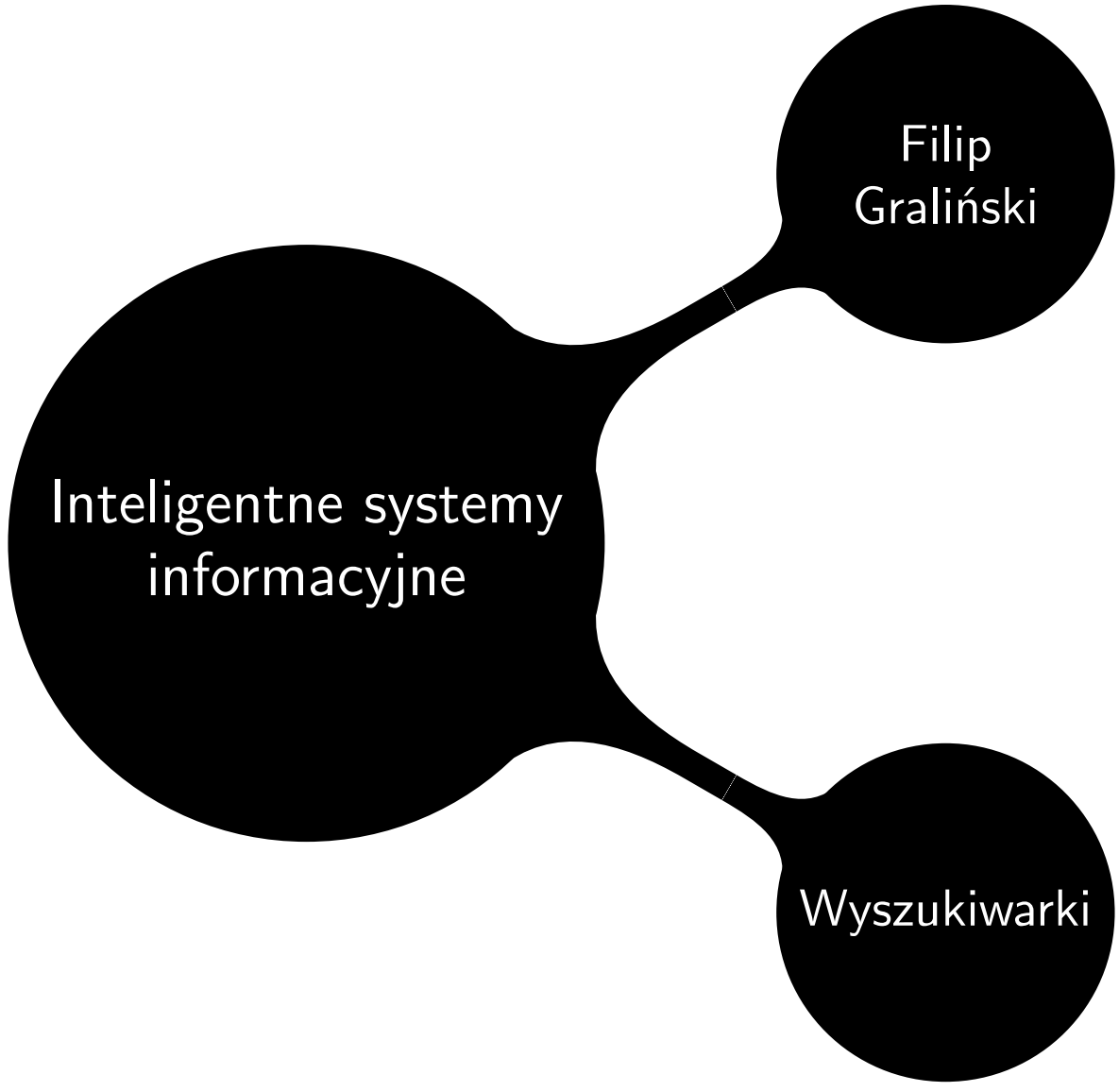
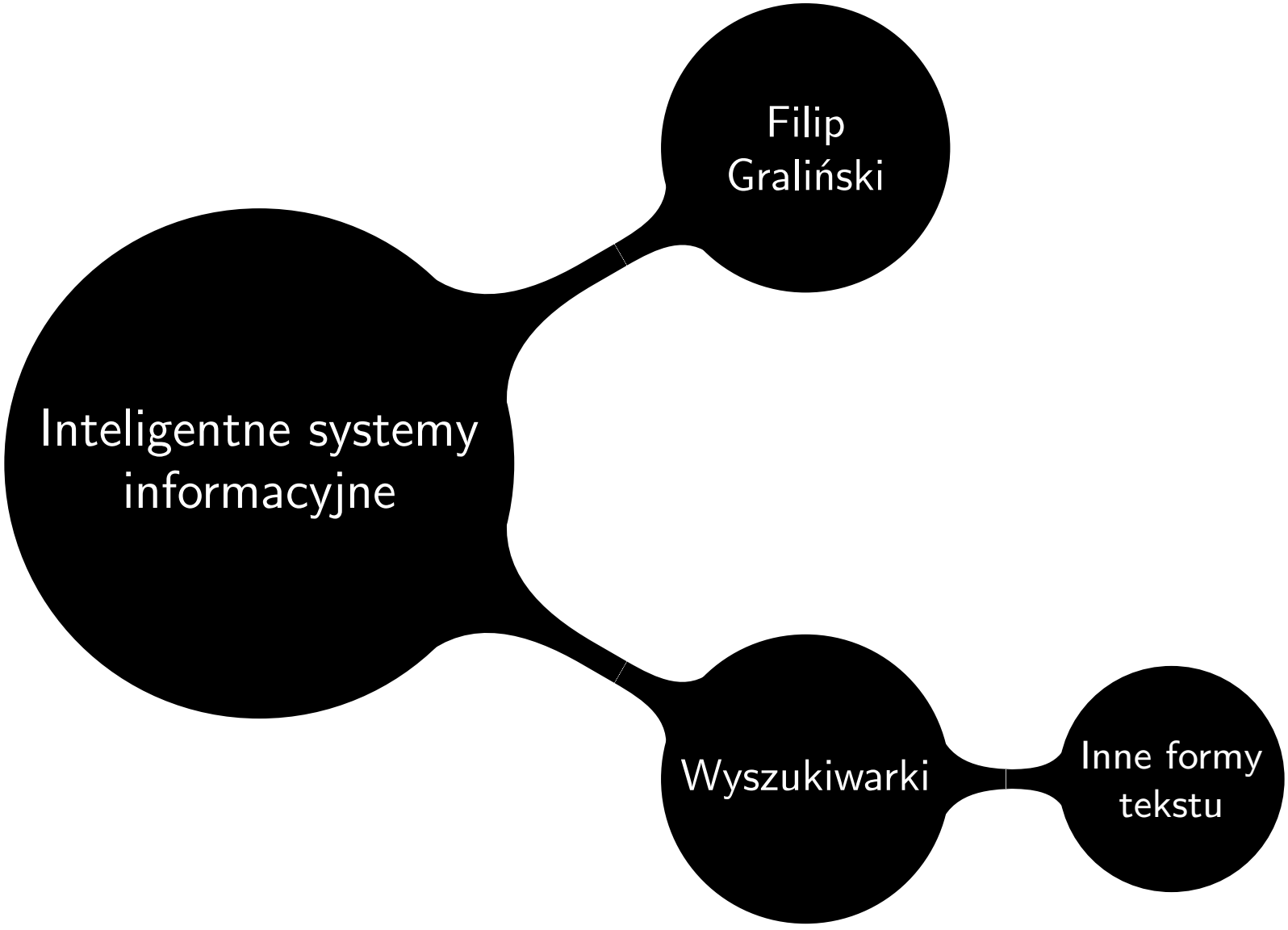
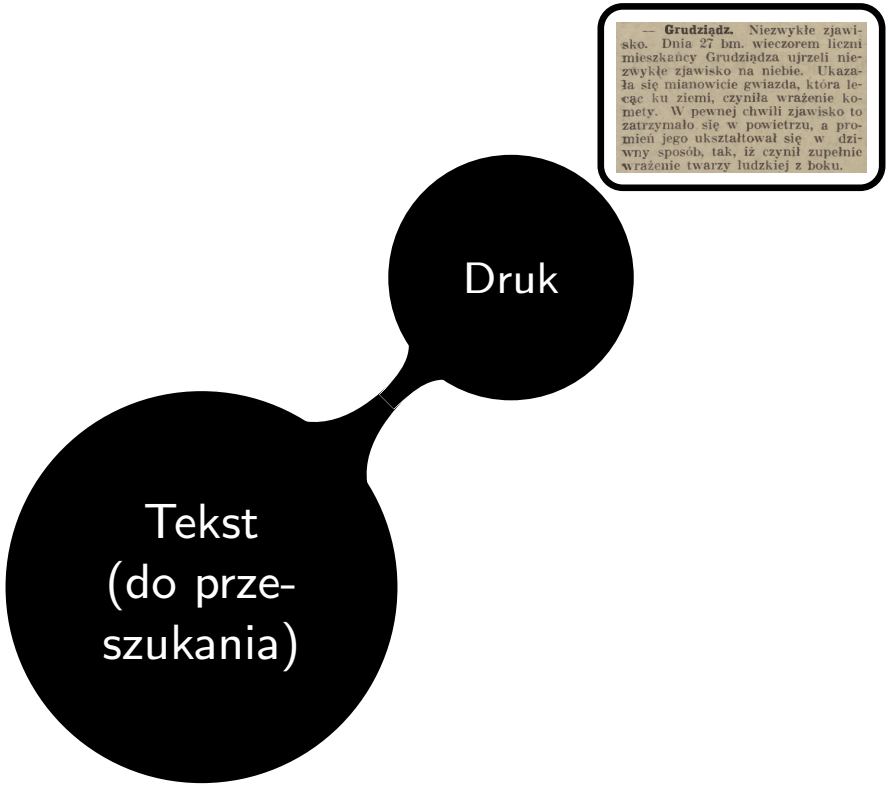
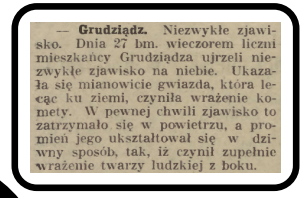

Filip
Graliński
Inteligentne systemy
informacyjne
Wyszukiwarki


Filip
Graliński
Inteligentne systemy
informacyjne
Wyszukiwarki
Inne formy
tekstu


Tekst
(do prze-
szukania)


— Grudziądz. Niezwykłe zjawisko. Dnia 27 bm. wieczorem liczni mieszkańcy Grudziądza ujrzeli niezwykłe zjawisko na niebie. Ukazała się mianowicie gwiazda, która lecąc ku ziemi, czyniła wrażenie komety. W pewnej chwili zjawisko to zatrzymało się w powietrzu, a promień jego ukształtował się w dziwny sposób, tak, iż czynił zupełnie wrażenie twarzy ludzkiej z boku.
Druk
Tekst (do przeszukania)


— Grudziądz. Niezwykłe zjawisko. Dnia 27 bm. wieczorem liczni mieszkańcy Grudziądza ujrzeli niezwykłe zjawisko na niebie. Ukazała się mianowicie gwiazda, która lecąc ku ziemi, czyniła wrażenie komety. W pewnej chwili zjawisko to zatrzymało się w powietrzu, a promień jego ukształtował się w dziwny sposób, tak, iż czynił zupełnie wrażenie twarzy ludzkiej z boku.
Druk
Tekst (do przeszukania)
Pismo odręczne

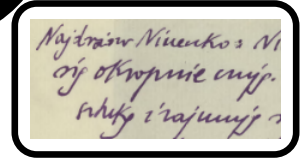

— **Grudziądz.** Niezwykłe zjawisko. Dnia 27 bm. wieczorem liczni mieszkańcy Grudziądza ujrzeli niezwykłe zjawisko na niebie. Ukazała się mianowicie gwiazda, która lecąc ku ziemi, czyniła wrażenie komety. W pewnej chwili zjawisko to zatrzymało się w powietrzu, a promień jego ukształtował się w dziwny sposób, tak, iż czynił zupełnie wrażenie twarzy ludzkiej z boku.

Druk

Tekst (do przeszukania)

Mowa

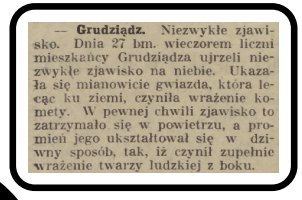
Pismo odręczne



조선글
한글

Język obcy

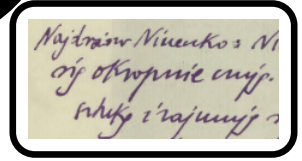
Druk


— **Grudziądz.** Niezwykłe zjawisko. Dnia 27 bm. wieczorem liczni mieszkańcy Grudziądza ujrzeli niezwykłe zjawisko na niebie. Ukazała się mianowicie gwiazda, która lecąc ku ziemi, czyniła wrażenie komety. W pewnej chwili zjawisko to zatrzymało się w powietrzu, a promień jego ukształtował się w dziwny sposób, tak, iż czynił zupełnie wrażenie twarzy ludzkiej z boku.

Tekst
(do prze-
szukania)

Mowa

Pismo
odręczne

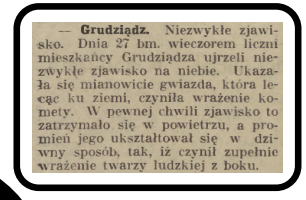


조선글
한글

Tekst (do przeszukania)

Język obcy

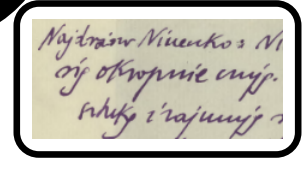
Druk


— **Grudziądz.** Niezwykłe zjawisko. Dnia 27 bm. wieczorem liczni mieszkańcy Grudziądza ujrzeli niezwykłe zjawisko na niebie. Ukazała się mianowicie gwiazda, która lecąc ku ziemi, czyniła wrażenie komety. W pewnej chwili zjawisko to zatrzymało się w powietrzu, a promień jego ukształtował się w dziwny sposób, tak, iż czynił zupełnie wrażenie twarzy ludzkiej z boku.

Texting

Mowa

Pismo odręczne



Forma zapytania czy przeszukiwanych treści?

	forma.	
	zapytania	treści
druk		
pismo odręcz-ne		
mowa		
język obcy		
texting		

Forma zapytania czy przeszukiwanych treści?

	forma.	
	zapytania	treści
druk	—	przeszukiwanie historycznych materiałów
pismo odręczne		
mowa		
język obcy		
texting		

Forma zapytania czy przeszukiwanych treści?

	forma.....	
	zapytania	treści
druk	—	przeszukiwanie historycznych materiałów
pismo odręczne	interfejs urządzeń mobilnych (?)	przeszukiwanie rękopisów (??)
mowa		
język obcy		
texting		

Forma zapytania czy przeszukiwanych treści?

	forma.	
	zapytania	treści
druk	—	przeszukiwanie historycznych materiałów
pismo odręczne	interfejs urządzeń mobilnych (?)	przeszukiwanie rękopisów (??)
mowa	*voice search*	*spoken content retrieval*
język obcy		
texting		

Forma zapytania czy przeszukiwanych treści?

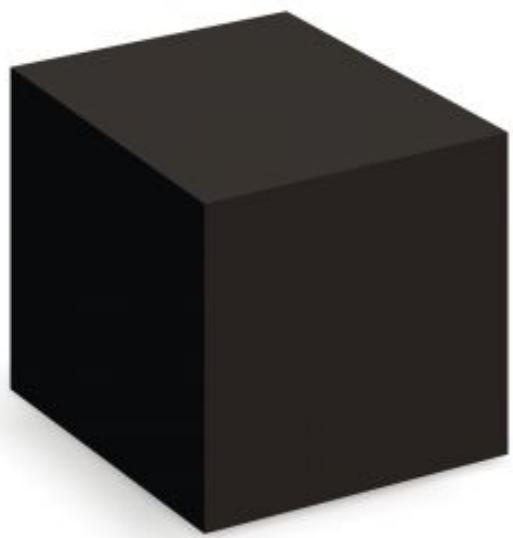
	forma.	
	zapytania	treści
druk	—	przeszukiwanie historycznych materiałów
pismo odręczne	interfejs urządzeń mobilnych (?)	przeszukiwanie rękopisów (??)
mowa	*voice search*	*spoken content retrieval*
język obcy	—	*cross-language information retrieval*
texting		

Forma zapytania czy przeszukiwanych treści?

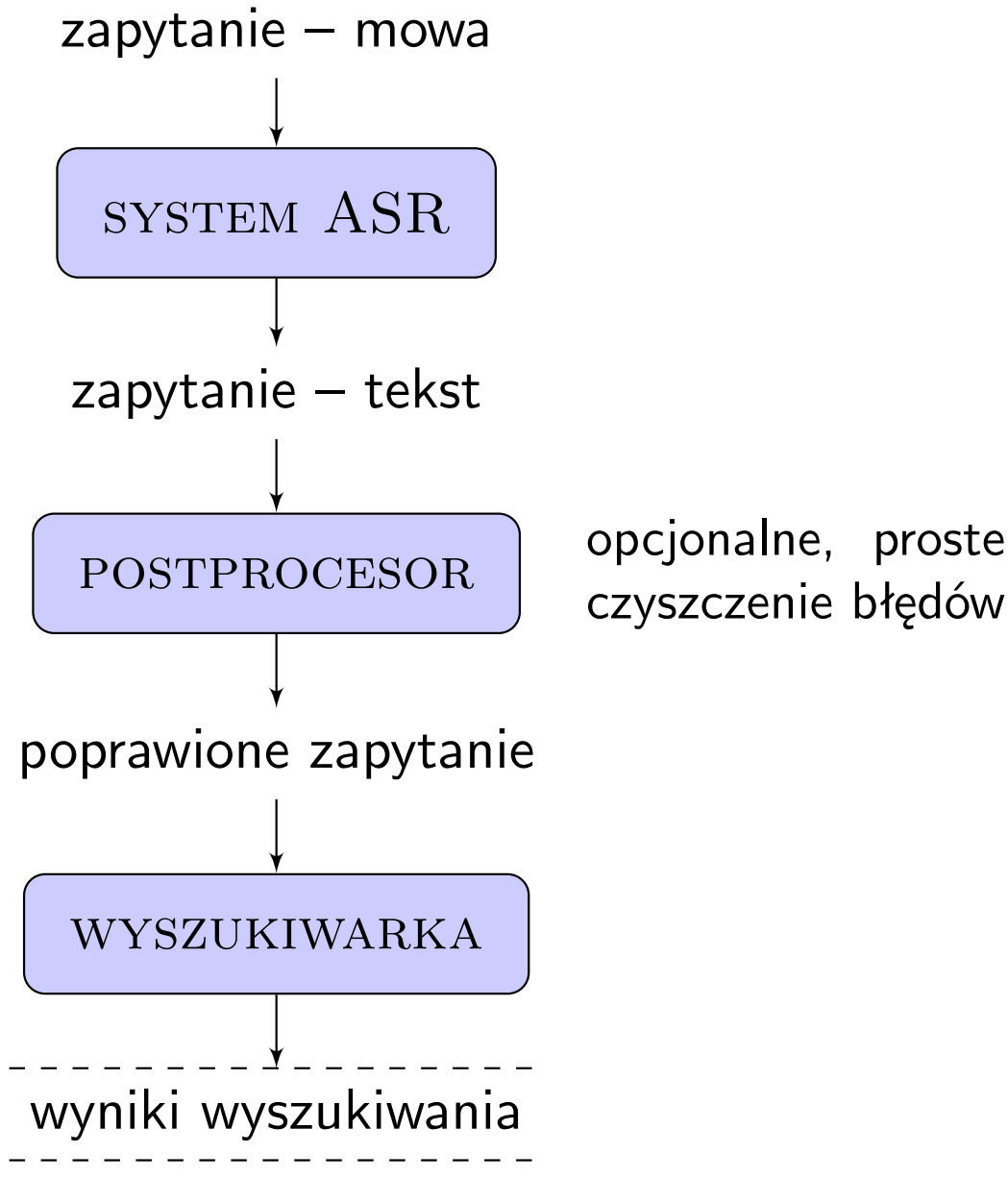
	forma.	
	zapytania	treści
druk	—	przeszukiwanie historycznych materiałów
pismo odręczne	interfejs urządzeń mobilnych (?)	przeszukiwanie rękopisów (??)
mowa	*voice search*	*spoken content retrieval*
język obcy	—	*cross-language information retrieval*
texting	interfejs urządzeń mobilnych	przeszukiwanie Twittera, chatów, SMS-ów etc.

Jak wyszukiwać w innych formach tekstu?

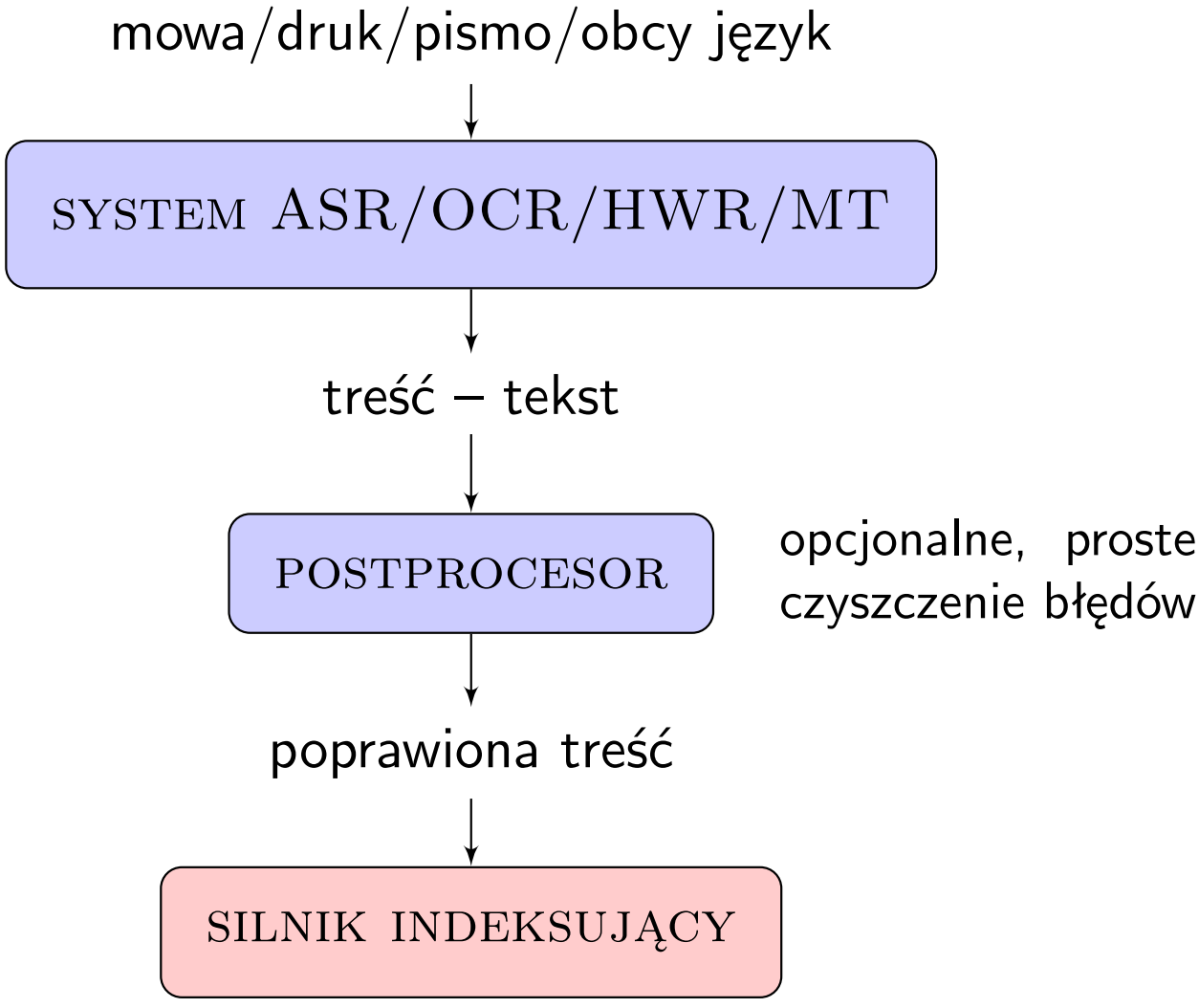
Rozwiązanie 1. Najprościej traktować OCR/HWR/ASR/MT jako czarną skrzynkę, która zwraca zwykły tekst.



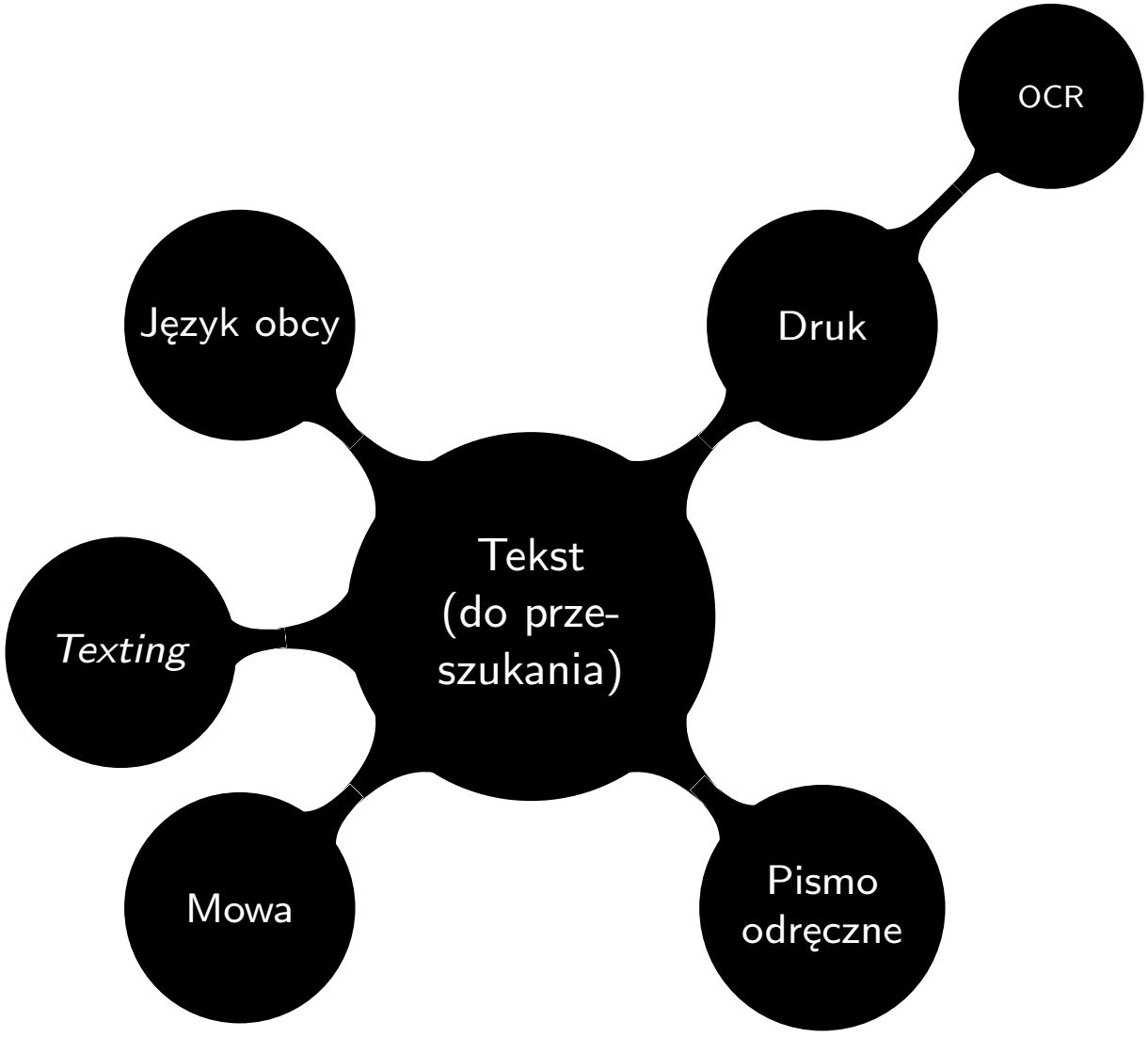
Przez czarną skrzynkę przepuszczamy zapytanie...



. . . albo indeksowane treści

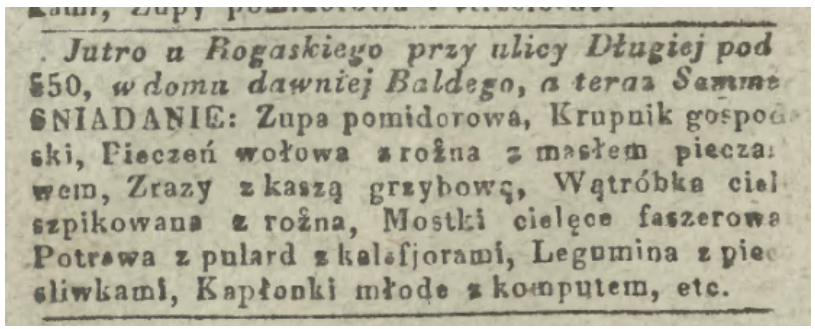


Przyjrzyjmy się bliżej czarnym skrzynkom...



OCR = *Optical Character Recognition*

Trudności:

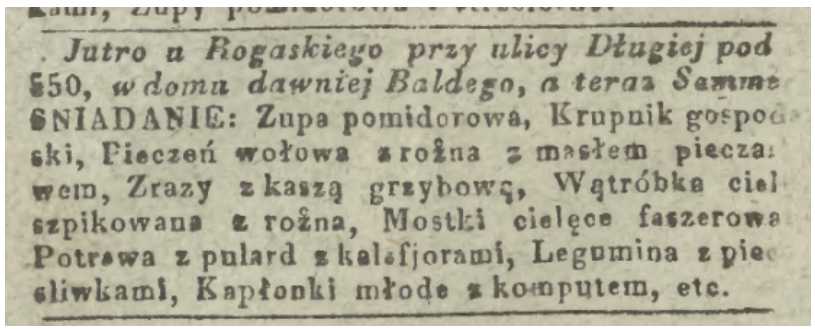

. Jutro u Bogaskiego przy ulicy Długiej pod
S50, w domu dawniej Baldego, a teraz Samme
ŚNIADANIE: Zupa pomidorowa, Krupnik gospo
ski, Pieczeń wołowa z rożna z masłem piecza
wem, Zrazy z kaszą grzybową, Wątróbka ciel
szpikowana z rożna, Mostki cielęce faszerowa
Potrawa z pulard z kalafjorami, Legumina z pie
śliwkami, Kapłonki młode z kompotem, etc.

Jutro a Bogaskiego przy ulicy Długiej pod S50, w domu dawniej B aide go, a teraz Sammt ŚNIADANIE: Zupa pomidorowa, Krupnik gospoó ski, Pieczeń wołowa z rolna 5 mssłetn piecza; wetn, Zrazy z kaszą grzybowy, Wątróbka ciel ¿ szpikowana Z rożna, Mostki cielęce faszerowp Potrawa z pulard % kalafjorami, Legnmina z pie śliwkami, Kapłonki młode z komputera, etc.

OCR = *Optical Character Recognition*

Trudności:

- zniekształcone strony, kleksy, pleśń, dżemik itd.



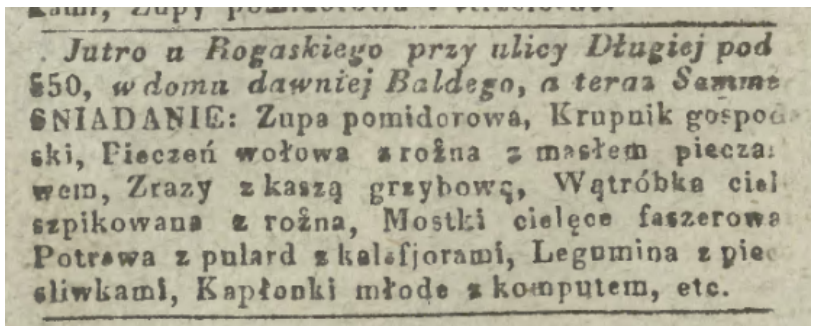
. Jutro u Rogaskiego przy ulicy Długiej pod
S50, w domu dawniej Baldego, a teraz Samme
ŚNIADANIE: Zupa pomidorowa, Krupnik gospo
ski, Pieczeń wołowa z rożna z masłem piecza
wem, Zrazy z kaszą grzybową, Wątróbka ciel
szpikowana z rożna, Mostki cielęce faszerowa
Potrawa z pulard z kalafjorami, Legumina z pie
śliwkami, Kapłonki młode z kompotem, etc.

Jutro a Bogaskiego przy ulicy Długiej pod S50, w domu dawniej B aide go, a teraz Sammt ŚNIADANIE: Zupa pomidorowa, Krupnik gospoó ski, Pieczeń wołowa z rolna 5 mssłetn piecza; wetn, Zrazy z kaszą grzybowy, Wątróbka ciel ¿ szpikowana Z rożna, Mostki cielęce faszerowp Potrawa z pulard % kalafjorami, Legnmina z pie śliwkami, Kapłonki młode z komputera, etc.

OCR = *Optical Character Recognition*

Trudności:

- zniekształcone strony, kleksy, pleśń, dżemik itd.
- rozpoznanie struktury strony (kolumny, ilustracje itd.)

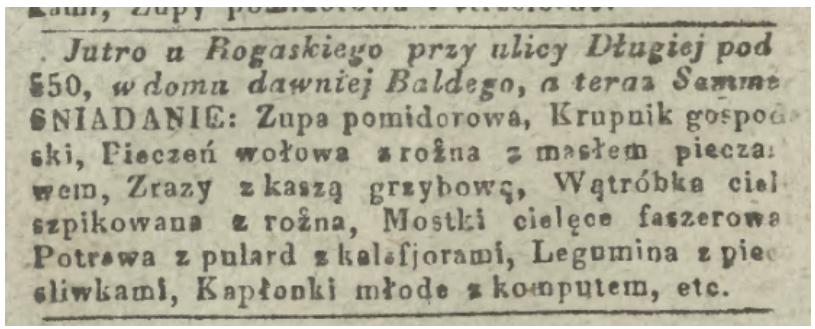

. Jutro u Rogaskiego przy ulicy Długiej pod
S50, w domu dawniej Baldego, a teraz Samme
ŚNIADANIE: Zupa pomidorowa, Krupnik gospo
ski, Pieczeń wołowa z rożna z masłem piecza
wem, Zrazy z kaszą grzybową, Wątróbka ciel
szpikowana z rożna, Mostki cielęce faszerowa
Potrawa z pulard z kalafjorami, Legumina z pie
śliwkami, Kapłonki młode z kompotem, etc.

Jutro a Bogaskiego przy ulicy Długiej pod S50, w domu dawniej B aide go, a teraz Sammt ŚNIADANIE: Zupa pomidorowa, Krupnik gospoó ski, Pieczeń wołowa z rolna 5 mssłetn piecza; wetn, Zrazy z kaszą grzybowy, Wątróbka ciel ¿ szpikowana Z rożna, Mostki cielęce faszerowp Potrawa z pulard % kalafjorami, Legnmina z pie śliwkami, Kapłonki młode z komputera, etc.

OCR = *Optical Character Recognition*

Trudności:

- zniekształcone strony, kleksy, pleśń, dżemik itd.
- rozpoznanie struktury strony (kolumny, ilustracje itd.)
- zmieniające się konwencje typograficzne, kroje czcionek itd.
- zmieniające się konwencje ortograficzne

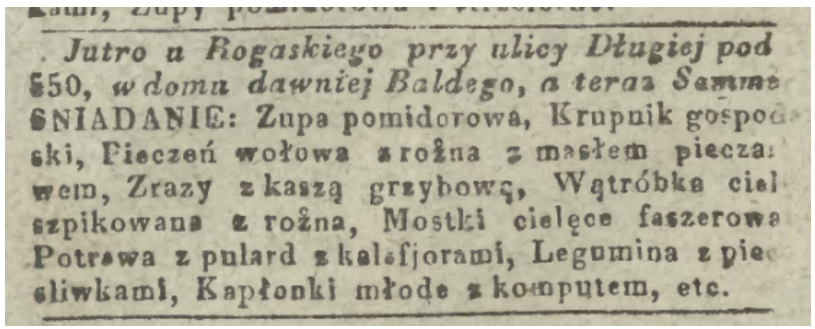

. Jutro u Rogaskiego przy ulicy Długiej pod
S50, w domu dawniej Baldego, a teraz Samme
ŚNIADANIE: Zupa pomidorowa, Krupnik gospo
ski, Pieczeń wołowa z rożna z masłem piecza
wem, Zrazy z kaszą grzybową, Wątróbka ciel
szpikowana z rożna, Mostki cielęce faszerowa
Potrawa z pulard z kalafjorami, Legumina z pie
śliwkami, Kapłonki młode z komputem, etc.

Jutro a Bogaskiego przy ulicy Długiej pod S50, w domu dawniej B aide go, a teraz Sammt ŚNIADANIE: Zupa pomidorowa, Krupnik gospoó ski, Pieczeń wołowa z rolna 5 mssłetn piecza; wetn, Zrazy z kaszą grzybowy, Wątróbka ciel ¿ szpikowana Z rożna, Mostki cielęce faszerowp Potrawa z pulard % kalafjorami, Legnmina z pie śliwkami, Kapłonki młode z komputera, etc.

OCR = *Optical Character Recognition*

Trudności:

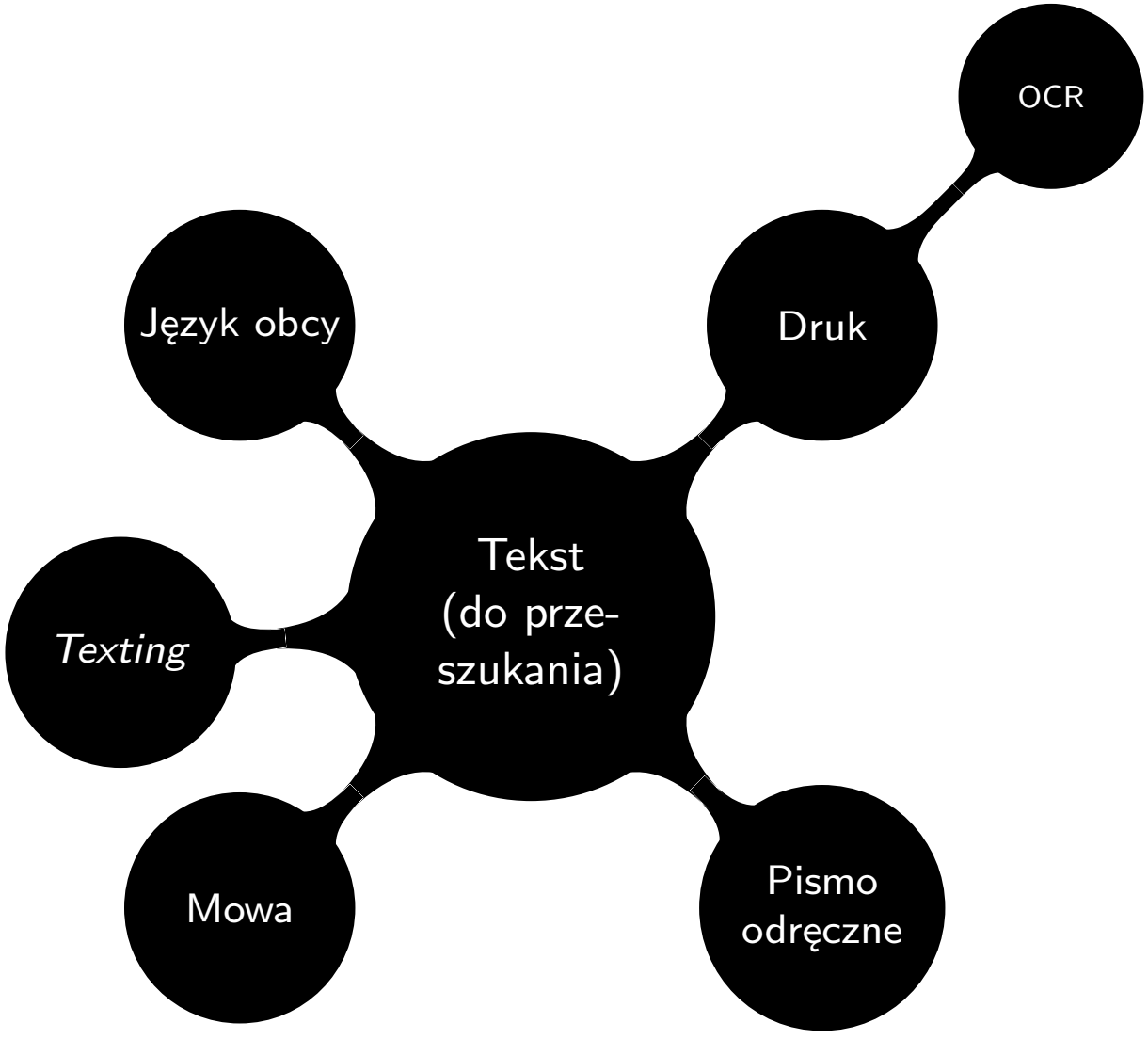
- zniekształcone strony, kleksy, pleśń, dżemik itd.
- rozpoznanie struktury strony (kolumny, ilustracje itd.)
- zmieniające się konwencje typograficzne, kroje czcionek itd.
- zmieniające się konwencje ortograficzne


. Jutro u Rogaskiego przy ulicy Długiej pod
S50, w domu dawniej Baldego, a teraz Samme
ŚNIADANIE: Zupa pomidorowa, Krupnik gospo
ski, Pieczeń wołowa z rożna z masłem piecza
wem, Zrazy z kaszą grzybową, Wątróbka ciel
szpikowana z rożna, Mostki cielęce faszerowa
Potrawa z pulard z kalafjorami, Legumina z pie
śliwkami, Kapłonki młode z kompotem, etc.

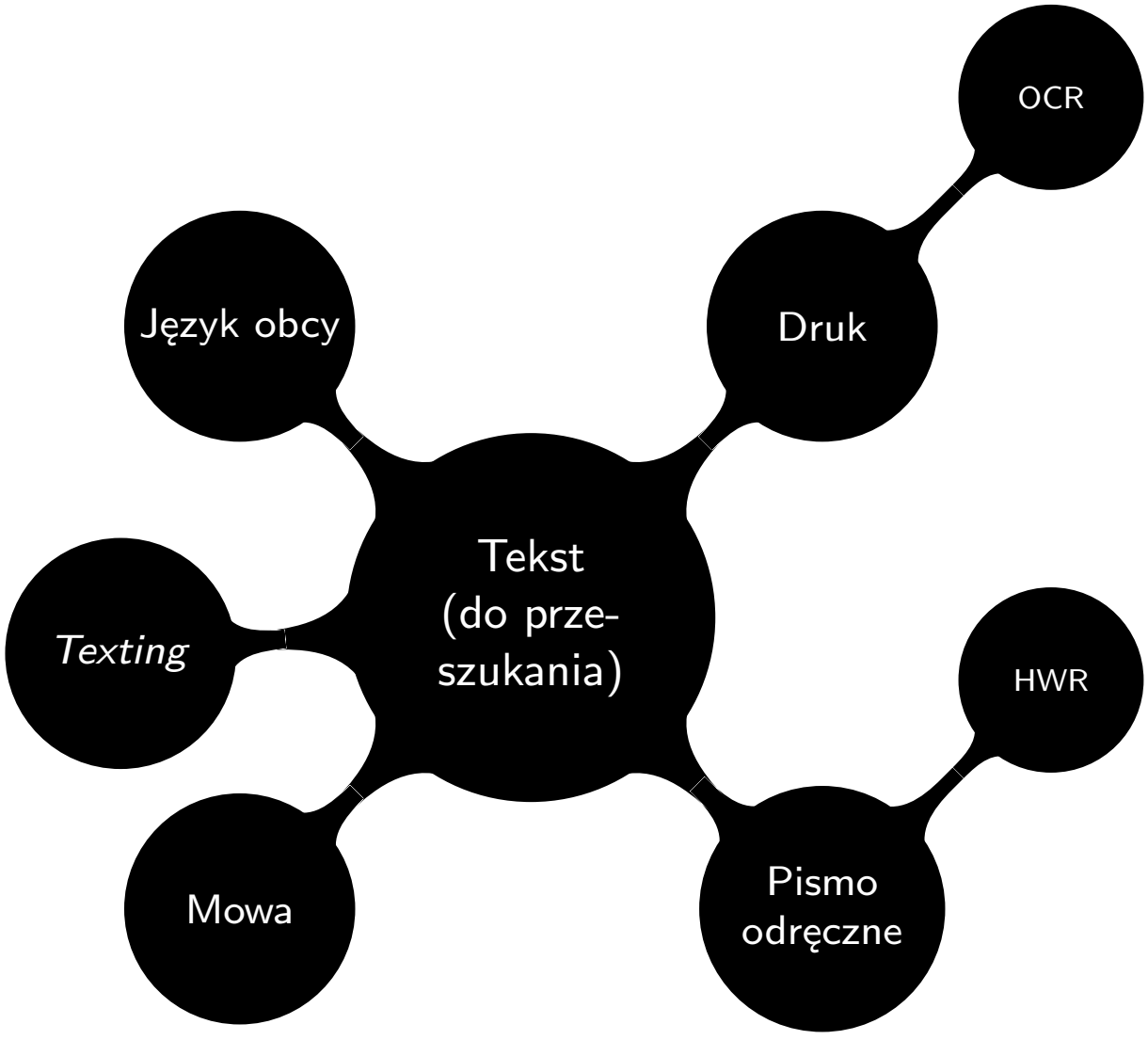
Jutro a Bogaskiego przy ulicy Długiej pod S50, w domu dawniej B aide go, a teraz Sammt ŚNIADANIE: Zupa pomidorowa, Krupnik gospoó ski, Pieczeń wołowa z rolna 5 mssłetn piecza; wetn, Zrazy z kaszą grzybowy, Wątróbka ciel ¿ szpikowana Z rożna, Mostki cielęce faszerowp Potrawa z pulard % kalafjorami, Legnmina z pie śliwkami, Kapłonki młode z komputera, etc.

Przydatne narzędzia: unpaper (usuwa niektóre zniekształecenia), tesseract (OCR), ocrodju (OCR-owanie plików DjVu)

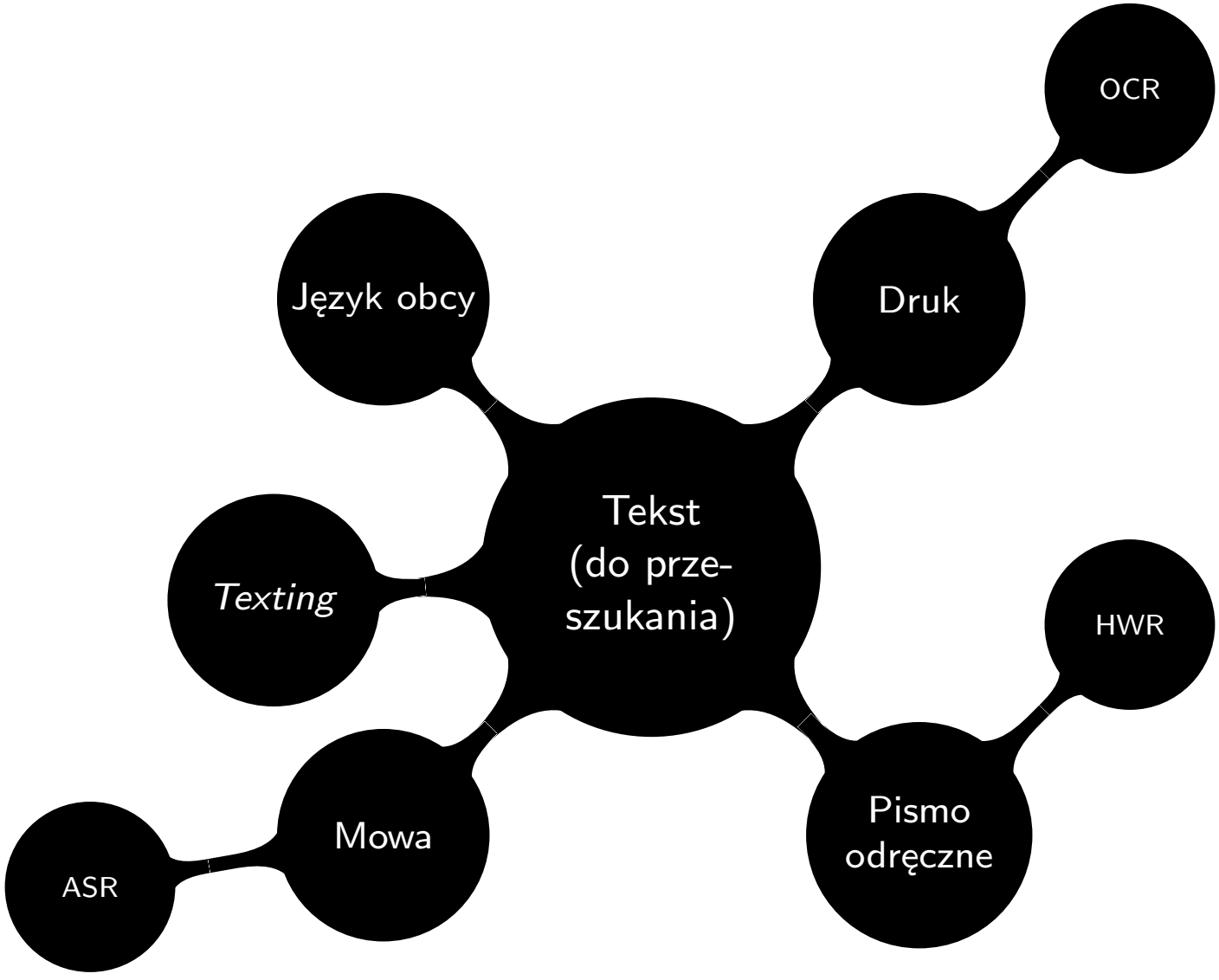
Przyjrzyjmy się bliżej czarnym skrzynkom...



Przyjrzyjmy się bliżej czarnym skrzynkom...



Przyjrzyjmy się bliżej czarnym skrzynkom...



Rozpoznawanie mowy (ASR, *Automatic Speech Recognition*)

Trudności:

- różnice w wymowie (indywidualne, dialekty, płeć itd.)

Rozpoznawanie mowy (ASR, *Automatic Speech Recognition*)

Trudności:

- różnice w wymowie (indywidualne, dialekty, płeć itd.)
- szumy tła i urządzenia, niedoskonałości mikrofonu itd.

Rozpoznawanie mowy (ASR, *Automatic Speech Recognition*)

Trudności:

- różnice w wymowie (indywidualne, dialekty, płeć itd.)
- szumy tła i urządzenia, niedoskonałości mikrofonu itd.
- nieregularności pisowni (nazwy własne obcojęzyczne!)

Rozpoznawanie mowy (ASR, *Automatic Speech Recognition*)

Trudności:

- różnice w wymowie (indywidualne, dialekty, płeć itd.)
- szumy tła i urządzenia, niedoskonałości mikrofonu itd.
- nieregularności pisowni (nazwy własne obcojęzyczne!)
- wymaga dużej mocy obliczeniowej

Rozpoznawanie mowy (ASR, *Automatic Speech Recognition*)

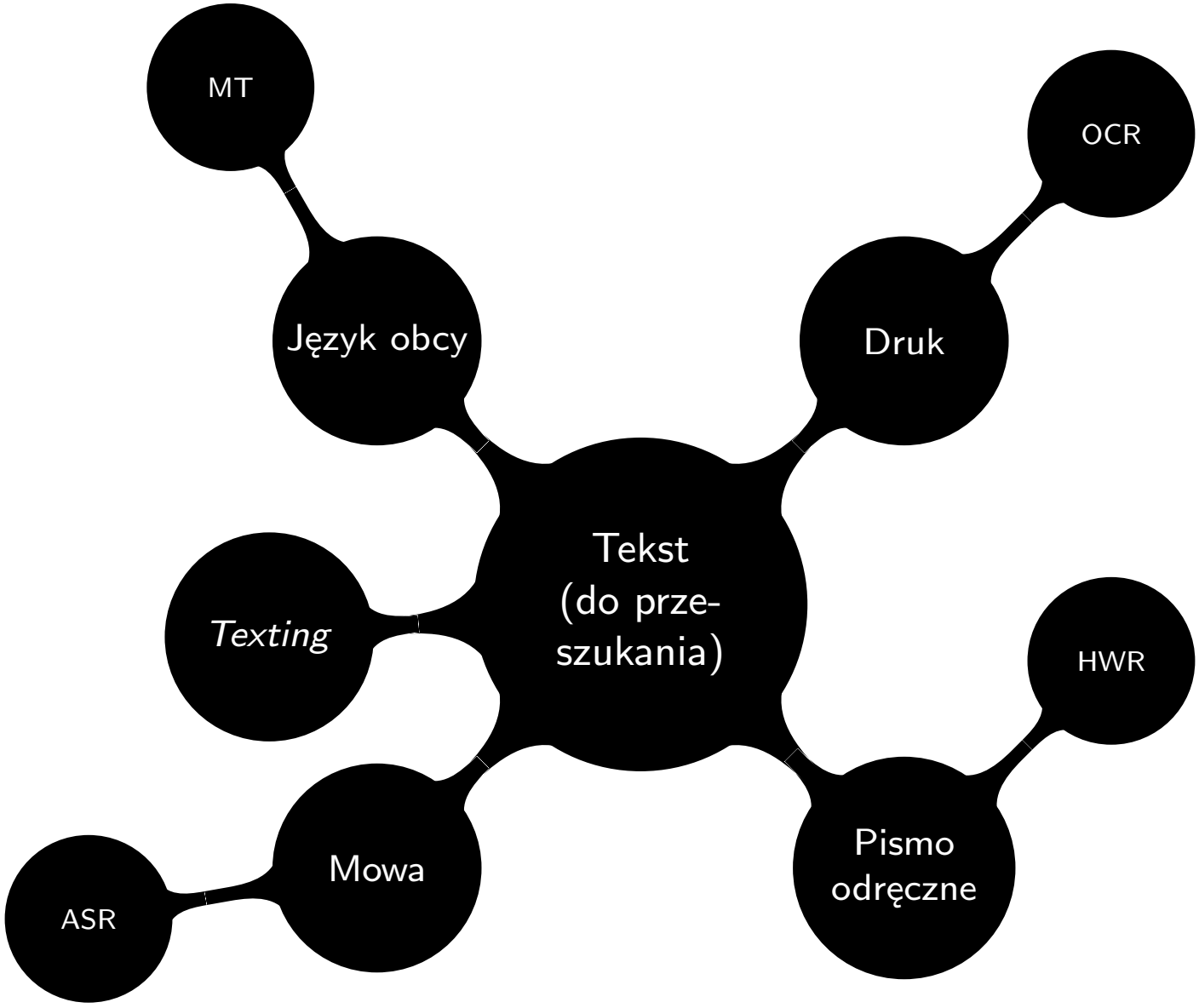
Trudności:

- różnice w wymowie (indywidualne, dialekty, płeć itd.)
- szumy tła i urządzenia, niedoskonałości mikrofonu itd.
- nieregularności pisowni (nazwy własne obcojęzyczne!)
- wymaga dużej mocy obliczeniowej

Przydatne narzędzia:

- do tworzenia systemów ASR: CMU Sphinx, Kaldi
- Google Speech API (pokątnie)

Przyjrzyjmy się bliżej czarnym skrzynkom...



Tłumaczenie maszynowe (MT, *machine translation*)

Dlaczego trudne?

- niejednoznaczności w tekście oryginalnym
  - homonimy: *powieść*, *Łódź*, *nie*, *lecz*, *albo*
  - co do część mowy: *komputerowi*, *pożółkły*, *last*
  - składniowa: *I saw a girl with a telescope*
  - semantyczna: *mysz*, *kanał*, *jack*, *table*
- niejednoznaczności między językami
  - *palec* = *finger* czy *toe* (*digit*?)
  - *rzeka* = *fleuve* czy *riviere*?
  - *siostra* = *meimei* czy *jiejie* (*jiemei*?)
  - *zobaczyłem dziewczynę* = *I saw a girl* czy *I saw the girl*?

A *texting*?

- przywracanie diakrytyków (*diacritic restoration*)
- specyficzne skróty (*LOL*, *4 U*, *3maj się*)
- specyficzne konwencje (np. hasztagi)

Zajrzyjmy do wnętrza czarnych skrzynek...

- ASR i MT – model zaszumionego kanału (*noisy channel model*)

Zajrzyjmy do wnętrza czarnych skrzynek...

- ASR i MT – model zaszumionego kanału (*noisy channel model*)
  - mowa to zniekształcona forma tekstu

Zajrzyjmy do wnętrza czarnych skrzynek...

- ASR i MT – model zaszumionego kanału (*noisy channel model*)
  - mowa to zniekształcona forma tekstu
  - tekst np. koreański jest tylko dziwnie zniekształconą formą tekstu polskiego

Zajrzyjmy do wnętrza czarnych skrzynek...

- ASR i MT – model zaszumionego kanału (*noisy channel model*)
  - mowa to zniekształcona forma tekstu
  - tekst np. koreański jest tylko dziwnie zniekształconą formą tekstu polskiego
- HWR i OCR – model zaszumionego kanału ma pewne zastosowanie, ale istotne jest przetwarzanie w 2 wymiarach

Model zaszumionego kanału

tłumaczenia / mowa

$$P(t) = P(s)P(t|s)$$

ale nas interesuje $P(s|t)$!

Piękno i wierność

$$P(e|f) = \frac{P(f|e)P(e)}{P(f)}$$

Piękno i wierność

$$P(e|f) = \frac{P(f|e)P(e)}{P(f)}$$

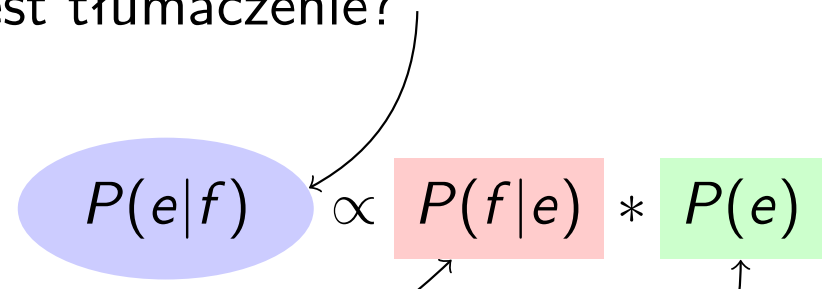
- Jak dobre jest tłumaczenie?
- Wierność

$$P(e|f) \propto P(f|e) * P(e)$$

Piękno i wierność

$$P(e|f) = \frac{P(f|e)P(e)}{P(f)}$$

- Jak dobre jest tłumaczenie?



- Wierność
- Piękno

$$\tilde{e} = \arg\max_{e \in e^*} P(f|e)P(e)$$

Modele języka

- $p(f|e)$ będzie szacowane za pomocą modelu tłumaczenia (*translation model*)
- $p(e)$ będzie szacowane za pomocą modelu języka (docelowego) (*target language model*)

Modele języka

- $p(f|e)$ będzie szacowane za pomocą modelu tłumaczenia (*translation model*)
- $p(e)$ będzie szacowane za pomocą modelu języka (docelowego) (*target language model*)

Modele języka

- $p(f|e)$ będzie szacowane za pomocą modelu tłumaczenia (*translation model*)
- $p(e)$ będzie szacowane za pomocą modelu języka (docelowego) (*target language model*)

Model języka przypisuje prawdopodobieństwa napisom.

Jeśli M ma być modelem języka polskiego, oczekiwalibyśmy, że dla napisów:

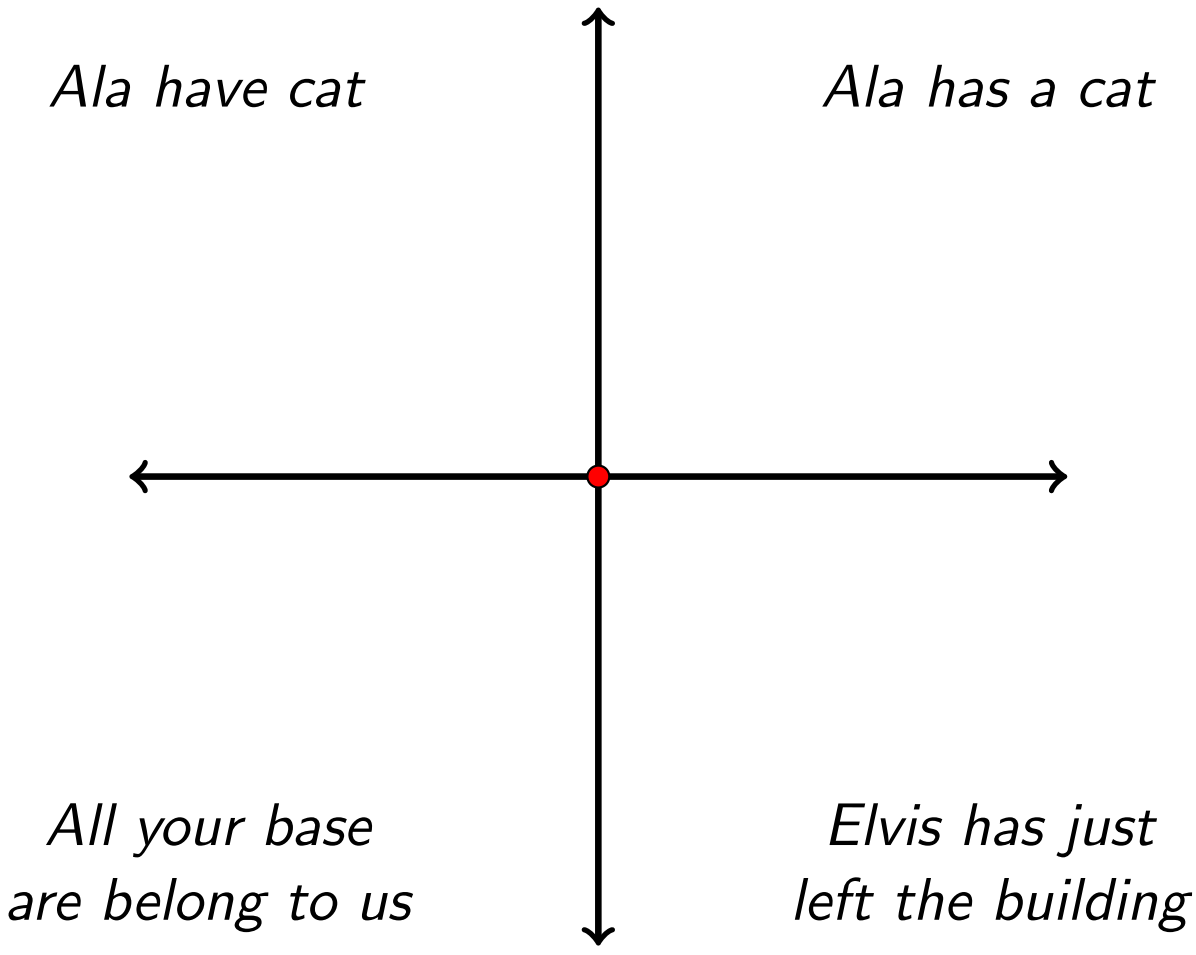
z_1 *W tym stanie rzeczy pan Ignacy coraz częściej myślał o Wokulskim.*
z_2 *Po wypełniony zbiornik pełny i należne kwotę, usłyszała w attendant*
z_3 *xxxxyźźźźźit backspace hooooooooooop x y z*

zachodzić będzie:

$$P(z_1|M) > P(z_2|M) > P(z_3|M)$$

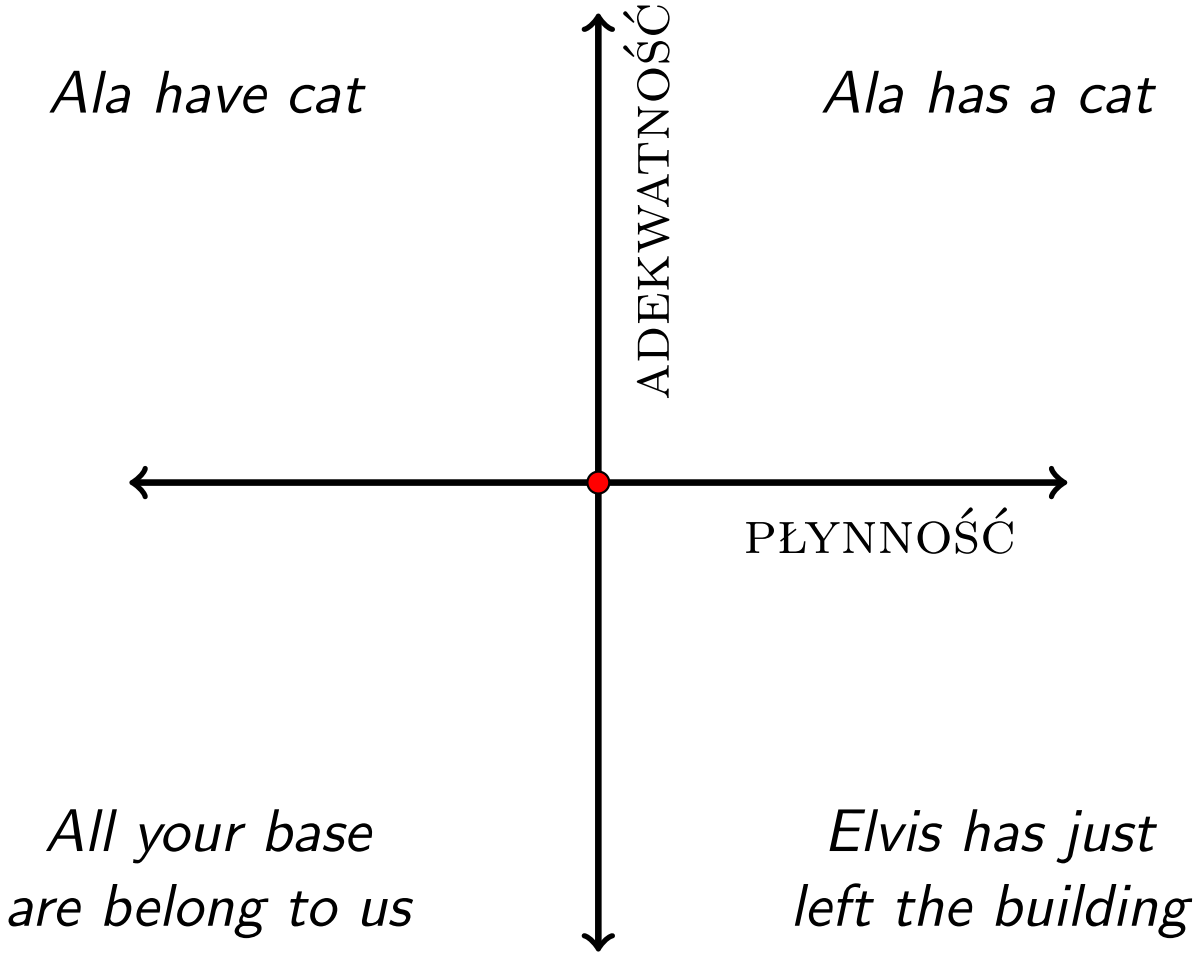
Dwa wymiary tłumaczenia

Ala ma kota = ...

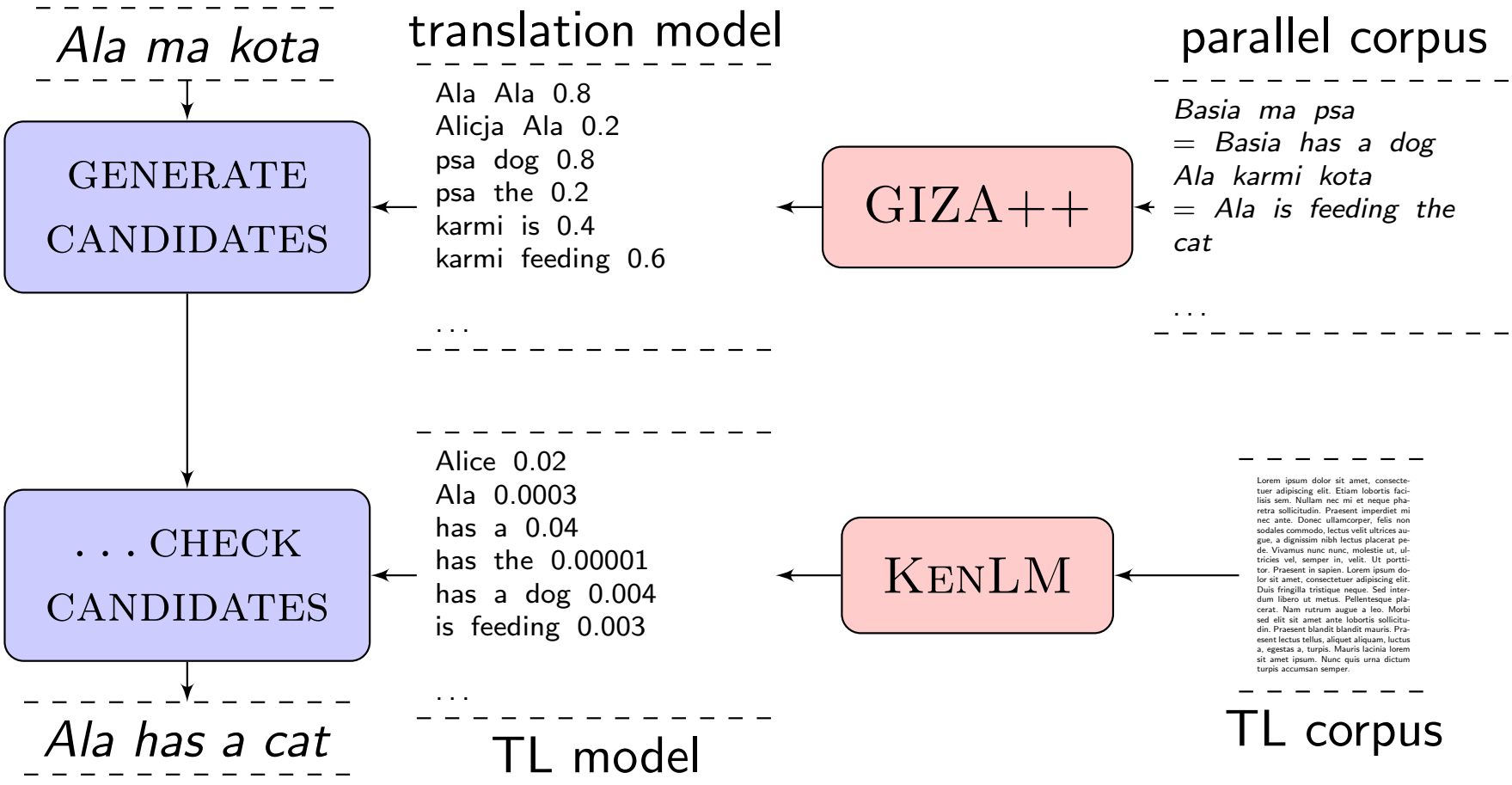


Dwa wymiary tłumaczenia

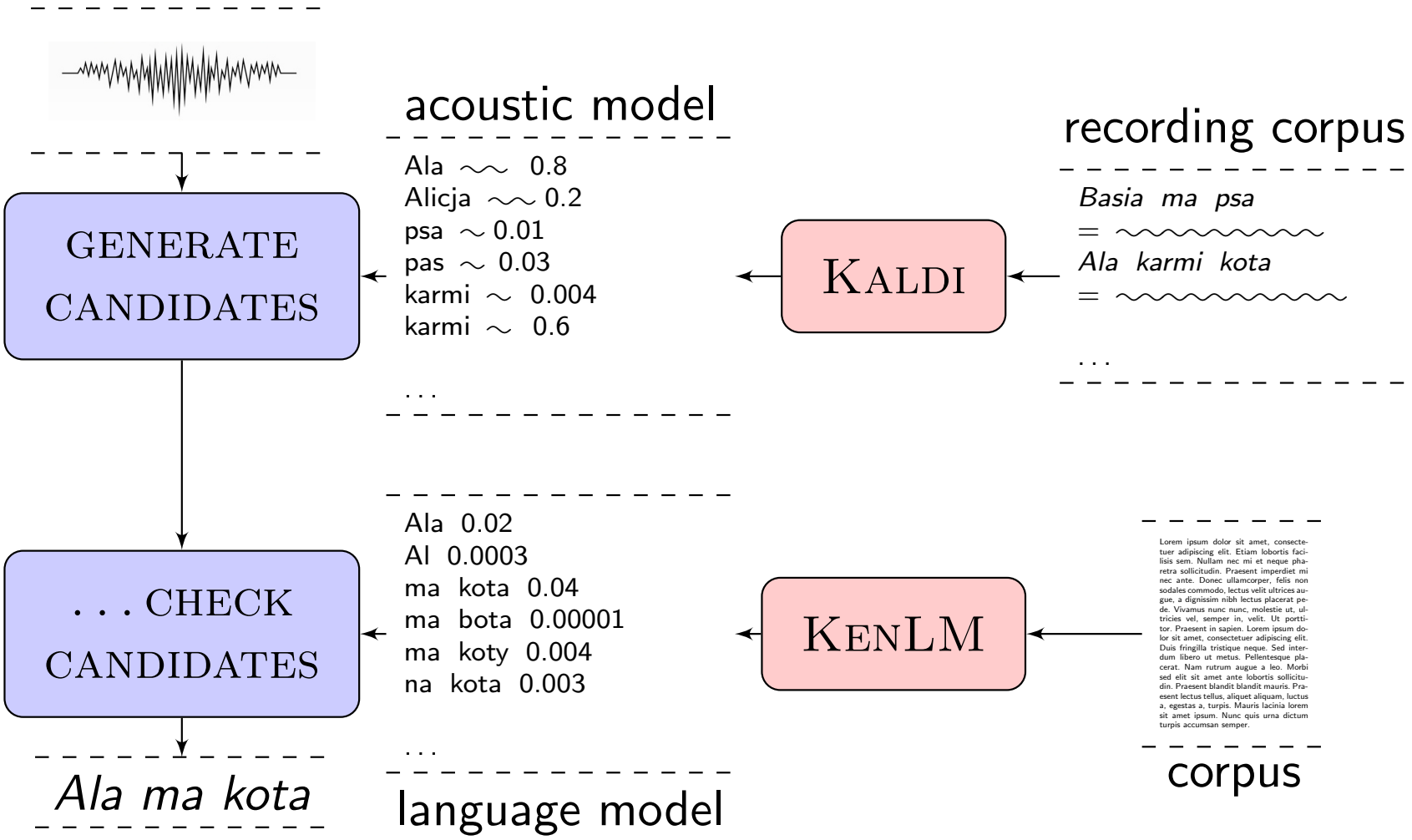
Ala ma kota = ...



Statystyczne tłumaczenia maszynowe



ASR


acoustic model
Ala ∼ 0.8
Alicja ∼ 0.2
psa ∼ 0.01
pas ∼ 0.03
karmi ∼ 0.004
karmi ∼ 0.6
...
recording corpus
Basia ma psa
=
Ala karmi kota
=
...
KALDI
GENERATE CANDIDATES
Ala 0.02
Al 0.0003
ma kota 0.04
ma bota 0.00001
ma koty 0.004
na kota 0.003
...
language model
KENLM
corpus
... CHECK CANDIDATES
Ala ma kota

Jak wyszukiwać w innych formach tekstu?

Rozwiązanie 1. Najprościej traktować OCR/HWR/ASR/MT jako czarną skrzynkę, która zwraca zwykły tekst.

Rozwiązanie 2. Brać pod uwagę *n-best list*, nie tylko pierwszą odpowiedź.

lista *n* najlepszych (*n-best list*)

Przykładowa *4-best list* (wyjście z ASR-a):

	możliwa interpretacja	**prawdopodobieństwo**
1.	*wojewódzki narodowy o powołanie komisji ochrony przyrody krajowej*	0.5219596
2.	*rejowa wojewódzki narodowy o powołanie komisji ochrony przyrody krajowej*	
3.	*prawa wojewódzki narodowy o powołanie komisji ochrony przyrody krajowej*	
4.	*wojewódzki narodowy o powołanie komisji ochrony przyrody krajowe*	

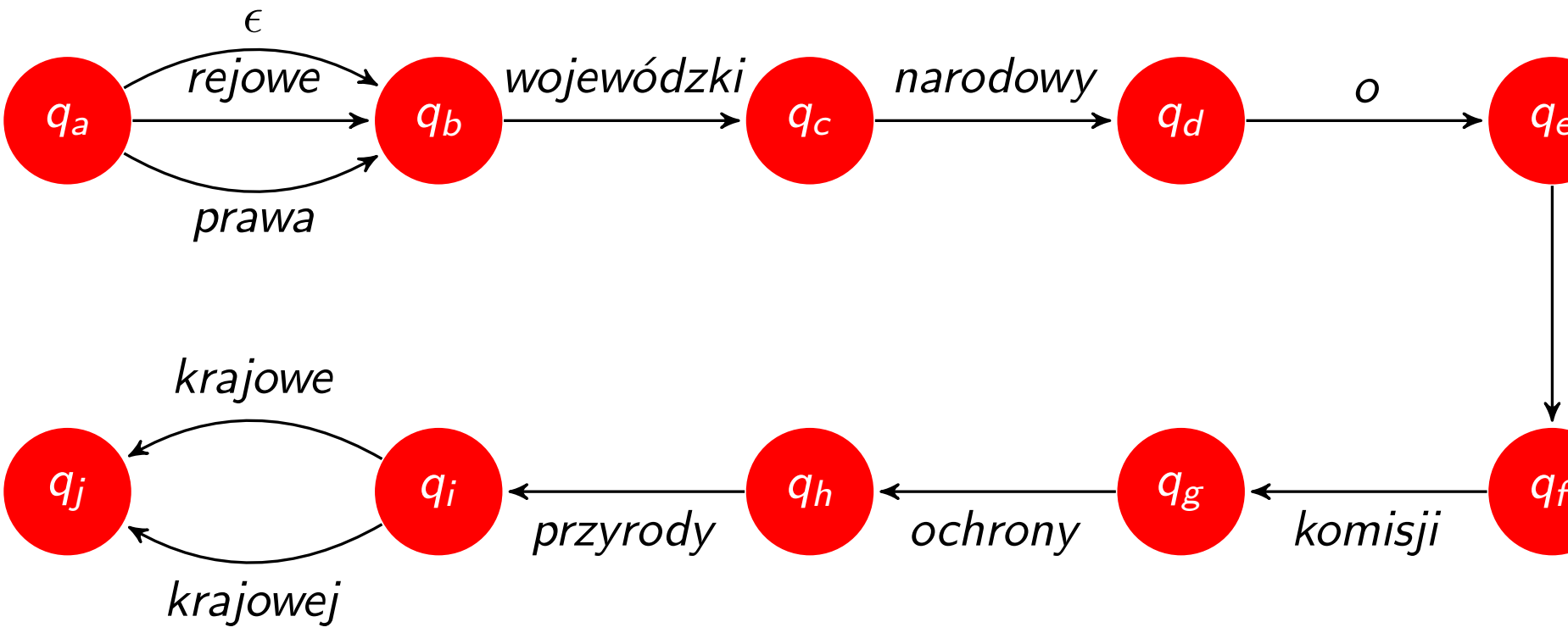
Jak wyszukiwać w innych formach tekstu?

Rozwiązanie 1. Najprościej traktować OCR/HWR/ASR/MT jako czarną skrzynkę, która zwraca zwykły tekst.

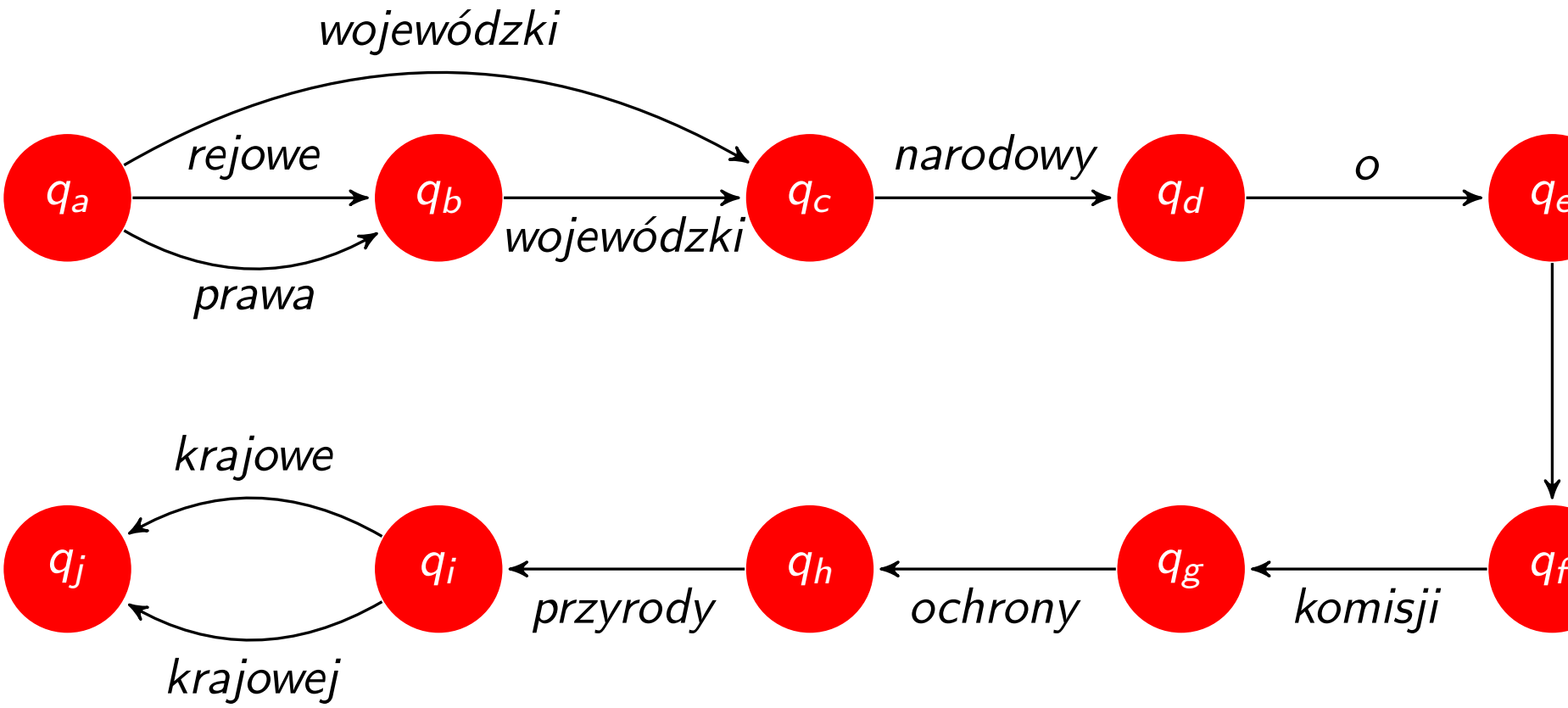
Rozwiązanie 2. Brać pod uwagę *n-best list*, nie tylko pierwszą odpowiedź.

Rozwiązanie 3. . . . albo bardziej zaawansowane sposoby reprezentacji wyjścia – „kiełbaski” albo kraty.

„Kiełbaski” (*word confusion networks*)



Kraty (*lattices*)



Jak wyszukiwać w innych formach tekstu?

Rozwiązanie 1. Najprościej traktować OCR/HWR/ASR/MT jako czarną skrzynkę, która zwraca zwykły tekst.

Rozwiązanie 2. Brać pod uwagę *n-best list*, nie tylko pierwszą odpowiedź.

Rozwiązanie 3. . . . albo bardziej zaawansowane sposoby reprezentacji wyjścia – „kiełbaski” albo kraty.

Rozwiązanie 4. Ściśle sprząc modele OCR/HWR/ASR/MT i wyszukiwania

Przykład: wyszukiwanie lokalne (*local search*)

$$\hat{L} = \arg\max_{L} P(L)P(C|L)P(Q|L)P(Q|X)$$

- L – miejsce
- C – gdzie przebywa użytkownik
- Q – zapytanie (tekst)
- X – sygnał dźwiękowy

Jak wyszukiwać w innych formach tekstu?

Rozwiązanie 1. Najprościej traktować OCR/HWR/ASR/MT jako czarną skrzynkę, która zwraca zwykły tekst.

Rozwiązanie 2. Brać pod uwagę *n-best list*, nie tylko pierwszą odpowiedź.

Rozwiązanie 3. . . . albo bardziej zaawansowane sposoby reprezentacji wyjścia – „kiełbaski” albo kraty.

Rozwiązanie 4. Ściśle sprząc modele OCR/HWR/ASR/MT i wyszukiwania. . .

Rozwiązanie 5. . . . i trenować wszystkie moduły dyskryminacyjnie.

Trenowanie dyskryminacyjne – na przykład ASR jest trenowany pod kątem jak najlepszych wyników wyszukiwarki, której jest częścią, nie zmniejszenia swojej stopy błędów.