



Językowe bazy danych

Eksploracja, interpretacja, analiza

Katarzyna Klessa
klessa@amu.edu.pl





Plan warsztatów

- Podstawowe pojęcia, terminologia „wokół” baz danych:
 - dane surowe vs. korpusy językowe
 - dane vs. metadane
 - dane, metadane vs. wielokrotny użytek zasobów
 - nagrania mowy: transkrypcja – segmentacja – anotacja
- Elementy projektowania i tworzenia korpusów nagrań mowy:
 - etapy prac
 - rodzaje, cechy i wymagania względem korpusów nagrań mowy
 - czynniki wpływające na sposoby gromadzenia i opisu danych oraz wybór technologii
- Elementy języka zapytań SQL

Podstawowe pojęcia

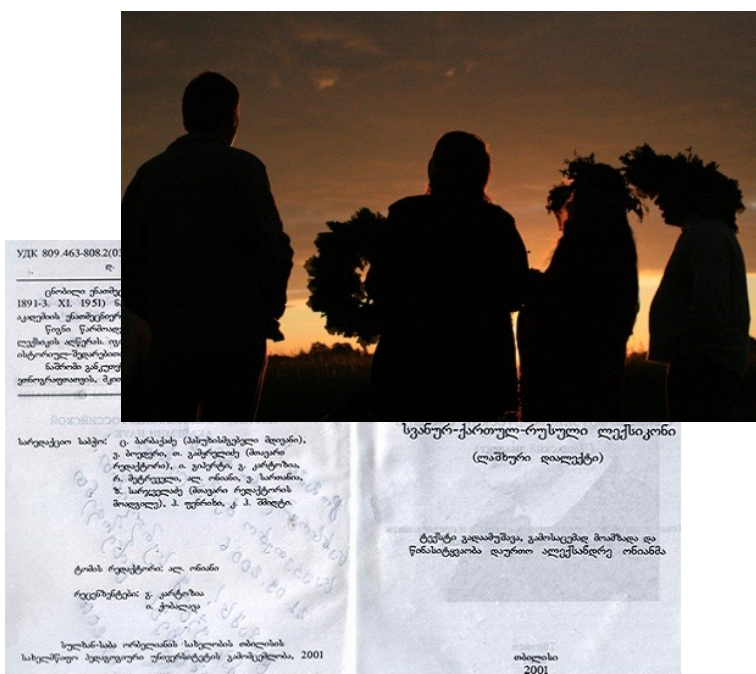


Podstawowe pojęcia?



Dane surowe i korpusy językowe

Dane surowe (raw/primary data) – dane bez obróbki, systematyzacji, nieprzetworzone lub przetworzone w niewielkim stopniu



Zdjęcia: languagesindanger.eu



Dane surowe i korpusy językowe

- **Korpus** (ang. *corpus*, *corpora*; z łac. *ciało*, **określona** całość)
 - dane zebrane zwykle wg założonego z góry klucza, w określonym celu
 - **uporządkowana** kolekcja tekstów pisanych, transkrypcji wypowiedzi i / lub powiązanych z nimi nagrań mowy
 - zwykle w trakcie projektowania korpusu planuje się określoną jego **strukturę** oraz **sposób opisu** danych



Dane Metadane

– metadane – dane o danych

- » dla nagrań mowy – dane osobowe mówcy, wiek, płeć, jego stan psychiczny, dane środowiska nagraniowego, mikrofonów, osób nagrywających, innych osób obecnych przy nagraniu, innych czynników...
- » dla danych tekstowych – dane jak np. w katalogu bibliotecznym, informacje o sposobach kolekcjonowania danych...
- » załączniki – np. zgody na wykorzystanie danych
- » uwaga: rozróżnienie nie jest jednoznaczne, jeśli korpus jest wielokrotnego użytku metadane mogą stać się danymi i na odwrót

Metadane i “re-usability”

Rozbudowany system metadanych może przyczynić się do większej użyteczności korpusu, także w obszarach poza pierwotnym docelowym obszarem zastosowań – np. w psychologii, socjologii, terapii języka i mowy, technologii informatycznej i in.

Kluczowa rola opisu i dokumentacji zasobów.

Aspekty współdzielenia danych i komercyjne.



Transkrypcja, segmentacja, anotacja

- Różnice w definiowaniu tych

Dialekt mazowiecki - Mazury

Tekst gwarowy — Biała Piska 1

PDF PRINT EMAIL

Justyna Garczyńska

Nagranie: Julia Pikacz, Anna Godziuk

Przepisanie: Monika Kresa

Opracowanie: Justyna Garczyńska

Informator: Helena Born, zamieszkała w Białej Piskiej, ur. 20.12.1933 r. w Kaliszkach. Rodzice pochodzili z Kaliszek, pracowali w majątku dziedzica, rodzina mieszkała w czworakach dworskich. Ukończyła sześć klas szkoły podstawowej niemieckiej. Pracowała w Piszu w fabryce drewna, a następnie w przedszkolu jako pomoc.

O wojnie



żeńską formę liczebnikową dwie zastąpioną przez formę męską dwa

Urodziłam się w Kaliszkach i tam żyłam do czasu wojny. I po wojnie..., jak widać, to nazad z pegieerzu z Kaliszka to nam ten pan kazał take wozi, przyszykować drabiniaste, nakładli słomy, tygo

Przykład: archiwum.gwarypolskie.uw.edu.pl



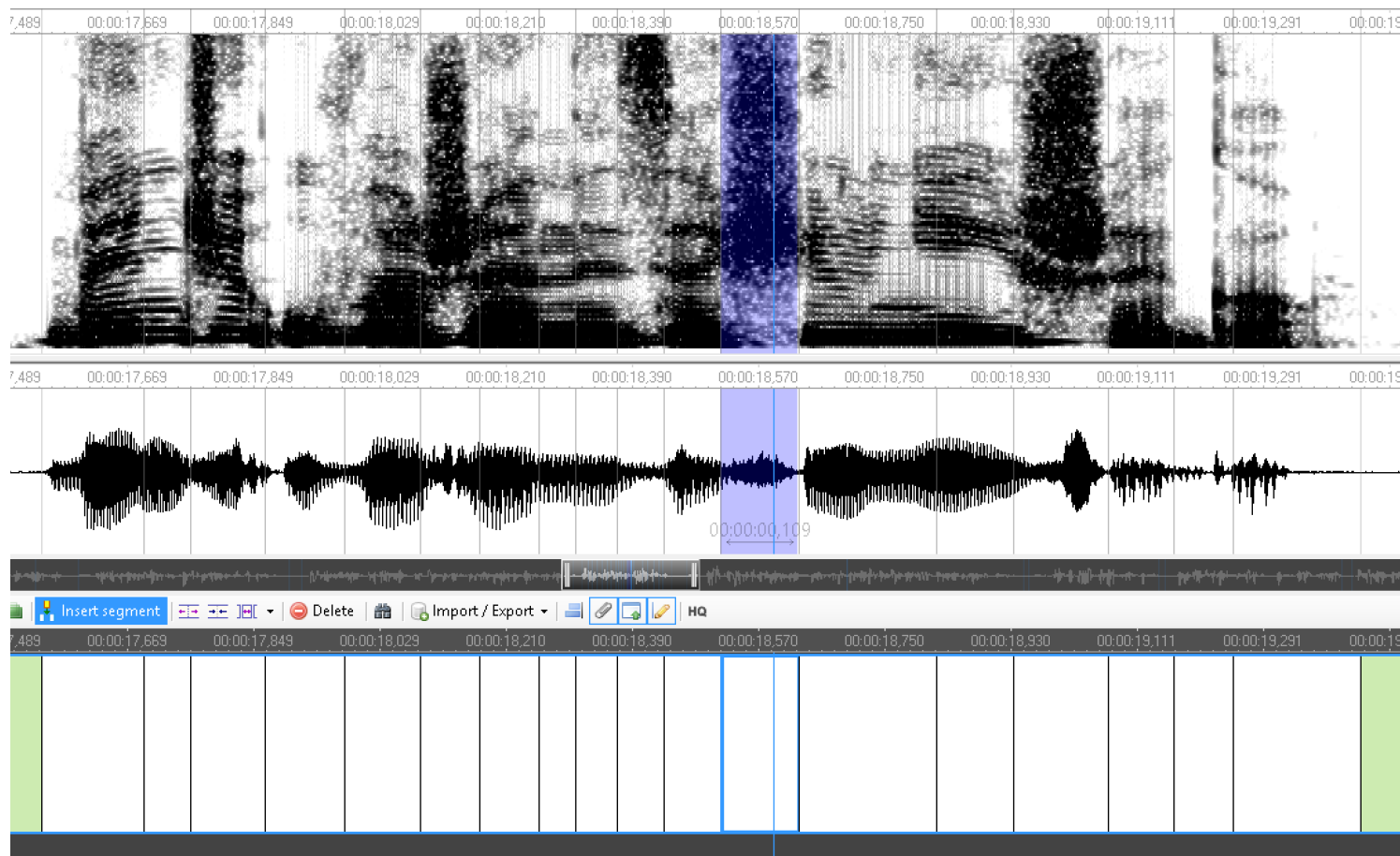
trullsenglish.weebly.com



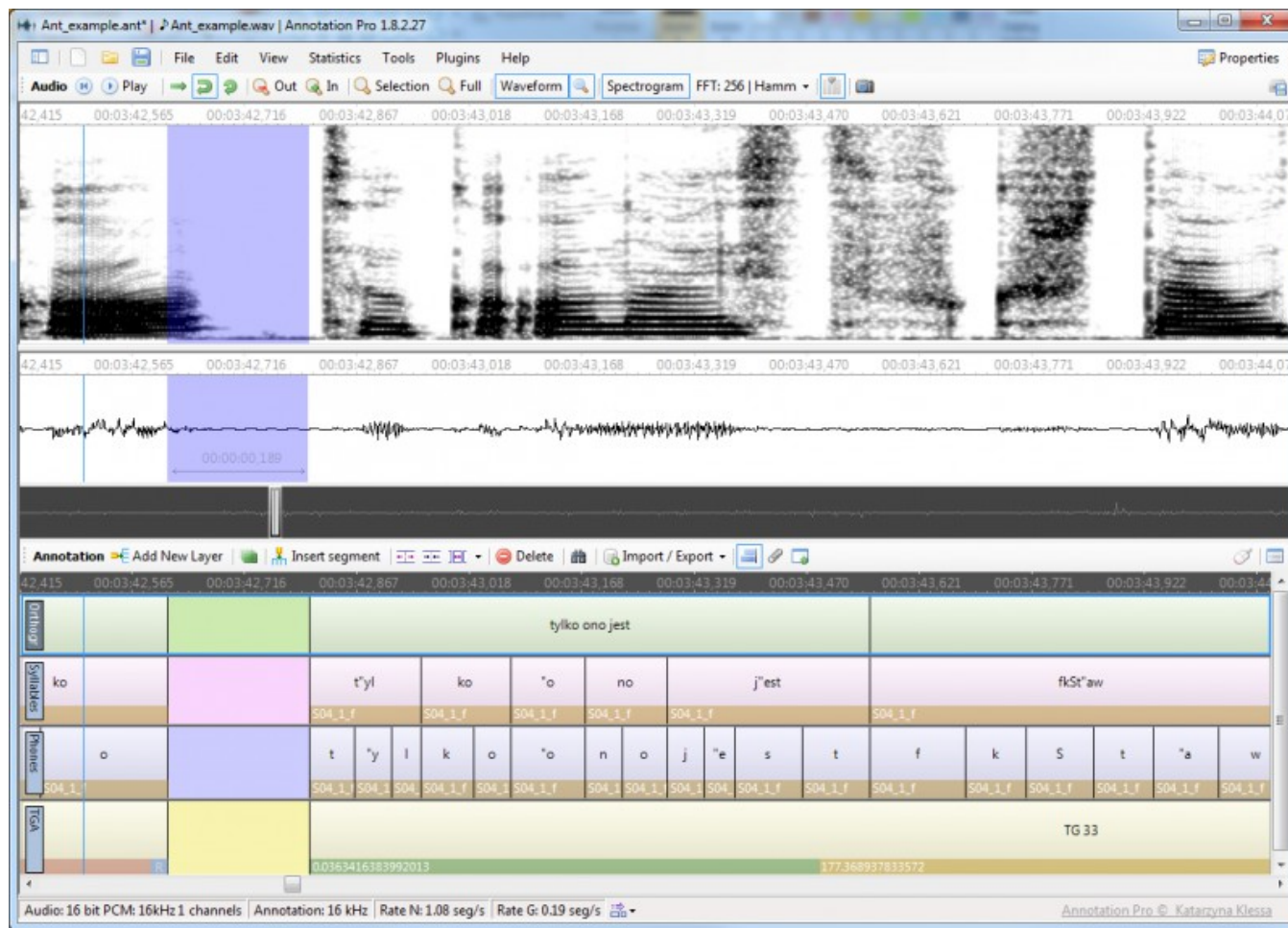
Transkrypcja, segmentacja, anotacja

pe`vnego ,razu || puw`notsni °vjatr i,swontse || spje>tfaliçe || `ktoznix
jest cil,nejfi || `vwaçene pje°xodziw °drogoŵ jaciçe >tfwovjek || ovi`penti
°ftçepwi ,pwaſtf || `umu ,viliçe °vjents || zetenznix °kturi `pjerfi zmuçi
pfexodzon,tsego || abi`zdjow o`kritçe || `bejdze uva°zani začilnej,ſego ||
puw`notsni °vjatr °zatſow od`razu °doptç s`tsawej ,çiwi || aleim`motsnej
>dow || tim`čilnej po°družni o°tulawçe ,fpwaſtf || `vreſtçe puw°notsni
°vjatr daw„spokuj || `ftedi °swontse za`tfewo pfi,gzevaç, || af'xfile
'pužnej po'družni zdjow,pwaſtf || 'ften >sposup || puw`notsni °vjatr `muçaw
>pfiznatç || ze`swontse jestčil°nejſe od„nego

Transkrypcja, **segmentacja**, anotacja



Transkrypcja, segmentacja, anotacja





Transkrypcja, segmentacja, anotacja

Wiele poziomów i rodzajów anotacji. Anotacja tekstu z wykorzystaniem narzędzi komputerowych (nie wymagająca informacji o powiązanym sygnale mowy), np. w znaczeniu tworzenia opisów morfologicznych (*interlinear glossing*); możliwa w Annotation Pro, zob. też program TypeCraft (on-line).

język	przykład	
polski	<i>Jan kupił książki.</i> <i>Czy Jan kupił książki?</i>	
maybrat [29]	<i>ana m-amao Kumurkek a</i> 3pl 3-iść Kumurkek q	'Czy oni idą do Kumurkek??'
mono [30]	<i>Charley=ũa? mia-pł</i> Charley= q iść-perf	'Czy Charley już wyszedł?' (pytanie oznaczamy klityką dołączaną do pierwszego wyrazu)

Przykład tabeli: inne-jezyki.amu.edu.pl

Transkrypcja, segmentacja, anotacja

- Różnice w definiowaniu tych pojęć, wieloznaczność w zależności od dziedziny stosowania

strona główna	interaktyw. mapa	księga wiedzy
dla nauczyciela	co możesz zrobić	o projekcie

- obecne użycie zob. też np. *Księga wiedzy*, rozdz. 10 (languagesindanger.eu), sekcja 'Przetwarzanie i analiza danych'
- kontekst dokumentacji języków:

<http://languagesindanger.eu/lets-revise/>



Elementy projektowania i tworzenia korpusów nagrań mowy



Wybrane etapy prac

Proszę wypełnić ankietę on-line [tutaj](#)



Tworzenie i wykorzystanie zasobów
językowych

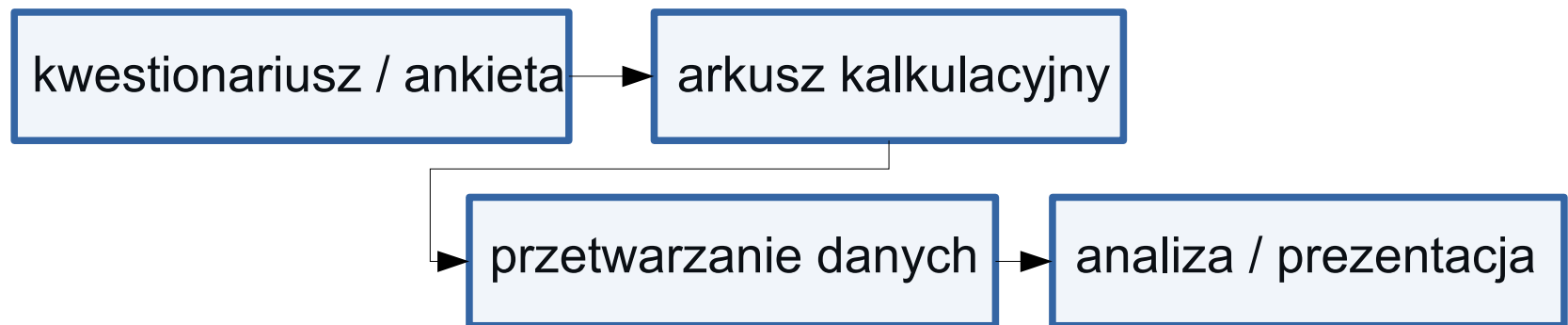
Wybrane etapy prac

- ☐ wyszukiwanie informacji, np. analiza stanu badań, przeszukiwanie tekstów źródłowych
- ☐ projektowanie ankiety osobowej uczestnika nagrań (metadane) ✓
- ☐ projektowanie scenariusza sesji nagraniowej ✓
- ☐ planowanie zasad dostępu do danych, współdzielenia
- ☐ dobór technologii do zarządzania danymi ✓
- ☐ dobór sprzętu i oprogramowania do nagrań ✓
- ☐ zgoda na nagrania - projekt, nadzorowanie, konsultacje prawne ✓
- ☐ rekrutacja mówców
- ☐ zbieranie danych do kwestionariusza osobowego, ankiety ✓
- ☐ prowadzenie nagrań mowy ✓
- ☐ nadzorowanie wypełnienia ankiety osobowej uczestnika nagrań
- ☐ zarządzanie bazą danych, serwerem
- ☐ tworzenie kopii bezpieczeństwa danych
- ☐ transkrypcja fonetyczna nagrań ✓
- ☐ segmentacja nagrań (podział na głoski, sylaby wg czasu występowania w nagraniu, 'time-aligned') ✓
- ☐ zarządzanie anotacją nagrań ✓



Ankieta on-line – zbieranie danych

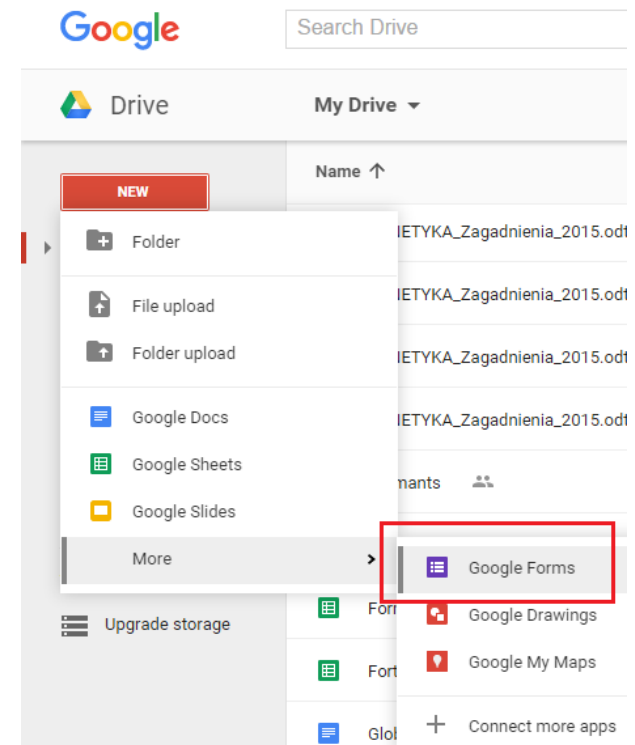
- Ankieta on-line jako jedna z przydatnych technik na etapie zbierania (meta)danych
 - » Wersje papierowe, elektroniczne, on-line
 - » Wady i zalety





Ankieta on-line – zbieranie danych

- Istnieje kilka darmowych narzędzi do tworzenia ankiet, np. Google Forms, ankietka.pl





Zgoda na nagrania

Zgoda na nagrania głosu i ich wykorzystanie - można jej udzielić w formie pisemnej (optymalnie) lub ustnej, jeśli zostanie ona zarejestrowana. Warto sprawdzić aktualne przepisy przed rozpoczęciem nagrań.

Tekst zgody powinien być jasny i wolny od specjalistycznej terminologii. Sporządź tekst tak, żeby podpisując zgodę, uczestnik nagrania:

- potwierdzał, że dobrowolnie zgadza się na udział w sesji nagraniowej;

mógł wyrazić osobną zgodę na:

- nagrywanie we wszystkich trybach (audio / wideo);
- użycie danych do celów badawczych;
- publikację danych;
- archiwizację danych.

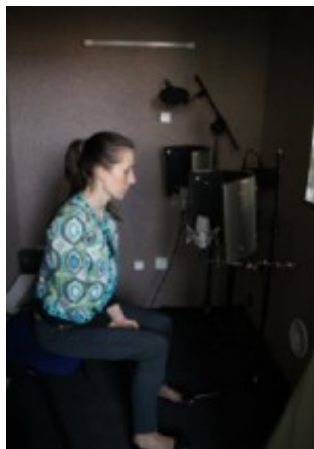
Pomyśl nad stworzeniem dwóch kopii formularza dla siebie i osoby, która ma być nagrywana. Pamiętaj o podpisach!



Scenariusze nagraniowe

- Rejestracja zachowań językowych
- Wiedza o języku, świadomość języka
- Wypowiedzi:
 - Spontaniczne
 - Quasi-spontaniczne, wywoływane, zadaniowe
 - Czytane (teksty ciągłe, jednostki izolowane)
 - Dialogi, monologi

Scenariusze nagraniowe



Zdjęcia:
languagesindanger.eu



Autentyczność danych w korpusie “Paradoks obserwatora” - Observer's Paradox

Celem badań językoznawczych prowadzonych w danej społeczności ma być ustalenie, jak ludzie rozmawiają ze sobą wtedy, gdy nikt ich systematycznie nie obserwuje. Ale takie dane możemy pozyskać tylko przez systematyczną obserwację.

Labov (1972) *Sociolinguistic Patterns*.



Cel: Konstrukcja zbioru bodźców do treningu niemowląt



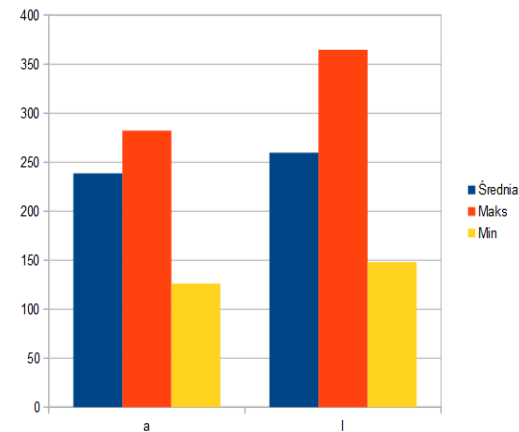


Mowa wywoływana, skierowana do dzieci i dorosłych - IDS vs. ADS korpus NeuroPerKog

Wypowiedzi izolowane (sylaby, krótkie słowa/logatomy) realizowane w sposób neutralny oraz jako skierowane do dzieci - instrukcja + zdjęcie jako sposób wywołania specyficznej realizacji wypowiedzi



Photo: Piotr Ciuchta, freeimages.com





Cel: badanie emocji w głosie





Portrety emocjonalne, aktorskie imitacje emocji: Paralingua Emo

Emotion portrayals, por. GEMEP Geneva corpus (Scherer et al.)

Wypowiedzi neutralne leksykalnie, co najmniej jedna fraza, wyrazy o
wysokiej częstości występowania w codziennym użyciu, wszyscy mówcy –
te same teksty wypowiedzi

Emocje podstawowe i emocje o zróżnicowanym nacechowaniu i stopniu
pobudzenia (*valence, activation*)

Teraz wszystko rozumiem.

Dzisiaj jest poniedziałek.

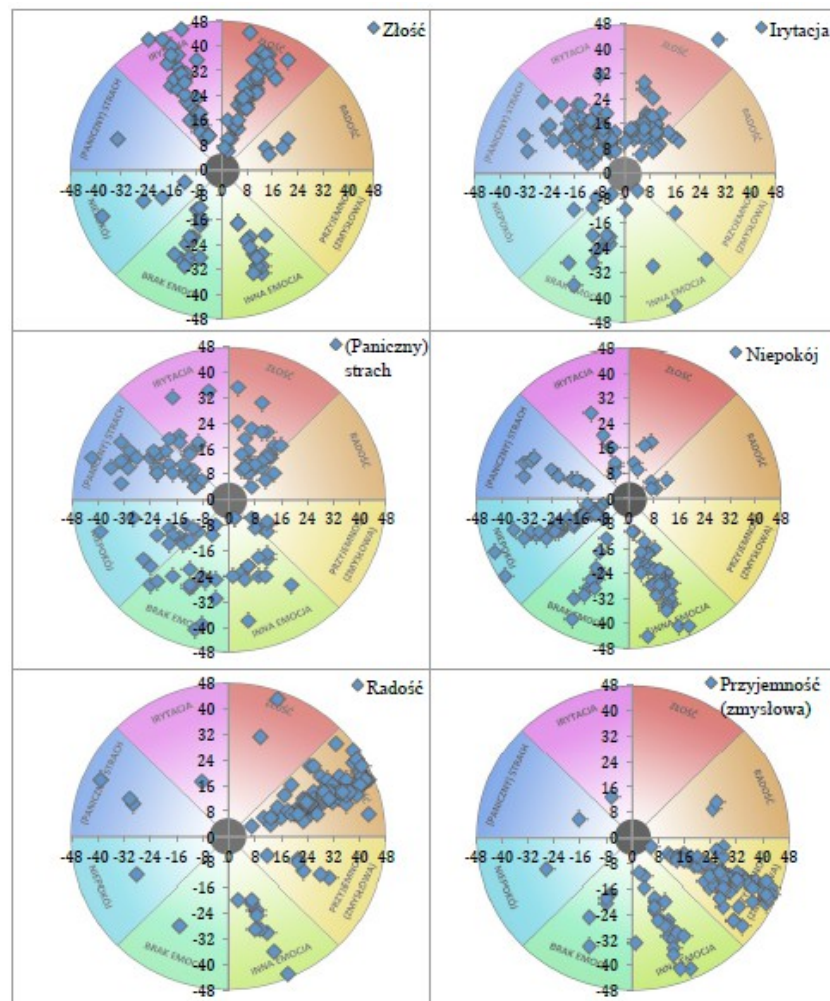
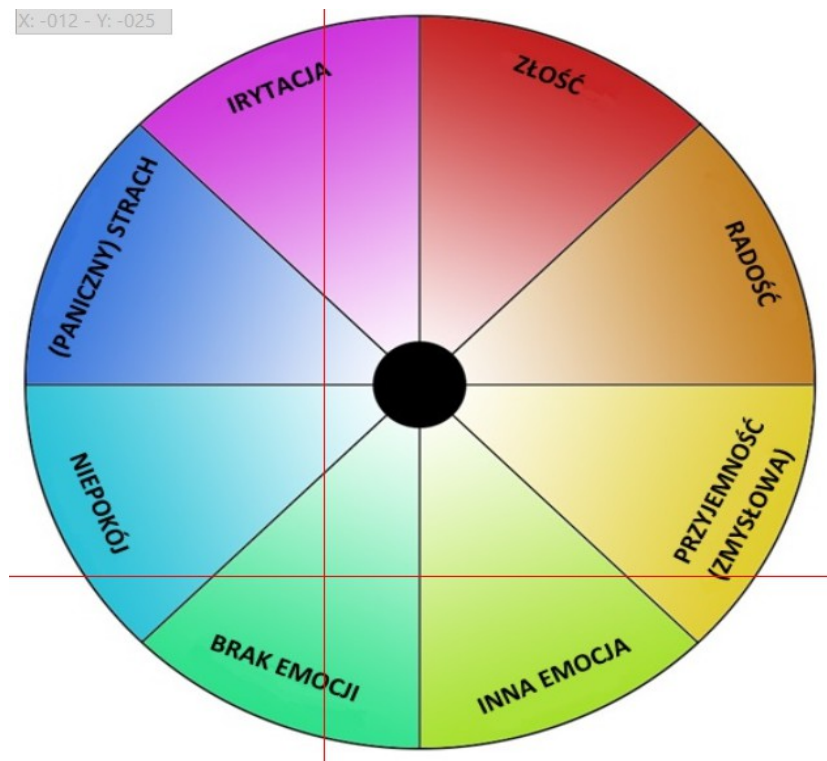
Od rana pada deszcz.

Powiedział, że nie się nie stało.

Jedziemy na wycieczkę do Grecji.

Portrety emocjonalne: Paralingua Emo

X: -012 - Y: -025



z: Głowacka, K. (2016), praca mgr UAM



Cel: cechy mowy spontanicznej





Mowa spontaniczna: baza JST / CREST

Baza mowy spontanicznej i ekspresywnej JST/CREST.

Baza złożona z szeregu podzestawów, z których jeden zebrano tak, że ochotnikom zamontowano małe przenośne urządzenia do nagrywania. Urządzenia rejestrowały spontaniczną mowę tych osób podczas codziennych czynności (w pracy, domu czy szkole) przez stosunkowo długi okres czasu (np. kilka miesięcy).



Mowa spontaniczna: język opiekuńczy - mikrokorpus UAM

4 matki mówiące do swoich dzieci, niemowląt <1 r.ż
Okoliczności: karmienie, przewijanie, kładzenie do snu i in.
codzienne czynności

Zarządzanie? Metadane?



Zdjęcie: poster „Design, structure, and preliminary analyses of a speech corpus of infant directed speech (IDS) and adult directed speech (ADS).” SLE 2015, dzięki uprzejmości M.Oleśkiewicz-Popiel



Mowa spontaniczna a kwestie techniczne, prawne i etyczne

Zarządzanie?
Metadane?
Kwestie etyczne
Ochrona danych osobowych





Cel: rejestracja zachowań językowych w różnych stylach wypowiedzi



[illegible]

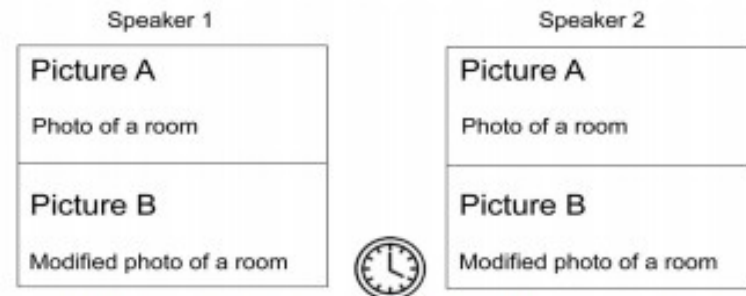


Dialogi zadaniowe: Paralingua Dial

Kilka etapów sesji
nagraniowej – różne
zadania: czytanie tekstu,
4 zadania dialogowe,
ponowne czytanie tekstu
Rozmowa przez telefon
komórkowy, brak
kontaktu wzrokowego
Nagrania za pomocą
kamer, mikrofonów i
telefonów



. Recording scenario Dial 1A session settings (a different picture for each speaker), no time restrictions.

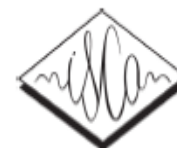


Recording scenario Dial 1B session settings (the same picture for both speakers). Time restrictions applied.



Cel: pozyskanie wypowiedzi swobodne i kontrolowane fraz identycznych pod względem leksykalnym, składniowym

INTERSPEECH 2015



The Acoustics of Word Stress in English as a Function of Stress Level and Speaking Style

Anders Eriksson, Mattias Heldner

Department of Linguistics, Stockholm University, Sweden
anders.eriksson@ling.su.se, mattias.heldner@ling.su.se



Polski Axio-Corpus





Archiwa językowe, dokumentacja języków

- Cele gromadzenia danych na potrzeby archiwizacji i dokumentacji oraz badań naukowych (fonetycznych, psycholingwistycznych, socjologicznych...) mogą być różne, ale wiele metod i technik pokrywa się
- Archiwa językowe np.
 - » <http://www.dialektologia.uw.edu.pl/index.php>
 - » <http://dobes.mpi.nl/>
 - » <https://www.soas.ac.uk/library/archives/>
 - » <http://inne-jezyki.amu.edu.pl/Frontend/>
- Wykorzystanie danych na potrzeby popularyzacji wiedzy:
<http://pl.languagesindanger.eu/map-html/pl/>



Reprezentatywność korpusów

Reprezentatywność – zakres w jakim korpus informuje o zjawiskach zachodzących w całej populacji; w przypadku korpusów językowych – w języku lub jego specjalizowanym podzbiorze

- **Odróżnia korpusy od archiwów językowych** *(jednak por. **korpusy** mowy spontanicznej, jak określić ich reprezentatywność?)
- **Możliwość generalizacji, wyciągania uogólnionych wniosków**



Korpusy ogólnego przeznaczenia vs. specjalizowane

Specjalizowane – selekcja materiału dostosowana do określonych potrzeb – a zatem łatwiejsza

- wnioski wyciągane na podstawie danych mogą tylko dotyczyć tego wybranego wycinka języka, którego dotyczy specjalizacja korpusu

Ogólnego przeznaczenia – z założenia dotyczą większości zjawisk w języku, powinny zakładać szczegółowy plan struktury korpusu, aby umożliwić jego późniejsze zrównoważenie, różne typy tekstów, stylów wypowiedzi, zróżnicowani mówcy

- Możliwość większego uogólnienia wniosków wysnuwanych na podstawie danych



Korpusy statyczne vs. dynamiczne

Statyczne – zbiór danych językowych dający informację na temat elementów systemu językowego w określonym momencie

- łatwiej kontrolować zrównoważenie zawartości korpusu
- użyteczność może być bardziej ograniczona (dezaktualizacja)

Dynamiczne

- umożliwiają obserwację zmienności zjawisk językowych w czasie
- trudniejsze w utrzymaniu, droższe
- trudniejsza kontrola zrównoważenia i jakości danych



Użytkownicy korpusów językowych

Jedna lub więcej grup docelowych

Grupa docelowa a sposób udostępniania prezentacji danych

bazy przeznaczone do badań naukowych nad sygnałem mowy,
badania podstawowe a stosowane
archiwa językowe dla specjalistów i popularyzatorskie
bazy dla technologii: synteza, rozpoznawanie mowy i mówcy

specjaliści z dziedziny technologii mowy

akustycy

informatycy

lekarze

terapeuci

językoznawcy

socjologowie

psycholingwiści

fonetycy

historycy

uczniowie i studenci

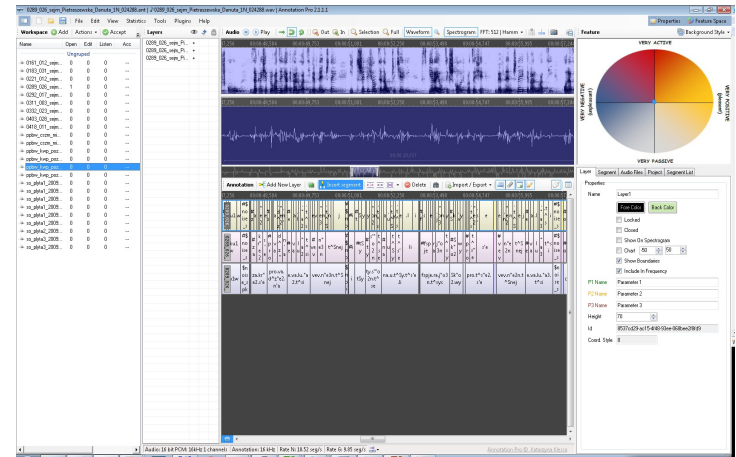
“zwykli” użytkownicy języka

osoby zainteresowane rewitalizacją języków



Systematyzacja danych w korpusach - wybór technologii

Nazwa	Data modyfikacji	Typ	Rozmiar
20150629_01_HU_ILKU_KAKL_0065_SYLL_CV_iloczas_nDA.ant	2016-01-29 13:12	Annotation Pro File	2 KB
20150629_01_HU_ILKU_KAKL_0065_SYLL_CV_iloczas_nDA.wav	2016-01-29 13:12	WAV	259 KB
20150629_01_HU_ILKU_KAKL_0065_SYLL_CV_iloczas_nDA_EEG.wav	2016-01-29 14:12	WAV	76 KB
20150629_01_HU_ILKU_KAKL_0066_SYLL_CV_iloczas_fDO.ant	2016-01-29 13:12	Annotation Pro File	2 KB
20150629_01_HU_ILKU_KAKL_0066_SYLL_CV_iloczas_fDO.wav	2016-01-29 13:12	WAV	259 KB
20150629_01_HU_ILKU_KAKL_0066_SYLL_CV_iloczas_fDO_EEG.wav	2016-01-29 14:12	WAV	90 KB
20150629_01_HU_ILKU_KAKL_0067_SYLL_CV_iloczas_da.ant	2016-01-29 13:12	Annotation Pro File	2 KB
20150629_01_HU_ILKU_KAKL_0067_SYLL_CV_iloczas_da_EEG.wav	2016-01-29 13:12	WAV	259 KB
20150629_01_HU_ILKU_KAKL_0067_SYLL_CV_iloczas_da_EEG.wav	2016-01-29 14:11	WAV	73 KB
20150629_01_HU_ILKU_KAKL_0068_SYLL_CV_iloczas_ta.ant	2016-01-29 13:12	Annotation Pro File	2 KB
20150629_01_HU_ILKU_KAKL_0068_SYLL_CV_iloczas_ta_EEG.wav	2016-01-29 14:11	WAV	63 KB
20150629_01_HU_ILKU_KAKL_0069_SYLL_CV_iloczas_vDO.ant	2016-01-29 13:12	Annotation Pro File	2 KB
20150629_01_HU_ILKU_KAKL_0069_SYLL_CV_iloczas_vDO.wav	2016-01-29 13:12	WAV	259 KB
20150629_01_HU_ILKU_KAKL_0069_SYLL_CV_iloczas_vDO_EEG.wav	2016-01-29 14:10	WAV	81 KB
20150629_01_HU_ILKU_KAKL_0070_SYLL_CV_iloczas_na.ant	2016-01-29 13:12	Annotation Pro File	2 KB
20150629_01_HU_ILKU_KAKL_0070_SYLL_CV_iloczas_na.wav	2016-01-29 13:12	WAV	259 KB
20150629_01_HU_ILKU_KAKL_0070_SYLL_CV_iloczas_na_EEG.wav	2016-01-29 14:10	WAV	72 KB
20150629_01_HU_ILKU_KAKL_0071_SYLL_CV_iloczas_thDA.ant	2016-01-29 13:12	Annotation Pro File	2 KB



Workspace-Export-20160201 172236.xlsx - LibreOffice Calc												
File	FilePart1	FilePart2	FilePart3	Layer	Start	StartInMilli	Duration	DurationInMilliseconds	Param1	Param2	Param3	Label
Asia_2_3362_0s_A.ant	Asia	2	3362	Phones	6228	141	1323	30				p
Asia_2_3362_0s_A.ant	Asia	2	3362	Phones	7551	171	1323	30				r
Asia_2_3362_0s_A.ant	Asia	2	3362	Phones	8874	201	1323	30				o
Asia_2_3362_0s_A.ant	Asia	2	3362	Phones	10197	231	1764	40				S
Asia_2_3362_0s_A.ant	Asia	2	3362	Phones	11961	271	3528	80				e
Asia_2_3362_0s_A.ant	Asia	2	3362	Phones	15489	351	1323	30				p
Asia_2_3362_0s_A.ant	Asia	2	3362	Phones	16812	381	3528	80				a
Asia_2_3362_0s_A.ant	Asia	2	3362	Phones	20340	461	1764	80				n
Asia_2_3362_0s_A.ant	Asia	2	3362	Phones	22104	501	3528	80				a



Systematyzacja danych w korpusach - wybór technologii

Wybór technologii – zależny od rodzaju i przeznaczenia korpusu

Optymalnie jest, gdy sposób systematyzacji danych dobrany jest już na etapie wstępnego projektowania tworzonego zasobu.

Różne sposoby systematyzacji korpusów językowych: np.

- bazy plikowe
- arkusze kalkulacyjne
- relacyjne bazy danych

» **Relacyjna baza danych** – zbiór danych zorganizowanych w postaci tabel powiązanych różnego rodzaju relacjami



Czy zbiór plików umieszczonych w folderze to baza danych?

Jeśli pojęcie 'baza danych' to dla nas po prostu jakikolwiek zbiór danych, to na powyższe pytanie możemy odpowiedzieć – 'tak'

- Ograniczenia - brak powiązań między poszczególnymi plikami, trudności w sortowaniu ze względu na różne kryteria, trudności we współpracy, gdy dane są edytowane przez różne osoby (kto edytuje, który plik, kiedy, co zostało zmienione?)



Czy zbiór plików umieszczonych w folderze to baza danych?

- Systemy wspomagające pracę na zbiorach plików, np. Git, <http://git-scm.com/>; darmowy hosting dla Git: <https://bitbucket.org/> albo <https://github.com/>
- Opcje wbudowane w programy pracujące na zbiorach plików - *Annotation Pro* - opcje dotyczące kolekcji plików (workspace, .antw) ułatwiają pracę na zbiorach plików i operacje zbiorowe na nich, a także - w ograniczony sposób - połączeń/relacji między nimi



Czy arkusz kalkulacyjny to baza danych?

- Podobnie jak zbiór plików w folderze, arkusz kalkulacyjny może pełnić funkcję dość prostej bazy danych
- W arkuszach kalkulacyjnych możemy do pewnego stopnia i stosunkowo łatwo kontrolować dane (w tym przypadku zawartość komórek), możemy m.in.: napisać formuły, stworzyć makro przetwarzające dane, tworzyć tabele przestawne itp. Nadal jednak ta forma przechowywania danych nie daje możliwości pełnego wykorzystania powiązań między danymi, ani jednoczesnej pracy wielu osób na tych samych danych z zachowaniem historii zmian, przypisywaniem uprawnień do edycji itp.



Warto - nie warto - trzeba?

Czy / kiedy warto szukać czegoś więcej niż zwykły zbiór plików czy arkusz kalkulacyjny?

- kiedy mamy 1 osobę pracującą nad 1 tekstem?
- kiedy mamy 1 osobę i wiele danych?
- kiedy mamy dużo osób i średnią ilość danych?
- kiedy mamy dużo osób i bardzo dużo danych?
- kiedy różne osoby pracują z różnych miejsc, zdalnie?



Warto - nie warto - trzeba?

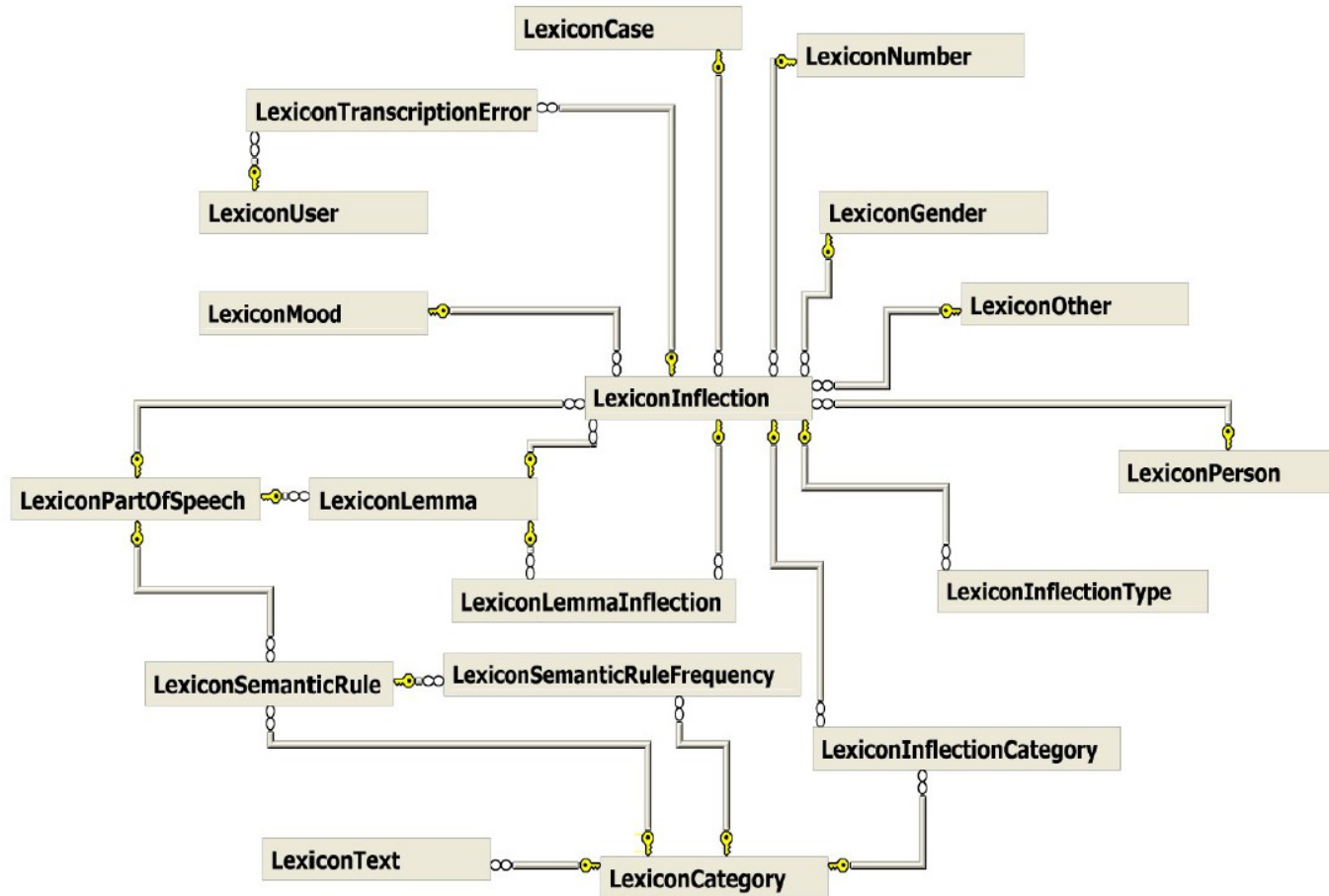
Czy / kiedy warto szukać czegoś więcej niż zwykły zbiór plików czy arkusz kalkulacyjny?

- kiedy mamy 1 osobę pracującą nad 1 tekstem - **nie warto**
- kiedy mamy 1 osobę i wiele danych - **warto**
- kiedy mamy dużo osób i średnią ilość danych - **warto**
- kiedy mamy dużo osób i bardzo dużo danych - **trzeba**
- kiedy różne osoby pracują z różnych miejsc, zdalnie - **warto**
czasem **trzeba** - zależnie od ilości danych (np. za stronę internetową z danymi stoi zwykle baza danych)



Relacyjna baza danych

- i relacyjny model rzeczywistości

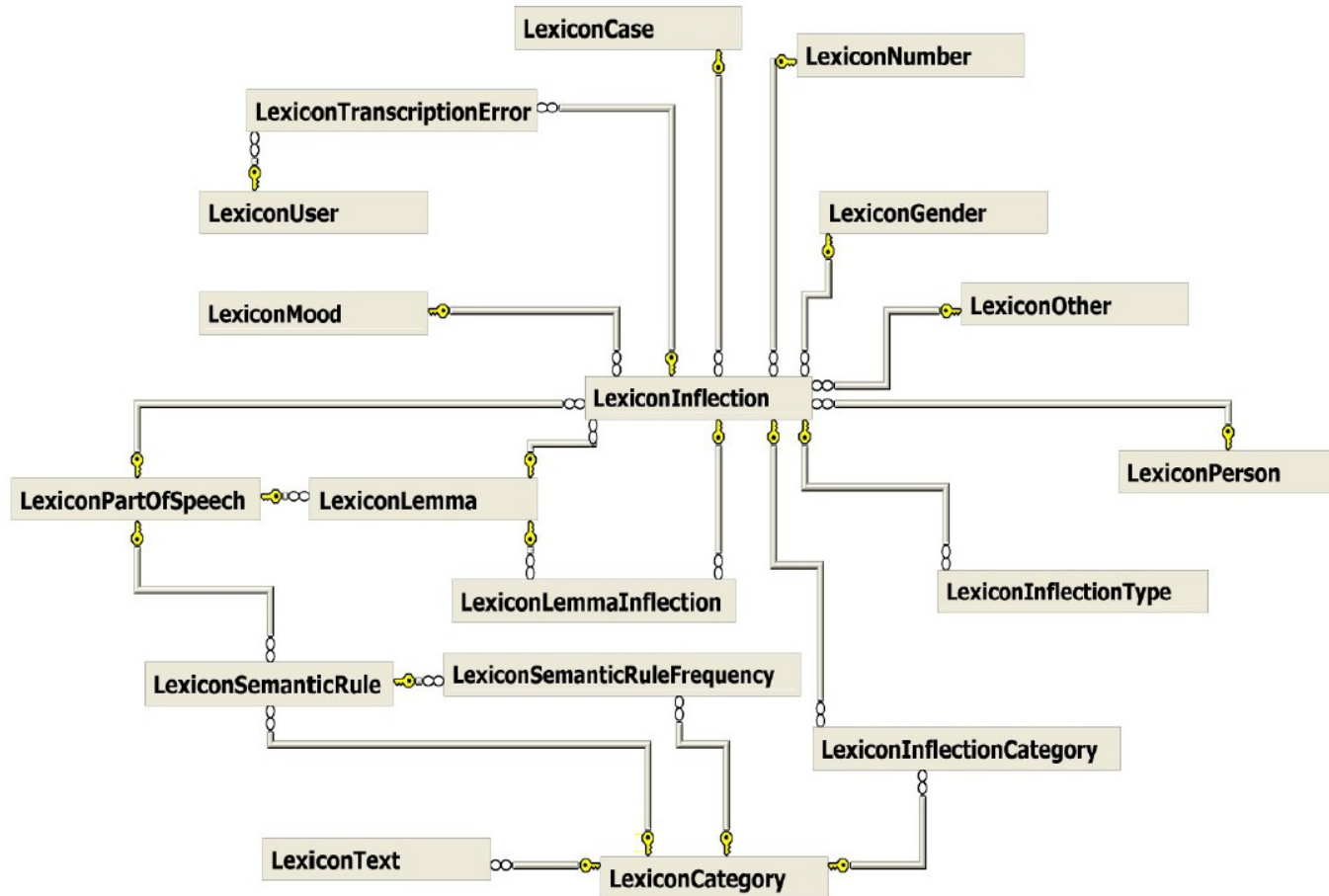


Relacyjna baza danych oparta jest na tzw relacyjnym modelu rzeczywistości: ma odzwierciedlać pewien fragment rzeczywistości, jej elementów oraz powiązań między nimi - zob. np. strukturę bazy - słownika obok.



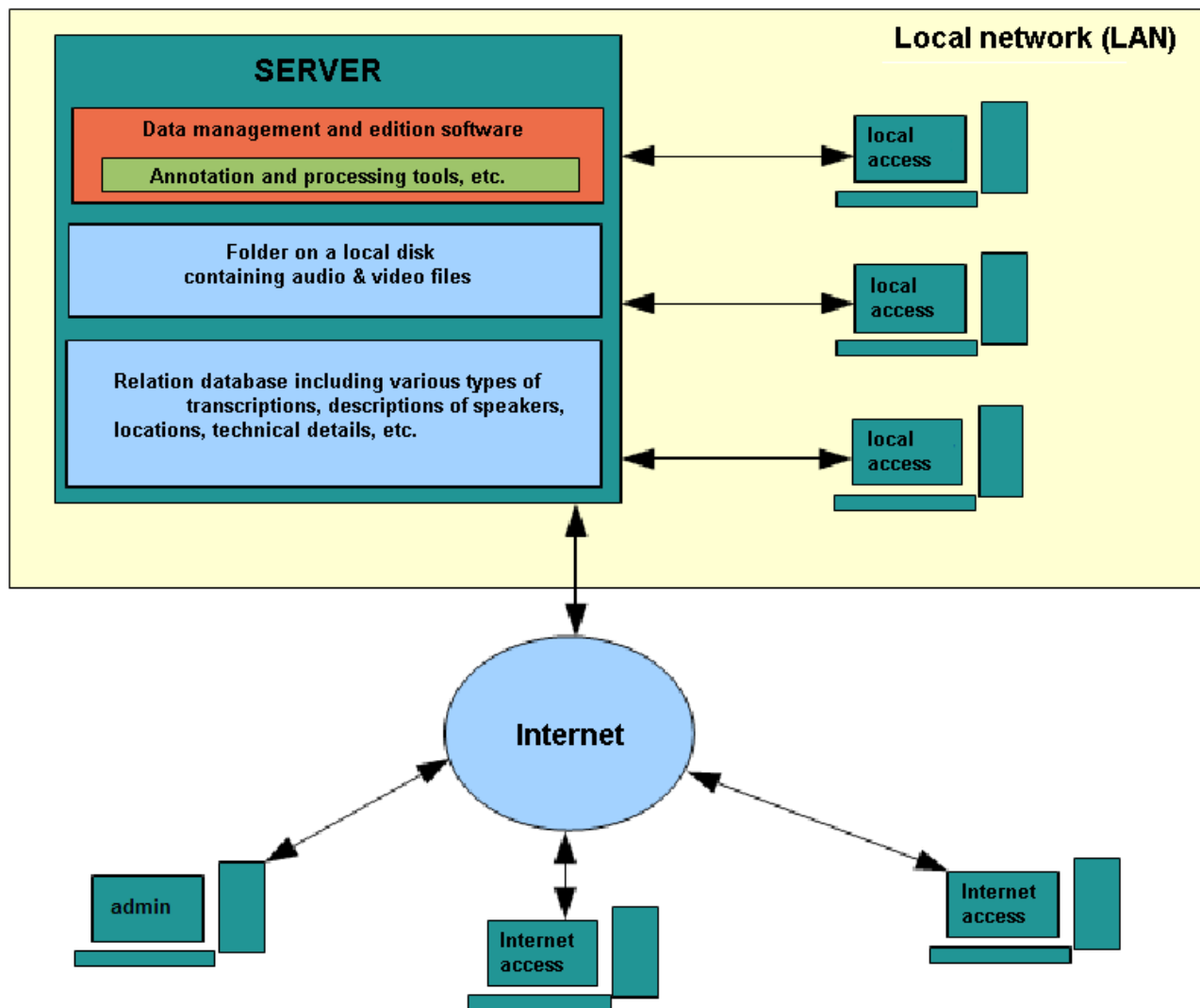
Relacyjna baza danych

- i relacyjny model rzeczywistości



Bazy relacyjne składają się z wielu powiązanych ze sobą tabel.

Jak to działa? - architektura klient - serwer





Jak nauczyć się rozmawiać z bazą danych?

1. Przede wszystkim **trzeba ją mieć** – instalacja *(narzędzia, np. MySQL).
2. Trzeba poznać **język obcy: SQL**
3. Z bazą relacyjną można komunikować się **różnymi kanałami:**
 - bezpośrednio zapytania do bazy (w języku SQL)
 - za pośrednictwem program(ów) do zarządzania / edycji danych (które wykonają polecenia SQL w naszym imieniu)
 - lokalnie lub przez Internet

Elementy SQL



SQL - nomenklatura

Jak mówi staropolskie przysłowie, w bazach danych nie używamy znaków diakrytycznych, np. polskich ó, ż i in., ponieważ **piszemy po angielsku...**

Używamy tylko znaków alfanumerycznych, czyli liter i cyfr.



SQL - nomenklatura MySQL

W MySQL utarł się zwyczaj pisania nazw małą literą i oddzielania słów podkreślnikiem. Tak więc tabela zawierająca studentów nazywa się `student`. Tabela zawierająca oceny studentów nazywa się `student_note`. Klucz główny piszemy dodając do nazwy tabeli `id`, czyli dla tabeli `student` będzie to `student_id` (niektóre środowiska wymagają, aby klucz zawsze nazywał się tylko `id`).



Lista słówek - do zapamiętania na początek

Z podstaw SQL:

CREATE - Tworzenie tabeli, czyli *Coś czego zwykle nie będę musiał/a robić ręcznie*

INSERT - Dodanie czegoś do tabeli

SELECT - Wyświetlenie, wybieranie danych

UPDATE - Zmodyfikowanie czegoś w tabeli

DELETE - Usunięcie z tabeli, czyli *Ratunku, nie ma cofnij!*

DROP - Usunięcie tabeli, czyli *Ratunku, wyrzuciłem/am wszystko!*

Kopie bezpieczeństwa



Lista słówek - do zapamiętania na początek

Z podstaw SQL:

CREATE - Tworzenie tabeli, czyli *Coś czego zwykle nie będę musiał/a robić ręcznie*

INSERT - Dodanie czegoś do tabeli

SELECT - Wyświetlenie, wybieralnie danych

UPDATE - Zmodyfikowanie czegoś w tabeli

DELETE - Usunięcie z tabeli, czyli *Ratunku, nie ma cofnij!*

DROP - Usunięcie tabeli, czyli *Ratunku, wyrzuciłem/am wszystko!*

Kopie bezpieczeństwa



MySQL, phpMyAdmin

<https://www.phpmyadmin.net/>

Dane do logowania na zajęcia:

<http://warsztaty2016.klessa.pl/>

Użytkownik: 01686302_0000005



Witamy w phpMyAdmin

Język - Language

Polski - Polish ▼

Login

Użytkownik:

Hasło:

Wykonaj



MySQL, phpMyAdmin

phpMyAdmin

Ostatnie Ulubione

01686302_0000005

- Nowy
- 2016
- 2017 F1-F2
- temp

information_schema

Serwer: localhost » Baza danych: 01686302_0000005 » Tabela: 2017 F1-F2

Przeglądaj Struktura SQL Szukaj Wstaw Eksport Import Więcej

⚠ Current selection does not contain a unique column. Grid edit, checkbox, Edit, Copy and Delete features are not available.

✓ Pokazano wiersze 0 - 24 (28172 ogółem, Wykonanie zapytania trwało 0.0007 sekund(y).)

`SELECT * FROM `2017 F1-F2``

☐ Profilowanie [[Edit inline](#)] [[Edytuj](#)] [[Wyjaśnij SQL](#)] [[Utwórz kod PHP](#)] [[Odśwież](#)]

1 > >> Liczba wierszy: 25 Filtrowanie wierszy: Szukaj w tej tabeli

+ Opcje

A	B	C	D	E	F	G	H	I	J	K	L
20141216_10_PL_ANSU_AGCZ_bajka_ET.ant	IDS	ANSU	story	PL	phones_f1	585284	36580	7360	460	"a	246
20141216_10_PL_ANSU_AGCZ_bajka_ET.ant	IDS	ANSU	story	PL	phones_f1	2756378	172273	2080	130	"a	999



MySQL, phpMyAdmin - dostosowanie struktury tabeli

← → ↻ warsztaty2016.klessa.pl/tbl_structure.php?db=01686302_0000005&table=2016&token=6fa16a2dd151c89a8235460439a656fc

Aplikacje St BB JIRA TB docs npk Transl y nowa Rozdzielczość bitowa amu IS17 GESP Gran r SPPAS Sch DB Txt

phpMyAdmin

Ostatnie Ulubione

01686302_0000005

- Nowy
- 2016
- 2017 F1-F2
- temp
- information_schema

Server: localhost » Baza danych: 01686302_0000005 » Tabela: 2016

Przeglądaj Struktura SQL Szukaj Wstaw Eksport Import Operacje Wywołaj

#	Nazwa	Typ	Metoda porównywania napisów	Atrybuty	Null	Ustawienia domyślne	Dodatkowo	Działanie
☐	1 id	int(11)			Nie	Brak	AUTO_INCREMENT	Zmień
☐	2 file_name	varchar(255)	latin2_general_ci		Nie	Brak		Zmień
☐	3 file_part1	varchar(100)	latin2_general_ci		Nie	Brak		Zmień
☐	4 file_part2	varchar(100)	latin2_general_ci		Nie	Brak		Zmień
☐	5 file_part3	varchar(100)	latin2_general_ci		Nie	Brak		Zmień
☐	6 layer	varchar(100)	latin2_general_ci		Nie	Brak		Zmień
☐	7 start_samples	decimal(10,2)			Nie	Brak		Zmień
☐	8 start_ms	decimal(10,2)			Nie	Brak		Zmień
☐	9 duration_samples	varchar(500)	latin2_general_ci		Nie	Brak		Zmień
☐	10 duration_ms	varchar(500)	latin2_general_ci		Nie	Brak		Zmień
☐	11 parameter1	varchar(500)	latin2_general_ci		Nie	Brak		Zmień
☐	12 parameter2	varchar(500)	latin2_general_ci		Nie	Brak		Zmień
☐	13 parameter3	varchar(500)	latin2_general_ci		Nie	Brak		Zmień



Zapytanie SELECT - wstęp

Zapytanie SELECT zawiera często elementy wymienione niżej. “Prosimy” bazę danych o informacje, mówiąc jej:

1. **Które kolumny** chcemy zobaczyć?
2. Z jakiej **tabeli**?
3. Które **wiersze**?
4. Czy **pogrupować** dane?
5. **Posortujmy** alfabetycznie!



Zapytanie SELECT - składnia

Składnia:

```
SELECT kolumny_do_pokazania as nowa_nazwa_kolumny  
from tabela as nazwa_tabeli  
where warunki_filtrowania  
group by kolumny po_których_grupować  
order by kolumny po_których_sortować  
having warunki_filtrowania
```

*(pola wyszarzone są opcjonalne)



Zapytanie SELECT - składnia

```
SELECT kolumny_do_pokazania as nowa_nazwa_kolumny  
from tabela as nazwa_tabeli  
where warunki_filtrowania  
group by kolumny_po_których_grupować  
order by kolumny_po_których_sortować  
having warunki_filtrowania
```

W powyższym zapytaniu polecenia języka SQL to **SELECT, FROM, AS, GROUP BY, ORDER BY**. “Kolumny do pokazania” należy wymienić jako listę kolumn, gdzie każda kolumna jest oddzielona przecinkiem, np.: kolumna1, kolumna2, kolumna3. “tabela” oznacza nazwę tabeli (lub kilku tabel), z których chcemy czytać dane.



Zapytanie SELECT - lista kolumn. Przykład 0. - podobno prościej się nie da:)

```
SELECT * FROM `files` WHERE 1
```

Wynik: wszystkie wiersze, wszystkie kolumny z tabeli 'files' (wher)

		id	file_name	file_part1	file_part2	file_part3	layer	start_samples	start_ms	duration_samples	duration_ms	parameter1	parameter2	parameter3	label
☐	Edytuj	Kopiuj	Usuń	4	Set2_SCV_mono_p...	Set2	SCV	mono	sounds	278608.00	17413.00	480	30		#S
☐	Edytuj	Kopiuj	Usuń	8	Set2_SCV_mono_p...	Set2	SCV	mono	sounds	278608.00	17413.00	480	30		#S
☐	Edytuj	Kopiuj	Usuń	9	Set2_SCV_mono_p...	Set2	SCV	mono	sounds	279088.00	17443.00	640	40		"y2



Zapytanie SELECT - lista kolumn. Przykład 1. - wybór jednej kolumny

```
SELECT label FROM `files` WHERE 1
```

Zapytanie wybiera kolumnę *label* z tabeli *files*.

UWAGA:

W naszym przypadku nazwa tabeli to 2017.

Wynik: lista wszystkich etykiet (z powtórzeniami)

label

#S

#S

"y2

.j

a

#s

y

.m



Zapytanie SELECT - lista kolumn. Przykład 2. - wybór kilku kolumn

```
SELECT label, duration_ms FROM `files` WHERE 1
```

Wynik: lista etykiet (*label*) i ich czasów trwania (w ms)

label	duration_ms
#S	30
#S	30
"y2	40
.j	40
a	115
#s	79
y	121
.m	47
e	125
.t	49



Zapytanie SELECT - lista kolumn. Przykład 3. - poglądowe ograniczenie wyników

```
SELECT * FROM files LIMIT 1
```

Wynik: zapytanie zwraca ograniczoną liczbę wierszy, tyle ile wpisujemy po LIMIT (tutaj 1)



Zapytanie SELECT – wybieramy unikalne wiersze

Przykład 4. - DISTINCT

```
SELECT distinct layer from files
```

Kolumna utworzona z połączenia dwóch kolumn ma teraz nową nazwę. Używamy do tego słowa kluczowego **AS** (od: 'alias') zaraz za taką kolumną i potem piszemy nową nazwę kolumny.

Wynik:

layer
sounds
syllables
words
comments
Quadrants ZScores
Quadrants Chart
Quadrants ZScores 2



Zapytanie SELECT – lista kolumn

Przykład 5. - nadajemy nazwę wybranej kolumnie

```
SELECT DISTINCT label as 'Etykieta' FROM `2017`
```

Kolumna wyświetlona na wydruku ma teraz nową nazwę. Używamy do tego słowa kluczowego **AS** (od: 'alias'), które wpisujemy zaraz za nazwą kolumny, a następnie w apostrofach podajemy nazwę nowej kolumny. Przydaje się to np. gdy potrzebne są wydruki demonstracyjne opisane w innych językach niż angielski, który zastosowaliśmy w nazewnictwie kolumn w bazie.

Wynik:

Etykieta
"a
t
n'
a
u
n



Zapytanie SELECT - lista kolumn. Przykład 6. - połączenie dwóch kolumn

```
.....  
SELECT concat(file_part3,' ',layer) FROM `files`
```

```
.....  
SELECT concat_ws(' ',file_part3,layer) FROM `files`
```

Pobieramy z tabeli `files` kolumny file_part3 i layer.
Dodatkowo pomiędzy nimi dodajemy ciąg znaków
(string) ze spacją.

*(w MySQL można to zrobić na 2 sposoby: za pomocą concat i concat_ws. Działanie w obu przypadkach jest podobne, z niewielką różnicą, zobacz np. tutaj:
<http://stackoverflow.com/questions/8809037/which-to-use-concat-concat-ws-in-mysql>).

Dzięki temu w wyniku mamy jedną nową kolumnę,
zamiast dwóch.

```
concat(file_part3,' ',layer)
```

mono sounds

mono sounds

mono sounds

mono sounds

mono sounds

mono sounds

mono sounds

mono sounds

mono sounds

mono sounds

mono sounds

mono sounds

mono sounds

mono sounds



Zapytanie SELECT - lista kolumn.

Przykład 7. - nadajemy nazwę połączonej kolumnie

```
select concat(layer, ' ', file_part3) as moja_nazwa from files
```

Kolumna utworzona z połączenia dwóch kolumn ma teraz nową nazwę. Używamy do tego słowa kluczowego **AS** (od: 'alias'), które wpisujemy po nazwie kolumny, a zaraz potem piszemy nową nazwę kolumny.

Wynik:

moja_nazwa
sounds mono
sounds mono
sounds mono
sounds mono
sounds mono
sounds mono
sounds mono
sounds mono
sounds mono
sounds mono
sounds mono



Zapytanie SELECT - lista kolumn. Przykład 8. - funkcje matematyczne

Na liście kolumn można używać funkcji, które umożliwiają nam manipulację zawartością kolumny. Na Funkcje matematyczne pozwalają na dokonanie obliczeń na podstawie zawartości kolumn zawierających dane liczbowe.



Zapytanie SELECT - lista kolumn. Przykład 8. - funkcje matematyczne: Avg

```
SELECT AVG(formant_mean) FROM `2017` WHERE label ='a'
```

Zapytanie zwraca średnią wartość z kolumny
formant_mean dla wierszy z etykietą transkrypcji /a/.
Wynik:

```
AVG(formant_mean)
```

```
1157.109018227489300
```

A co zwróci nam poniższe zapytanie?

```
SELECT label, AVG(formant_mean) FROM `2017` GROUP by `label`
```





Zapytanie SELECT - lista kolumn. Przykład 8. - funkcje matematyczne: **Min**

```
SELECT label, MIN(formant_mean) FROM `2017` WHERE label = 'i'
```

Minimalna wartość liczbową w wybranej kolumnie.

Wynik:

+ Opcje

label	MIN(formant_mean)
i	219.30044546429



Zapytanie SELECT - lista kolumn.

Przykład 8. - funkcje matematyczne: **Max**

```
SELECT label, MAX(formant_mean) FROM `2017` WHERE label = 'i'
```

Minimalna wartość liczbową w wybranej kolumnie.
Wynik:

+ Opcje

label	MAX(formant_mean)
i	2654.000000000000



Zapytanie SELECT - lista kolumn. Przykład 9. - funkcje tekstowe

Funkcje tekstowe są kolejnym przykładem funkcji szczególnych przydatnych w obróbce danych tekstowych w językowych bazach danych. Przykłady takich funkcji to:

LEN, LEFT, RIGHT, REVERSE, LOWER, UPPER,
TRIM, REPLACE



Zapytanie SELECT - lista kolumn. Przykład 9. - funkcje tekstowe: **LENGTH**

```
select length(file_name) from files
```

LENGTH - liczba znaków zawartości komórki.

Wynik:

length(file_name)
54
54
54
54
54
54
54
54



Zapytanie SELECT - lista kolumn.

Przykład 9. - funkcje tekstowe: **LEFT**, **RIGHT**

```
select left(file_name, 4), right(file_name, 4) from files
```

LEFT, RIGHT - zwraca określoną liczbę znaków licząc od lewej lub od prawej, czyli np. 3 pierwsze literki lub 3 ostatnie.

Wynik:

left(file_name, 4)	right(file_name, 4)
Set2	.ant
Set2	.ant
Set2	.ant
Set2	.ant
Set2	.ant



Zapytanie SELECT - lista kolumn.

Przykład 9. - funkcje tekstowe: **REVERSE**

```
select label, reverse(label) from files
```

REVERSE - odwraca kolejność liter w tekście w komórce tabeli.

- Pytanie: do czego może być przydatna ta funkcja?

Wynik:

label	reverse(label)
#S	S#
#S	S#
"y2	2y"
.j	j.
a	a
#s	s#



Zapytanie SELECT - lista kolumn.

Przykład 9. - funkcje tekstowe: **LOWER**, **UPPER**

```
select lower(layer), upper(layer) from files
```

Lower, Upper - zamieniają wszystkie litery w wyniku selekcji na małe albo na wielkie.

Wynik:

<code>lower(layer)</code>	<code>upper(layer)</code>
sounds	SOUNDS
sounds	SOUNDS
sounds	SOUNDS
sounds	SOUNDS



Zapytanie SELECT - lista kolumn. Przykład 9. - funkcja tekstowa: **TRIM**

```
select trim(layer) from files
```

TRIM – usuwa zbędne spacje przed/po tekście.

Wynik:

*(w naszym konkretnym przykładzie nic się nie zmienia, ale funkcja ta może być bardzo przydatna, np do 'czyszczenia' danych)

trim(layer)

sounds

sounds

sounds

sounds



Zapytanie SELECT - lista kolumn.

Przykład 9. - funkcje tekstowe: **REPLACE**

```
select distinct layer, replace(layer, 'sounds', 'sound') from files
```

REPLACE zamienia fragment zawartości kolumny na inny tekst, czyli może też zamienić wybraną literę inną literą, albo usunąć wybraną literę lub ciąg liter (znaków).

Wynik:

layer	replace(layer, 'sounds', 'sound')
sounds	sound
syllables	syllables
words	words
comments	comments
Quadrants ZScores	Quadrants ZScores
Quadrants Chart	Quadrants Chart
Quadrants ZScores 2	Quadrants ZScores 2



Linki

Wybrane linki (niektóre ściśle, a inne luźniej powiązane z tematyką zajęć:)

- <http://www.w3schools.com/sql/>
- <https://www.phpmyadmin.net/>
- <http://dev.mysql.com/doc/refman/5.0/en/string-functions.html>
- <http://pl.languagesindanger.eu/book-of-knowledge/language-documentation/>
- <http://www.gwarypolskie.uw.edu.pl/>
- www.inne-jezyki.amu.edu.pl
- www.annotationpro.org
- <http://stackoverflow.com/>
- <http://sldr.org/>
- <http://nkjp.pl/>
- <http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=intro>
- <http://phoible.org/>
- <http://wals.info/>
- <http://glottolog.org/>
- <http://www.buckeyecorpus.osu.edu/>



Dziękuję za uwagę!

klessa@amu.edu.pl
annotationpro.org
katarzyna.klessa.pl