Tekstowe bazy danych

Jakub Wilk

Wydział Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego

12 kwietnia 2007 r.

Definition

- **Tekst** to komputerowy zapis symboli języka naturalnego.
- Baza danych jest **tekstowa** gdy:
  - przechowuje głównie tekst oraz
  - tekst jest głównym przedmiotem zapytań.

Tekstowe bazy danych — różnorodność i wspólnota

- Rodzaje tekstowych baz danych:
  - wyszukiwarki internetowe;
  - archiwa gazet lub czasopism, biblioteki elektroniczne;
  - systemy informacji prawnej;
  - encyklopedie, słowniki, twory encyklopediopodobne;
  - korpusy tekstu.
- Cechy wspólne:
  - użytkownikami są ludzie;
  - tekst jest przedmiotem zapytań i odpowiedzi;
  - język zapytań:
    - deklaratywny,
    - prosty (tzn. czytelny dla człowieka),
    - zwykle o małej sile wyrazu,
    - przeważają zapytania selektywne;
  - zapisy są stosunkowo rzadkie, dominuje odczyt;
  - odpowiedź: dokument pasujący do zapytania (+ „uzasadnienie").

Przeszukiwanie tekstu a przeszukiwanie napisów

- Operujemy nie na znakach ale na *słowach*.
  Foka nie bywa *konfokalna*.
- Takie same słowa mogą różnić się napisowo.
  Łatwo pomylić *fokę* z inną *foką*.
- Dopuszczamy niedokładne odpowiedzi.
  Angielskie *seal* może oznaczać zarówno *fokę* jaki i *uchatkę* (np. *lwa morskiego*).
- Dopuszczamy niedokładne zapytania.
  Taaaaaaaką fokę dzisiaj widziałem. Co za okropne déjà vu.
- Tekst ma strukturę:
  - wielopoziomową,
  - często nieoczywistą,
  - zazwyczaj niejawną.

Jawne oznaczenie struktury *powinno* znacząco ułatwić przetwarzanie (w tym przeszukiwanie) tekstu.

Example

- *Corpus Encoding Standard for XML* (*XCES*).
- *TEI P5*:
  - wyróżnienia: `foreign`, `emph`, `distinct`;
  - cytaty: `q`, `quote`, `cit`, `soCalled`;
  - korekta: `choice`, `sic`, `corr`, `reg`, `orig`, `gap`, `unclear`, `add`, `del`;
  - odwołania: `rs`, `name`;
  - liczby i miary: `num`, `measure`, `date`, `time`.
- DocBook:
  - `abbrev`, `acronym`, `emphasis`, `foreignphrase`, `quote`, `wordasword`.
- *OASIS Open Document Format for Office Applications* (*ODF*).

Możliwości XQFT

- Przeszukiwanie zawartości tekstowej elemntów jak również atrybutów.
- Spójniki logiczne $\neg$, $\vee$, $\wedge$ i *ale nie* w zapytaniach:
  - `/book ftcontains "usability" not in "usability testing"`
- Kontrola kolejności składników zapytania:
  - `/book/title ftcontains ("web site" && "usability") ordered`
- Kontrola odległości pomiędzy składnikami zapytania:
  - `/book ftcontains "usability" && "testing" same paragraph`
  - `/book ftcontains "web" && "site" && "usability" distance at most 2 words`
- Specyfikacja ilości wystąpień:
  - `/book ftcontains "usability" occurs at least 2 times`

Możliwości XQFT, cd.

- Przeszukiwanie jest zawsze niewrażliwe na granice elementów.
- Przeszukiwanie z uwzględnieniem morfologii słów.
- Przeszukiwanie wspomagane tezaurusem (słownikiem, taksonomią).
- Pomijanie *stop words* w zapytaniach.
- Wyszukiwanie niewrażliwe na znaki diakrytyczne.
- Wyszukiwanie niewrażliwe na wielkość liter.
- Znaki wieloznaczne w zapytaniach.
- Pomijanie zawartości niektórych elementów:
  - `chapter ftcontains "users can be tested at any computer workstation or in a lab" without content .//footnote`
- *Scoring*.

XQFT — implementacje i alternatywy

Dostępne implementacje:

- *GalaTex*:
  ⟨`http://www.galaxquery.com/galatex/`⟩.

Alternatywy:

- *DB2 Net Search Extender*;
- *Quark DB* (*TeXQuery*):
  ⟨`http://www.cs.cornell.edu/database/quark/`⟩.

Definition

Korpus —

1. «ciało człowieka lub zwierzęcia oprócz głowy i kończyn»;
2. «10 pt»;
3. «zasadnicza część czegoś»;
4. «główna część budowli»;
5. «centralna część budynku»;
6. «nawowa część kościoła»;
7. «główna część, na której oparta jest całość jakiegoś urządzenia, przyrządu itp.»;
8. «jednostka taktyczna składająca się z kilku dywizji lub brygad»;
9. «grupa żołnierzy mających taki sam stopień wojskowy»;
10. «teksty, dane itp. zgromadzone ze względu na swą reprezentatywność, stanowiące podstawę do analizy naukowej».

Definition

- **Korpus** — adnotowany zbiór tekstów o dużym rozmiarze: pewnym partiom tekstu towarzyszą dane, które mogą stanowić kryterium wyszukiwania.
- Adnotacja:
  - *morfoskładniowa* — słowo $\mapsto$ opis gramatyczny;
  - *składniowa* — zdanie $\mapsto$ struktura składniowa zdania;
  - *strukturalna* — podział na rozdziały, akapity, zdania, itp.
- Korpus jest **zrównoważony** jeśli obejmuje różne typy języka w proporcjach odpowiadających stopniowi ich rozpowszechnienia wśród użytkowników języka.

Zastosowania korpusów

- leksykografia — projektowanie słowników;
- lingwistyka typologiczna;
- przetwarzanie języka naturalnego (materiał treningowy);
- nauka języków obcych (i nieobcych).

Korpusy dostępne w Internecie

- *Korpus IPI PAN*:
  - $\approx$ 250 mln segmentów,
  - ⟨`http://korpus.pl/`⟩;
- *Korpus Słownika Frekwencyjnego*:
  - $\approx$ 0,5 mln segmentów,
  - ⟨`http://korpus.pl/`⟩;
- *Korpus Języka Polskiego Wydawnictwa Naukowego PWN*:
  - wersja bezpłatna — $\approx$ 7,5 mln słów,
  - wersja płatna — $\approx$ 40 mln segmentów,
  - ⟨`http://korpus.pwn.pl/`⟩;
- *Český národní korpus*:
  - ⟨`http://ucnk.ff.cuni.cz/`⟩;
- *British National Corpus*:
  - $\approx$ 100 mln słów,
  - ⟨`http://www.natcorp.ox.ac.uk/`⟩;
- oraz mnóstwo innych.

Etapy tworzenia korpusu (na przykładzie Korpusu IPI PAN)

1. Dobór i pozyskiwanie tekstów (i praw autorskich).
2. Konwersja do jednolitego formatu (*XCES*):
   - 2.1 konwersja automatyczna, następnie
   - 2.2 ręczna weryfikacja i korekta.
3. Znakowanie morfoskładniowe:
   - 3.1 Podział tekstu na segmenty.
   - 3.2 Utworzenie możliwych opisów gramatycznych każdego segmentu.
   - 3.3 Dezambiguacja.
4. Przekształcenie do postaci ostatecznej:
   - 4.1 Konwersja do zwartej postaci binarnej.
   - 4.2 Zbudowanie indeksów.

Klasy gramatyczne (≈ części mowy)

- rzeczownik:
  - `subst` (*profesorowie*),
  - deprecjatywny — `depr` (*profesory*),
  - ciało obce nominalne `xxs` (*l'Hospital*);
- ciało obce luźne — `xxx` (*bene*);
- liczebnik:
  - główny — `num` (*pięciu*),
  - zbiorowy — `col` (*pięcioro*);
- przymiotnik:
  - `adj` (*polski*),
  - przyprzymiotnikowy — `adja` (*polsko-niemiecki*),
  - poprzyimkowy — `adjp` (*po polsku*);
- przysłówek odprzymiotnikowy/stopniowalny — `adv` (*polsko brzmiący*);
- zaimek:
  - nietrzecioosobowy — `ppron12` (*ja*),
  - trzecioosobowy — `ppron3` (*on*),
  - *siebie* — `siebie`;

Klasy gramatyczne (≈ części mowy), cd.

- „czasownik”:
  - forma nieprzeszła — `fin` (*czytam*),
  - forma przyszła *być* — `bedzie` (*będę*),
  - aglutynant *być* — `aglt` (*czytaliśmy*),
  - pseudoimiesłów — `praet` (*czytaliśmy*),
  - rozkaźnik — `impt` (*czytaj*),
  - bezosobnik — `imps` (*czytano*),
  - bezokolicznik — `inf` (*czytać*),
  - imiesłów przyszły współczesny — `pcon` (*czytając*),
  - imiesłów przyszły uprzedni — `pant` (*przeczytawszy*),
  - odsłownik — `ger` (*czytanie*),
  - imiesłów przymiotnikowy czynny — `pact` (*czytający*),
  - imiesłów przymiotnikowy bierny — `ppas` (*czytany*),
  - czasownik typu *winien* — `winien`,
  - predykatyw — `pred` (*można*);
- przyimek — `prep` (*pod*);
- spójnik — `conj` (*lub*);
- kublik — `cub` (*też*, *jutro*).

Kategorie gramatyczne

- liczba: `sg`, `pl`;
- przypadek: `nom`, `gen`, `dat`, `acc`, `inst`, `loc`, `voc`;
- rodzaj: `m1`, `m2`, `m3`, `f`, `n`;
- osoba: `pri`, `sec`, `ter`;
- stopień: `pos`, `comp`, `sup`;
- aspekt: `imperf`, `perf`;
- zanegowanie: `aff`, `neg`;
- akcentowość: `akc` (*tobie*), `nakc` (*ci*);
- poprzyimkowość: `praep` (*niego*), `npraep` (*jego*);
- akomodacyjność: `congr` (*dwaj*), `rec` (*dwóch*);
- aglutynacyjność: `nagl`, `agl`;
- wokaliczność: `wok` (*ze sobą*), `nwok` (*z tobą*).

Wieloznaczność leksykalna

Example

Wieloznaczność formy *tusz*:

forma podstawowa	opis gramatyczny	znaczenie
tusz	`subst:sg:nom:m3`	I «farba wodna» II «prysznic, natrysk» III «fanfara» IV «w szermierce: trafienie»
	`subst:sg:acc:m3`	
tusza	`subst:pl:gen:f`	«ubite zwierzę rzeźne»
tuszyć	`impt:sg:sec:imperf`	*daw.* «spodziewać się czegoś»

Wieloznaczność składniowa i semantyczna

Example

Wieloznaczność zdania *Widziano ją pijaną.*:

słowo	forma podstawowa	opis gramatyczny
widziano	*widzieć*	`imps:imperf`
ją	*on*	`ppron3:sg:acc:f:ter:_:npraep`
pijaną	*pijany*	`adj:sg:acc:f:pos`
		`adj:sg:inst:f:pos`
	pijać	`ppas:sg:acc:f:imperf:aff`
		`ppas:sg:inst:f:imperf:aff`

Wieloznaczność składniowa i semantyczna

Example

Wieloznaczność zdania *Widziano ją pijaną.*:

słowo	forma podstawowa	opis gramatyczny
widziano	*widzieć*	`imps:imperf`
ją	*on*	`ppron3:sg:acc:f:ter:_:npraep`
pijaną	*pijany*	`adj:sg:acc:f:pos`
		`adj:sg:inst:f:pos`
	pijać	`ppas:sg:acc:f:imperf:aff`
		`ppas:sg:inst:f:imperf:aff`

Wieloznaczność składniowa i semantyczna

Example

Wieloznaczność zdania *Widziano ją pijaną.*:

słowo	forma podstawowa	opis gramatyczny
widziano	*widzieć*	`imps:imperf`
ją	*on*	`ppron3:sg:acc:f:ter:_:npraep`
pijaną	*pijany*	`adj:sg:acc:f:pos`
		`adj:sg:inst:f:pos`
	pijać	`ppas:sg:acc:f:imperf:aff`
		`ppas:sg:inst:f:imperf:aff`

Wieloznaczność składniowa i semantyczna

Example

Wieloznaczność zdania *Widziano ją pijaną.*:

słowo	forma podstawowa	opis gramatyczny
widziano	*widzieć*	`imps:imperf`
ją	*on*	`ppron3:sg:acc:f:ter:_:npraep`
pijaną	*pijany*	`adj:sg:acc:f:pos`
		`adj:sg:inst:f:pos`
	pijać	`ppas:sg:acc:f:imperf:aff`
		`ppas:sg:inst:f:imperf:aff`

Wieloznaczność składniowa i semantyczna

Example

Wieloznaczność zdania *Widziano ją pijaną.*:

słowo	forma podstawowa	opis gramatyczny
widziano	*widzieć*	`imps:imperf`
ją	*on*	`ppron3:sg:acc:f:ter:_:npraep`
pijaną	*pijany*	`adj:sg:acc:f:pos`
		`adj:sg:inst:f:pos`
	pijać	`ppas:sg:acc:f:imperf:aff`
		`ppas:sg:inst:f:imperf:aff`

Wieloznaczność składniowa

Example

Wieloznaczność zdania *Ma mama ma mamałygę.*:

słowo	forma podstawowa	opis gramatyczny
ma	*mieć*	`fin:sg:ter:imperf`
	mój	`adj:sg:nom:f:pos`
mama	*mama*	`subst:sg:nom:f`
ma	*mieć*	`fin:sg:ter:imperf`
	mój	`adj:sg:nom:f:pos`
mamałygę	*mamałyga*	`subst:sg:acc:f`

Wieloznaczność składniowa

Example

Wieloznaczność zdania *Ma mama ma mamałygę.*:

słowo	forma podstawowa	opis gramatyczny
ma	*mieć*	`fin:sg:ter:imperf`
	mój	`adj:sg:nom:f:pos`
mama	*mama*	`subst:sg:nom:f`
ma	*mieć*	`fin:sg:ter:imperf`
	mój	`adj:sg:nom:f:pos`
mamałygę	*mamałyga*	`subst:sg:acc:f`

Wieloznaczność składniowa

Example

Wieloznaczność zdania *Ma mama ma mamałygę.*:

słowo	forma podstawowa	opis gramatyczny
ma	*mieć*	`fin:sg:ter:imperf`
	mój	`adj:sg:nom:f:pos`
mama	*mama*	`subst:sg:nom:f`
ma	*mieć*	`fin:sg:ter:imperf`
	mój	`adj:sg:nom:f:pos`
mamałygę	*mamałyga*	`subst:sg:acc:f`

Język zapytań

- *zapytanie* → *zapytanie-główne ograniczenie*
- *zapytanie-główne* → wyrażenie regularne nad *zapytanie-proste*
- *zapytanie-proste* → [*wyrażenie*]
- *wyrażenie* →
  - (*wyrażenie*) |
  - `!` *wyrażenie* |
  - *wyrażenie* `&` *wyrażenie* |
  - *wyrażenie* `|` *wyrażenie* |
  - *atrybut operator specyfikacja modyfikatory*
- *atrybut* →
  - `orth` | `base` | `pos` | `tag` |
  - `nmb` | `cas` | `gnd` | `per` | `deg` | `asp` | `neg` | `acm` | `acn` | `ppr` | `agg` | `vcl`
- *operacja* → `=` | `!=` | `==` | `!==` | `~` | `!~` | `~~` | `!~~`
- *specyfikacja* → wyrażenie regularne
- *modyfikatory* → ε | `/i` | `/x` | `/ix`

Język zapytań, cd.

- *ograniczenie* → *ograniczenie-strukturalne ograniczenia-metadanych*
- *ograniczenie-strukturalne* → ε | `within s` | `within p`
- *ograniczenie-metadanych* → ε | `meta` *m-wyrażenie*
- *m-wyrażenie* →
  - `(`*m-wyrażenie*`)` |
  - `!`*m-wyrażenie* |
  - *m-wyrażenie* `&` *m-wyrażenie* |
  - *m-wyrażenie* `|` *m-wyrażenie* |
  - *m-atrybut m-operator m-specyfikacja m-modyfikatory*
- *m-atrybut* →
  - `autor` | `tytuł` | `data_powstania` | `styl` | `medium` |
  - `wydawca` | `miejsce_wydania` | `data_wydania` | `data_pierwszego_wydania`
- *m-operacja* → `=` | `!=` | `<` | `<=` | `>` | `>=`
- *m-specyfikacja* → wyrażenie regularne
- *m-modyfikatory* → ε | `/I` | `/X` | `/IX`

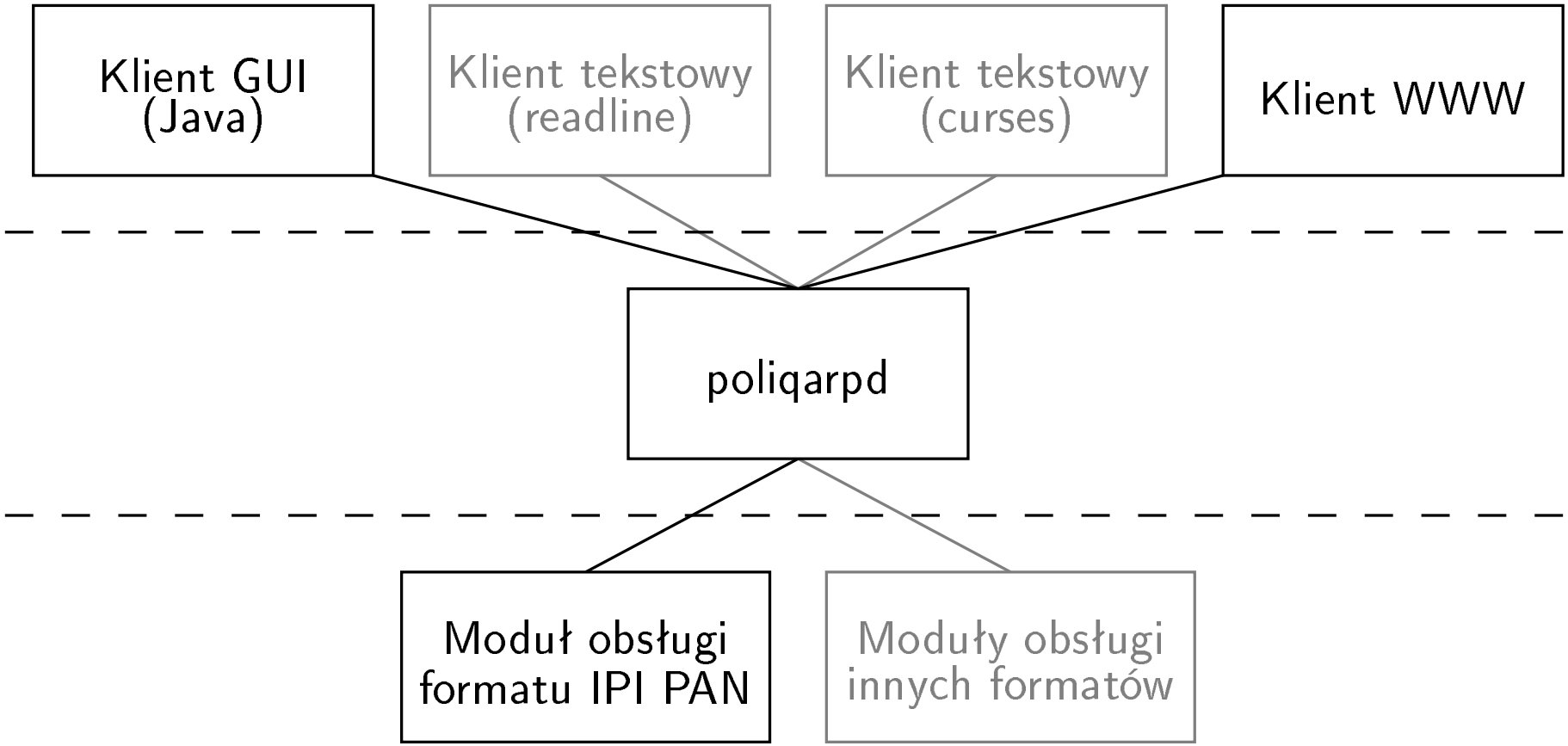
Język zapytań i język odpowiedzi — przykład

Zapytanie:

```
[pos=adj]^[pos=subst & case=gen & number=pl & orth=".*[iy]j"]
```

1999 kilka	*przemiłych atrakcyj*	*: 1*
	[przemiły:adj:pl:gen:f:pos] [atrakcja:subst:pl:gen:f]	
dość mi	*tych ceremonij*	*. -*
	[ten:adj:pl:gen:f:pos] [ceremonia:subst:pl:gen:f]	
, od	*takich delicyj*	*rozum może*
	[taki:adj:pl:gen:f:pos] [delicja:subst:pl:gen:f]	
festonów i	*barchanowych draperyj*	*, szczerzyła*
	[barchanowy:adj:pl:gen:f:pos] [draperia:subst:pl:gen:f]	
do "	*dyskretnych funkcyj*	*",*
	[dyskretny:adj:pl:gen:f:pos] [funkcja:subst:pl:gen:f]	
się do	*mechanicznych funkcyj*	*! Ja*
	[mechaniczny:adj:pl:gen:f:pos] [funkcja:subst:pl:gen:f]	
i tym	*podobnych galanteryj*	*- po*
	[podobny:adj:pl:gen:f:pos] [galanteria:subst:pl:gen:f]	

Architektura systemu Poliqarp 1.0



Podstawowe struktury danych

- *wektor* — sekwencja rekordów ustalonego rozmiaru;
- *słownik* — sekwencja rekordów o zmiennym rozmiarze; 2 lub 3 pliki:
  - *obraz słownika* — skonkatenowane elementy słownika,
  - indeks nr-elementu $\mapsto$ miejsce w słowniku — wektor offsetów kolejnych elementów,
  - (opcjonalnie) indeks element $\mapsto$ miejsce w słowniku — tablica haszująca.

Binarny format korpusu

- *słownik form literalnych* + 2 indeksy: *a fronte* i *a tergo*;
- *słowniki form podstawowych*;
- *słownik znaczników morfoskładniowych*;
- *słowniki interpretacji*:
  - elementami są ciągi interpretacji;
  - interpretacja:
    - forma podstawowa (20 bitów),
    - znacznik morfoskładniowy (12 bitów);
- *obraz korpusu*:
  - wektor segmentów;
  - segment:
    - czy przed segmentem występuje spacja? (1 bit),
    - forma literalna (21 bitów),
    - interpretacje ujednoznacznione (21 bitów),
    - interpretacje wieloznaczne (21 bitów).
- metadane;
- indeksy odwrotne.

Korpus IPI PAN w liczbach

- Obraz korpusu:
  - 255,5 mln segmentów;
  - 1941,9 MiB.
- Wielkości słowników:

słownik	# elementów (tys.)	% limitu	rozmiar obrazu (MiB)	rozmiar indeksu (MiB)
form literalnych	1 396 832	60,0	21,69	5,33
form hasłowych	775 476	73,9	11,75	2,95
interpretacji	1 579 275	75,3	12,21	6,02
znaczników morfoskładniowych	1 282	31,3	4,56	0,03

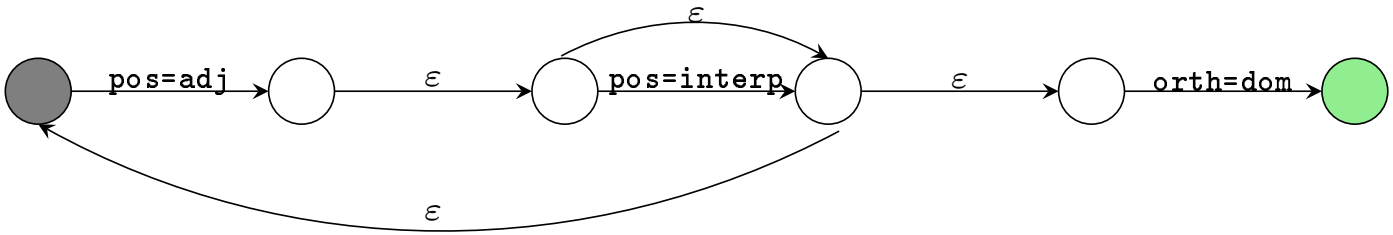
Reprezentacja zapytań prostych

- *wyrażenie* — reprezentowane jako drzewo:
  - symbole &, | lub ! w węzłach wewnętrznych,
  - wyrażenia proste w liściach,
  - wyliczenie: rekurencyjnie;
- *wyrażenie proste*:
  - opisuje podzbiór słów pewnego słownika,
  - reprezentowane przez parę: słownik + ciąg bitów,
  - kompilacja: długa,
  - wyliczenie: w czasie stałym i bez zaglądania do słowników!

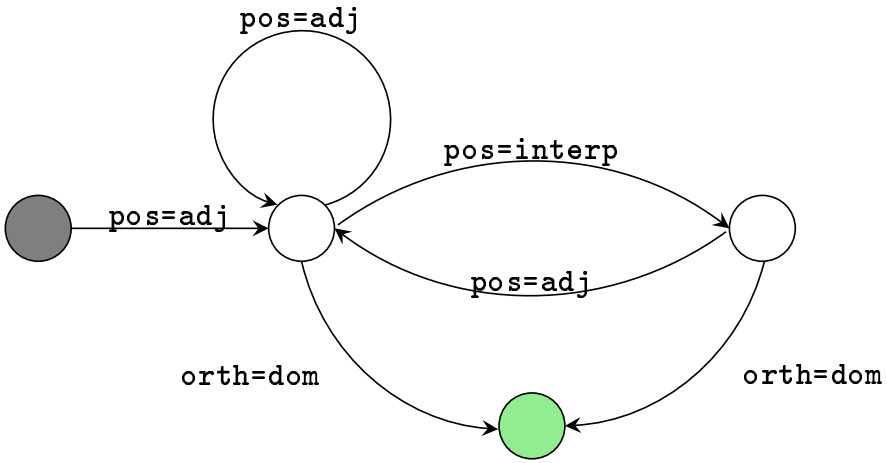
Reprezentacja zapytań złożonych

Example

1. Zapytanie: `([pos=adj][pos=interp]?)+[orth=dom]`
2. Automat niedeterministyczny z ε-przejściami:



3. Automat niemal deterministyczny:



Ograniczenie przestrzeni poszukiwań

- explicite w zapytaniu;
- implicite, w przypadku prostych zapytań — indeksy odwrotne:
  - pierwszy pomysł: obiekt ze słownika $\mapsto$ nry segmentów w obrazie;
  - lepszy pomysł: obiekt ze słownika $\mapsto \left\lfloor \frac{\text{nry segmentów w obrazie}}{k} \right\rfloor$
    - k — duże ale nie za duże (1024),
    - reprezentacje zbiorów można skompresować,
    - efekt: indeksy odwrotne to 18% rozmiaru obrazu korpusu.

Google jako narzędzie do badań nad językiem

(Eksperyment przeprowadzony 3 kwietnia 2007 r.)

Example

- `"w ogóle"`: $\approx$ 1,7 mln wyników;
- `wogóle`: $\approx$ 3,4 mln wyników.

Example

- `microsoft`: $\approx$ 621 mln wyników;
- `windows`: $\approx$ 759 mln wyników;
- `microsoft windows`: $\approx$ 651 mln wyników.

- Rafał T. Prinke: *Fontes ex machina. Komputerowa analiza źródeł historycznych*
- *XQuery 1.0 and XPath 2.0 Full-Text* (W3C Working Draft)
  ⟨http://www.w3.org/TR/xquery-full-text/⟩
- Sihem Amer-Yahia, Chavdar Botev, Jochen Dörre, Jayavel Shanmugasundaram: *XQuery Full-Text extensions explained*
  ⟨http://www.research.ibm.com/journal/sj/452/amer.html⟩
- Janusz S. Bień: *Aparat pojęciowy wybranych systemów przetwarzania tekstów polskich*
- Marcin Woliński: *System znaczników morfosyntaktycznych w korpusie IPI PAN*
  ⟨http://nlp.ipipan.waw.pl/~wolinski/morfeusz/znakowanie.pdf⟩
- Adam Przepiórkowski: *Korpus IPI PAN, wersja wstępna*
  ⟨http://nlp.ipipan.waw.pl/~adamp/Papers/2004-corpus/book_pl.pdf⟩
- Daniel Janus: *Metody przeszukiwania dużych korpusów tekstów*
  ⟨http://korpus.pl/~nathell/praca.pdf⟩