Corpus Linguistics

Katarzyna Klessa

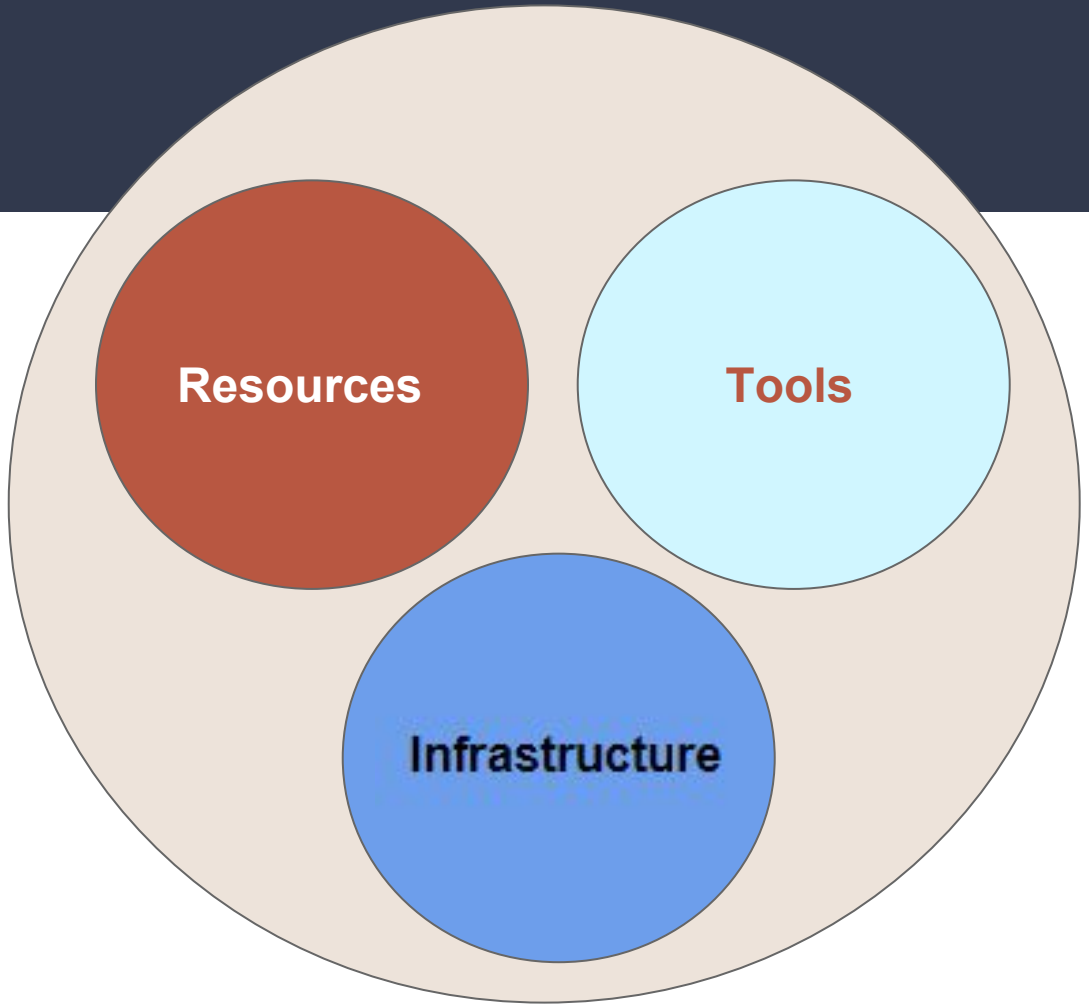
Empirical Linguistics & Language Documentation

CLARIN, CLARIN-PL: resources & tools (1)

What is CLARIN? CLARIN centres

Basic concepts

- Language resources
- Language tools
- Language infrastructure

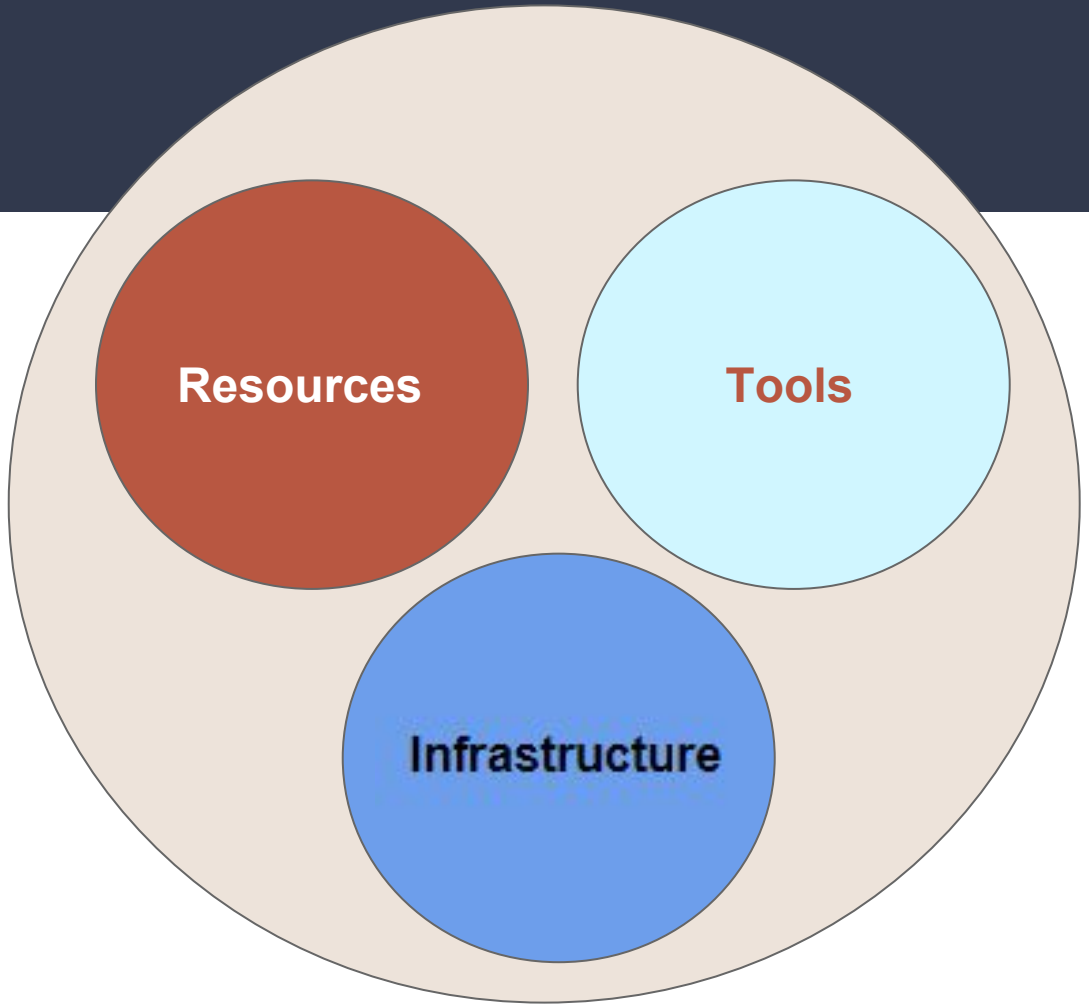


Slides based on the work by Maciej Piasecki et al, e.g.: http://clarin-pl.eu/wp-content/uploads/2017/01/SM-intro-part1.pdf, http://clarin-pl.eu/pl/category/clarin-pl/

Basic concepts

Language technology

- Language resources
- Language tools
- Language infrastructure



Slides based on the work by Maciej Piasecki et al, e.g.: http://clarin-pl.eu/wp-content/uploads/2017/01/SM-intro-part1.pdf, http://clarin-pl.eu/pl/category/clarin-pl/

CLARIN: Central Services

https://www.clarin.eu/

Samsung R&D Institute
Invit. Lecture
2017-01-17
CLARIN-PL

CLARIN portal

Get an example-based impression of what's currently available

Depositing services

Store language resources in a sustainable repository at a CLARIN centre

Virtual Language Observatory

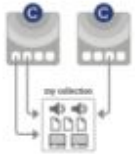
Discover language resources using a faceted browser or a map

Easy access to protected resources

Get easy access to protected resources, with your institutional username and password.

Web services and applications

Explore and analyze language data with a wide variety of tools



Virtual Collections

Create your own digital bookmarks, ideal for citing data sets.

Language Resource Inventory

Submit and access information about

Content Search (prototype)

Search different corpora with a single search

Consulting Services

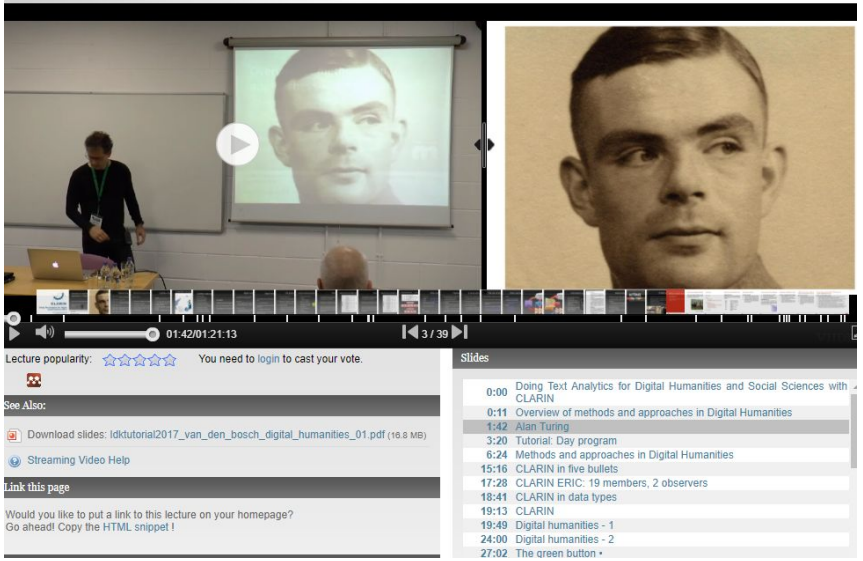
Searching for a specific data set or

Slajd: Maciej Piasecki http://clarin-pl.eu/wp-content/uploads/2017/01/SM-intro-part1.pdf

CLARIN knowledge sharing, video lectures

Since 2016, we have been building an on-line library of talks and tutorials from our training and academic events on the VideoLectures.NET portal, which is an award-winning free and open access educational videolectures repository. You can currently choose between 58 videos that are synchronised with the slides from our 6 international events: the 5th CLARIN Annual Conference, the CLARIN-PLUS workshops on Oral History Archives, Newspaper Collections, Parliamentary Records and Social Media Data, and the CLARIN tutorial on Text Analytics for Digital Humanities and Social Sciences.

https://www.clarin.eu/content/clarin-for-researchers



http://videolectures.net/ldktutorial2017_van_den_bosch_digital_humanities/

CLARIN-PL: clarin-pl.eu

Current tasks:

http://clarin-pl.eu/en/what-are-we-working-on/

Tasks	Headlines	Partner Centers
A1		
A2	Construction of Language Technology Centre	Wrocław University of Technology
A3	Long-term archiving of digital data	Polish-Japanese Institute of Information Technology
A4	Polish Speech Recording Corpus for Training and Evaluation	Polish-Japanese Institute of Information Technology
A5	Corpus of (Press) Articles Published Between 1945 and 1954.	University of Wrocław
A6	The Corpus of Conversation Recordings	University of Lodz
A7	Parallel Polish-English Text Corpus	University of Lodz
A8	Polish, Bulgarian and Russian Text Corpus	Institute of Slavic Studies, Polish Academy of Sciences
A9	The Corpus of Polish and Lithuanian Texts	Institute of Slavic Studies, Polish Academy of Sciences

Example 1: Słowa dnia

- Words of the day - frequency of words based on RSS feeds from 7 popular news services (about counting the frequencies: here)
  Topics of the day - FRAZEO
  Ćwiczenie:
- Let's see the words and topics of the day from today, 31 December 2017, and 20 July 2017. What tendencies can we see?
- Trends for any phrase typed in by the user: http://frazeo.pl/trends

Example 2 - Paralela & SlopeQ for BNC

Paralela: http://paralela.clarin-pl.eu

- SlopeQ for the BNC:
  - Instruction http://pezik.pl/doku.php?id=slopeq
  - Search engine: http://pelcra.clarin-pl.eu/SlopeqBNC/#home
- POS Tagset: http://www.natcorp.ox.ac.uk/docs/c5spec.html

Example 2 – Paralela: exercises

1. How many occurrences of the phrase *before class* do we find depending on switching on/off the ‘word order’ option? How many do we find when setting the slop factor to 1, 2? In what kind of sources/genre/medium categories does it occur? Compare it to the number and areas of occurrences of the phrases *not significant* and *I don’t care.*
2. Find the occurrences of the noun *corpus* preceded by any adjective, article or an -ing form of a lexical verb. Do the same for the word *corpora* - compare the outputs.

Example 3: Spokes Conversational Speech

Documentation:

http://pelcra.pl/docs/doku.php?id=spokes_documentation

- Let's search the Spokes data for the phrase *chce mi się spać* using various settings of slop and word order constraint. Check the results also with regard to categories of speaker gender, sex, and others. See also the visualisations in Facets section.
- Now, let's take a look at the phrase *nie chce mi się* - what can you observe now? What can be the reasons? :)

Słowosieć

PL WordNet On-line:

- http://plwordnet.pwr.wroc.pl/wordnet/
- Semantic similarity function

Discussion and teamworking:

Possible advantages of using WordNets, compare Polish wordnet to other similar tools, primarily Princeton WordNet (e.g., search for the same concept and compare results).

Princeton WordNet

Psycholexicography as the starting point

The initial idea of Princeton WordNet was to provide an aid to use in searching dictionaries **conceptually**, rather than merely alphabetically

WordNet can be said to be **a dictionary based on psycholinguistic principles** (Miller et al.).

*WordNet interlinks not just word forms—strings of letters—but specific **senses** of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet **labels the semantic relations** among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity.*

http://wordnet.princeton.edu/wordnet/

About Princeton WordNet

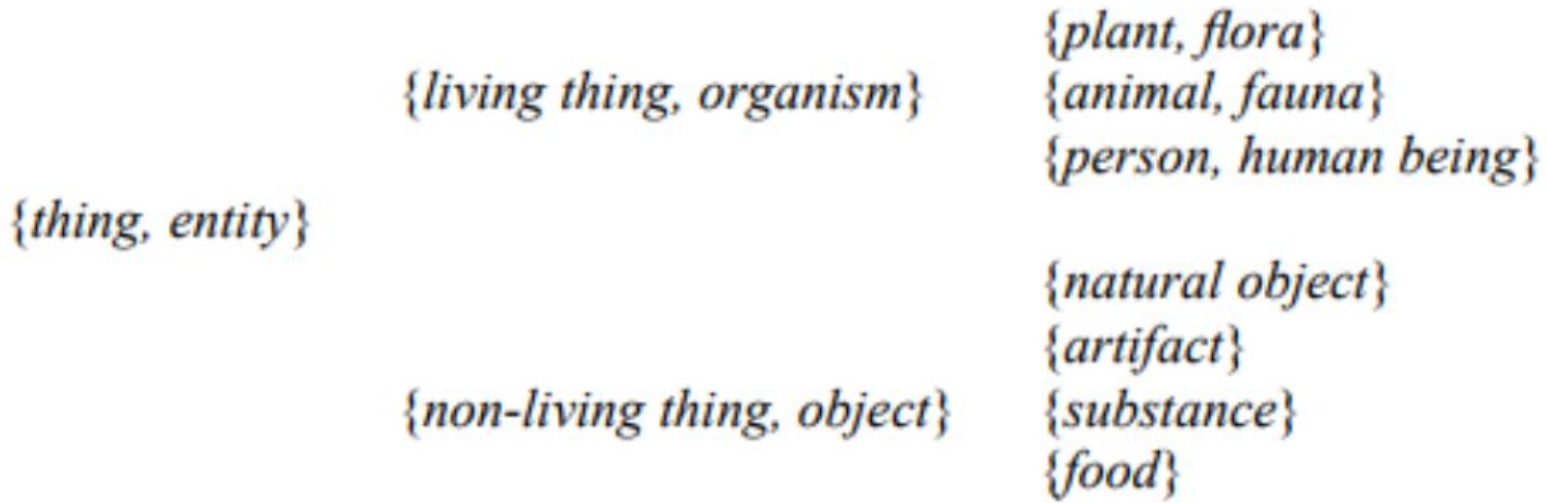
- **Structure** (synonymy) - synonyms grouped into sets (*synsets*)
- **Relations** - hyperonymy, hyponymy (super-subordinate relations), antonymy (for adjectives), meronymy

- **Cross POS** relations (nouns, verbs, adjectives and adverbs), “morphosemantic” links for words sharing a stem with the same meaning: observe (verb), observant (adjective) observation, observatory (nouns)

About Princeton WordNet

- **Structure** (synonymy) - synonyms grouped into sets (*synsets*)
- **Relations** - hyperonymy, hyponymy (super-subordinate relations), antonymy (for adjectives), meronymy (part-whole), hierarchies (verbs at the bottom most specific: communicate-talk- whisper)
- **Cross POS** relations (nouns, verbs, adjectives and adverbs), “morphosemantic” links for words sharing a stem with the same meaning: observe (verb), observant (adjective) observation, observatory (nouns)

About Princeton WordNet



Hyponymic relations between several words denoting different kinds of tangible things

Global WordNet Association

A free, public and non-commercial organization that provides a platform for discussing, sharing and connecting wordnets for all languages in the world.

http://wordnetcode.princeton.edu/5papers.pdf

- **Maximum overlap and compatibility** across wordnets in different languages, while at the same time, allow for the distributive development of wordnets in the world, each wordnet being a language specific structure and lexicalization pattern
- Two main approaches have been followed for building wordnets:
  - **Expand** approach: translate the synsets in the Princeton WordNet to your own language, take over the relations from Princeton and revise;
  - **Merge** approach: define synsets and relations in your own language and then align your wordnet with the Princeton WordNet using equivalence relations;

Polish Wordnet – Słowosieć

plWordNet:

- The **largest** wordnet in the world and is still growing.
- Connected with Princeton WordNet -> **PL-EN** dictionary
- Annotation of affective/**emotional** markedness

Słowosieć

PL WordNet On-line:

- http://plwordnet.pwr.wroc.pl/wordnet/
- Semantic similarity function

Discussion and teamworking:

Possible advantages of using WordNets, compare Polish wordnet to other similar tools, primarily Princeton WordNet (e.g., search for the same concept and compare results).