Pracownia badawcza
Lingwistyka formalna i komputerowa

Adam Przepiórkowski

INSTYTUT PODSTAW INFORMATYKI
POLSKIEJ AKADEMII NAUK
ul. Jana Kazimierza 5, 01-248 Warszawa

UNIWERSYTET
WARSZAWSKI

K2 UW
27 listopada 2018

Lingwistyka formalna i komputerowa

Zakres pracowni: **badanie języka** (cel naukowy, nie inżynieryjny) **metodami formalnymi i komputerowymi**, chętnie w odniesieniu do *psycholingwistyki* i *neurokognitywistyki*.

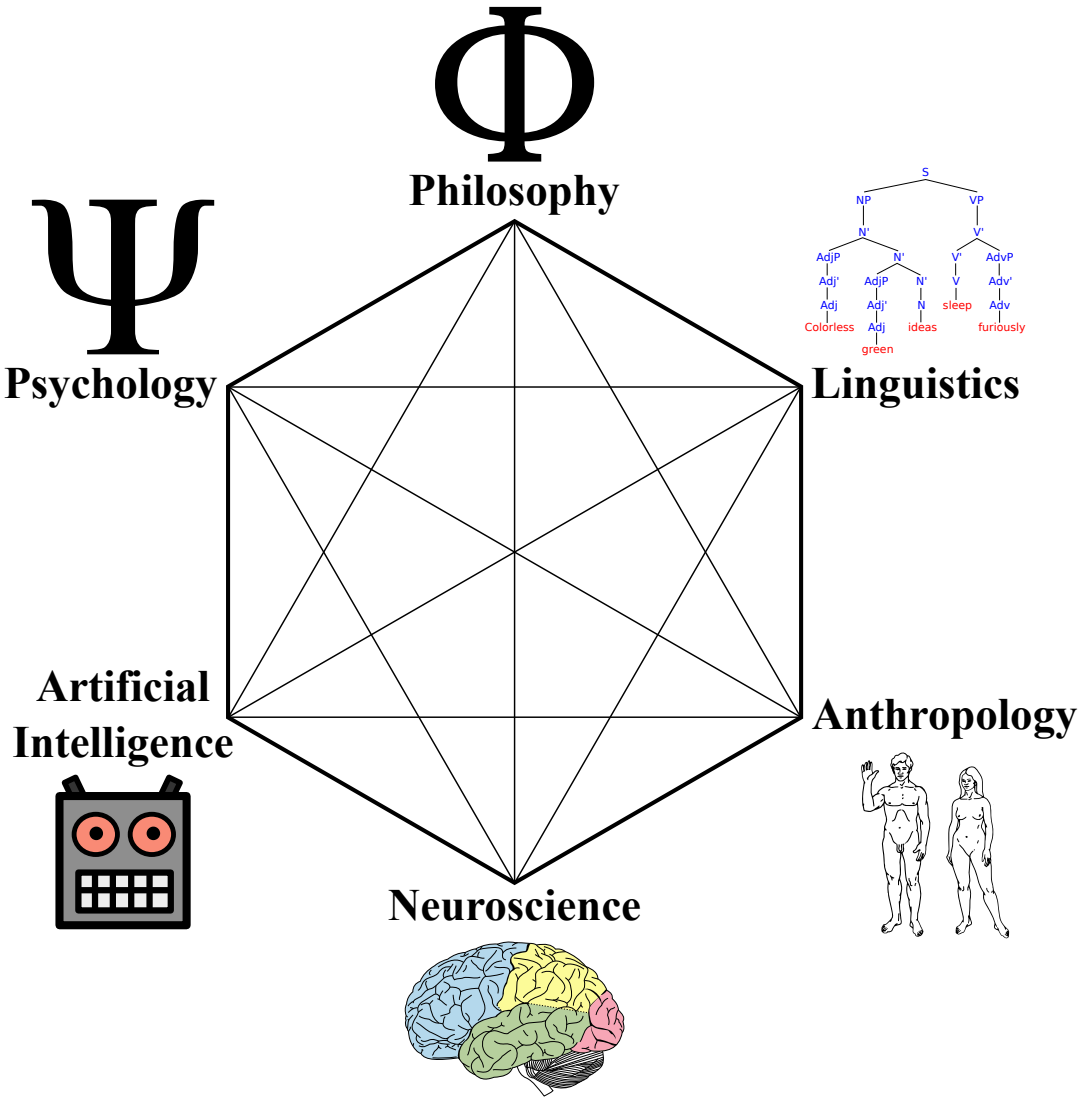
Dlaczego **lingwistyka jest istotna dla kognitywistyki**?

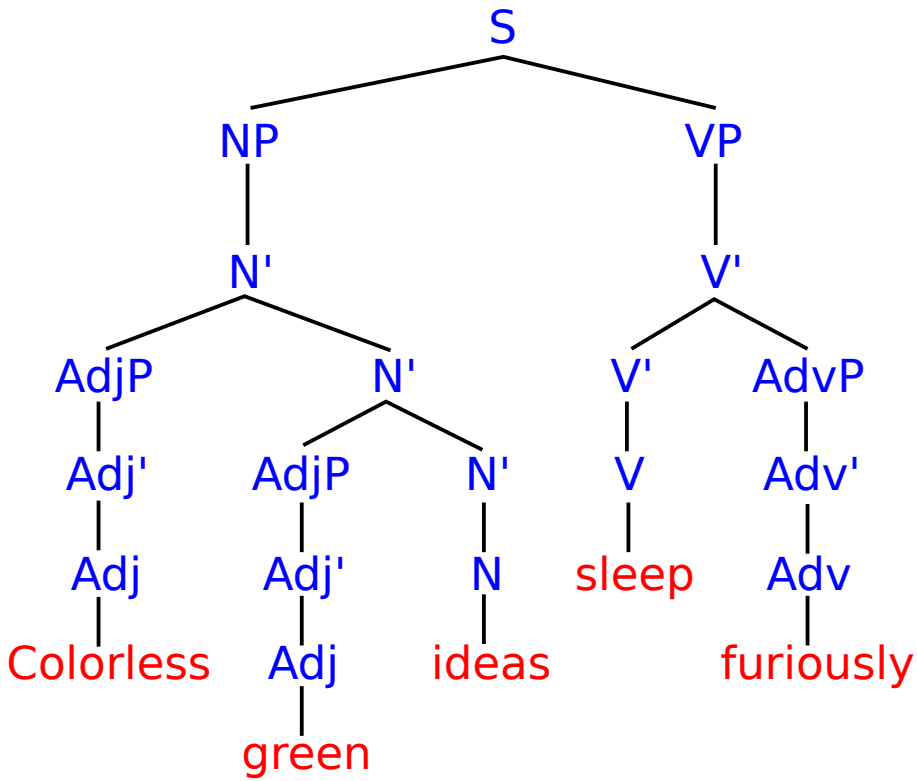
- **umysł** jest kluczowym pojęciem kognitywistyki,
- **znaczenia**, **pojęcia** i **myśli** są kluczowe dla umysłu,
- **język** pozwala na **uchwycenie pojęć i myśli**,
- badanie **znaczeń wyrażonych za pomocą języka naturalnego** jest więc kluczowe dla kognitywistyki.

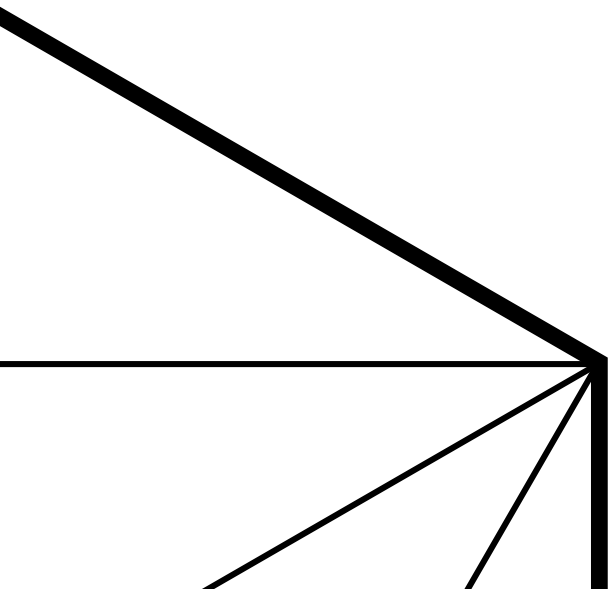
Kolejny powód:

- język jest w zasadzie **wyłączną cechą człowieka**,
- a więc badanie **jak działa język** jest też badaniem **jak umysł człowieka różni się od umysłu innych zwierząt**.

Wikipedia o kognitywistyce







Linguistics

Ogólny zarys badań pracowni

Główne akcenty:

- nacisk na **język polski** (ale badanie polszczyzny jest też badaniem zdolności językowych w ogóle!)
- **składnia i semantyka**

Dwa główne przenikające się **nurty**:

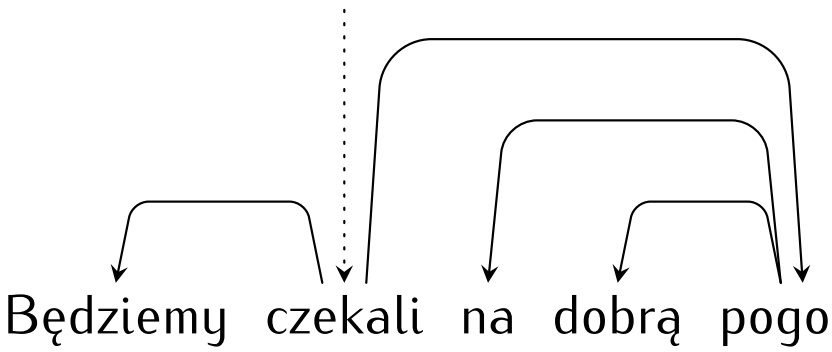
- 1. całościowy opis **składni języka polskiego** (Patejuk, Przepiórkowski):
  - **teoretyczny** (liczne publikacje, m.in. w proceedingsach międzynarodowej konferencji Lexical Functional Grammar)
  - **implementacyjny** – prezentacja POLFIE: `http://clarino.uib.no/iness/xle-web`
- 2. opis **walencji** – wymagań składniowych i semantycznych – **w języku polskim** (Hajnicz, Przepiórkowski, Patejuk)
  - **zasobowy** – prezentacja Walentego: `http://walenty.ipipan.waw.pl/`
  - **teoretyczny** (w tym kwestia odróżniania argumentów od modyfikatorów)
  - **implementacyjny** (w ramach opisu składni i semantyki)
- także 3. **semantyka formalna** (Przepiórkowski) i **lingwistyczna i automatyczna analiza metafor** (Hajnicz)

Aktywny wątek – reprezentacje zależnościowe

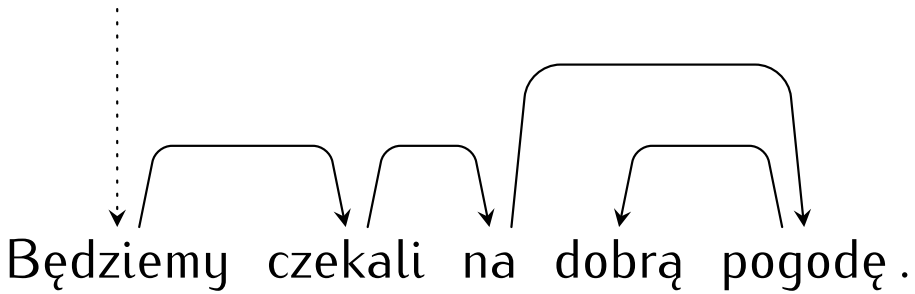
Ten wątek jest **na pograniczu przetwarzania języka** i lingwistyki formalnej i komputerowej.

Reprezentacje zależnościowe dla: *Będziemy czekali na dobrą pogodę*.

Universal Dependencies (UD):



Surface-Syntactic UD:



- Które sensowniejsze **(psycho)lingwistycznie**? (Wiele teorii zależnościowych i schematów znakowania korpusów...)
- Które lepsze **w przetwarzaniu języka**? Na których można wytrenować lepsze parsery?

*Praca w pewnej mierze **programistyczna** (skryptowa).*

Koordynacja – **konstrukcje współrzędnie złożone**, np.:

- **Przyjdź i pozamiataj**.
- Widzę **Janka i Marysię**.
- **Janek i Marysia** spotkali się wczoraj.
- Zwiedzałem Amerykę **Północną i Południową**.
- Rozmawia z **Jankiem i chyba Marysią**.
- **Janek dał Marysi kwiaty a Tomek czekoladki**.
- Zdziwiła go **ta podwyżka i że podniesiono cenę tak drastycznie**.
- Dajcie **wina i całą świnię**!
- **Kto, kogo i kiedy** zaprosił?
- Lubię **Toma i Jerry'ego i Scooby-Doo**.
- Spotkałem się z **Jankiem i Marysią**. vs
- Spotkałem się z **moim lekarzem i najlepszym przyjacielem**.

Zadanie 1: **Badania korpusowe nad koordynacją**:

- **częstość** różnych typów koordynacji (i inne statystyki)
- typowe **konteksty** różnych typów koordynacji

Praca trochę lingwistyczna, trochę mniej programistyczna, trochę mniej statystyczna.

Zadanie 2: **Badania składniowo-słownikowe nad koordynacją**:

- **Walenty** – możliwość koordynacji różnych typów fraz w ramach jednej pozycji składniowo-semantycznej
- należy wydobyć z Walentego takie informacje – **przetwarzanie pliku XML**
- policzyć **statystyki** koordynacji różnych typów składniowych i semantycznych

Praca trochę programistyczna, trochę mniej lingwistyczna, trochę mniej statystyczna.

Walencja – opisuje **argumenty** (nie: modyfikatory).

Załóżmy, że potrafimy wskazać w zdaniu **podrzędniki** danego czasownika (ogólniej: predykatu). **Które z nich wyrażają jego argumenty, a które są modyfikatorami?**

- Janek {przeczytał / położył} książkę **na fotelu**.
- Janek {odłożył / potraktował} książkę **niedbale**.
- Janek {czekał / przeczekał} **całą godzinę**.

W literaturze **parami niezgodne kryteria**, np.:

- niektóre **argumenty** są obligatoryjne (nieusuwalne), wszystkie modyfikatory są opcjonalne (usuwalne),
- frazy rzeczownikowe to **argumenty**, przysłówkowe to **modyfikatory**,
- argumenty określają uczestników stanu czy zdarzenia, zaś **modyfikatory** określają jego okoliczności.

Badanie sensowności i operacyjności kryteriów odróżniania argumentów od modyfikatorów, szczególnie kryterium obligatoryjności:

- jakie są **rodzaje i źródła obligatoryjności**?
- składniowe, np.:
  - Janek już zjadł.
  - ?Janek już pożarł.
- semantyczne, np.:
  - **A**: Janek przyjechał. **B**: Skąd? **A**: Nie wiem.
  - **A**: Janek przyjechał. **B**: Dokąd? ?**A**: Nie wiem.
- pragmatyczne, np.:
  - Ten dom został zbudowany {wczoraj / przez znaną firmę / niedbale...}.
  - ?Ten dom został zbudowany.
- jak **zależą od kontekstu**? np.:
  - **A**: Kto jest gotowy do wyjścia? **B**: Janek już zjadł.
  - **A**: Kto jest gotowy do wyjścia? ?**B**: Janek już pożarł.
  - **A**: Co się stało z jabłkiem, które tu leżało? **B**: Janek pożarł.

Praca lingwistyczna (może trochę psycholingwistyczna?).

Aktywizujący się wątek – walencja dla ludu

Walenty:

Warstwa składniowa:

`NATAPIROWAĆ: perf:`
`subj{np(str)} + obj{np(str)} + {np(dat)} + {np(inst)}`

Warstwa semantyczna (w przybliżeniu):

`{CZŁOWIEK} + {WŁOSY} + {CZŁOWIEK} + {GRZEBIEŃ, SZCZOTKA DO WŁOSÓW}`

Jaka postać informacji dla nie-lingwisty?

Może tak?

Ktoś NATAPIROWAŁ włosy komuś grzebieniem lub szczotką.

Praca UX i programistyczna, w znacznie mniejszym stopniu lingwistyczna. (Kontynuacja licencjatu sprzed paru lat).

Zapraszam

Zapraszam do udziału w pracowni!

Proszę o zgłoszenia **do końca tego semestru**.