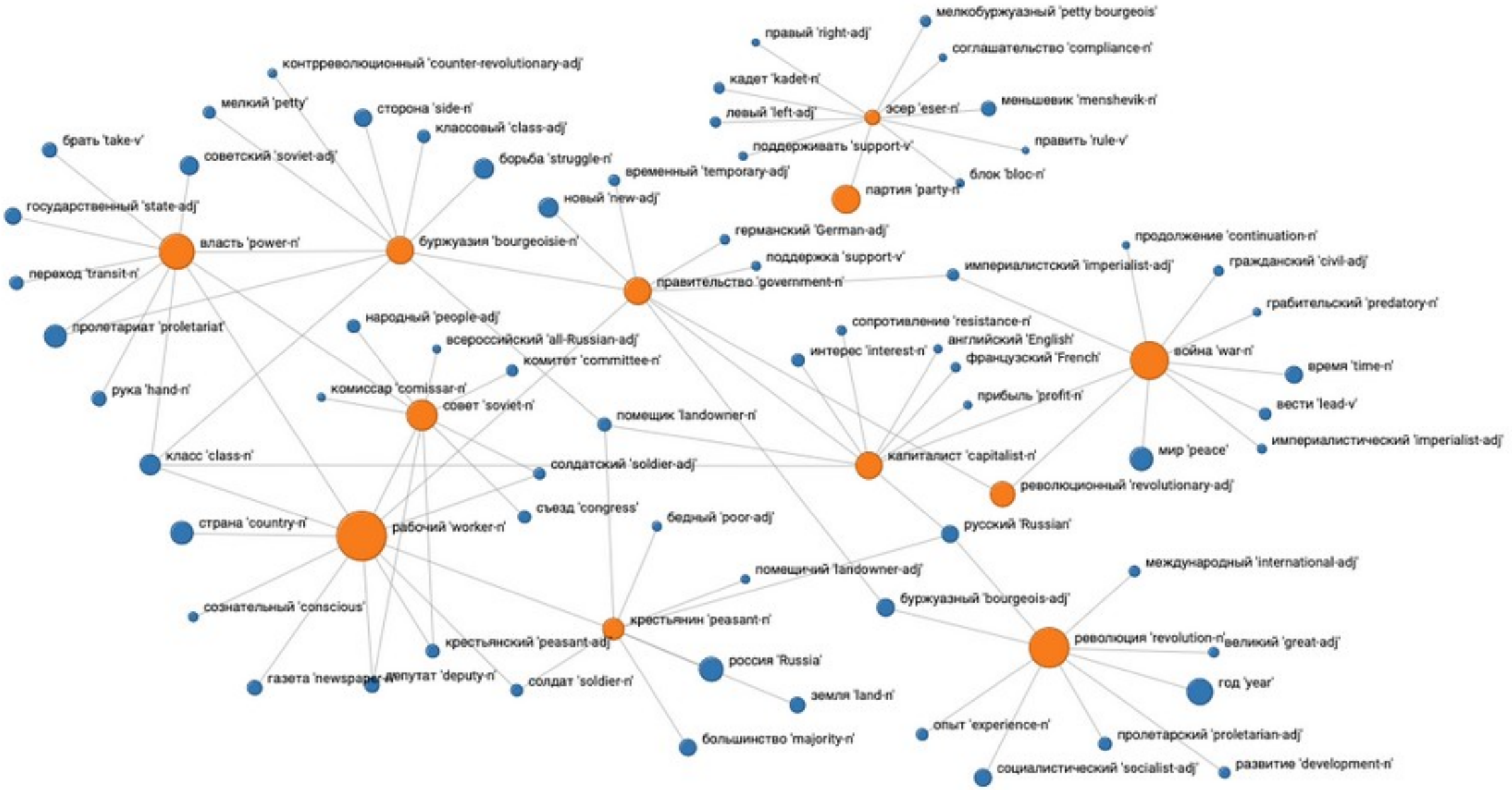
Tampereen yliopisto
Tampere University

Semantic Network: Revolutions 1917



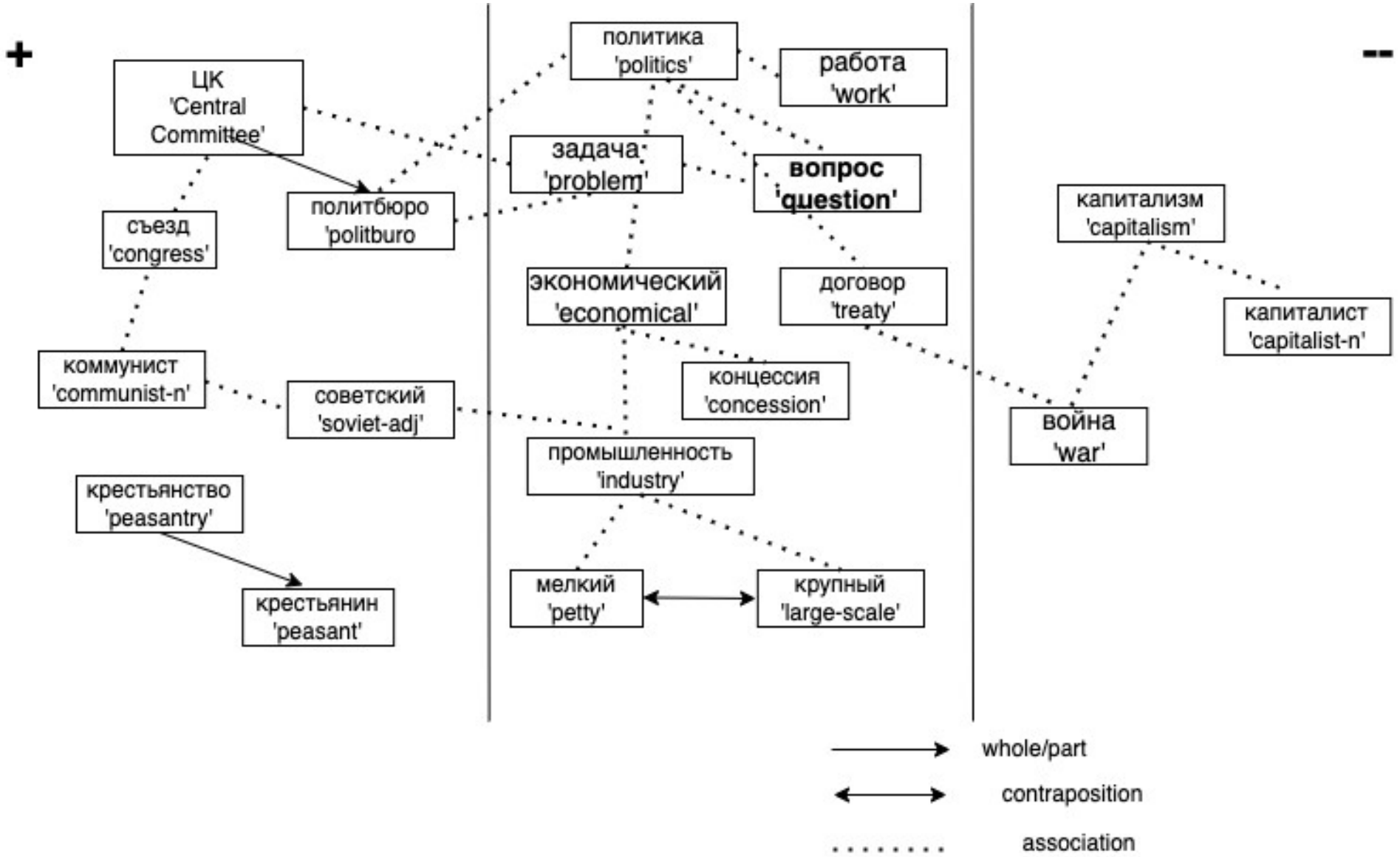
Tampereen yliopisto
Tampere University

Collocation network: Revolutions 1917



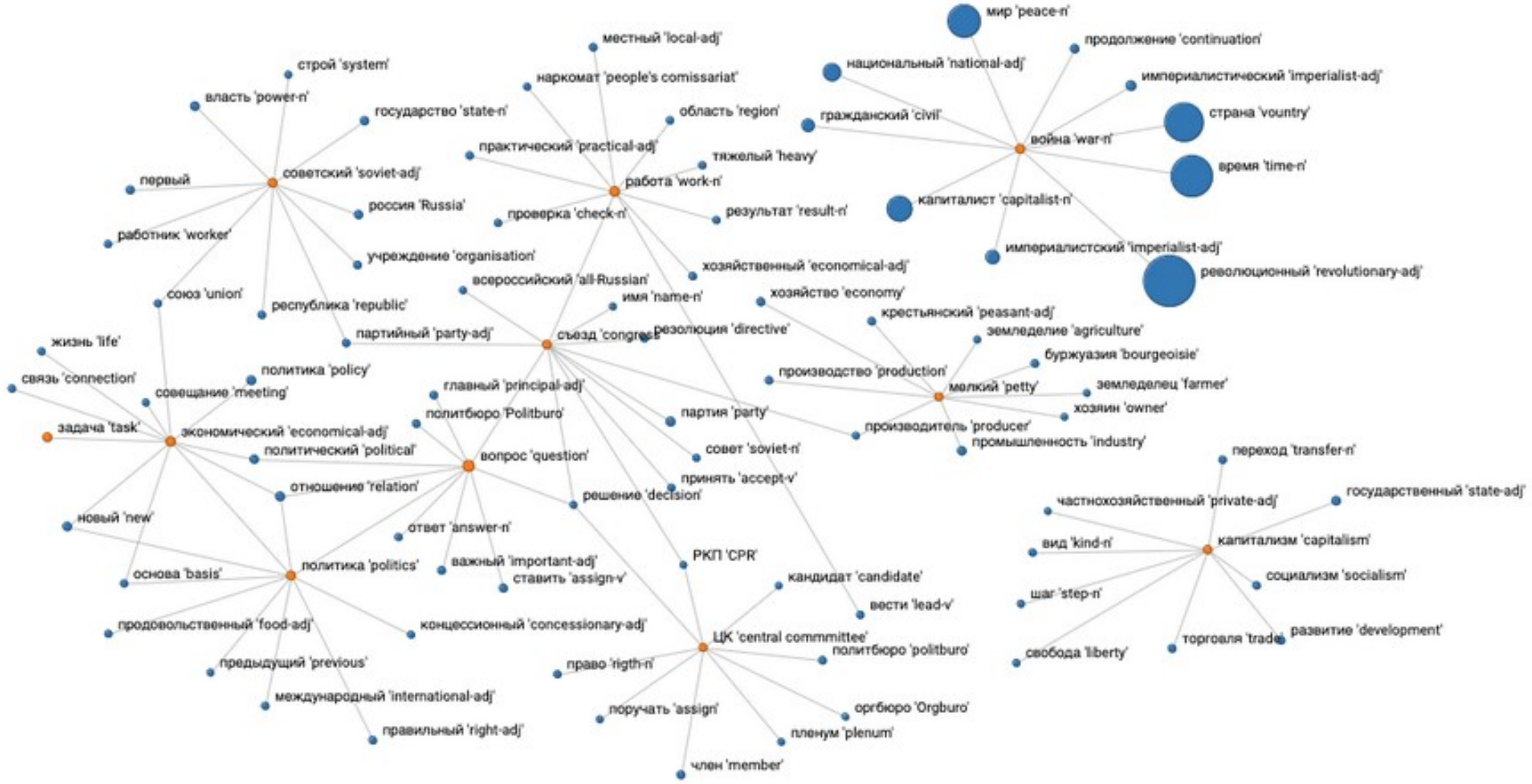
Tampereen yliopisto
Tampere University

Semantic network: building socialism



Tampereen yliopisto
Tampere University

Collocation network: building socialism



Results

- The methods used in the study help to find the most important words of the discourse and to reveal connections between them.
- It is possible to show the evolution of the discourse.
- The study of Lenin's works shows how he turns from abstract political economy to revolutionary propaganda and after revolution to the statesman's work.

Not exactly what I expected

- Where is *демократия* 'democracy'? *свобода* 'liberty'? *диктатура* 'dictatorship'?
  - They are not the top 20 keywords
- Where are metaphors?
- Where is labeling the enemies?
- Where are the irony and sarcasm?
- Where are the swearwords?
  - They are not high frequency words
  - Other methods should be used to detect these features

Tampereen yliopisto
Tampere University

References

- Baker, Paul. 2006. Using Corpora in Discourse Analysis. Continuum Discourse Series. London and New York: Continuum.
- Escobar, M. & Martinez-Uribe L. 2020. Network Coincidence Analysis: The netCoin R Package . Journal of Statistical Software , 93:11, . doi: 10.18637/jss.v093.i11
- Fairclough, Norman. 2010. Critical Discourse Analysis : The Critical Study of Language, Taylor & Francis Group.
- Gillings, Mathew and Gerlinde Mautner and Paul Baker. 2023. Corpus-assisted Discourse Studies. Cambridge University Press.
- Sketch Engine 2015. Statistics used in Sketch Engine. https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf
- Vine, Bernadette. 2023. Understanding Discourse Analysis. London and New York: Routledge.

Tampereen yliopisto
Tampere University

Thank you!