

Leksykografia elektroniczna

Praktyki, metody, narzędzia

Krzysztof Nowak

Pracownia Łaciny Średniowiecznej

Instytut Języka Polskiego PAN



Plan wykładu

- **empiryczne badania nad użytkownikiem słownika**

- po co?
- jak?

- **zbieranie danych**

- ankieta, wywiad, kwestionariusz
- analiza logów
 - przypadek: ANW
 - przypadek: Wiktionary
 - laboratorium: analiza logów polskiego Wikisłownika
- eye-tracking
 - Tono 2011
 - Lew et al. 2013
 - laboratorium

Cel badań

Lew & de Schryver (2014): „The aim of dictionary user research is to study how human users interact with dictionaries with the aim of making this interaction **more effective** (improving success), **more efficient** (faster), and **more satisfying** (pleasant to use)”



Co zmieniły słowniki
elektroniczne w technice i
efektach badań?

Empiryczne badania nad użytkownikiem (Koplenig 2014)

- 1. Pytanie badawcze**
- 2. Operacjonalizacja**
- 3. Model badawczy**
- 4. Zbieranie danych**
- 5. Analiza danych**

Empiryczne badania nad użytkownikiem (Koplenig 2014)

1. Pytanie badawcze

- co?
- dlaczego?
- jak?
- kiedy?
- **ex. czy różne grupy użytkowników mają różne upodobania względem słownika internetowego?**
 - zmienna zależna: preferencja
 - zmienna niezależna: zawód, wykształcenie itd..

2. Operacjonalizacja

- jak mierzyć zmienne?
- **ex. wiek, płeć, wykształcenie**
- **ex. doświadczenie użytkownika słownika**
 - użycie słownika każdego dnia

Empiryczne badania nad użytkownikiem (Koplenig 2014)

3. Model badawczy

Wymiar czasowy: typ

- **poprzeczne**

- próbka zbierana w tym samym czasie (nie śledzi zmiany)
- ex. ankietowanie użytkowników słownika i porównanie grup

- **podłużne**

- trend
 - różne próbki na przestrzeni czasu
 - ex. śledzenie trendu użycia słownika na podstawie logów
- panel
 - ta sama próbka na przestrzeni czasu
 - ex. rezultat uczenia się używania słownika

Wymiar metodologiczny: struktura

- **eksperymentalne**

- warunki
 - grupa kontrolna
 - losowy przydział
 - manipulowanie zmiennymi niezależnymi
- pozwala na wnioskowanie przyczynowe

- **quasi-eksperymentalne**

- nielosowy przydział do warunków (np.. efekt jakiejś akcji przed i po)
- ex. efektywność nowych funkcji

- **ex-post-facto**

- nielosowy przydział + brak manipulowania (porównanie grup wg istniejących wcześniej cech, np. wpływ czynników na edukację itp.)
- ex. wpływ kompetencji językowych na odbiór słownika

Empiryczne badania nad użytkownikiem (Koplenig 2014)

4. Zbieranie danych

- **metody nieprzeszkadzające (bez świadomości uczestnika)**
- **metody przeszkadzające**

5. Analiza danych

- **reprezentatywność próbki**

Metody (Welker 2013; Bergenholz et al. 2013)

Ankieta, wywiad

- używane już w pierwszych badaniach
- problemy
 - informatorzy: szczerść, pamięć
 - kłopotliwa generalizacja
 - ograniczenie wyboru w pytaniach zamkniętych
 - większość badań dot. użycia słowników opartych o technikę ankietową nie spełnia żadnych wymogów metodologicznych
 - przypadkowy dobór próby (*convenience sampling*, Lew 2015)

Testy

- umiejętność użycia
- efekty użycia (np. pół klasy czyta ze słownikiem, pół bez; uczniowie piszą esej ze słownikiem jedno- lub dwujęzycznym)

Protokoły użytkowników

- rodzaje
 - pisemne
 - dzienniki
 - użycie w określonych okolicznościach (np.. pisanie eseju, tłumaczenie, czytanie) - po lub w trakcie
 - ustne
 - Think-Aloud Protocols (komentarz towarzyszy użyciu)
 - retrospektywne protokoły ustne

Analiza logów (1)

Analiza logów pozwala na określenie:

- a) jakie słowa powinny znaleźć się w słowniku elektronicznym?
- b) jakie są szczególnie często poszukiwane formy ortograficzne?
- c) wyszukiwań słów nieistniejących czy niepoprawnej pisowni

Zalety

- obiektywizm

Wady

- nie daje informacji o motywacji i funkcji słownika (kodowanie, dekodowanie), sytuacji wyszukiwania itd.
- ograniczony w analizie zachowania pojedynczego użytkownika (Lew 2015)
- konieczna obróbka informacji (wyszukiwania robotów, dostęp z instytucji itp.)
- prywatność

Analiza logów: rodzaje informacji

Logi różnią się ilością i rodzajem informacji

- IP
- przeglądarka
- sytuacja
- pola wyszukiwania używane przez użytkownika
- poszukiwany ciąg
- kryteria wyszukiwania
- data i czas
- wyniki wyszukiwania
- kilkanaście linki w haśle
- kopiowane informacje
- źródła zewnętrzne

Tiberius & Niestadt 2015

Studium użycia ANW (1)

Materiał

- logi Algemeen Nederlands Woordenboek
- logi zbierano od 2009 r., a zawierają informacje o wyszukiwaniach i przeglądanych hasłach

Informacje o użytkownikach

- średnio 20 tys. sesji miesięcznie (więcej w okresach „promocji”)
- 65% Holandia, 20% Belgia, pozostałe: Surinam, Antyle Holenderskie

Tiberius & Niestadt 2015

Studium użycia ANW (2)

Techniczne

- niewielka przewaga starszych przeglądarek i systemów
- przewaga PC nad urządzeniami mobilnymi (2012!)

Sesja

- 60% użytkowników czyta jedno hasło podczas sesji; 20% - dwa i więcej
- większość sesji trwa do 10 s
- tylko 30% przeprowadza wyszukiwanie (Google!)

Strategie wyszukiwania

- mimo istnienia 4 możliwych sposobów dostępu najpopularniejszy jest tradycyjny (słowo → znaczenia)

Muller-Spitzer et al. 2015

Analiza logów Wiktionary

Źródła

- logi niemieckiego Wiktionary bez informacji o IP użytkowników

Selekcja danych

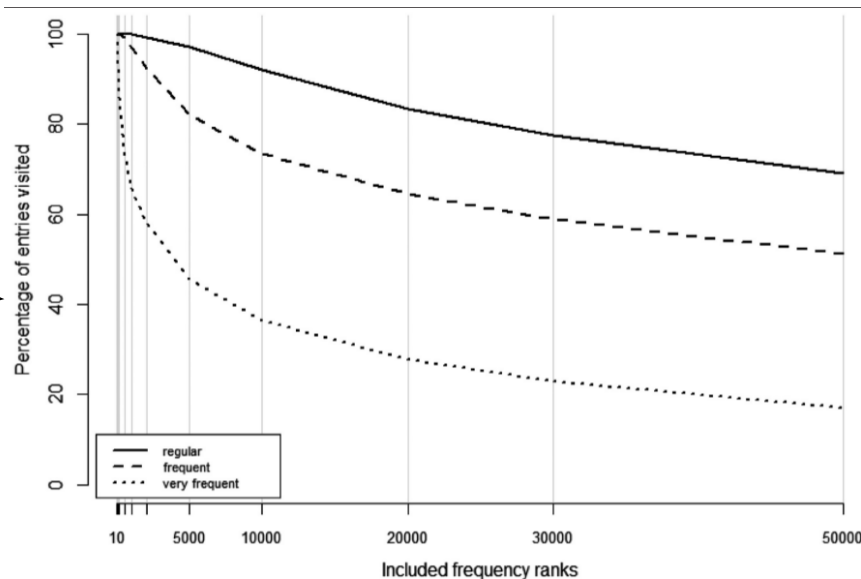
- wybrano słowa, dla których dysponowano informacją o frekwencji i które sprawdzano częściej niż raz na milion wyszukiwań

Indeks:Polski_-_Imiona	29	674315
gimbaza	23	130801
Wikisłownik:Strona_główna	22	140048
Indeks:Arabski_-_Najpopularniejsze_słowa	21	3164014
Kategoria:Polskie_przysłowia	20	271164
cojones	19	89028
qui_pro_quo	19	97750
mężczyzna	18	167504
chuj	15	192291
Indeks:Polski_-_Gwara_więzienna	14	664833
Indeks: Słownik_języka_śląskiego	14	1114854
kawał_z_brodą	14	77952
sarkazm	14	116110
Indeks:Polski_-_Zwi%C4%85zki_frazeologiczne	12	797506
Indeks:Włoski_-_Imiona	12	131775
Słownik_języka_polskiego	12	62151
dupa	12	122130
gniazdo	11	107600
hipokryta	11	79340
mandżur	11	63184

Muller-Spitzer et al. 2015

Ilość wizyt a frekwencja w korpusie

Ilość odwiedzin a frekwencja w korpusie?



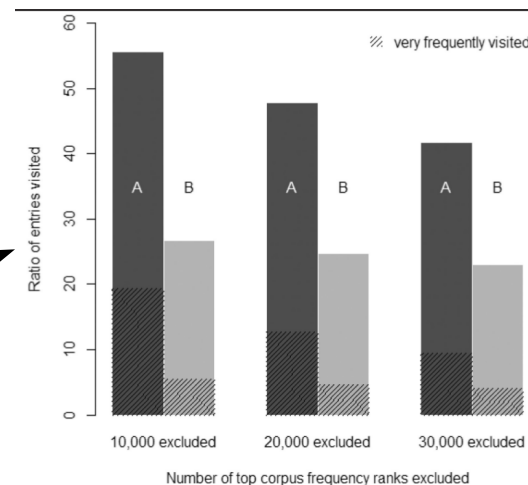
źródło: Muller-Spitzer et al. 2015

Ilość odwiedzin (regularne, częste, bardzo częste) w symulowanych słownikach zawierających wyrazy o frekwencji oznaczonej na osi X

ile wyrazów należy uwzględnić?

Ilość odwiedzin	Procent odwiedzonych stron
1	65,44
2-10	26,75
11-49	6,70
50-500	1,10
500+	0,01

Dystrybucja odwiedzin w niemieckim Wiktionary (źródło: Muller-Spitzer et al. 2015)



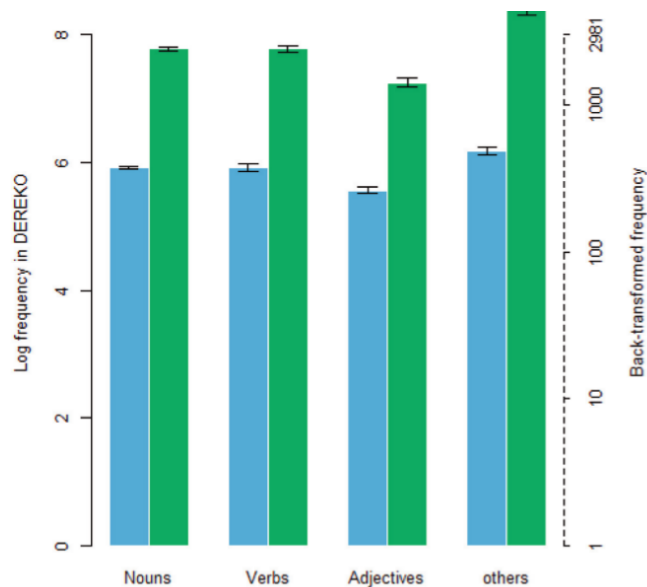
(A – słownik bez pierwszych 10, 20 i 30 tys. najczęstszych wyrazów; B – słownik losowy)

Muller-Spitzer et al. 2015

Ilość wizyt a inne czynniki

Polisemia

Hipoteza: wyrazy polisemiczne będą częściej poszukiwane

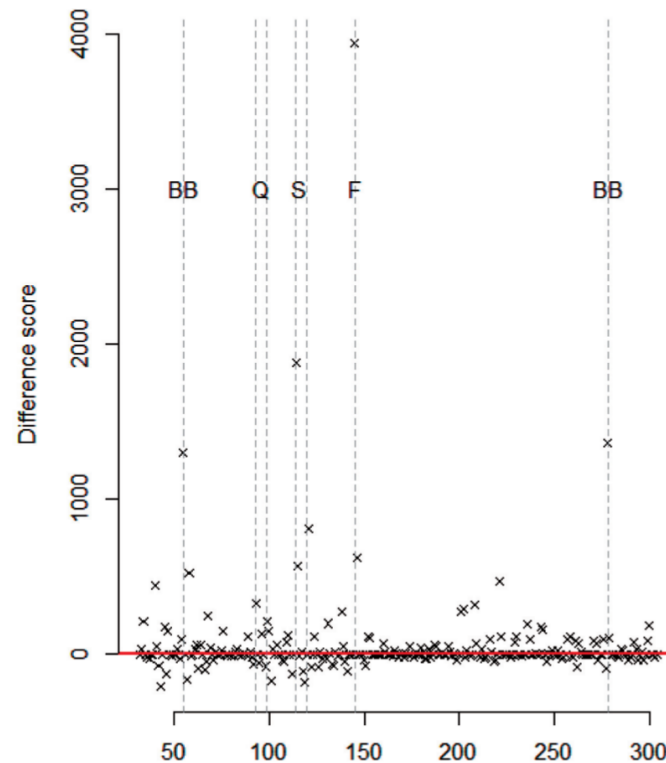


Wniosek: wyrazy polisemiczne są częściej wyszukiwane

Zastrzeżenie: polisemiczność jest maską frekwencji?

Okresowa popularność

Cel: identyfikacja wyrazów okresowo popularnych



Założenie
ilość wizyt =
trend + „szum”
(„szum” jest tu
poszukiwaną
wielkością)

Laboratorium: Analiza logu Wikisłownika

1. Otwórz plik logu za pomocą Libre Office Calc, MS Excel itd.

<http://tinyurl.com/lexilogs>

2. Posortuj dane wg częstości odwiedzin

3. Przeanalizuj najczęściej wyszukiwane słowa i wyrażenia:

a) czego szukają użytkownicy? jakie wnioski z tego powinni wyciągnąć twórcy słowników?

b) czy potrafisz uzasadnić popularność wyrazów?

Obserwacja: *eye tracking* (Tono 2011)

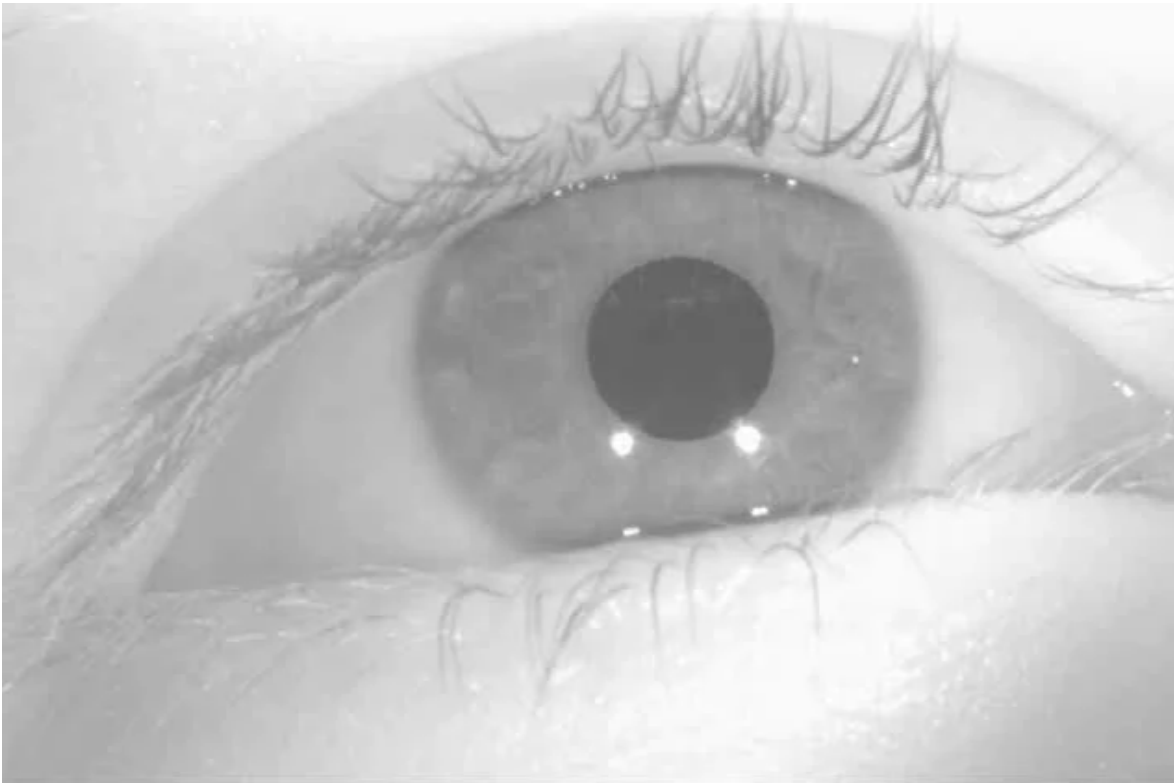
Cel

- analiza procesu korzystania ze słownika przez uczących się angielskiego (zaawansowanych i początkujących)
- analiza wpływu języka (jedno- lub dwujęzyczny słownik), elementów mikrostruktury (pozycja definicji, obecność menu) na użycie słownika

Technika

- śledzenie ruchu gałek ocznych: spojrzenie a uwaga (używane w psychologii poznawczej, ergonomii itd.)
- „spreparowane” hasła słowników + zdanie prezentowane na ekranie PC 8 badanym

Ruchy gałek ocznych



Fiksacje (Tono 2011; Lew 2013: dla cichego czytania : 225 ms) – utrzymanie spojrzenia (koresponduje z przetwarzaniem bodźców: Lew et al. 2013)

Ruchy sakkadowe (skokowe) - ruch oczu pomiędzy dwoma fiksacjami

(źródło wideo: Creative Commons, użytkownik: InkassoSchroeder)

Tono 2011: przykład wpływu signposts

A

make¹ verb

past tense and past participle made

1 PRODUCE [transitive] to produce something, for example by putting the different parts of it together: *I'm going to show you how to make a box for your tools. | A family of mice had made their nest in the roof. | She made her own wedding dress. | The company has been making quality furniture for over 200 years. | They met while they were making a film. Make a list of all the things you need. make somebody something He made her a toy horse, using just some straw and bamboo twigs. | be made from something Paper is made from wood. | be made (out) of something a shirt made of silk | make something from/out of something She's very good at making things from old scraps of material. | Japanese-made/English-made etc (=produced in Japan etc)*

2 DO [transitive] used with some nouns to say that someone does something: *Anyone can make a mistake. | I can't make a decision just yet. | I need to make a quick phone call. | You could have made more effort to talk to him. | He made no attempt to apologize. | Could I make a suggestion? | There are a few points I'd like to make. | The police were called but no arrests were made. | I suppose we should make a start on cleaning this room. | Stop making such a fuss!*

3 COOK [transitive] to cook or prepare food or drink: *When was the last time you made a cake? | John was making breakfast in the kitchen. | Who's going to make the tea? | make somebody something I'll make you some sandwiches.*

4 CAUSE [transitive] to cause something to happen, or cause a particular state or condition: *Its beautiful beaches make this a highly popular area with tourists. | It was this movie which made him a star. | His attitude made him very unpopular with colleagues. | The photo makes her look much older than she really is. | make somebody/ something do something I like him because he makes me laugh. | make something difficult/easy/possible etc The use of computers has made it possible for more people to work from home. | make something the best/worst/most expensive etc Over 80,000 people attended, making it the biggest sporting event in the area. | The President has made it clear that he is not going to change his mind. |*

5 FORCE [transitive] to force someone to do something: *make somebody do something My parents always make me do my homework before I go out. | be*

7 make the meeting/the party/Tuesday etc spoken to be at date or time: *I'm sorry, I can't make Friday as I have to go to the next meeting.*

8 ACHIEVE [transitive] to succeed in achieving a particular position or result: *He made the deadline.*

9 GET MONEY [transitive] to earn or get money: *The plan could cost more than you need to make it work. | He made more than you in life was to make a living (to earn money) from writing children's books. | He's made a fortune (a lot of money) selling computers on the Internet. | make something How easy is it to make money out of garden?*

10 HAVE A QUALITY [linking verb] to have the qualities that are necessary for a particular purpose: *Cardboard makes a comfortable bed for a kitten.*

11 make it/that something spoken used to correct said: *Can we have two cups of coffee, please? No, make it three.*

12 make do to to manage with the things that you have, but not really enough: *I hardly had any food in the house so I had to make do with what I had.*

13 make your someone to hear you, understand you, or know that you are there: *I had to make myself heard above the music.*

14 BE A TOTAL [linking verb] to be a particular amount when added together: *Two and two make four. | There are nine people coming, plus the ones who are already there.*

fiksacja
sakada

B

make¹ verb

past tense and past participle made

1 [transitive] to produce something, for example by putting the different parts of it together: *I'm going to show you how to make a box for your tools. | A family of mice had made their nest in the roof. | She made her own wedding dress. | The company has been making quality furniture for over 200 years. | They met while they were making a film. Make a list of all the things you need. make somebody something He made her a toy horse, using just some straw and bamboo twigs. | be made from something Paper is made from wood. | be made (out) of something a shirt made of silk | make something from/out of something She's very good at making things from old scraps of material. | Japanese-made/English-made etc (=produced in Japan etc)*

2 [transitive] used with some nouns to say that someone does something: *Anyone can make a mistake. | I can't make a decision just yet. | I need to make a quick phone call. | You could have made more effort to talk to him. | He made no attempt to apologize. | Could I make a suggestion? | There are a few points I'd like to make. | The police were called but no arrests were made. | I suppose we should make a start on cleaning this room. | Stop making such a fuss!*

3 [transitive] to cook or prepare food or drink: *When was the last time you made a cake? | John was making breakfast in the kitchen. | Who's going to make the tea? | make somebody something I'll make you some sandwiches.*

4 [transitive] to cause something to happen, or cause a particular state or condition: *Its beautiful beaches make this a highly popular area with tourists. | It was this movie which made him a star. | His attitude made him very unpopular with colleagues. | The photo makes her look much older than she really is. | make somebody/ something do something I like him because he makes me laugh. | make something difficult/easy/possible etc The use of computers has made it possible for more people to work from home. | make something the best/worst/most expensive etc Over 80,000 people attended, making it the biggest sporting event in the area. | The President has made it clear that he is not going to change his mind. |*

5 [transitive] to force someone to do something: *make somebody do something My parents always make me do my homework before I go out. | be*

7 make the meeting/the party/Tuesday etc spoken to be able to go to something that has a particular date or time: *I'm sorry, I can't make Friday as I have to go to the next meeting.*

8 [transitive] to succeed in achieving a particular position or result: *He made the deadline.*

9 [transitive] to earn or get money: *The plan could cost more than you need to make it work. | He made more than you in life was to make a living (to earn money) from writing children's books. | He's made a fortune (a lot of money) selling computers on the Internet. | make something How easy is it to make money out of garden?*

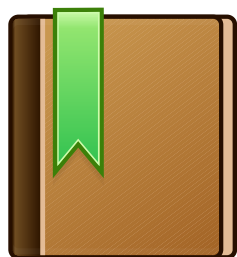
10 [linking verb] to have the qualities that are necessary for a particular purpose: *Cardboard makes a comfortable bed for a kitten.*

11 make it/that something spoken used to correct said: *Can we have two cups of coffee, please? No, make it three.*

12 make do to to manage with the things that you have, but not really enough: *I hardly had any food in the house so I had to make do with what I had.*

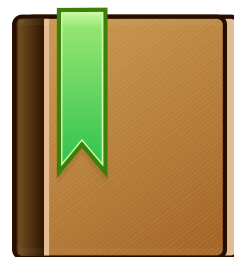
źródło: Tono 2011

Laboratorium: interpretacja zapisu



A

- high
- + signposts



B

- high
- - signposts

Tonno 2011: wybrane wnioski

- **30 % wyszukiwań zakończyło się niepowodzeniem**
 - potrzeba nauczania techniki korzystania ze słownika
 - niepowodzenie $\neq$ strata czasu
- **złożony związek pomiędzy elementami wspomagającymi (menu, znaki) a biegłością użytkowników**
 - menu przydatne tylko użytkownikom niezaawansowanym
 - konieczność treningu wobec konserwatyzmu użytkowników
- **niewielka różnica w użyciu słownika jedno- i dwujęzycznego**

Lew et al. 2013: wybór znaczenia w słowniku dwujęzycznym

Wskazówki na tarczy jego zegarka wskazywały godzinę dwunastą.

tarcz|a

1. (rycerska) shield; **osłaniać się tarczą przed ciosami przeciwnika** to protect oneself from the enemy's blows with a shield
2. (policyjna) (riot) shield
3. (z cyframi, liczbami) dial; **tarcza zegarka** (na rękę) the dial of a. on the watch; (stojącego) the face of the clock; **tarcza kompasu** a compass dial; **tarcza telefoniczna** a. *telefonu* a telephone dial

tarczyc|a

f Anat. thyroid (gland);
nadczynność/niedoczynność tarczycy an
overactive/underactive thyroid

tarczycow|y adi. Anat. thyroid attr.; **gruczoł tarczycowy**
the thyroid (gland)

Taren|t m (*G Tarentu*) Geog. Taranto

targ

Materiał

Spreparowane dla
czytelności
ekscerpty ze
słowników polsko-
angielskich wraz
ze zdaniem
prezentującym
słowo w kontekście

Cel

- jak użytkownicy poruszają się po
polisemicznych hasłach słownika L1
(pl) → L2 (en) przy tłumaczeniu L1 → L2

Uczestnicy

- studenci zaawansowani (B2-C1) i
niezaawansowani (A2-B1)

Słowa

Znane uczestnikom, ale w
rzadszych znaczeniach, np.
poślizg 'opóźnienie'.
Średnia ilość sensów: 6
Średnia pozycja sensu: 4

<https://www.youtube.com/watch?v=tpLUkKN3AWE>

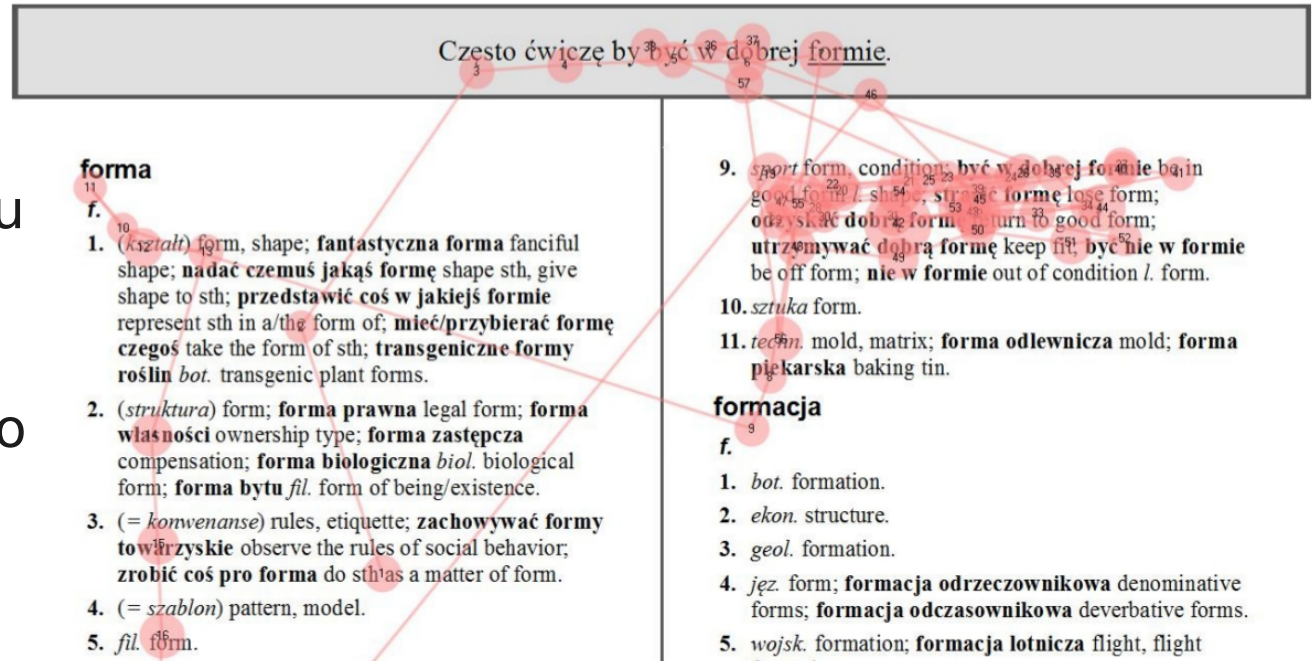
Lew et al. 2013: wybrane rezultaty (1)

Czas fiksacji

- nie różnił się dramatycznie od czasu w lekturze (możliwość ekstrapolacji?)
- 23% czasu poświęcono na elementy wskazujące

Typowy przebieg

1. Lektura zdania (koncentracja na podkreślonym wyrazie). 2. Przegląd wyrazów hasłowych (*formacja* → *forma*). 3. Systematyczne przeglądanie wszystkich sensów. 4. Po dotarciu do właściwego znaczenia (tu: 9) rozpoczyna się jego pogłębiona lektura (szczeg. wyrażenie *być w dobrej formie*). 5. Powrót do zdania dla potwierdzenia i z powrotem. 6. Lektura pozostałych zdań i przegląd pozostałych znaczeń.



Lew et al. 2013: wybrane rezultaty (2)



źródło: <https://vimeo.com/59560367>

Obserwacje

- w przeciwieństwie do niezaawansowanych zaawansowani użytkownicy czytali często wszystkie sensy, nawet jeśli odnaleźli sens właściwy
- lektura słownika: góra → dół, ale na dole wzrok przenosi się do sąsiedniej kolumny i skanowanie postępuje dół → góra
- krótkie hasła nie dawały wcale wyższej skuteczności
- niemal zawsze sens niepoprawny znajdował się przed poprawnym
- pogrubienia i elementy wskazujące przyciągały w sposób znaczący uwagę uczestników

Nie zdążyliśmy powiedzieć o...

- **makro- i mikrostrukturze słowników elektronicznych**
- **crowdsourcingu w leksykografii (Obserwatorium Językowe)**
- **hybrydyzacji słowników**
- **terminologicznych bazach danych**
- **wizualizacji w słownikach (Theron)**
- **promocji słowników (Biesaga)**
- **LOD (Cimiano)...**