LINGUISTICA COPERNICANA
Nr 14 / 2017

http://dx.doi.org/10.12775/LinCop.2017.011

Zygmunt Vetulani

Uniwersytet im. Adama Mickiewicza w Poznaniu
Wydział Matematyki i Informatyki

Informatyczne technologie języka – wyzwania dla polskiej lingwistyki

Na przykładach z badań własnych

Słowa kluczе: lingwistyka komputerowa; technologie języka; zasoby językowe; systemy z kompetencją językową

Keywords: computational linguistics; language technologies; language resources; systems with language competence

1. Wstęp

Postęp technologiczny ma obecnie charakter globalny i – jak się wydaje – nieodwracalny, choć prędkość i zakres tego postępu nie rozkładają się równomiernie. Poza spektakularnymi zmianami w zakresie technologii „twardych" i środowiska, w którym żyjemy, nie mniej szokujące przemiany dotyczą „miękkiej" sfery zjawisk socjalnych i kulturowych. W tym zakresie czołową rolę odgrywają nauki i technologie związane z komunikacją interpersonalną, ze szczególnym uwzględnieniem języka naturalnego. Ich rozwój pozostaje w ścisłej korelacji z procesami globalizacji, a co więcej te procesy warunkuje. W ostatnim ćwierćwieczu XX wieku stało się jasne, że dalszy postęp cywilizacyjny zależeć będzie w dużym stopniu od doskonałości tech-

nik informatycznych związanych z językiem naturalnym, zaś brak technologii i zasobów językowych będzie groził wykluczeniem całych społeczności[1].

Zdarzeniem przełomowym była konferencja zapamiętana jako Grosseto Workshop (On Automating the Lexicon) zorganizowana przez A. Zampollego, N. Calzolari i D. Walkera w roku 1986 (Walker et al. 1994). Wizjonerska działalność, którą prowadził Antoni Zampolli, doprowadziła do zrozumienia, że dalszy rozwój dyscypliny zmierzający do przewartościowania celów w kierunku wykorzystania badań podstawowych w użytecznych aplikacjach jest uwarunkowany budową zasobów elektronicznych w postaci, między innymi, „korpusów tekstów pisanych lub mówionych, słownikowych baz danych, gramatyk" (Zampolli 1996), ale także standardów, formalizmów oraz narzędzi wykorzystujących te zasoby.

W niniejszym artykule, bez pretensji do wyczerpania zagadnienia, zasygnalizujemy wybrane wyzwania technologiczne stojące w tym zakresie przed językoznawstwem polskim.

2. Wyzwania: w kierunku Społeczeństwa Informacji

To, że przepływ informacji w obrębie danej społeczności jest niezbędny do jej rozwoju, jest oczywistością, a do świata nasyconego technologią informacyjną dzieci przyzwyczajają się niemal od niemowlęctwa. Środowisko technologiczne człowieka w obszarach wysokiej cywilizacji staje się interaktywne, a formy tej interakcji coraz powszechniej zakładają wykorzystanie języka naturalnego jako najbardziej przyjaznego wobec człowieka medium komunikacyjnego. Wizja, która zaczyna się realizować, jest wizją świata zapełnionego urządzeniami wyposażonymi w kompetencję językową kompatybilną z kompetencją językową człowieka. Wykształciła się ona w literackiej wyobraźni człowieka sporo czasu przed powstaniem technologii umożliwiających jej realizację poprzez modelowanie kompetencji myślowej i językowej człowieka. Są one zależne od języka i rodzaju aplikacji i muszą uwzględniać zmienność otoczenia technologicznego, a w szczególności jego szybkie tempo. O ile do niedawna otoczenie to było informacyjnie puste i składało się ze statycznych, nieaktywnych artefaktów, to teraz staje się nasyconym informa-

[1] Stanowisko to zostało jasno wyartykułowane w środowisku konferencji LREC (Language Resources and Evaluation) i inicjatyw Komisji Europejskiej (FlaReNet, META-NET).

cją rozszerzeniem środowiska naturalnego ze swoją własną autonomią. Jego elementy, takie jak internet, zdają się mieć własną tożsamość, niezależną od poszczególnych ludzi czy organizacji. Technologie języka naturalnego służą umożliwieniu komunikacji człowieka ze środowiskiem technologicznym przy pomocy języka. Zapewnienie środków dla komunikacji językowej przy szybko zmieniającym się otoczeniu technologicznym człowieka jest źródłem kolejnych wyzwań dla technologii języka naturalnego oraz lingwistyki jako podstawy tych technologii[2]. Z niektórymi z nich przyszło się nam skonfrontować na miarę naszych potrzeb i celów. Omówione w tym artykule przypadki nie wyczerpują zakresu problematyki podejmowanej w Polsce, lecz, jak nam się zdaje, stanowią wartościową ilustrację różnorodności i złożoności przeszkód na drodze do tworzenia systemów przetwarzania języka polskiego o potencjale praktycznym.

3. Kilka podstawowych definicji

– *Informatyczne technologie języka naturalnego (HLT)* to technologie, które operują na danych językowych w postaci elektronicznej i służą do wytworzenia produktów informatycznych wykorzystujących te dane.
– *Przetwarzanie języka naturalnego* (ang. *Natural Language Processing*) – dziedzina, która zajmuje się automatyzacją analizy, rozumienia, tłumaczenia i generowania języka naturalnego przez komputer.
– *Lingwistyka komputerowa* – dziedzina, która tworzy modele komputerowe języka nadające się do projektowania i realizacji programów komputerowych przetwarzających język naturalny.
– *Przemysł lingwistyczny* (ang, *Language Industry*) – dziedzina przemysłu produkująca artefakty wymagające w istotnym stopniu przetwarzania języka naturalnego oraz narzędzia do tej produkcji. To pojęcie wymaga dodatkowego komentarza, gdyż zostało ono wylansowanie i upowszechnienie stosunkowo niedawno (w końcu lat 80. XX w.) przez Antonio Zampollego, który stwierdził, że zapotrzebowanie na narzędzia wykorzystujące wiedzę lingwistyczną spowodowało zmianę metodologii badań podstawowych w zakresie językoznawstwa przez odejście od „przeważa-

[2] Patrz także: *Joint Panel on Technology for Linguistics, Linguistics for Technology* (Vetulani Z. 2005).

jącej w latach siedemdziesiątych i pierwszej połowie lat osiemdziesiątych tendencji do testowania hipotez językoznawczych na podstawie niewielkiej liczby danych o (rzekomo) krytycznym znaczeniu" (Zampolli 1996). Tworząc podstawy przemysłu lingwistycznego, Zampolli przyczynił się do powstania nowej generacji wyzwań dla przezwyciężenia szeregu dotychczasowych barier.

4. Prace własne zespołu[3]

Złożoność i różnorodność wyzwań[4], z którymi konfrontowani są twórcy systemów z kompetencją językową zilustrujemy na przykładzie wybranych prac własnych zespołu w zakresie lingwistyki komputerowej (i dziedzin pokrewnych) (od 1985 do teraz). Przykładowe projekty to:

– Systemy z kompetencją językową
  • gramatyki POLINT
– Zasoby leksykalne i ich formalny opis
  • POLEX, CEGLEX i GRAMLEX
  • kolokacje werbo-nominalne
– Formalizmy gramatyczne i ich zastosowanie
  • formaty danych leksykalnych POLEX, GRAMCODE
– Ontologie lingwistyczne
  • PolNet i leksykon – gramatyka języka polskiego

4.1. Początkowe prace nad systemami z kompetencją językową

Początek naszych prac nad systemami z kompetencją językową miał miejsce w sytuacji braku dostępu do zasobów lingwistycznych języka polskiego (słowników, gramatyk) w formie umożliwiającej bezpośrednie zastosowanie w aplikacjach. Tym niemniej w latach osiemdziesiątych i dziewięćdziesiątych stan wiedzy lingwistycznej w odniesieniu do wielu języków, w tym języka polskiego, znacznie odbiegał od sytuacji krytykowanej przez pionierów

[3] Zakład Lingwistyki Informatycznej i Sztucznej Inteligencji na Wydziale Matematyki i Informatyki UAM.

[4] „Złożoność i różnorodność wyzwań" ilustrujemy w tym artykule pracami zespołu realizującego skomplikowane systemy lingwistyki komputerowej. Termin „ilustracja" należy rozumieć dosłownie – nie pretendujemy do uznania przeglądu za reprezentatywny dla całego spectrum problemów i wyzwań.

przetwarzania języka naturalnego lat 50. ubolewających nad brakiem przydatnych informatykom opracowań lingwistycznych[5]. W szczególności nieocenionym źródłem wiedzy w zakresie morfologii i składni polskiej okazały się artykuły zebrane w tzw. „żółtej gramatyce języka polskiego PWN" (Urbańczyk, red., 1984)[6].

GRAMATYKI POLINT i SYSTEM POLINT-112-SMS

Pomyślne rozwijanie systemów z kompetencją językową stało się możliwe dzięki rozpoczętym przez nas pracom nad opisem gramatycznym języka polskiego, nadającym się do wykorzystania informatycznego w algorytmach parsujących, tj. w algorytmach przeprowadzających analizę składniową, będącą etapem przygotowawczym w procesie ustalania znaczenia jednostek tekstu[7].

Doprowadziły one do gramatyk POLINT (rozwijanych od 1985) inspirowanych gramatykami systemu ORBIS[8] opracowanego dla języków angielskiego i francuskiego przez A. Colmerauea i R.Kittredge'a (1982) w PROLOGU. Dla języka polskiego były to wówczas prace pionierskie[9]. Pierwsze programy POLINT (Vetulani Z. 1988), zorientowane na modelowanie dialogów pytanie-odpowiedź, nie były tworzone z bezpośrednim zamiarem uzyskania aplikacji praktycznych, lecz w celu wykazania potencjału aplikacyjnego w zakresie pokrycia językowego. Potencjał ten został pozytywnie zweryfikowany w konfrontacji z materiałem empirycznym w postaci korpusu dialogów wygenerowanych w drodze eksperymentu (Vetulani Z. 1990),

---

[5] Patrz np. Silvio Ceccato w (Vetulani Z. 2006).

[6] Stanisław Urbańczyk był redaktorem naukowym całości, zaś autorami rozdziałów w tomach *Składnia* i *Morfologia* Maciej Grochowski, Renata Grzegorczykowa, Krystyna Kallas, Stanisław Karolak, Krystyna Kowalik, Roman Laskowski, Alicja Orzechowska, Jadwiga Puzynina, Zuzanna Topolińska, Henryk Wróbel.

[7] Należy w tym miejscu zaznaczyć, że prace prowadzono przy założeniach upraszczających, a mianowicie kompozycyjności i obliczalności języka. Założenia te są przedmiotem dyskusji wśród filozofów języka i lingwistów co najmniej od okresu oświecenia, lecz wydają się niezbędne dla tworzenia informatycznie użytecznych precyzyjnych modeli języka, zakreślając granice deterministycznego modelowania języka.

[8] Mało znany jest fakt, że gramatyki systemu ORBIS były przykładem pierwszej zapewne implementacji idei leksykonu-gramatyki (informacja składniowa była powiązana z czasownikiem).

[9] Pierwsze prace prowadzone były niezależnie od podobnych, szybko niestety zaniechanych, prac S. Szpakowicza (1983), J. Bienia czy W. Lubaszewskiego.

a ostatecznie potwierdzony w aplikacji POLINT-112-SMS (Vetulani Z. et al. 2010). Systemy POLINT generowały reprezentację znaczenia pytania przekazywaną do dalszego przetwarzania (wyliczenie odpowiedzi). Prace te doprowadziły do opracowania metody parsingu heurystycznego wykorzystującego identyfikację cech tekstu (świadectw formalnych), pozwalających ograniczyć zbiór hipotez składniowych odnoszących się do analizowanego elementu (a tym samym zmniejszyć złożoność obliczeniową analizy).

System POLINT-112-SMS był aplikacją powstałą w ramach projektu finansowanego w latach 2006–2010 przez MNiSzW[10] zorientowaną na potrzeby zarządzania kryzysowego poprzez wspomaganie komunikacji pomiędzy informatorami a sztabem kryzysowym. Został on pomyślany jako „inteligentny" bajpas komunikacyjny umożliwiający uniknięcie zatorów komunikacyjnych (Vetulani Z., Osiński J. 2017). Zakładał użycie języka naturalnego w obustronnej komunikacji między użytkownikiem-człowiekiem a pośredniczącym systemem informatycznym. Główne funkcje realizowane przez wykonany prototyp systemu (POLINT-112-SMS) sprowadzały się do:

a) pobierania informacji przekazywanej przez informatora tekstem (SMS),
b) przetwarzania (zakładającego rozumienie człowieka przez system) informacji tekstowej celem zbudowania spójnego modelu sytuacji/zdarzenia w czasie rzeczywistym, a także do integracji wiedzy pochodzącej z różnych źródeł (fragmenty informacji pochodzącej od różnych informatorów mogą nie spotkać się w modelu tradycyjnym),
c) automatycznej kontroli spójności, oceny wiarygodności źródła informacji
d) możliwości przejęcia przez system pewnych funkcji operatora, jak np. zadawanie pytań i sterowanie dialogiem (np. w sytuacji zatoru komunikacyjnego, niedostatku operatorów, natłoku zgłoszeń itd.).

Prace wymagały pokonania wielu problemów teoretycznych i praktyczno-informatycznych na poziomie morfologii, składni, semantyki, pragmatyki, analizy struktury dialogu, logiki pragmatycznej, formalizmów opisu języka i danych językowych. Tym samym stanowiły dobrą ilustrację trudności typowych przy modelowaniu kompetencji językowej w wymiarze odpowiadającym potrzebom praktycznym (choć niewyczerpującą[11]). Nb. większość

[10] „Technologie przetwarzania tekstu polskiego zorientowane na potrzeby bezpieczeństwa publicznego" – grant MNiSzW nr R00 028 02/2006-2010.

[11] W szczególności nasze prace nie dotyczyły analizy mowy. Tym niemniej jest dla nas jasne, że technologie mowy i technologie tekstu jako komplementarne będą coraz czę-

z tych doświadczeń można przenieść na szybko rozwijającą się dziedzinę tworzenia interfejsów w języku naturalnym do urządzeń i systemów informatycznych (w tym gier komputerowych czy wirtualnych asystentów socjalnych jak Amazon Lex).

W tym wypadku wyzwaniem nowatorskim było nie tyle stworzenie nowych podstaw teoretycznych, ile sprowadzenie istniejącego dorobku teoretycznego do postaci umożliwiającej wykorzystanie informatyczne. Powstały prototyp był aplikacją w skali rzeczywistej, w przeciwieństwie do licznych przykładów akademickich o charakterze ilustracyjnym.

Poniższy przykład daje pewną orientację w uzyskanej kompetencji językowej (z tym, że rzeczywiste pokrycie językowe znacznie wykracza poza ten przykład).

Tab. 1. Interakcja pomiędzy agentami w trakcie sesji systemu POLINT-112-SMS

Informator: Kowal i Wolski należą do bojówki. System POLINT-112-SMS: Zrozumiałem I: Osoba, która nosi czarne spodnie, jest niebezpieczna. Kowal i Wolski atakują policjantów kamieniami. S: Zrozumiałem Analityk: Kto atakuje policjantów? S: Kowal. Wolski. Brak dalszych odpowiedzi!

POLEX, GRAMLEX i CEGLEX

Już początkowe prace nad parsingiem struktur zdaniowych języka polskiego szybko pokazały, że obrany kierunek jest obiecujący dla zastosowań. Przekonały nas, że dalszy postęp ku rzeczywistym aplikacjom wymagającym budowy właściwych modeli języka naturalnego jest niemożliwy bez sprostania wyzwaniom wyartykułowanym w programie Antonio Zampollego,

ściej przenikać się w zastosowaniach praktycznych. W Polsce mamy długą tradycję w zakresie rozwijania technologii mowy. Np. obszarem od dawna prowadzonych prac była fonologia formalna i synteza mowy, gdzie mieliśmy wybitnych pionierów już w latach siedemdziesiątych (W. Jassem, T. Batóg, M. Steffen-Batogowa). Przetwarzanie mowy jest dziedziną, która w ostatnim właśnie okresie nabrała wyjątkowego znaczenia praktycznego, a to ze względu na wzrastające zainteresowanie przemysłu. W tym zakresie działa intensywnie kilka ośrodków, jak np. AGH, Politechnika Poznańska, IPI PAN, PJATK (dawniej PJSTK), UAM i inne.

a mianowicie krytycznego braku podstawowych narzędzi, począwszy od słowników morfologicznych, leksykonów gramatycznych w postaci nadającej się do wykorzystania w programach komputerowych. Realizacja trzech projektów POLEX, GRAMLEX i CEGLEX była próbą podjęcia tego wyzwania dla języka polskiego.

Głównym celem projektu POLEX (Vetulani Z. et al. 1998a)[12] było wypełnienie luki w istniejących zasobach elektronicznych w zakresie słowników morfologicznych języka polskiego. W projekcie wzięto pod uwagę dotychczasowy dorobek tradycyjnej morfologii opisowej języka polskiego, z tym że istniejące wówczas opisy i klasyfikacje, szczególnie dla kategorii rzeczownika (oparte na metodologii "reguła-wyjątki") na ogół nie były wystarczająco rygorystyczne dla bezpośrednich zastosowań w systemach informatycznych. Ponadto przy całkowitym braku reprezentatywnego korpusu tekstów polskich i listy frekwencyjnej w celu ustalenia pokrycia leksykalnego za surogat bazy empirycznej wzięto trzytomowy słownik Mieczysława Szymczaka (Szymczak, red., 1978–1981) obejmujący ok. 100 000 leksemów. Tym samym POLEX obejmuje w zasadzie kompletne słownictwo standardowego języka polskiego, dziedzicząc jednak pewne niedostatki źródła (po uzupełnieniach ok. 120 000 haseł) (brak wulgaryzmów i słownictwa żargonowego, niedostatek słownictwa regionalnego i potocznego). Zaletą tego projektu było opracowanie precyzyjnego, interpretowanego maszynowo (ale również czytelnego dla człowieka) formatu kodowania leksemów, identycznego dla wszystkich części mowy. Hasła słownikowe mają następującą postać:

FORMA_PODST.;LISTA_TEMATÓW;KOD_FLEKSYJNY;DISTRYBUCJA_TEMATÓW

Kod paradygmatyczny (= kod fleksyjny) obejmuje pełną informację odnośnie fleksji, w szczególności listę końcówek właściwych dla wszystkich pozycji paradygmatycznych. Informacja jest pełna, jednoznaczna, a klasy fleksyjne są tak skonstruowane, że nie ma potrzeby rozważania wyjątków. Przykładowo hasła słownikowe dla dwóch wariantów fleksyjnych słowa *frajer* wyglądają następująco:

[12] 1994–1997: „POLEX – Polska Leksykalna Baza Danych" /"POLEX – lexical data base for Polish"/ (Grant: KBN8S50301007; UAM); zasób można pozyskać przez stronę http://www.islrn.org/resources/identify_name/, względnie kontaktując się przez vetulani@amu.edu.pl.

frajer; frajer, frajerz; N110; 1;1-5,9-13;2:6-8,14
frajer; frajer, frajerz; N110; 1;1-5,8-14;2:6-7

Mając do dyspozycji tabelę końcówek oraz rozmieszczenie tematów[13], generowanie wszystkich form fleksyjnych jest trywialne, a algorytm generowania jest ten sam dla wszystkich części mowy. Prace projektowe wykazały, że możliwe jest uzyskanie prostego opisu morfologicznego nieprzewidującego wyjątków i dzięki temu niezwykle wygodnego w aplikacjach informatycznych. W końcowej fazie projektu POLEX jego wstępne wyniki były wykorzystane w europejskim projekcie GRAMLEX (COPERNICUS Project 621) (realizowanym w latach 1995–1998. Cztery główne tematy badań w ramach projektu GRAMLEX dotyczyły: 1) słowników, 2) podstawowych narzędzi software'owych, 3) zaawansowanych narzędzi wykorzystujących słownik, 4) zastosowania w inżynierii języka. Głównym wynikiem prac jest słownik morfologiczny języka polskiego w formacie GRAMCODE (ponad 22 500 haseł atestowanych w korpusie) i zależne od niego narzędzia (lematyzator, generator form fleksyjnych, generator konkordancji) (Vetulani Z. et al. 1998b). Projekt GRAMLEX w części wykorzystywał wyniki równolegle uzyskiwane w projekcie europejskim CEGLEX Copernicus 1032, realizowanym w latach 1995–1996 (Vetulani Z. 2000), który miał na celu przetestowanie generycznego modelu danych gramatycznych opracowanego w ramach projektu GENELEX dla języków zachodnioeuropejskich (francuski, angielski, niemiecki, włoski, ...) dla języków Europy Centralnej: czeskiego, polskiego i węgierskiego. Model wykorzystywał formalizm SGML, w którym informacja lingwistyczna jest przypisywana danym poprzez znaczniki (tags) i atrybuty SGML. Zakładał on istnienie abstrakcyjnego uniwersum złożonego z „elementów" mających charakterystyczny dla siebie zestaw atrybutów. Pojęcie „elementu" odpowiada klasycznemu pojęciu kategorii. Trzy warstwy modelu CEGLEX/GENELEX zostały skonfrontowane z danymi pochodzącymi z badanych języków z wynikiem pozytywnym. Ponadto dla języka polskiego zaproponowane zostały pewne rozszerzenia w stosunku do początkowego

[13] Ważnym założeniem znacznie upraszczającym tworzenie słownika i jego wykorzystanie było przyjęcie, że „tematem jest to, co nie jest końcówką". Definicja ta odbiegała od tej przyjmowanej w tradycyjnej morfologii opisowej, lecz okazała się bardzo dogodna obliczeniowo.

modelu GENELEX, w szczególności w zakresie opisu zjawisk fleksyjnych, zgodnie z wynikami uzyskanymi w projekcie POLEX.

```
<!--CEGLEX Polish Morphology Description Examples-->
<Simple_mu              id=”smu-fra-1”
                        appellation=”rzeczownik rodzaju męskoosobowego
                                    ‘frajer’ odmiana 1 ”
                        autonomy=”YES”
                        category=”NOUN”
                        subcategory=”COMMON”>
 <Graph_mu              current_nb=”0”
                        prefferred=”YES”
                        inflection=”infl-fra-1”>
                        % parametr infl-fra-1 zawiera informację o końcówkach
                                                        i dystybucji tematów
      <Spelling>frajer</Spelling>     % forma podstawowa
      <Gstem            current_nb=”0”>
          <Spelling>frajer</Spelling>% temat
      </Gstem>
      <Gstem            current_nb=”1”>
          <Spelling>frajerz</Spelling>  % temat
      </Gstem>
      <Graph_mu
</Simple_mu>
```

Choć pierwsza faza projektu POLEX, umożliwiająca praktyczne wykorzystywanie jego wyników w projektach, została ukończona w roku 1997, to zasób podstawowy został poważnie powiększony podczas realizacji projektu POLINT-112-SMS, a kolejna aktualizacja i rozszerzenie prowadzone jest obecnie.

4.2. Leksykon-gramatyka

Wśród priorytetów, o których pisał Zampolli, ważne miejsce zajmują gramatyki formalne w postaci wygodnej do wykorzystania informatycznego. O ile przydatność gramatyk w formie zbioru reguł (i wyjątków od nich) znana była od dawna, to były one adresowane z oczywistych względów do ludzi i w związku z tym nie musiały spełniać rygorów właściwych współczesnym potrzebom przemysłu lingwistycznego. Pionierskimi w tym zakresie

były prace Maurice Grossa nad leksykonem-gramatyką języka francuskiego (Gross 1975) (potem także innych). Gross – inspirowany pracami Harrisa – opracował formalizm gramatyczny w postaci leksykonu (w formie tablic syntaktycznych), gdzie opis leksemów predykatywnych (w pierwszym rzędzie czasowników prostych i złożonych) zawiera informację składniową w postaci schematów zdań elementarnych, w których dane słowo hasłowe może występować. W tym samym mniej więcej czasie (lata 70.) zbliżony pomysł realizował dla języka polskiego Kazimierz Polański w monumentalnym dziele *Generatywny słownik czasowników polskich*, którego poszczególne tomy (5) ukazywały się w latach 1980–1992. Przydatność tego podejścia mieliśmy okazję zweryfikować już w pierwszych implementacjach gramatyk POLINT. Charakteru leksykonu-gramatyki zaczął nabierać system PolNet wraz z włączeniem do niego synsetów czasownikowych i kolokacji werbo-nominalnych (od roku 2012). Poniżej przytaczamy (za (Vetulani Z., Vetulani G. 2012) przykład opisu prostego synsetu czasownikowego.

POS: v ID: 3441
Synonyms: {pomóc:1, pomagać:1}
Definition: "wziąć (brać) udział w pracy jakiejś osoby, aby ułatwić jej tę pracę"
VALENCY:

- Agent(N)_Benef(D)
- Agent(N)_Benef(D) Action('w'+NA(L))
- Agent(N)_Benef(D) Manner
- Agent(N)_Benef(D) Action('w'+NA(L)) Manner

Usage: Agent(N)_Benef(D); "Pomogłam jej."
Usage: Agent(N)_Benef(D) Action('w'+NA(L)); "Pomogłam jej w robieniu lekcji."
Usage: Agent(N)_Benef(D) Manner Action('w'+NA(L)); "Chętnie pomogłam jej w lekcjach."
Usage: Agent(N)_Benef(D) Manner;"Chętnie jej pomagałam."
Semantic_role: [Agent] {człek:1, człowiek:1, homo sapiens:1, istota ludzka:1, zwierzę:2, ...} (*{man:1,...,animal:2,...}*)
Semantic_role: [Benef] {człek:1, człowiek:1, homo sapiens:1, istota ludzka:1, zwierzę:2, ...} (*{man:1,...,animal:2,...}*)
Semantic_role: [Action] {czynność:1}
Semantic_role: [Manner] {CECHA_ADVERB_JAKOŚĆ:1}

4.3. Ontologia leksykalna PolNet

Odpowiedzią na zidentyfikowaną (w ramach projektu POLINT-112-SMS) potrzebę reprezentacji wiedzy o świecie było przystąpienie (2006)[14] do tworzenia ontologii leksykalnej PolNet[15], będącej siecią leksykalną (leksykalną bazą danych) typu wordnet (Vetulani Z 2014). O ile koncepcja i struktura bazy wzorowana jest na rozwiązaniach systemu Princeton WordNet (Miller & Feldbaum, 2007), o tyle PolNet tworzony jest od podstaw przez zespół informatyków i leksykografów metodą tradycyjną w oparciu o istniejące słowniki języka polskiego w celu zachowania odwzorowanej w języku polskim konceptualizacji[16].

Baza PolNet jest zbudowana z klas synonimów i relacji między tymi klasami. Klasy synonimów (synsety) reprezentują pojęcia identyfikowalne w języku naturalnym, dzięki czemu PolNet można wykorzystywać jako ontologię. System ten jest budowany w oparciu o wysokiej jakości klasyczne słowniki języka polskiego i badanie dostępnych korpusów językowych (IPI PAN Corpus (Przepiórkowski 2004) i małe korpusy dziedzinowe). Tworzenie zasobu odbywa się przyrostowo, poczynając od słownictwa o dużej częstości użycia i słów ważnych[17].

O ile początkowe prace nad PolNet'em prowadzone były w kierunku systemu o strukturze zbliżonej do Princeton WordNet i mającego służyć jako ontologia w naturalny sposób powiązana z językiem, to w miarę czasu projekt ten, pod wpływem prac Grossa, Polańskiego i Colmerauera, ewoluował w kierunku leksykonu-gramatyki przez włączanie czasowników prostych,

---

[14] Niezależnie, mniej więcej w tym samym okresie, rozpoczął się wrocławski projekt realizowany w znacznej mierze metodami automatycznymi zwany Słowosiecią (później plWordNet) (Piasecki et al. 2016).

[15] Projekt „PolNet – Polish WordNet" został rozpoczęty w ramach grantu MNiSzW nr R00 028 02/2006-2010.

[16] Wiele spośród podobnych projektów dla różnych języków polega na tłumaczeniu oryginalnego WordNet'u na dany język bądź na automatycznym generowaniu synsetów. Obie te metody obciążone są zwiększonym ryzykiem zagubienia konceptualizacji właściwej dla języka przedmiotowego.

[17] Odstępstwem od tej zasady poczynionym ze względów metodologicznych, po to, aby umożliwić wczesne testowanie rozwijanego zasobu w aplikacjach, dla których musi zostać spełniony warunek kompletności leksykalnej, było uwzględnienie terminologii właściwej tym aplikacjom.

a potem także złożonych. Ewolucja ta zbiegła się w czasie z postępem prac nad opracowaniem sformalizowanego słownika wyrażeń złożonych o charakterze kolokacji werbo-nominalnych zapoczątkowanych w latach 90-tych XX wieku przez G. Vetulani[18].

Pierwsze udostępnienie bazy PolNet v1.0 w roku 2012 obejmowało przede wszystkim rzeczowniki (i pierwsze czasowniki, jeszcze bez informacji składniowej). W tym też okresie rozpoczęło się włączanie do bazy PolNet kolokacji werbo-nominalnych. Aktualnie udostępniany[19] (rozszerzony o kolokacje werbo-nominalne, rejestry) PolNet v3.0 jest zarejestrowany jako zasób o numerze ISLRN 944-121-942-407-9. PolNet v3 ma wiele cech leksykonu gramatycznego (Vetulani Z., Vetulani G. 2014) przede wszystkim poprzez powiązanie z synsetami czasownikowymi informacji syntaktycznej, co zbliża go do koncepcji leksykalnych baz danych typu FrameNet (Fillmore 2002). Informacja składniowa powiązana z synsetami czasownikowymi umożliwia konstrukcję heurystyk istotnie polepszających efektywność parsingu (Vetulani Z. et al. 2010).

Tab. 2. Rozwój bazy PolNet

	PolNet 0.1 (2009)	**PolNet 1.0 (2011)**	**PolNet 2.0 (2013)**	**PolNet 3.0 (2016)**
Rzeczowniki	10 629	11 700	11 700	12 011
Czasowniki proste	---	1 500	1 500	3 645
Kolokacje	---	---	1 200	1 908

5. Stan aktualny

Na sytuację inżynierii języka i lingwistyki w Polsce można patrzeć w miarę optymistyczne, lecz z pewnymi zastrzeżeniami:

– język polski jest bardzo dobrze opisany lingwistycznie[20],

[18] (Vetulani G. 2000, 2012).

[19] Pozyskanie zasobów przez http://www.islrn.org/resources/identify_name/ lub http://ltc.amu.edu.pl/polnet/ (ISLRN – International Standard Language Resource Number).

[20] Doskonałe opisy języka polskiego do lat osiemdziesiątych XX wieku z reguły nie były budowane z przeznaczeniem do zastosowań informatycznych. Przejawem tego zjawiska było tworzenie systemów reguł zawierających liczne wyjątki łatwo interpretowane

– istnieją podstawowe tradycyjne zasoby bardzo dobrej jakości (słowniki, tezaurusy, gramatyki),

ale…

– ciągle odczuwa się niedostatek ogólnodostępnych (tj. nieobwarowanych barierami w dostępie) zasobów cyfrowych,
– o ile za główne wyzwanie, któremu w dużej mierze udało się sprostać (mimo problemów dostępności), był niedostatek zasobów lingwistycznych w postaci przystosowanej do przetwarzania informatycznego (wstępnie przetworzonych korpusów tekstowych i głosowych, słowników elektronicznych, ontologii leksykalnych), to ciągle – mimo intensywnie prowadzonych prac – niezaspokojone są potrzeby w zakresie algorytmów do rozwiązywania wielu zasadniczych problemów składniowych, semantycznych i pragmatycznych (np. odnośnie do zjawisk anafory, metafory czy metonimii, a w zakresie analizy mowy – do zjawisk paralingwistycznych i nielingwistycznych),
– środowisko użytkowników technologii jest zachowawcze i chętnie wybiera narzędzia proste, sprawdzone, lecz niekoniecznie odpowiadające stanowi wiedzy o języku(ach),
– środowisko lingwistyczne jest zachowawcze w tym sensie, iż rzadko podejmuje problemy antycypujące wymogi przemysłu lingwistycznego.

Niezależnie od dwóch ostatnich obserwacji wypada stwierdzić, że
– w Polsce powstało lub umocniło się w czasie ostatnich 20 lat szereg silnych środowisk technologii języka naturalnego[21],
– ponadto istnieje szereg ośrodków z małymi zespołami,
– silne ośrodki technologiczne pojawiają się tam, gdzie prowadzone są badania podstawowe (niekoniecznie związane bezpośrednio z zastosowaniami),

przez człowieka, lecz nieprzystosowanych do konstruowania na ich podstawie algorytmów odwołujących się do wiedzy lingwistycznej. Z problemem tego rodzaju spotkaliśmy się przy budowaniu elektronicznego słownika morfologicznego dla języka polskiego (Vetulani Z. et al. 1998a).

[21] Przede wszystkim Krakowskie, Poznańskie, Śląskie i Warszawskie /kolejność alfabetyczna/.

– prowadzone są badania we wszystkich głównych nurtach HLT, choć nie zawsze są wspierane działalnością naukową w zakresie lingwistyki teoretycznej[22].

6. Podsumowanie

Wymienione w tym artykule wybrane prace, zarówno podstawowe w zakresie technologii związanych z językiem polskim, jak i aplikacje, utwierdzają nas w przekonaniu, że zaistniały możliwości realizowania projektów o znaczeniu praktycznym pod warunkiem dostępu do zasobów inżynierii języka. Wynika to zarówno z prac specyficznie związanych z językiem polskim i prowadzonych w Polsce, jak też z prac o charakterze ogólnym. Wśród tych ostatnich należy wymienić wyniki w zakresie automatycznego rozpoznania mowy („z mowy do tekstu")[23]. W tym wypadku doskonałe wyniki

[22] W miejsce szczegółowego merytorycznego przeglądu wybitnych osiągnięć środowiska polskiego w zakresie informatycznych technologii języka, co wymagałoby obszernej monografii, opinię o intensywności prac prowadzonych w Polsce można sobie wyrobić studiując listę krajowych ośrodków reprezentowanych na konferencjach LTC w ciągu minionego dziesięciolecia (Language and Technology Conference, Poznań, lata 2007–2015). Wśród nich są (w kolejności alfabetycznej): Gliwice (Wydział Automatyki, Elektroniki i Informatyki Politechniki Śląskiej), Kraków (Wydział Fizyki, Astronomii i Informatyki Stosowanej UJ, Wydział Informatyki, Elektroniki i Telekomunikacji AGH, Wydział Zarządzania i Komunikacji Społecznej UJ), Olsztyn (Wydział Humanistyczny Uniwersytetu Warmińsko-Mazurskiego), Poznań (Poznańskie Centrum Komputerowo-Sieciowe, Wydział Anglistyki UAM, Wydział Elektrotechniki i Informatyki Politechniki Rzeszowskiej, Wydział Elektryczny Politechniki Poznańskiej, Wydział Fizyki UAM, Wydział Informatyki Politechniki Poznańskiej, Wydział Informatyki i Gospodarki Elektronicznej Uniwersytetu Ekonomicznego, Wydział Matematyki i Informatyki UAM, Wydział Neofilologii UAM), Warszawa (Instytut Podstaw Informatyki PAN, Instytut Języka Polskiego PAN, Wydział Neofilologii UW), Wrocław (Instytut Informatyki Stosowanej Uniwersytetu Wrocławskiego, Wydział Informatyki i Zarządzania Uniwersytetu Wrocławskiego). (Listę tą należy traktować orientacyjnie. Została ona skompilowana w oparciu o deklaracje afliacyjne wskazujące na jednostki, których nazwy niejednokrotnie ulegały zmianom ze względów organizacyjnych.) W wymienionym powyżej okresie w LTC brało udział 176 autorów (ok.18% ogólnej liczby), którzy wygłaszali odczyty zaklasyfikowane do następujących grup tematycznych: analiza i przetwarzanie tekstu, aplikacje, formalizmy, komunikacja międzyagentowa, morfologia, ontologie i wordnet, parsing, przetwarzanie mowy, semantyka komputerowa, tłumaczenie maszynowe, wyszukiwanie i ekstrakcja informacji w tekstach, zasoby językowe.

[23] Przetwarzanie mowy nie stanowiło przedmiotu naszych prac i nie jest też przedmiotem analizy niniejszego artykułu. Stanowi jednak obiekt prac w szeregu jednostkach

osiągnięto w laboratoriach korporacji Google, gdzie udało się pokonać trudności, z którymi borykało się wiele zespołów narodowych w licznych krajach, w tym także w Polsce. Jednocześnie należy pamiętać, że wyzwania stawiane i pokonywane przez lingwistów i informatyków mają ze swojej natury charakter otwarty. Oznacza to, że osiągnięty sukces ma zazwyczaj „ograniczony czas ważności", a wyniki rzeczowe, w postaci zasobów danych różnego rodzaju (słowniki, korpusy, gramatyki itp.), wymagają ciągłej konserwacji i uaktualniania, jako że języki naturalne są przedmiotem procesów ewolucyjnych i ulegają ciągłym przekształceniom na wszystkich poziomach.

Zebrane w ciągu ubiegłych 30 lat doświadczenia i obserwacje, potwierdzają przewidywania Antonio Zampollego adresowane do narodowych dysponentów środków i decydentów w zakresie polityki naukowej co do konieczności tworzenia informatycznie przetwarzalnych zasobów językowych siłami własnymi zainteresowanych krajów, będących końcowymi beneficjentami technologii. Uwaga ta jest szczególnie istotna wobec drastycznego ograniczenia środków europejskich na wsparcie prac w kierunku rozwiązywania problemów specyficznych dla języków innych niż angielski.

Bibliografia

Colmerauer A., Kittredge R., 1982, ORBIS, *COLING 1982*.

Fillmore Ch.J., Baker C.F., Sato H., 2002, The FrameNet Database and Software Tools. LREC 2002 Proceedings, ELRA/ELDA Paris, 1157–1160.

Gross M., 1975, *Méthodes en syntaxe*, Paris: Hermann.

Polański K. (red.), 1980–1992, *Słownik syntaktyczno-generatywny czasowników polskich*, vol. 1–4, Ossolineum: Wrocław; vol. 5, Kraków: Instytut Języka Polskiego PAN.

Marasek K. , Gubrynowicz R., 2008, Design and Data Collection for Spoken Polish Dialogs Database, *LREC 2008 Proceedings,* ELRA/ELDA, Paris, 185–189.

Miller G.A., Felbaum Ch.. 2007, WordNet then and now. *Language Resources and Evaluation 41(2)*, 209–214.

w Polsce oraz jest silnie reprezentowane na konferencjach LTC. Zainteresowanych specyfiką tych prac odsyłamy do monografii (Ziółko B.. Ziółko M. 2011). Patrz też (Wagner et al. 2015) oraz (Marasek et al. 2014).

Piasecki M., Burdka Ł., Maziarz M., Kalinski M., 2016, Diagnostic Tools in plWordNet Development Process, w: *Lecture Notes in Computer Science 9561,* Springer, 255–273.

Przepiórkowski A., 2004, *The IPI PAN Corpus. Preliminary version*, Warszawa: Instytut Podstaw Informatyki PAN.

Szpakowicz S., 1983, *Formalny opis składniowy zdań polskich*, Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.

Szymczak M. (red.), 1978–1981, *Słownik języka polskiego*, t. 1–3, Warszawa: Państwowe Wydawnictwo Naukowe.

Urbańczyk St. (red.), 1984, *Gramatyka współczesnego języka polskiego. Składnia. Morfologia*, Warszawa: Państwowe Wydawnictwo Naukowe.

Vetulani G., 2000, *Rzeczowniki predykatywne języka polskiego. W kierunku syntaktycznego słownika rzeczowników predykatywnych*, Poznań: Wydawnictwo Naukowe UAM.

Vetulani G., 2012, *Kolokacje werbo-nominalne jako samodzielne jednostki języka. Syntaktyczny słownik kolokacji werbo-nominalnych języka polskiego na potrzeby zastosowań informatycznych. Część I.*, Poznań: Wydawnictwo Naukowe UAM.

Vetulani Z., 1990, *Corpus of consultative dialogues. Experimentally collected source data for* AI applications, Poznań: Adam Mickiewicz University Press.

Vetulani Z., 1988, PROLOG Implementation of an Access in Polish to a Data Base, *Studia z automatyki XII*, 5–23.

Vetulani Z., 2000, *Electronic Language Resources for POLISH: POLEX, CEGLEX and GRAMLEX*, w: M. Gavrilidou et al. (eds.), Second International Conference on Language Resources and Evaluation, Athens, Greece, 30.05.–2.06.2000, (Proceedings), Paris: ELRA, 367–374.

Vetulani Z., 2005, *Joint Panel on Technology for Linguistics, Linguistics for Technology*, w: Human Language Technologies as a Challenge for Computer Science and Linguistics, Proc. of Language and Technology Conference, April 21–23, 2005, Poznań, Poland, Poznań: Wyd. Poznańskie, pp. XXVI–XXX.

Vetulani Z., 2006, Tradition and New Challenges for the HLT Community, *Studia Informatica,* 1/2(7), 161–177.

Vetulani Z., 2014, PolNet – Polish WordNet., w: *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2011. Revised Selected Papers. LNAI 8387,* Berlin Heidelberg: Springer-Verlag, 408–416.

Vetulani Z., Martinek J., Obrębski T., Vetulani G., 1998b*, Dictionary Based Methods and Tools for Language Engineering*, Poznań: Adam Mickiewicz University Press.

Vetulani Z., Marcinak J., Obrębski J., Vetulani G., Dabrowski A., Kubis M., Osiński J., Walkowska J., Kubacki P., Witalewski K., 2010, *Zasoby języko-*

we i technologie przetwarzania tekstu. POLINT-112-SMS jako przykład aplikacji z zakresu bezpieczeństwa publicznego, Adam Mickiewicz University Press: Poznań.

Vetulani Z., Osiński J., 2017. Intelligent Tnformation Bypass for More Efficient Emergency Management, *CMST 23(2)*, 105–123.

Vetulani Z., Vetulani G., 2014, Through Wordnet to Lexicon Grammar, w: Fryni Kakoyianni Doa (ed.). *Penser le lexique grammaire: perspectives actuelles*, Paris : Editions Honoré Champion, 531–543.

Vetulani Z., Walczak B., Obrębski T., Vetulani G., 1998a, *Unambiguous coding of the inflection of Polish nouns and its application in electronic dictionaries – format POLEX*, Poznań: Adam Mickiewicz University Press.

Wagner A., Bachan J., Klessa K., Demenko G., 2015, Przegląd wybranych aspektów analizy prozodii mowy spontanicznej na potrzeby technologii mowy, *Prace Filologiczne LXVI,* 271–298.

Walker D., Zampolli A., Calzolari N. (eds.), 1994, *Automating the lexicon: research and practice in a multilingual environment*, Oxford: OUP

Ziółko B., Ziółko M., 2011, *Przetwarzanie mowy,* Kraków: Wydawnictwa AGH.

Zampolli A. 1996, Współpraca międzynarodowa w dziedzinie LR; *Informatyka*, Nr 3, s. 34–37.

Human Language Technologies as a Challenge for Polish Linguistics. Illustrated by the Author's Own Research

(summary)

In order to exemplify complexity and diversity of problems that language engineers are faced with we present selected works in the field of human language technologies that have been done within research projects of the Department of Computer Linguistics and Artificial Intelligence at the Adam Mickiewicz University in Poznań over the last 30 years. These are first of all contributions in creation of language resources including lexicons, grammars, Polish wordnet, as well as creation of systems with emulated language competence. Our aim is to illustrate – via our contributions – a number of challenges facing today's linguistics of the Polish language. We also intend to bring the reader's attention to the fact that many of these challenges is – and will continue to be – still valid. This overview does not pretend to completeness. In particular, the very important area of speech processing is passed over.