Analiza parametryczna tekstu a translacja maszynowa – wybrane zagadnienia

Łukasz Karpiński
Uniwersytet Warszawski

Abstract
This article focuses on selected aspects of parametric description of text in the context of machine translation enhancement. The initial part attempts to examine conventional communication chain "sender-message-recipient" accepting that machine translator can serve as the (indirect) recipient. In order to assure maximum accuracy of the sender's intentions determined by algorithm programme, the present author suggests a preliminary analysis of a given text and – on the basis of the resulting numerical value the programme – attempts to embark on translation process. A crucial part of this translation method is a comprehensive lexicographic database, including appropriate microstructure parameters, linguistically verified according to a textual corpus.

Keywords: communication, machine translator, text parametrisation, database, text assessment markers

Abstrakt
Artykuł poświęcony jest wybranym zagadnieniom parametrycznego opisania tekstów w kontekście usprawnienia translacji maszynowej. Na początku omówiono klasyczny układ komunikacyjny „nadawca-komunikat-odbiorca" z założeniem, iż odbiorcą (pośrednim) może być translator maszynowy. Aby algorytm programu jak najdokładniej określił właściwe intencje nadawcy autor proponuje wykonywanie wstępnej analizy

danego tekstu i na podstawie otrzymanych wartości liczbowych program przystępował do realizacji tłumaczenia. Niezbędnym elementem takiego procesu translacji jest również właściwie opracowana baza danych leksykograficznych, zawierająca odpowiednie parametry mikrostruktury oraz lingwistycznie zweryfikowana na podstawie korpusu tekstowego.

Słowa klucze: komunikacja, translator maszynowy, parametryzacja tekstu, baza danych, wskaźniki wartościowania tekstu

W tradycyjnym podejściu produkt językowy sam w sobie nie jest traktowany jako praktyczna wiedza ludzka, a jedynie sposób jej przekazania. Umieszczając w kodzie komunikacyjnym myśl, którą chce się wyrazić, tworzy się zjawisko, składające się ze strony znaczeniowej i strony formalnej. Odbiorca powstałego komunikatu określa rodzaj (typ) kodu, w którym on występuje, a następnie rozpoczyna analizę jego strony znaczeniowej, porównując i przetwarzając wejściowe dane z posiadanymi abstrakcyjnymi konceptami i cechami dystynktywnymi. Koncept jest przy tym rozumiany jako abstrakcyjna jednostka myślenia, składająca się z czynników określających jego pełne wyobrażenie i znaczenie zależnie od wewnętrznych asocjacji mentalnych człowieka, a cecha dystynktywna – jako abstrakcyjny pojedynczy element myślowy, składający się na jedno charakterystyczne wyobrażenie, mogący posiadać funkcję znaczenia, emocji, rozumienia, sposobu przetwarzania oraz łączenia z innymi wyobrażeniami, czyli cząstka składowa konceptu.

W procesie komunikacji wg klasycznego modelu występują: nadawca, informacja zawarta w danym kodzie i odbiorca, a całe zdarzenie zachodzi w szerszej, otaczającej je rzeczywistości. G.Miller definiuje komunikowanie jako transmisję informacji z jednego miejsca do drugiego i wyróżnia przy tym pięć klas (cyt. za Nęcki 20):

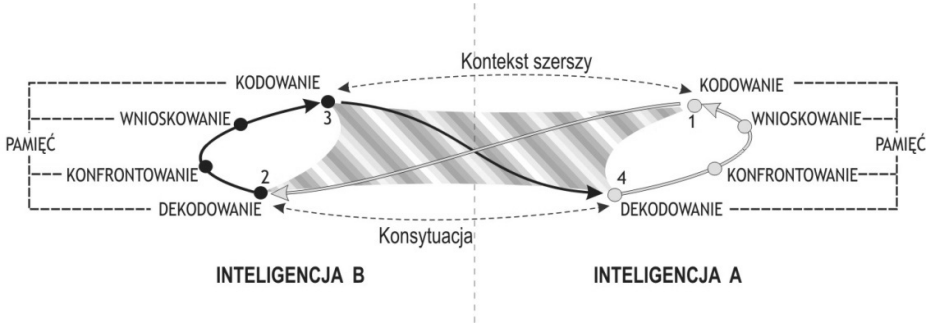
1. źródło – osoba, która wytwarza przekaz;
2. emisor – biologiczny system lub sztuczne urządzenie, które przetwarza informację w jakąś formę energii możliwą do przesłania – informacji „czystej" przesłać nie można;
3. kanał – środek, dzięki któremu pokonuje się dystans czasowy i przestrzenny między nadawcą i odbiorcą;
4. receptor – system zmieniający sygnały emitowane w formie pewnej energii na powrót w informację;
5. cel – odbiorca lub grupa odbiorców, do których przekaz ten był wysłany.

Powyższy układ komunikacyjny stanowi swego rodzaju dydaktyczną podstawę dla rozwijania i formułowania doskonalszych teorii komunikacyjnych, uwzględniających wpływ czynników ekstrajęzykowych, badanych przez matematykę, logikę, psychologię, medycynę, fizykę, itd. Przykładowo, w trakcie wydarzenia komunikacji językowej (WKJ) mogą występować znaki przekazywane poprzez wyraz twarzy, gestykulację, ułożenie ciała (kod kinezyczny), sygnały wpisane w zachowania kształtujące przestrzeń i odległość między mówiącym i słuchaczem (kod proksemiczny) i sygnały przekazywane przy pomocy wielkości liter i interpunkcji (kod graficzny). Kody te układają się w specyficzne systemy znakowe, rejestrowane w pamięci obok środków stricte językowych (Mikołajczuk 32–33).

Łącząc przedstawione założenia komunikacyjne z szerokimi możliwościami informatyki, szczególnie w zakresie emisji i recepcji, otrzymać można wielopłaszczyznowe WKJ, w którym mogą uczestniczyć nie tylko organizmy zbudowane w oparciu o związki węgla (człowiek), ale również twory techniczne oparte o układy krzemowe (komputer). Obydwa typy uczestników wykazują duże analogie w procesach działania. Pamięć ludzka stanowi obecnie niedoścignionу wzór kojarzenia, obciążony przez to pewną zawodnością. Technika informatyczna tworzy z kolei coraz pojemniejsze rodzaje sztucznej pamięci. Obecnie maszyny przy pomocy odpowiedniego oprogramowania są w stanie przetworzyć światło i dźwięk na odpowiedni rodzaj kodu (cyfrowego), przy pomocy którego zapisywane są informacje. Dalszym etapem jest stworzenie oprogramowania, przy użyciu którego maszyna będzie mogła dokonywać nowych wpisów do komórek pamięci poświęconym zakodowanym pojęciom. Postęp, jaki następuje w tej dziedzinie jest tak szybki, iż to, co do niedawna było fantastyką naukową jest obecnie rzeczą powszednią.

Jeden z aspektów komunikacji poświęcony jest udoskonalaniu translatorów maszynowych, a więc odpowiednich algorytmów, które w oparciu o posiadane dane informacyjne mają za zadanie przetworzyć otrzymaną informację w kodzie J1 na kod J2, uwzględniając jak najwięcej czynników gramatycznych, leksykalnych, stylistycznych, znaczeniowych i in.

Najważniejszym elementem WKJ jest proces rozumienia (w przypadku istot żywych), który został określony jako konfrontowanie pozyskanych danych z tymi posiadanymi w zasobach pamięci. Ten element stanowi obecnie główny problem informatyczny, gdyż wymaga stworzenia samowystarczalnego oprogramowania, które pozwoli maszynie na wytwarzanie i komunikowanie odpowiednich danych. Modelową wymianę informacji można w powyższym kontekście przedstawiać następująco:



Rys. 1. Model uniwersalny WKJ
Inteligencja A zawiera źródło i emisor poprzez określony kanał,
Inteligencja B – to docelowy odbiorca i receptor nastawiony na określony kanał
(schemat własny)

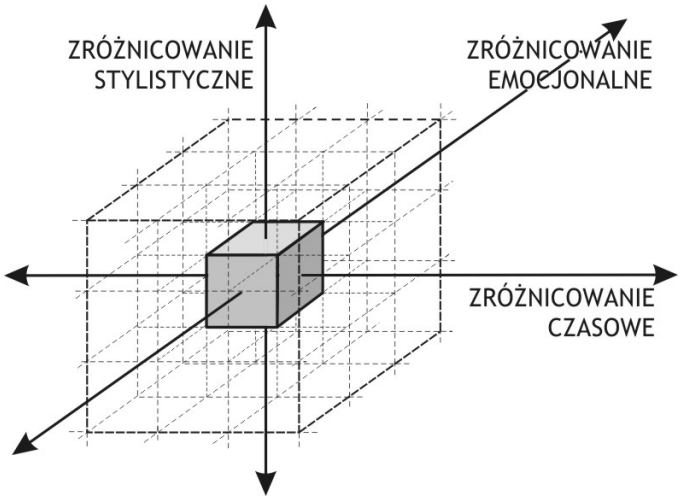
W tym kontekście słownik przyszłości, oparty na elektronicznej bazie danych przekształci się wręcz w informatyczny neurotezaurus, na bazie którego sztuczna inteligencja będzie w stanie dokonywać logicznej komunikacji. Perspektywy rozwoju leksykografii zmierzają w związku z tym w kierunku opracowywania złożonych baz danych, zawierających w sobie uporządkowane obszerne dane na temat każdego pojęcia (tj. swego rodzaju odpowiednik dużej liczby parametrów mikrostruktury słownika), dodatkowo odwołujące się do szerokiego zaplecza korpusów tekstowych oraz sieci Internet.

Zgodnie z tym, co przedstawiono wyżej postęp idzie w kierunku integracji w jednym produkcie wszystkich poziomów języka – od formy graficznej poprzez głosową aż do poziomu formalnej reprezentacji wiedzy. Koszt tak gwałtownego postępu jest jednak bardzo wysoki, co powoduje, że elektroniczne korpusy i leksykony dobrej jakości stworzono jedynie dla największych języków świata (Pawłowski 83). Tworzone we wspomniany sposób lingwistyczne modele językowe są podstawą przekładu maszynowego. Uogólniając, można stwierdzić, iż efektywność działania tych modeli jest bezpośrednio uzależniona od zdefiniowania odpowiednio dużej liczby tzw. wskaźników wartościowania tekstu. Przypisanie tym parametrom właściwych, opartych na podstawach znaczeniowych analizowanych wyrażeń wartości liczbowych będzie istotnym etapem, zbliżającym językoznawców i informatyków do skonstruowania doskonałego translatora maszynowego. Przykładowo wyodrębnienie i odpowiednie wartościowanie takich relacji jak np. *stary – nowy, osłabienie – nasilenie, rozpoczęcie – zakończenie, warunek – efekt, apatia – ekspresja* i in. pozwoli sztucznej inteligencji dokładniej określić przy użyciu odpowiedniego algorytmu najbliższy odpowiednik znaczeniowy w innym języku.

Do zilustrowania proponowanej parametryzacji tekstu posłużą podstawowe cechy doboru ekwiwalentów jak:

- indeks czasowy, określający dla jakiego okresu w rozwoju języka dany ekwiwalent był charakterystyczny lub jakie są prognozowane tendencje językowe (inaczej można go opisać jako relację na linii archaizm – futuryzm);
- indeks emocjonalny, rozumiany jako wskaźnik nacechowania wypowiedzi znaczeniem pozytywnym lub negatywnym w stosunku do określanego pojęcia;
- indeks stylistyczny, określający sposób kształtowania tekstu w zależności od sytuacji komunikacyjnej, w jakiej ma być użyty, celu oraz funkcji tekstu.

Model pojęcia, posiadającego swoiste pole znaczeniowe, opisywane ww. indeksami przedstawiono na poniższym rysunku. Centralny punkt pola, zobrazowany na poniższym modelu jako szary sześcian, stanowi pojęcie względnie neutralne pod względem stylistycznym, emocjonalnym, a także jest aktualnie używane we współczesnym języku ogólnym. Odnosząc się do metaforycznego wyobrażenia, jest to pudełko, w którym znajduje się jedno określenie, którego omawiane indeksy mają neutralną (zerową) wartość. Każdy inny sześcian zawiera w sobie kolejne określenie, którego przynajmniej jeden z indeksów ma wartość zmodyfikowaną.



Rys. 2. Model pola znaczeniowego.

Rolą tłumacza jest rozpoznanie w tekście oryginału właściwego szerokiego kontekstu i bieżącej bliskiej konsytuacji oraz zastosowanie w tekście przekładu odpowiednio nacechowanego ekwiwalentu. Przykładowo, dane pole znaczeniowe pojęcia *kobieta* może

zawierać m.in. określenia: *niewiasta, białogłowa, połowica, dziewczę, babsko, dziewucha, wiedźma, baba, kobita, słaba płeć, dama, pani, płeć nadobna, dziewczyna, facetka, babka, laska, sztuka* itd., a jego odpowiednik w języku angielskim – *woman* np.: *matron, missus, damsel, mistress, miss, hag, bimbo, witch, softer sex, finer sex, dame, madame, lady, ol`lady, comely lady, girl, madmoiselle, butch, hottie, chick, babe* itd. O ile tłumacz w procesie kształcenia, na podstawie własnej wiedzy językowej czy wreszcie praktyki zawodowej jest w stanie, przy użyciu odpowiedniej strategii translatorskiej podać najbliższy znaczeniowo i najlepszy kontekstowo ekwiwalent, o tyle algorytm programowy translatora maszynowego musi mieć podane pewne wyjściowe wartości liczbowe (parametry), wg których przypisze do podanego wyrażenia odpowiedni ekwiwalent. Innymi słowy, posiadane przez program parametry skierują algorytm programu do odpowiedniego „pudełka" z ekwiwalentem.

Idąc dalej tym tropem, komercyjne programy tłumaczące radzą sobie względnie dobrze z szykiem zdania czy rozpoznaniem fleksji, natomiast automatyczny dobór ekwiwalentu pod względem stylistycznym, czy znaczeniowym pozostawia jeszcze wiele do życzenia. Wsparciem dla algorytmów maszynowych ma stać się więc zestaw wskaźników wartościowania tekstu. Wprowadzone zakresy wartości tych wskaźników, opracowane zawczasu na próbkach tekstów (lub korpusach tekstowych różnych rodzajów tekstów) będą stanowiły podstawę wyboru przez program tłumaczący odpowiedniego ekwiwalentu.

Podsumowując wyniki badań prowadzonych przez autora, na obecnym etapie można założyć, iż podstawowy zestaw wskaźników wartościowania tekstu pozwalający w mechaniczny sposób przydzielić przez algorytm translatora maszynowego odpowiedni ekwiwalent, może składać się z 24 parametrów (indeksów), podzielonych na 4 grupy tematyczne, obecnie określane jako: a) układ czasoprzestrzenny; b) intensywność oddziaływania; c) stan i jakość i d) wyrażane relacje. Oprócz nich wstępna analiza statystyczna tekstu przed zasadniczym procesem translacyjnym może pozwolić np. na wstępne rozpoznanie stylistyki tekstu. Poniżej w tabeli 1 przedstawiono zestawienie wartości częstości użycia wyrazów z początku list frekwencyjnych zróżnicowanych stylistycznie tekstów[1]:

Dane te nie są jeszcze w pełni jednoznaczne, choć różnice pomiędzy tekstami prasowymi i językiem literackim a prawnym językiem specjalistycznym, szczególnie w pierwszych częściach list frekwencyjnych są powtarzalne i zauważalne. Język ogólny charakteryzuje się jednak częstym użyciem spójników, wykrzykników i modulantów. Uaktywnienie w programie analizującym opcji pomijania na liście frekwencyjnej spój-

[1] Wszystkie analizy frekwencyjne i statystyczne wykonano autorskim programem „Pantext" 2.0. Cytowane dane pochodzą z raportów końcowych generowanych po przeprowadzeniu analizy danego tekstu.

ników i innych niesamodzielnych części mowy daje rezultaty, które można uznać za bardziej miarodajne i różnicujące[2]:

Tabela 1.

rodzaj tekstu	% wyrazów, licząc od początku listy frekwencyjnej:[2]				
	10	20	30	40	50
blog znawcy cygar –	47,57%	60,52%	68,80%	75,20%	79,35%
B.Prus „Kamizelka" –	50,48%	63,85%	72,22%	78,84%	82,36%
B.Prus „Grzechy dzieciństwa"	63,48%	74,75%	81,18%	85,59%	88,99%
artykuł prasowy 1 –	48,20%	61,61%	69,96%	76,98%	80,83%
artykuł prasowy 2 –	38,67%	50,58%	60,96%	66,52%	72,15%
artykuł prasowy 3 –	38,39%	51,06%	58,99%	64,89%	70,79%
artykuł naukowy – medycyna	54,51%	68,55%	76,52%	82,06%	86,19%
referat z językoznawstwa –	46,84%	60,64%	69,28%	76,92%	80,88%
ustawa o bankowości –	77,78%	87,47%	92,43%	95,23%	96,95%
ustawa o telekomunikacji –	72,88%	84,33%	90,09%	93,43%	95,60%
ustawa o oświacie –	71,14%	83,06%	89,00%	92,68%	95,10%

Tabela 2.

rodzaj tekstu	% wyrazów, licząc od początku listy frekwencyjnej,					
	10	20	30	40	50	odrzucono
blog znawcy cygar -	**21,67%**	30,33%	37,45%	41,07%	44,68%	**37,51%**
B.Prus „Kamizelka" –	**19,85%**	28,10%	34,27%	38,07%	41,18%	**43,35%**
B.Prus „Grzechy dzieciństwa"	**26,46%**	35,11%	40,45%	44,64%	47,39%	**42,20%**
artykuł prasowy 1 –	**23,85%**	32,38%	38,96%	43,33%	46,62%	**36,87%**
artykuł prasowy 2 –	**20,26%**	30,26%	36,84%	41,89%	46,86%	**28,22%**
artykuł prasowy 3 –	**24,66%**	34,81%	40,04%	45,26%	50,48%	**23,40%**
artykuł naukowy – medycyna	**35,17%**	45,31%	51,4 %	56,30%	59,35%	**29,55%**
referat z językoznawstwa –	**28,62%**	39,88%	46,76%	52,64%	56,08%	**26,64%**
ustawa o bankowości –	**52,58%**	60,56%	64,89%	67,32%	68,84%	**28,43%**
ustawa o telekomunikacji –	**50,23%**	60,02%	65,04%	68,02%	69,39%	**25,95%**
ustawa o oświacie –	**47,66%**	57,45%	62,54%	65,75%	67,93%	**27,67%**

[2] Ubocznym spostrzeżeniem przy prowadzeniu badań statystycznych jest to, iż opierając się na przedstawionych w tabeli 2 wartościach można rozstrzygać, które teksty traktować jako specjalistyczne, a które, jako pisane językiem ogólnym, co ma swoje przełożenie na finansowe rozstrzygnięcia na rynku tłumaczeniowym.

[3] W nagłówku tabeli występują progi procentowe (10, 20, 30, 40, 50), które wskazują procentowy udział jednostek wyrazowych w ogólnej liczbie wszystkich wyrazów w analizowanym tekście (ilościowo). Wartości procentowe wypełniające tabelę to z kolei udział tych jednostek w całym tekście,

Wyliczenie przedziałów wartości charakterystycznych w oparciu o analizę większej grupy reprezentatywnych tekstów (porównywalnych objętościowo korpusach tekstów) stanowi jeden z elementów wskazujący algorytmowi maszynowemu optymalną i pożądaną ścieżkę wyboru ekwiwalentu.

Interpretując wartości podane w powyższej tabeli, można przyjąć, że teksty pisane w języku ogólnym, utwory literackie (proza) oraz część artykułów prasowych pisane są w stosunkowo podobnej manierze językowej. Charakteryzuje je stosunkowo duża powtarzalność wyrazów (i w związku z tym długa lista frekwencyjna) – pierwsze 10% wyrazów na liście frekwencyjnej stanowi ca 20–26% całej użytej leksyki, za to 36–45% tych tekstów stanowią wyrazy niesamodzielne. Z kolei teksty, które można uznać za tematycznie ukierunkowane (w domyśle specjalistyczne) ukazują pewną regułę jeśli chodzi o pierwsze 10% wyrazów na liście frekwencyjnej, które stanowią 50–55% całej użytej leksyki, a tylko 22–28% tych tekstów stanowią wyrazy niesamodzielne.

Ciekawym przykładem do interpretacji jest tekst prasowy nr 3, dla którego wskaźnik użycia pierwszych 10% wyrazów na liście frekwencyjnej (tj. 24,66%) plasuje go w grupie tekstów języka ogólnego. Jeśli porównać kolejne wartości wskaźników użycia dla 20%-50% wyrazów z listy frekwencyjnej, to wartości te zbliżają badany tekst do stylistyki tekstów specjalistycznych. Wskazuje na to również końcowy procent odrzuconych wyrazów niesamodzielnych – 23,40%.[4] Wskaźniki te odzwierciedlają faktyczną budowę artykułu, który poświęcony jest walucie – polskiemu złotemu przed denominacją, jednak pierwsza część artykułu to wspomnienia autora artykułu z czasów liceum, kiedy *za złotówki można było co najwyżej kupić ocet w sklepie, natomiast wszelki handel rozliczany w dolarach kwitł w szkolnym męskim klozecie*. Druga część artykułu przedstawia więcej danych faktograficznych i jest poświęcona jednemu tematowi – legalnym i nielegalnym formom wymiany walut. Ta monotematyczność znajduje swoje odzwierciedlenie w strukturze leksyki i jej opisie parametrycznym, co pokazuje możliwości rozpoznania przez algorytm maszynowy zmiany stylu i konwencji w różnych partiach badanego tekstu.

[4] Wg raportu programu „Pantext" pełny rozkład wartości dla „artykułu prasowego nr 3" wygląda następująco:

Pierwsze 10% wyrazen wystepuje na liscie 255 razy, co daje 24.66%; pierwsze 20% – 360 razy, tj. 34.82%; pierwsze 30% – 414 razy, tj. 40.04%; 40% – 468 razy, tj. 45.26% (odtąd przyrost nowej leksyki nie maleje); 50% – 522 razy, tj. 50.48%; 60% – 576 razy, tj. 55.71%; 70% – 630 razy, tj. 60.93%; 80% – 684 razy, tj. 66.15%; 90% – 738 razy, co daje 71.37%. Po analizie 100% wyrażeń na liście wyrazy występują 792 razy, co daje 76.60%, a włączone filtry tekstowe usunęły z listy 242 wyraz/ow,t.j. 23.40 %.

Badanie frekwencyjne tekstu znalazło praktyczne zastosowanie już wiele lat temu. Prawa Zipfa[5] w uogólnieniu odnoszą się do określenia szeregu prawidłowości językowych o charakterze kategorialnym i statystycznym, opisanych przez Zipfa za pomocą modeli funkcyjnych. W szczególności, Zipf zbadał związek pomiędzy:

- częstością wyrazów a ich pozycją na liście rangowej;
- częstością wyrazów a ich długością;
- częstością wyrazów a liczbą ich znaczeń;
- częstością wyrazów a ich wiekiem i pochodzeniem.

W literaturze lingwistycznej pojęcie „prawa Zipfa" kojarzone jest najczęściej tylko z pierwszą zależnością, opartą na powszechnie znanej prawidłowości, zgodnie z którą iloczyn rang i częstości słów z listy frekwencyjnej jest wartością stałą. W uproszczeniu prawo zakładało m.in., że chociaż pełny słownik jakiegokolwiek języka zawiera setki tysięcy słów, to na tysiąc słów najbardziej używanych, tzn. tych, które zajmują w słowniku częstościowym pierwsze tysiąc miejsc przypada około 80% tekstu (Lewin 28). Jest to bardzo istotny fakt z punktu widzenia tworzenia słowników frekwencyjnych i baz do przekładu maszynowego.

Kontynuując wątek list frekwencyjnych i baz tekstowych, wydawnictwa wprowadzające aktualnie na rynek nowe (lub aktualizowane) słowniki coraz dobitniej podkreślają, iż baza językowa w tych kompendiach została zaktualizowana i zweryfikowana przy pomocy korpusu językowego. Można pozytywnie odnosić się do tego typu chwytów marketingowych, gdyż zbadanie częstotliwości występowania poszczególnych wyrazów ma szczególne znaczenie choćby w procesie tworzenia słowników dydaktycznych, ponieważ pozwala wyodrębnić słownictwo podstawowe i/lub często stosowane, wprowadzane na początku procesu dydaktycznego, oraz wyrazy (wyrażenia) rzadziej stosowane, które w związku z tym można uwzględniać w późniejszych etapach nauki. Obliczenie częstotliwości wyrazów stało się więc jednym z najbardziej powszechnie wykorzystywanych narzędzi lingwistyki korpusowej.

Mianem korpusu można określić zbiór wytworów językowych, które zostały wybrane i uporządkowane według wyraźnych kryteriów, których zadaniem jest odwzorowanie jakiegoś języka. Do podstawowych kryteriów można zaliczyć (typologia por. Waliński):

1. Ilość – powinna być wystarczająca do przeprowadzenia miarodajnych badań; można przyjąć, iż przeciętny słownik zawiera około 10.000 określeń, średnia powtarzalność wyrazów w tekstach przytaczanych w niniejszym artykule to około 7, a zweryfikować znaczenie danego pojęcia można przy przynajmniej 20-krotnym wystąpieniu;

[5] Por. Lewin 30–32.

z tak przyjętych założeń wynika iloczyn, wskazujący iż korpus tekstowy dla słownika z 10.000 pojęć powinien zawierać jako niezbędne minimum 1.400.000 wyrazów.

2. Reprezentatywność – określenie to odnosi się do różnorodności języka, jaki powinien być reprezentowany w korpusie. Podstawowym celem tworzenia korpusu powinno być jak najbardziej odwzorowanie interesującego nas obszaru językowego w odpowiednich, naturalnych proporcjach, uwzględniających teksty oficjalne, popularno-naukowe, specjalistyczne, dydaktyczne oraz zapisy rozmów ustnych.
3. Jakość i autentyczność, rozumiana jako kategoria polegająca na jak najwierniejszym odtworzeniu naturalnej komunikacji językowej. W związku z tym postulatem zebrane teksty nie powinny być w żaden sposób wstępnie modyfikowane (zmiana szyku, stylistyki, nacechowania emocjonalnego) w celu zachowania oryginalnych intencji nadawców tych komunikatów, gdyż w ten sposób można analizować aktualne tendencje językowe już na etapie procesów myślowych.
4. Systematyzacja i unifikacja – zakłada zbieranie zasobów językowych w określonym formacie kodowania znaków (np. ASCII czy Unicode) oraz przyjmowanie wyraźnie oddzielonych znaczników, które można w każdej chwili odseparować od samego tekstu. Postuluje się przy tym wypracowanie pewnych symboli i oznaczeń informatycznych wspólnych dla wszystkich powstających korpusów.[6] Ważnym elementem badań jest wzajemna kompatybilność tekstów zapisywanych w popularnych edytorach tekstowych i powszechnie dostępnych formatach i analiz generowanych przez programy do analiz.
5. Udokumentowanie oznacza pełne i dokładne udokumentowanie zebranych w korpusie danych językowych (np. gramatycznych, funkcjonalnych czy etymologicznych), które są najczęściej przechowywane rozdzielnie od korpusu w formie bazy danych.
6. Format elektroniczny – ten postulat jest oczywisty, ponieważ zakłada możliwość przetwarzania zebranych tekstów za pomocą komputera.
7. Skończona wielkość – postulat ten zakłada określenie na początku projektu docelowej wielkości budowanego korpusu, odpowiednio do docelowych zastosowań. W momencie osiągnięcia zamierzonego rozmiaru powinien on zostać zamknięty, a jego wielkość nie powinna ulegać zmianie, co nie wyklucza włączenia zebranych zasobów językowych w przyszłości do innego korpusu.

---

[6] Program „Pantext" rozpoznaje naniesione w tekstach źródłowych odpowiednie znaczniki, służące np. do miejscowej graficznej wizualizacji tekstu. Inne znaczniki mogą służyć do oznaczenia kategorii gramatycznych danego pojęcia czy poziomu rejestru stylistycznego. Podstawowym formatem tekstu, akceptowanym przez program „Pantext" jest dokument tekstowy (*.txt). W tym samym formacie generowane są raporty statystyk i pracy analizatora.

Korpus tekstów musi być odpowiednio zrównoważony gatunkowo, chronologicznie, stylowo, terytorialnie i pod innymi względami, np. ze względu na wiek i płeć autorów. To właśnie założona uprzednio struktura oraz rodzaj wyszukiwarki różni korpusy naukowe od innych wielkich zbiorów tekstów, choćby internetowych archiwów gazet codziennych bądź ogólnych zasobów sieci.

Powyższe rozważania łączy wizja nowoczesnej leksykografii. Obecnie podstawą wszelkiej działalności wydawniczej jest jak najlepiej opracowana baza danych. Mikrostruktura takich baz, a więc budowa każdego rekordu bazy, zawierającego praktycznie wszystkie funkcjonujące w literaturze parametry, sprowadzi słownik w formie książkowej do dobranego tematycznie wydruku z bazy, a słownik elektroniczny – do interfejsu lub witryny internetowej, umożliwiającej dostęp do wykupionych zawczasu opcji dostępu. Ta sama baza danych posłuży również jako podstawa działania translatora maszynowego, a więc jedna baza stworzona według wszelkich zasad leksykografii może znaleźć kilka równoległych zastosowań. Wg J.Lukszyna i W.Zmarzer, a także autora parametry mikrostruktury mogą być pogrupowane na 12 zasadniczych typów: *rejestracyjne* (numer hasła, źródło, komentarz, data rejestracji), *formalne* (ortografia, transkrypcja, część mowy, warianty gramatyczne, skróty), *etymologiczne* (język źródłowy, język pośrednik, formy paralelne), *leksykalne* (ugrupowanie tematyczne, typ terminu, rejestr stylistyczny), *interpretacyjne* (różne typy definicji; def.predykatywna, skrócona def.intensjonalna, pełna def. intensjonalna, realna), *asocjacyjne* (synonimy, antonimy, homonimy, termin nadrzędny, podrzędny, pojęcia skojarzeniowe, pojęcia powiązane 4 rodzajami relacji semantycznych), *pragmatyczne* (neologizmy, terminy autorskie, terminy standardowe), *ilustracyjne* (diagramy, rysunki, wykresy), graficzne (liternictwo, symbole, wyróżnienia tekstowe), *ekwiwalencyjne* (dla odpowiedników w językach obcych), *użytkowe* (kolokacje: rzeczownikowe, czasownikowe, przymiotnikowe, frazeologizmy, a także teksty standardowe i korpusy tekstowe), *klasyfikacyjne* (określające położenie terminu w systemie językowym). Do każdego ekwiwalentu należy dodać jeden parametr zawierający informatyczny zapis różnych odcieni znaczeniowych, sygnalizowanych w niniejszym artykule. Całość, przy założeniu, iż ekwiwalenty będą opracowane dla 6 języków obcych, składa się na blisko 60 parametrów – elementów opisowych jednego rekordu. Dostosowanie bazy danych leksykograficznych do opisu modalności jednostek wyrazowych powoduje dodanie dodatkowych 24 parametrów (Lukszyn–Zmarzer 136, Karpiński „Zarys" 111).

Podsumowując powyższe rozważania warto wspomnieć, iż analiza frekwencyjno-terminologiczna danego korpusu tekstowego stanowi coraz powszechniej wykorzystywane narzędzie terminologiczne. Dzięki spisom wyrazów występujących w zbiorze tekstu wraz z przypisaną do każdego z nich częstotliwością występowania, można określić

częstotliwość poszczególnych użyć i wnioskować istnienie określonych reguł językowych w danych językach specjalistycznych. Listy frekwencyjne znajdują również zastosowanie przy tworzeniu słowników terminologicznych, mających za zadanie rejestrować dynamiczne zmiany w systemie terminologicznym. Na ich podstawie można ustalić udokumentowaną potrzebą zawodową normę zapożyczeń, skojarzyć zebrany korpus tekstowy z prototypem terminologicznym, wnioskować wpływ odpowiednich języków światowych na terminologię danej dziedziny czy też określić preferencje gramatyczne w konstruowaniu komunikatu w danym języku specjalistycznym w stosunku do norm danego języka narodowego. Na koniec wreszcie, dokładna analiza statystyczna każdego tekstu może dostarczyć podstawowych informacji (wskaźników liczbowych), które algorytm translatora maszynowego będzie mógł zastosować w procesie tłumaczenia maszynowego.

Bibliografia

Hetmański M. *Umysł a maszyny. Krytyka obliczeniowej teorii umysłu*. Lublin: UMCS, 2000.

Karpiński Ł. „Aby zrozumieć... – mechanizmy dyskursu specjalistycznego." *Linguodidactica*, t.X, Białystok: UwB WF, 2006. 69–82.

–-. *Wybrane założenia komputerowej analizy tekstów i gromadzenia danych*. Warszawa: KJS UW, 2009. 9–23.

–-. *Zarys leksykografii terminologicznej*, Warszawa: KJS UW, 2008.

Lewin J. „Znaki, język, matematyka". *Język, matematyka, cybernetyka*. Lewin J., Gastiew J., Rozanow J., eds. Warszawa: PWN, 1967. 7–57.

Lukszyn J. Zmarzer W. *Teoretyczne podstawy terminologii*. Warszawa: WLSiFW UW, 2001.

Mańczak W. *Problemy językoznawstwa ogólnego*. Wrocław, 1996

Mikołajczuk A. „O komunikacji zawodowej." *Praktyczna stylistyka*. *Komunikacja międzyludzka*. E.Bańkowska, A.Mikołajczuk, Warszawa: Książka i Wiedza, 2003.

Nęcki Z., eds. Kraków: Antykwa, 2000.

Pawłowski A. *Leksyka w lingwistyce kwantytatywnej i formalnej – przykład badań modelowych*. 28 grudnia 2008 <www.lingwistyka.uni.wroc.pl/~pawlowski/>

Waliński J. „Typologia korpusów oraz warsztat informatyczny lingwistyki korpusowej." B.Lewandowska-Tomaszczyk, ed, *Podstawy językoznawstwa korpusowego,* 2005. 27–41.