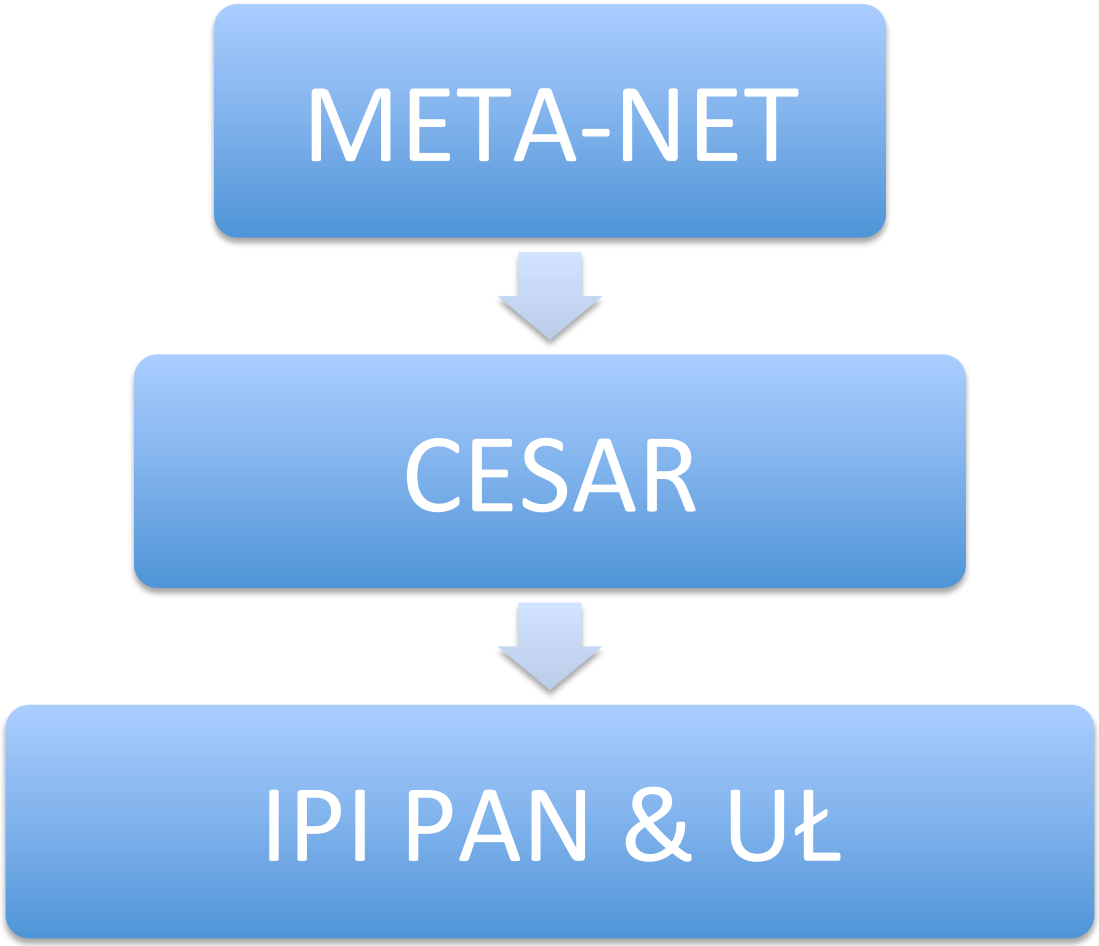
Polskie korpusy równoległe i zasoby wielojęzyczne w projekcie CESAR

Piotr Pęzik

Uniwersytet Łódzki

META-NET



Polskie zasoby w repozytorium META-NET

- Anotowane **korpusy równoległe** i korpusy języka mówionego
- Słowniki morfologiczne**, słowniki kolokacji**, leksykony
- Taggery, narzędzia NER

Zasoby wielojęzyczne

- Korpusy równoległe
- Słowniki kolokacji wygenerowane z NKJP i BNC
- Celem projektu CESAR jest zwiększenie ich dostępności (nowe zasoby, swobodne licencje) oraz interoperacyjności (standardy anotacji)

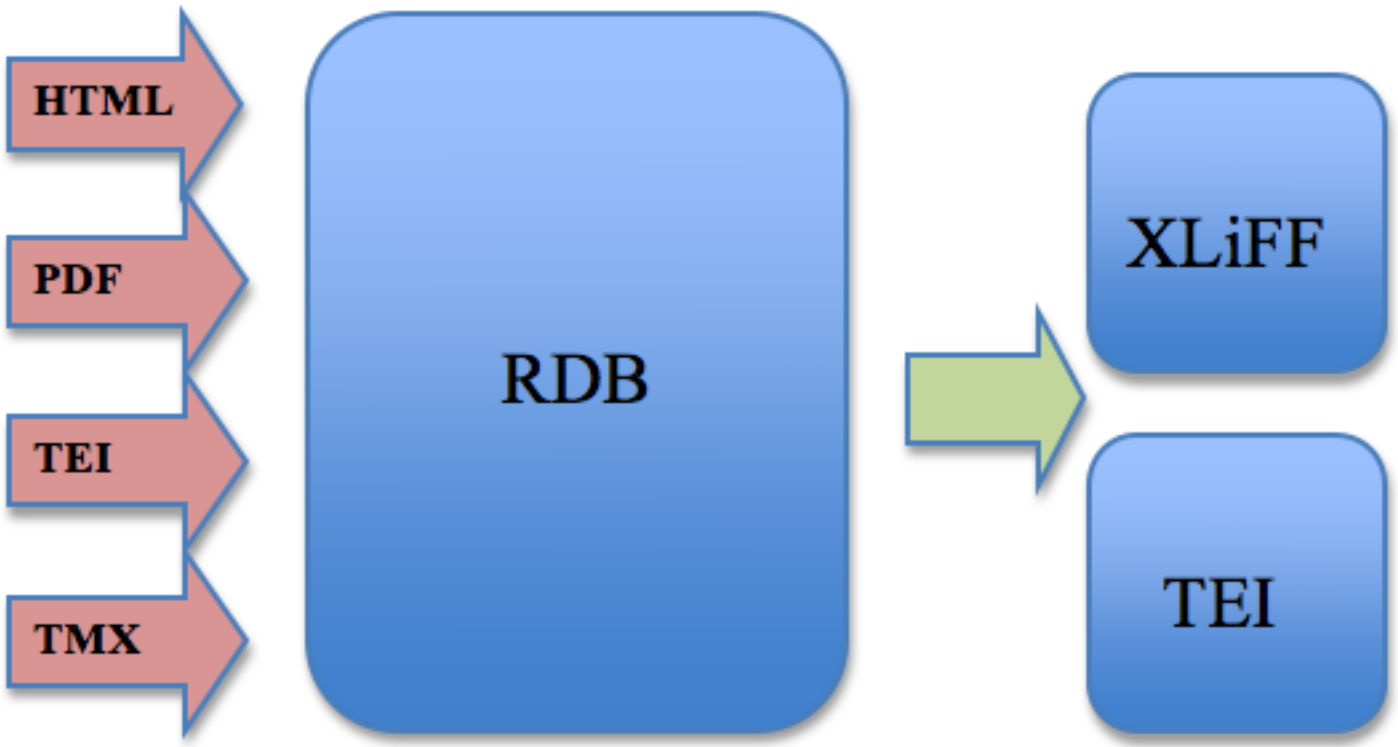
Polskie korpusy równoległe

- Stosunkowo dużo potencjalnych źródeł danych, ale mocno rozproszone
- Bardzo różnorodny poziom i formaty anotacji (od czystego tekstu do bogatej anotacji, np. TEI)
- Niska reprezentatywność gatunkowa (głównie rejestr tekstów prawnych i urzędowych oraz pamięci tłumaczeniowe oprogramowania)

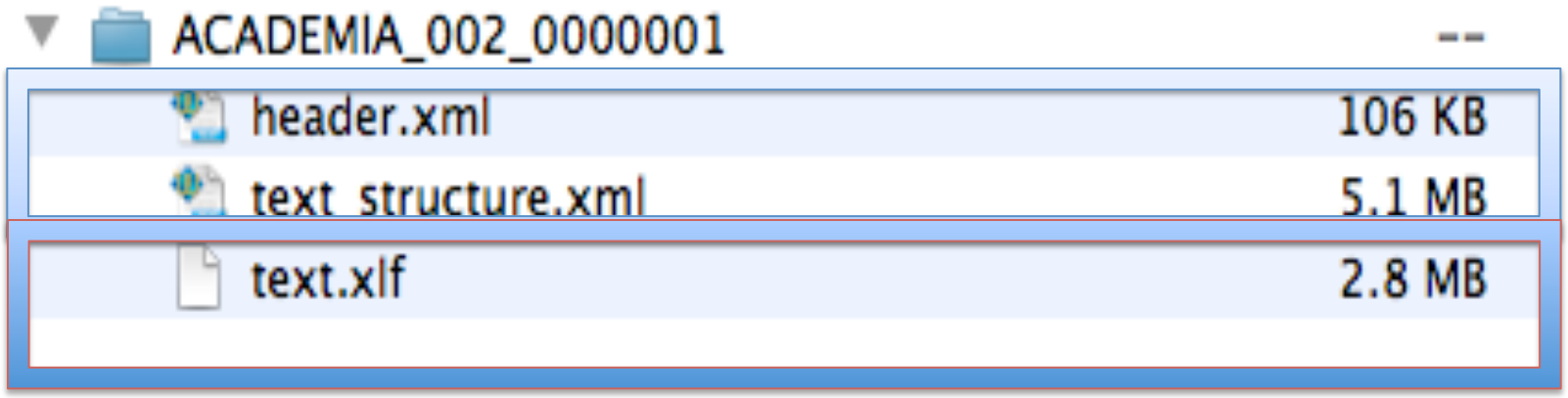
Korpusy równoległe

- W ramach projektu CESAR:
  - zwiększamy pulę polskich korpusów równoległych poprzez pozyskiwanie nowych danych
  - wzbogacamy anotację bibliograficzną i strukturalną istniejących korpusów
  - uwalniamy korpusy na możliwie swobodnych licencjach

Pozyskiwanie danych



Formaty dystrybucji



XLiFF, TMX, TEI (P5)

- TEI
  - Standard, bogaty zbiór wytycznych, dobra dokumentacja
  - Dobra obsługa poziomów anotacji strukturalnej, lingwistycznej i bibliograficznej
  - Możliwość anotacji niestandardowych przypadków ekwiwalencji tłumaczeniowej
  - Interoperacyjność?
- XLiFF
  - Nowy standard branży lokalizacyjnej
  - Lepsza niż w TMX możliwość anotacji formatowania dokumentu
  - Możliwość wykorzystania korpusów jako pamięci tłumaczeniowych w systemach CAT
- TMX
  - Obsługiwany przez większość CAT-ów
  - Stopniowo zastępowany przez XLiFF, choć nadal zdecydowanie popularniejszy

XLiFF

```
<?xml version="1.0" encoding="UTF-8"?>
<xliff xmlns:None="urn:oasis:names:tc:xliff:document:1.2" version="1.2">
  <file source-language="pl" target-language="en" datatype="plaintext"
  original="C20061212#79">
    <body>
      <!--...-->
      <trans-unit id="5">
        <source>Rośliny żyją dzięki światłu, choć różne gatunki wykorzystują je
        w odmienny sposób.</source>
        <target>All plants need light to stay alive, although various species
        harness light in different ways.</target>
      </trans-unit>
      <!--...-->
```

Nagłówek TEI P5

```
<?xml version="1.0" encoding="UTF-8"?>
<teiHeader xmlns:None="http://www.tei-c.org/ns/1.0" xmlns:pelcra="http://pel
  <fileDesc>
    <titleStmt>
      <title>Manual alignment of PAS Academia texts.</title>
    </titleStmt>
    <publicationStmt>
      <pubPlace>Łódź, Poland</pubPlace>
      <address>
        <addrLine>PELCRA</addrLine>
        <addrLine>Faculty of Philology</addrLine>
        <addrLine>University of Łódź</addrLine>
        <addrLine>al. Kościuszki 65</addrLine>
        <addrLine>90-514 Łódź</addrLine>
        <addrLine>Poland</addrLine>
        <addrLine>tel. (+48 42) 6655220</addrLine>
        <addrLine>
          <email>PELCRA</email>
        </addrLine>
        <addrLine>
          <ref target="http://pelcra.pl" n="www">http://pelcra.pl</ref>
        </addrLine>
      </address>
      <publisher>PELCRA</publisher>
      <distributor>PELCRA</distributor>
      <date when="2011-09-08T23:09:39">2011-09-08 23:09:39</date>
      <availability status="free">
        <p>public domain</p>
      </availability>
    </publicationStmt>
```

Nagłówek TEI P5

```
<bibl xml:id="bibl-38430">
    <ptr target="#text-38430"/>
    <relatedItem xml:lang="pl" type="original">
      <bibl>
        <title level="a">Na świetle i w mroku</title>
        <title level="j">Academia</title>
        <author>Jan Pilarski</author>
        <date type="acquired" when="2011-09-08T23:06:57">2011-09-08
        23:06:57</date>
        <ref type="display" target="ciemnosc_pilarski.pdf"/>
      </bibl>
    </relatedItem>
    <relatedItem xml:lang="en" type="translation">
      <bibl>
        <title level="a">Light After Darkness</title>
        <title level="j">Academia</title>
        <author>Jan Pilarski</author>
        <date type="acquired" when="2011-09-08T23:06:57">2011-09-08
        23:06:57</date>
        <ref type="display" target="darkness_pilarski.pdf"/>
      </bibl>
    </relatedItem>
  </bibl>
```

Tekst równoległy TEI P5

```
<text xml:id="text-38430" decls="#bibl-38430">
 <body>
 <div type="original" xml:lang="pl">
    <!--...-->
        <div type="alignment_unit" subtype="sentence" xml:id="div-6"
        corresp="#div-7" pelcra:alignment-score="0.999">
            <ab>
                <s xml:id="s-5863505" pelcra:seq="5">Rośliny żyją dzięki
                światłu, choć różne gatunki wykorzystują je w odmienny
                sposób.</s>
            </ab>
        </div>
    <!--...-->
</div>
<div type="translation" xml:lang="en">
      <!--...-->
      <div type="alignment_unit" subtype="sentence" xml:id="div-7"
      corresp="#div-6" pelcra:alignment-score="0.999">
        <ab>
            <s xml:id="p-5863574" pelcra:seq="5">All plants need light
            to stay alive, although various species harness light in
            different ways.</s>
        </ab>
    </div>
    <!--...-->
</div>
```

Dodatkowa anotacja segmentów tłumaczeniowych w TEI

```
<linkGrp>
    <link target="#div-11 #div-15" type="merge"/>
    <link target="#div-12 #div-16" type="split"/>
    <link target="#div-13 #div-17" type="complex"/>
    <link target="#div-14 #div-18" type="simple"/>
</linkGrp>
```

TEI vs XLiFF

- XLiFF ma umożliwiać zachowanie formatowania dokumentów
- XLiFF jest bardziej interoperacyjny?

- TEI zdecydowanie bardziej nadaje się do tworzenia anotowanych korpusów
- Interoperacyjność TEI?

Pierwsza transza polskich korpusów równoległych

Korpus	Pary języków	Poziom zrównoleglenia	Format źródłowy	Segmenty słów	Licencja
Academia PAN	1	Zdania (ręcznie)	PDF	320K PL 386K EN	CC-BY-NC
CORDIS	1 (5)	Zdania (automatycznie)	HTML	3 800K PL 4 150K EN	CC-BY
RAPID	1 (21)	Zdania (automatycznie)	HTML	3 250K PL 3 540K EN	CC-BY
JRC Acquis Communautaire	1 (21)	Zdania (automatycznie)	TEI	28 571K PL 32 447K EN	CC-BY

http://pelcra.pl/corpora
http://bach.ipipan.waw.pl/metashare/

Słowniki kolokacji

- Słownik wybranych kolokacji angielskich na podstawie korpusu BNC
- Słownik wybranych kolokacji polskich na podstawie NKJP
- Na razie są to słowniki kolokacji dwuwyrazowych, zasoby porównywalne, a nie równoległe
- W ramach CESAR-a zostaną udostępnione w postaci źródłowej
- Wersja demo BNC: http://212.191.73.200/ColosaurusWeb/Browser
- Zastosowania:
  - leksykograficzne, językoznawcze, kulturoznawcze
  - uzupełnienie leksykalnych baz danych o utrwalone relacje syntagmatyczne między wyrazami

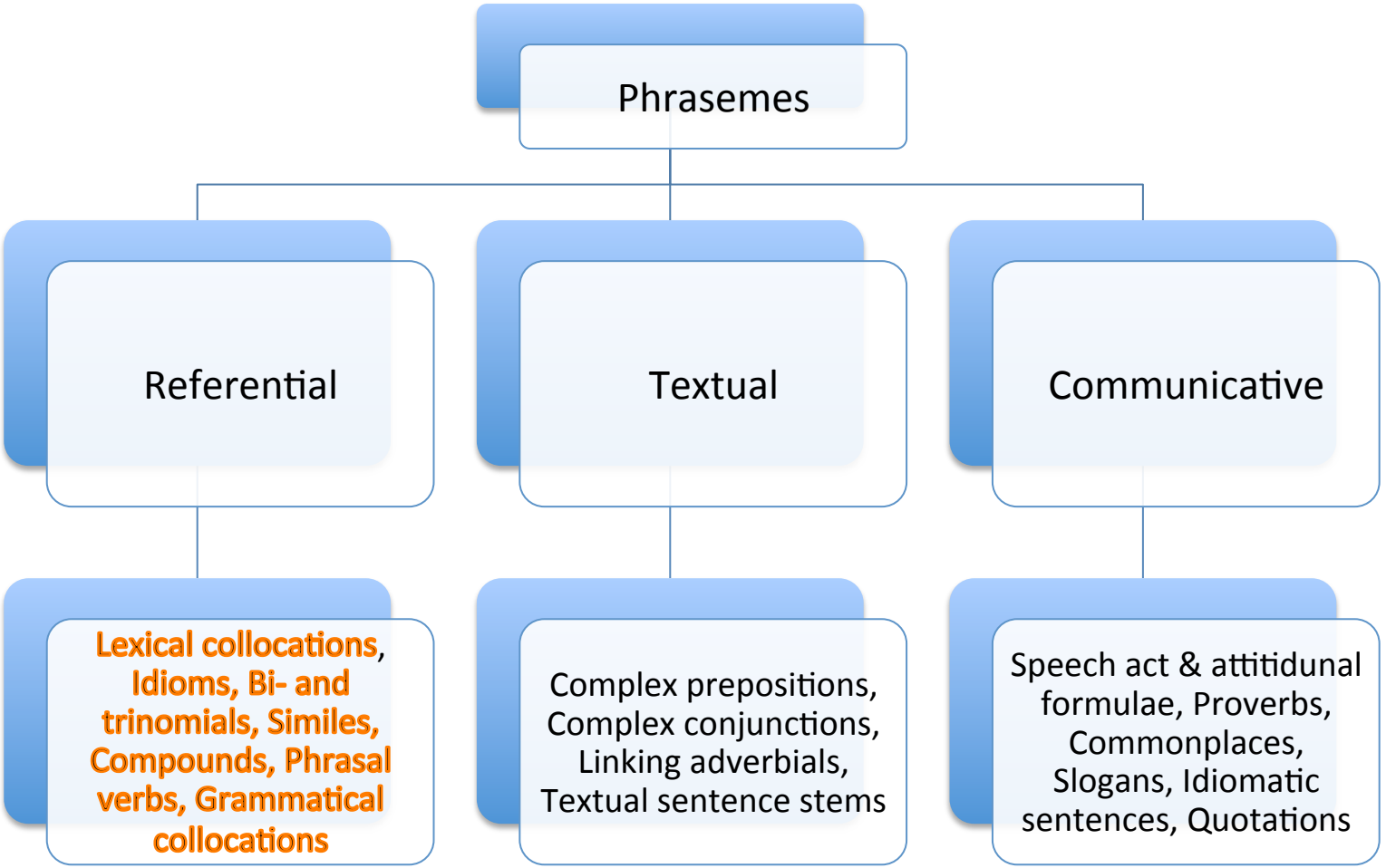
Słownik kolokacji a słowosieci

- Przymiotniki *frightening* i *alarming* znajdujemy w tym samym zbiorze synonimów angielskiego Wordnetu

- Tworzą one jednak różne związki frazeologiczne: http://tinyurl.com/7eleweu

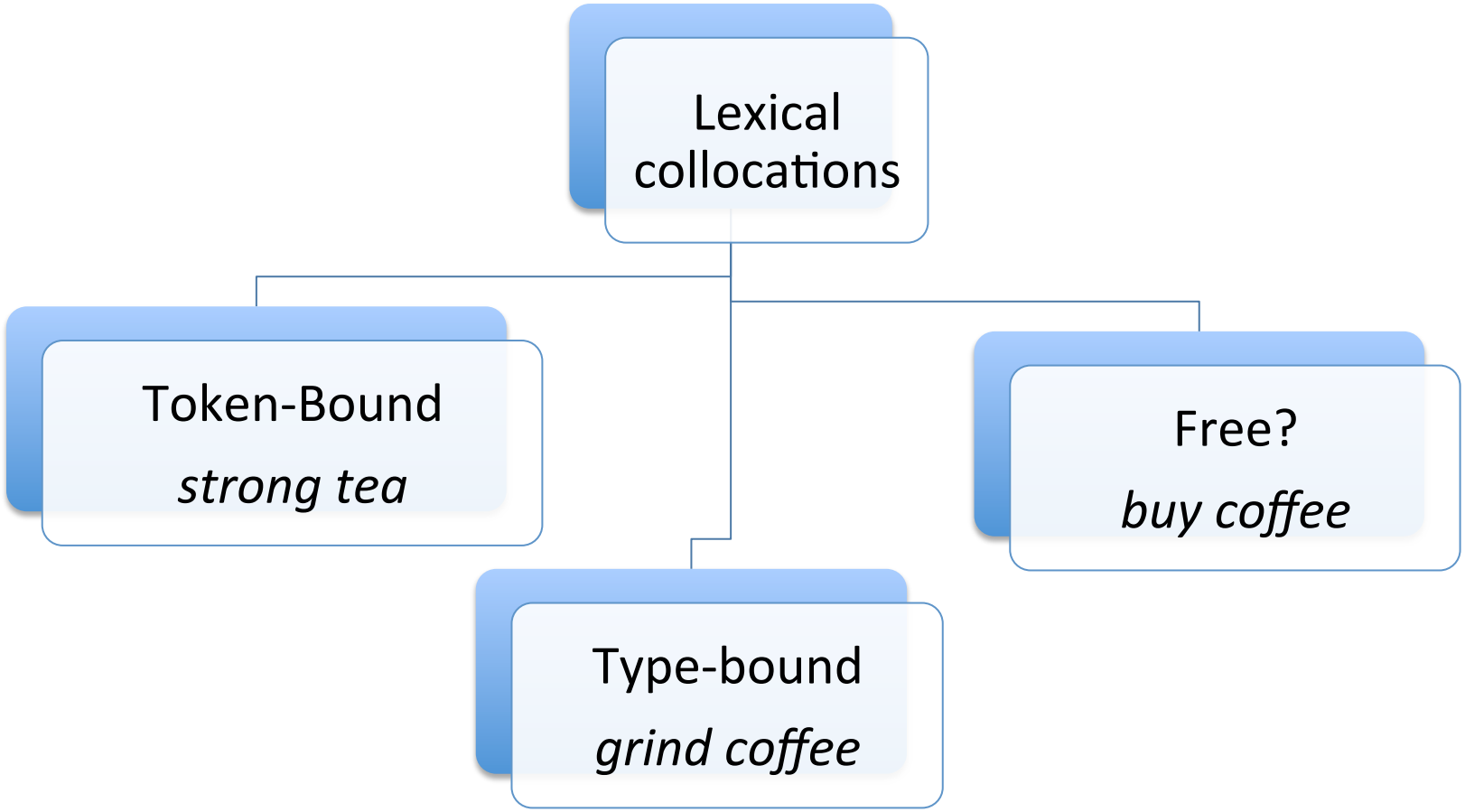
- Ich występowanie w tym samym zbiorze synonimów nie oznacza iż są całkowicie wymienne tekstowo

Spektrum frazeologiczne



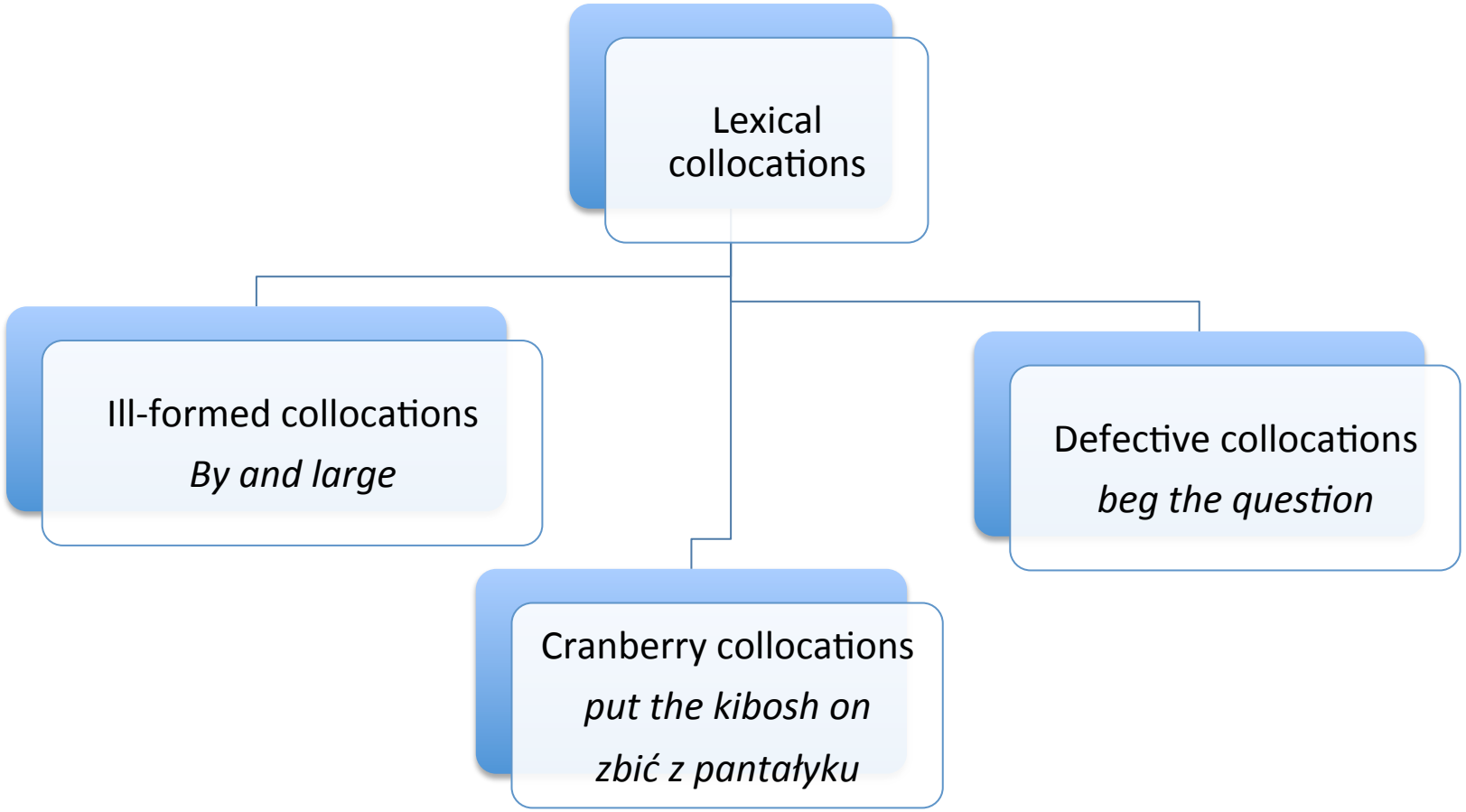
(Burger 1998), (Granger & Paquot 2004)

Typy kolokacji leksykalnych I



(Martin 2004)

Typy kolokacji II



(Moon 1998)

Badanie statusu frazeologizmu

- Cechy dystrybucyjne połączeń wyrazowych nie zawsze pozwalają wykryć jego status frazeologiczny
- „W różnych opracowaniach liczba kryteriów jednostki frazeologicznej waha się od 5-7 do niemal 20” (Chlebda 1991, 2003)
- Ważnym kryterium frazeologicznych jest jednak odtwarzalność i dlatego identyfikacja wybranych kolokacji na podstawie dużych korpusów jest możliwa

Ekstrakcja słowników kolokacji dwuwyrazowych

- Zakłada dostępność anotowanego na poziomie części mowy korpusu referencyjnego, np. BNC, czy też NKJP
- Reguły relacyjno-pozycyjne określające konteksty typu:
  - przymiotnik z rzeczownikiem w odległości 1 segmentu wyrazowego,
  - rzeczownik z czasownikiem w odległości 4 segmentów wyrazowych,
  - przysłówek z przymiotnikiem (1 segment) itd.

Na przykład dla każdego rzeczownika w korpusie BNC i NKJP rozpatrywane są konteksty:

1	2	3	4	5	…
AT0	AJ0	NN1	PRF	NP0	…
The	rapid	growth	of	ACET	…

Ekstrakcja słowników kolokacji dwuwyrazowych

- Reguły mają charakter pseudo-zależnościowy, tzn.:
  - wystąpienie rzeczownika w pobliżu czasownika nie musi oznaczać, że występuje między nimi zakładana zależność składniowa, co powoduje obniżenie precyzji (np.: koordynacja NKJP)
  - trudno przewidzieć wszystkie konfiguracje pozycji dla danej kolokacji, co powoduje obniżenie zwrotu (recall), np.
    - This type of religious **experience** is often **traumatic**.
    - Going to prison in unquestionably a **traumatic experience** .
- Być może lepszą metodą byłoby użycie zależnościowej anotacji składniowej, choć jej dokładność mogłaby zaważyć na wynikach
- Używane są ujednoznacznione morfologicznie, słownikowe formy wyrazów
- Ekstrakcja z korpusów BNC i NKJP

Kryteria dystrybucyjne

- Dla współwystąpień wyrazów w zadanych kontekstach wyliczane są:
  - Miary powiązania
  - Miary równomierności występowania

Miary powiązania

a ^ b	**a ^ ~b**
~a ^ b	**~a ^ ~b**

Miary powiązania

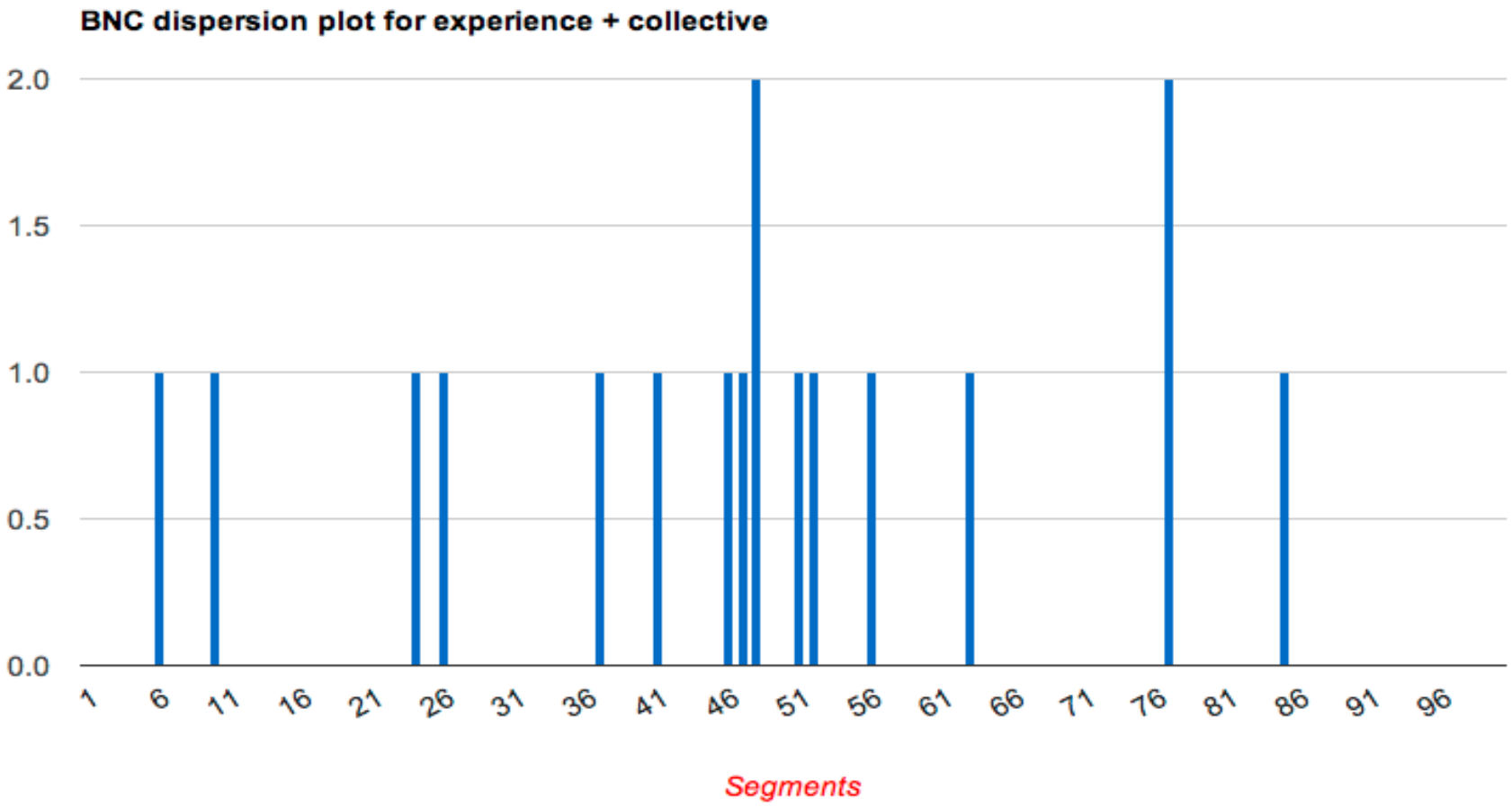
Experience									
#	**Collocate**	**POS**	**A**	**TTEST**	**MI3**	**G2**	**MI**	**LOGLOG**	**CHISQ**
1	personal	AJ%	306.0	12.99	18.47	384.2	1.95	16.16	661.08
2	past	AJ%	225.0	12.36	18.13	419.93	2.50	19.58	873.55
3	practical	AJ%	209.0	12.06	18.01	412.11	2.59	20.00	885.04
4	previous	AJ%	221.0	11.17	17.58	289.9	2.01	15.65	506.90
5	first-hand	AJ%	75.0	8.56	18.89	563.4	6.44	40.11	6396.3

Miary rozproszenia

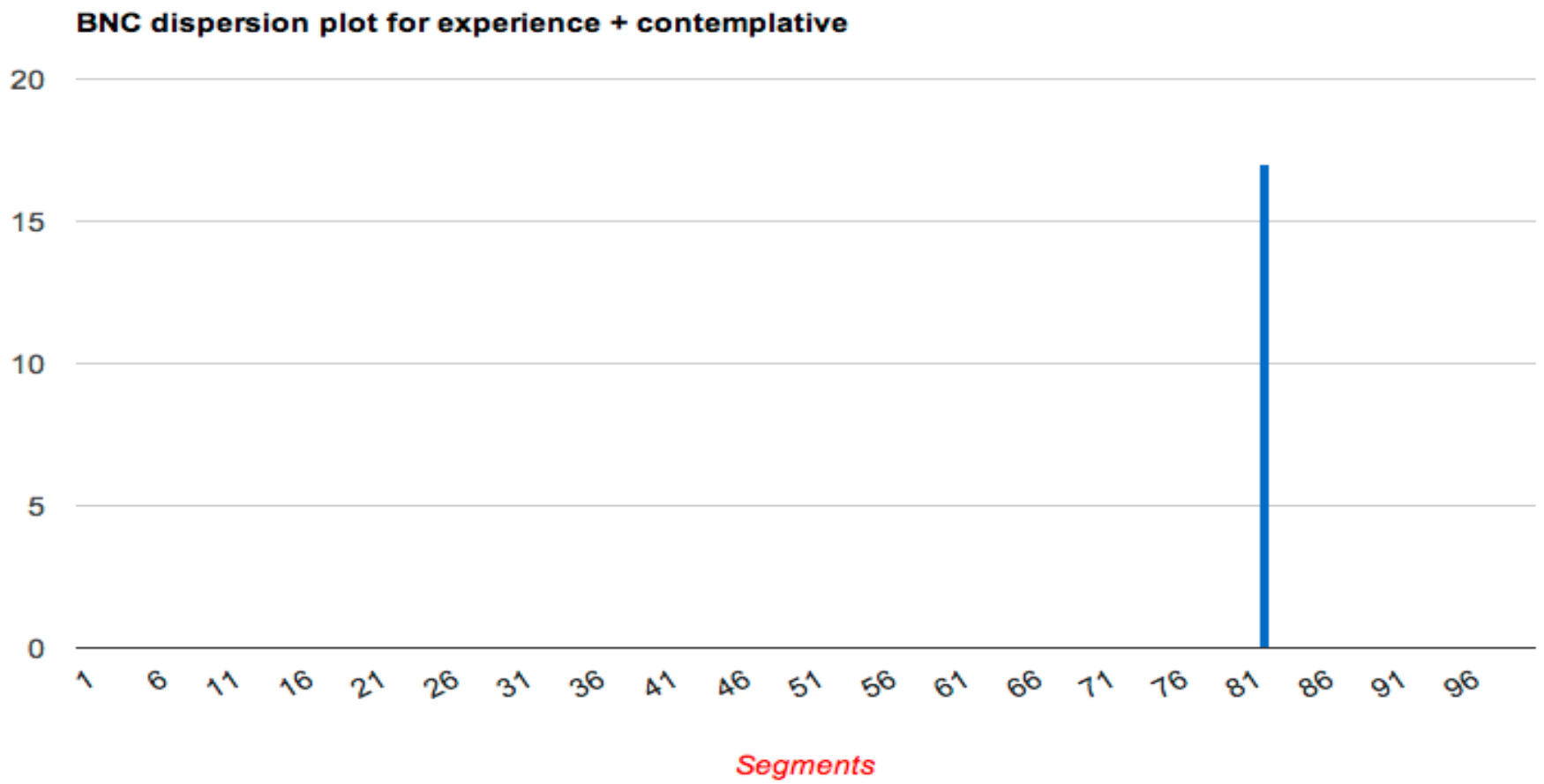
Experience							
#	**Collocate**	**POS**	**A**	**JD**	**R**	**AWT***	**fAWT***
1	personal	AJ%	306.0	0.91	85.0	527921	102.9
2	past	AJ%	225.0	0.91	81.0	666577	81.5
3	practical	AJ%	209.0	0.88	75.0	846144	64.2
4	previous	AJ%	221.0	0.89	71.0	967476	56.1
5	first-hand	AJ%	75.0	0.86	46.0	1982715	27.4

*(P. Savicky, J. Hlavacova 2003)

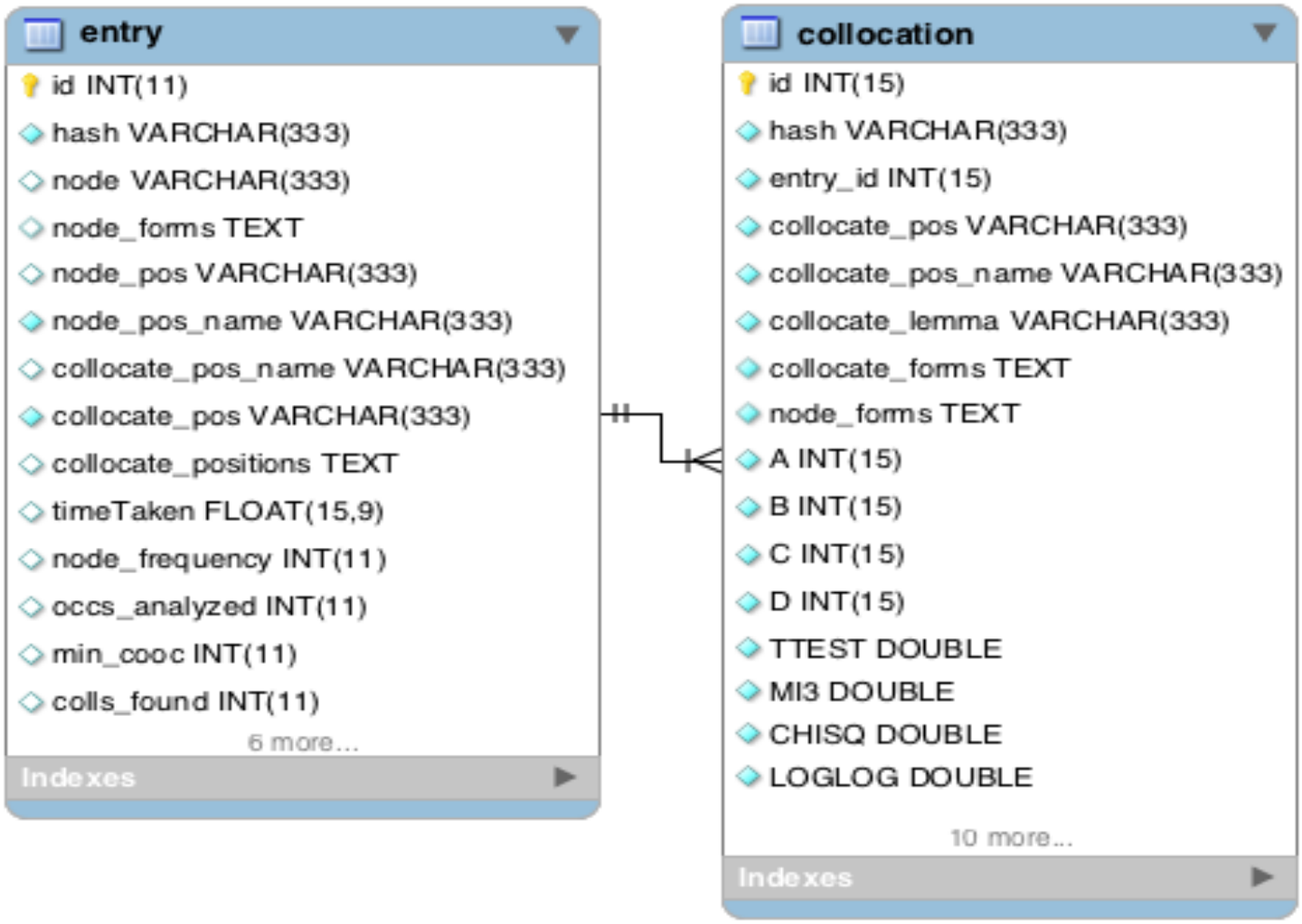
collective experience



contemplative experience



Podstawowe tabele słownika



Ewaluacja

- Trwają prace nad opracowaniem zbioru treningowego i testowego
- Nieostre kryteria związków frazeologicznych
- Zwrot (recall) trudny do ustalenia w dużych korpusach
- Precyzja dość wysoka
- Skuteczność jest różna dla różnych kombinacji kontekstów i zastosowanych miar, np. dla kolokacji czasownikowych ważna jest stosunkowo duża wartość T-test
- Potrzebne ustalenie optymalnych kombinacji wartości miar dla poszczególnych typów kolokacji

Identyfikacja kolokacji jako klasyfikacja współwystąpień

#	Feature	Value	Value	Value	Value
Instance features	rule	AJ+N	V+N	AJ+N	…
	collocate_pos	AJ%	N%	AJ%	…
	node_pos	N%	V%	N%	…
	A	44	20	10	…
	TTest	2.75	2.72	0.11	…
	MI	10.79	4.79	7.2	…
	JD	0.8	0.2	0.8	…
	RANGE	10	20	76	…
	N_C_Position	1	-2	1	…
	…	…	…	…	…
CLASS	Collocation	Y	N	Y	…

Zrównoleglanie słowników kolokacji

- Powiązanie słowników?
- Na razie dla kilkunastu tysięcy kolokacji
- Korpusy referencyjne i równoległe źródłem przykładów

1304319	physically	impossible	17
1298288	fizycznie	niemożliwy	99

<trans-unit id="88">

<source>Pogodzenie tych wszystkich obowiązków jest trudne, a czasem wręcz **fizycznie niemożliwe**, staram się jednak negocjować zobowiązania, dopasowywać terminy, i w miarę możliwości wysyłać młodszych kolegów.</source>

<target>It's hard to reconcile all these duties, and sometimes **physically impossible**, but I do my best to negotiate commitments, to adjust deadlines, and if possible send my younger colleagues.</target>

</trans-unit>

- Założenie użycia wielowyrazowych odpowiedników leksykalnych o podobnej składni
- Zrównoleglane kolokacje o potwierdzonym statusie w korpusach tekstów natywnych
- Zastosowania leksykograficzne, modele tłumaczenia maszynowego

Interfejs WWW

- Wersja próbna przegądarki kolokacji angielskich z BNC http://212.191.73.200/ColosaurusWeb/Browser.
- Dodatkowo: generator grafów kolokacyjnych: http://212.191.73.200/ColosaurusWeb/GraphGenerator
- Poprzednia wersja kolokacji polskich z NKJP (nieujednoznacznione części mowy) http://212.191.73.200/ASK
- Powstaje nowa wersja polskiego słownika

Harmonogram udostępniania zasobów

- I transza (listopad 2011): korpusy równoległe dostępne na licencjach CC-BY oraz CC-BY-NC
- II transza (czerwiec 2012) kolejne korpusy równoległe
- II transza (styczeń 2013) słownik kolokacji na podstawie BNC, słownik kolokacji na podstawie NKJP, korpusy równoległe

- http://pelcra.pl/corpora
- http://nlp.ipipan.waw.pl/metashare/