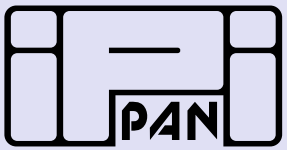
# Korpusomat.pl
# New functionalities and future development

Witold Kieraś Łukasz Kobyliński Maciej Ogrodniczuk
Michał Wasiluk Zbigniew Gawłowicz



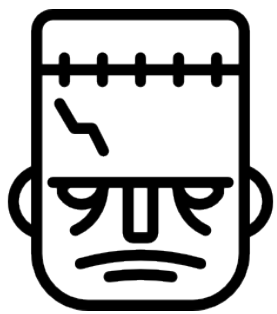

INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. Jana Kazimierza 5, 01-248 Warszawa

September 24th 2018, Prague

# What is Korpusomat?

Web application aimed at building

- automatically indexed and annotated
- searchable corpora
- from documents provided by the user,
- using modern NLP tools for Polish
- and equipped with intuitive user interface.



# For whom?

Mostly for linguists with limited technical skills:

- undergraduate and graduate students
- as well as individual reasearchers

interested in investigating

- sociolects in the Web,
- language of individual authors,
- language of niche and specialist press.

# Limitations

- Korpusomat has been around for nearly two years now.
- It gained some popularity but also quickly reached its limitations:
  - old search engine Poliqarp with no constant support,
  - no possibility to integrate other layers of annotation,
  - poorly scalable.
- A need for a significant reconstruction arose.
- Although the web interface remained the same, the majority of the components were either changed or at least updated.

KORPUSOMAT

NOWY KORPUS | MOJE KORPUSY | INFORMACJE | KONTAKT | PREFERENCJE | WYLOGUJ

# KORPUSOMAT

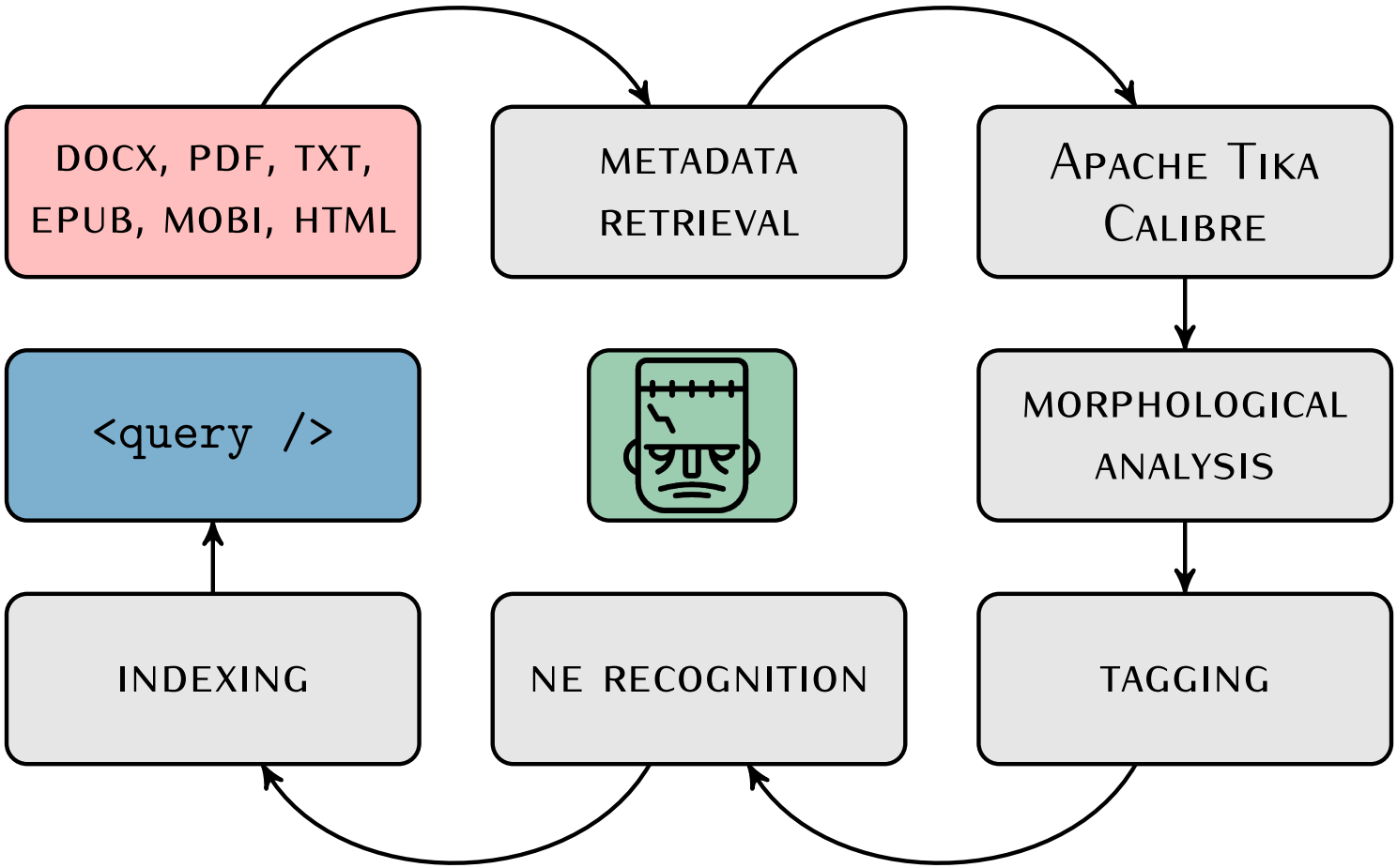
Utwórz własny korpus językowy

## MOJE KORPUSY JĘZYKOWE

## UTWÓRZ KORPUS

## PRZEŚLIJ OPINIĘ

# Workflow


DOCX, PDF, TXT,
EPUB, MOBI, HTML
METADATA
RETRIEVAL
Apache Tika
Calibre
<query />
MORPHOLOGICAL
ANALYSIS
INDEXING
NE RECOGNITION
TAGGING

# Morphosyntactic tagging

- Morfeusz 2 — morphological analyser [Woliński, 2014]
  - two custom dictionaries: SGJP and Polimorf,
  - frequent updates,
  - large portion of proper names added just recently.

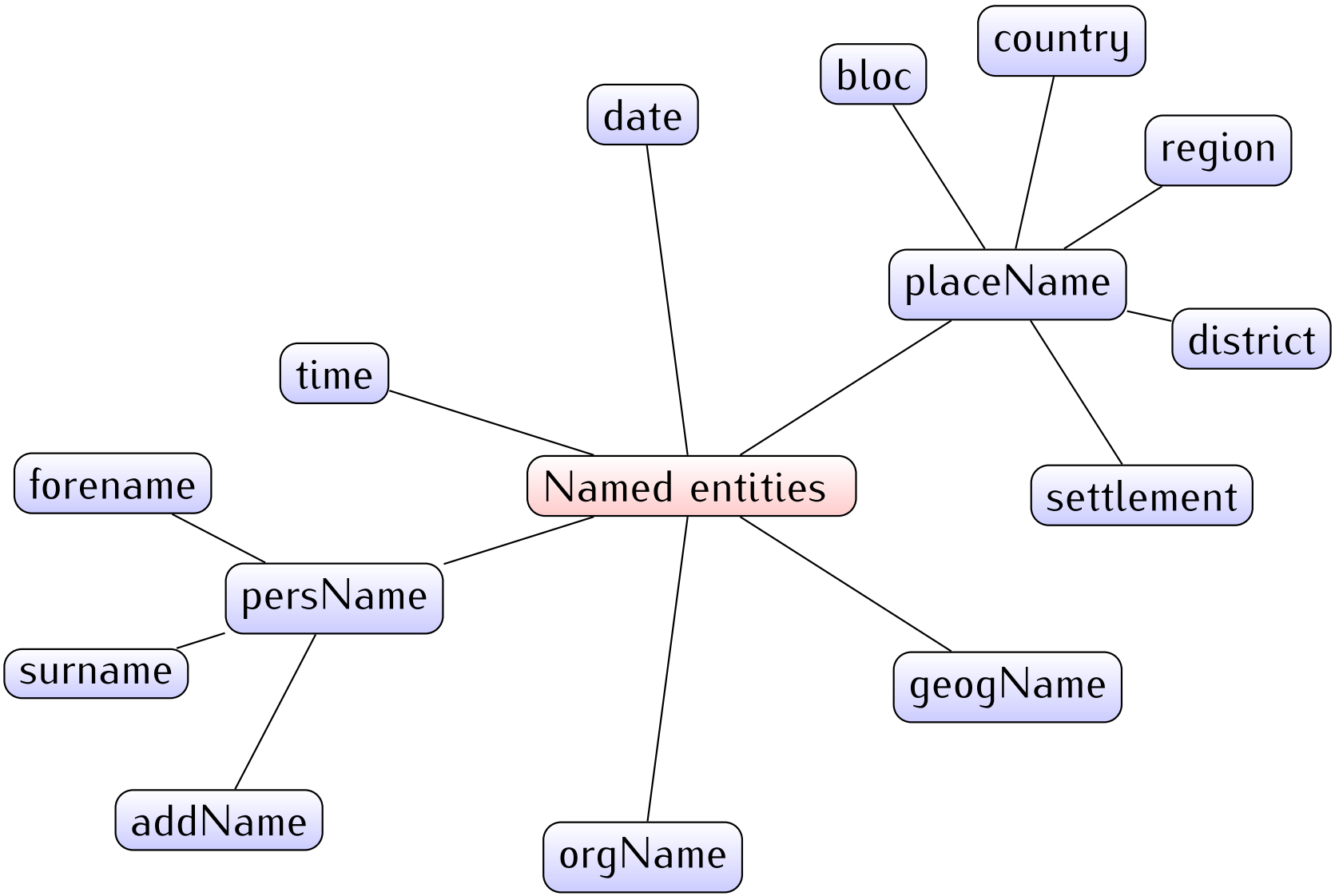
- Concraft 2 — morphosyntactic tagger [Waszczuk et al., 2018]
  - CRF-based,
  - tagging accuracy 92.6%,
  - well integrated with Morfeusz 2,
  - supports ambiguous segmentation,
  - performs sentencing.

# Named entities I

Liner2.6 — named entities classifier [Marcińczuk et al., 2013]

- CRF-based,
- classification model based on National Corpus of Polish,
- accuracy 81%,
- 3rd place in Poleval 2018 shared task,
- `https://github.com/CLARIN-PL/Liner2`

# Named entities II



# Multi Tier Annotation Search I

MTAS = Multi Tier Annotation Search [Brouwer et al., 2017]

- search engine created in Meertens Institute (Royal Netherlands Academy of Arts and Sciences),
- based on software developed by Apache Foundation:
  - Lucene — document search library,
  - Solr — server side application based on Lucene,
- actively developed, used in large projects such as Nederlab — $\approx$ 10 billion tokens large documents database,
- the development of MTAS is firmly situated in the CLARIN domain as part of the Dutch CLARIAH project.

# Multi Tier Annotation Search II

Crucial features:

- indexing of annotated documents and defining custom layers, including multi-word units,
- CQL (Corpus Query Language),
- incremental indexing — no need for processing the whole corpus after adding one document,
- scalability,
- filtering using metadata.

# Query examples

- `<ne="geogName" /> containing [pos="prep"]`
  - *Przy Agorze*
  - *U Huberta*
  - *Zamek Królewski na Wawelu*

# Query examples

- `<ne="geogName" /> containing [pos="prep"]`
  - *Przy Agorze*
  - *U Huberta*
  - *Zamek Królewski na Wawelu*

- `<ne="geogName" /> [pos="conj"] <ne="geogName" />`
  - *Europa Zachodnia i Skandynawia,*
  - *Bliski Wschód oraz Afryka Północna*
  - *Chmielna lub Nowogrodzka*

# Query examples

- `<ne="geogName" /> containing [pos="prep"]`
  - *Przy Agorze*
  - *U Huberta*
  - *Zamek Królewski na Wawelu*

- `<ne="geogName" /> [pos="conj"] <ne="geogName" />`
  - *Europa Zachodnia i Skandynawia,*
  - *Bliski Wschód oraz Afryka Północna*
  - *Chmielna lub Nowogrodzka*

- `[orth="A.*"][orth="M.*"] fullyalignedwith <ne="persName" />`
  - *Adam Małysz*
  - *Adam Michnik*
  - *Antoni Macierewicz*

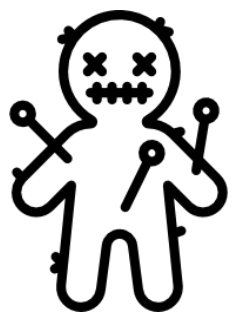
# Future functionalities

Near future:

- statistical tools for collocations,
- more visualizations,
- syntactic information based on dependecy parser.

Further plans:

- terminology extraction,
- more layers: sentiment analysis, word sense disambiguation,
- state-of-the-art NLP deep learning tools,
- platform for sharing users' custom corpora.



# Deployment scheme

So far Korpusomat served as a basis for deployment of various static corpora:

- Baroque Corpus o Polish
  `http://korba.pl`,
- Corpus of texts published between 1830 and 1918
  `http://korpus19.nlp.ipipan.waw.pl`,
- Corpus of Parliamentary Discourse
  `http://sejm.nlp.ipipan.waw.pl`,
- manually annotated gold-standard subcorpus of NCP
  `http://nkjp.nlp.ipipan.waw.pl/`

We hope for other successful deployments:

- integrated Diachronic Corpus of Polish?
- National Corpus of Polish 2.0?

# Conclusions

- Still a simple web application, but with some higher ambitions,
- with a mission for promoting NLP tools and techniques in Polish linguistic community.
- No intentions for developing new tools, focused rather on integrating and reusing the ones created in other projects.
- Feel free to test it at: `http://test.korpusomat.nlp.ipipan.waw.pl/`.
- Soon available at: `http://korpusomat.pl/`.

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

# References I


Brouwer, M., Brugman, H., and Kemps-Snijders, M. (2017).

MTAS: A Solr/Lucene based Multi Tier Annotation Search solution.
In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure*, number 136, pages 19–37. Linköping University Electronic Press, Linköpings universitet.

Marcińczuk, M., Kocoń, J., and Janicki, M. (2013).

*Liner2 – A Customizable Framework for Proper Names Recognition for Polish*, pages 231–253.
Springer Berlin Heidelberg, Berlin, Heidelberg.

Waszczuk, J., Kieraś, W., and Woliński, M. (2018).

Morphosyntactic disambiguation and segmentation for historical Polish with graph-based conditional random fields.
In *Text, Speech, and Dialogue 21th International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings*, Lecture Notes in Computer Science. Springer International Publishing.

Woliński, M. (2014).

Morfeusz reloaded.
In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavik, Iceland. ELRA.