Porównania 1 (26), 2020
DOI: 10.14746/por.2020.1.16

Inez Okulska
Uniwersytet im. Adama Mickiewicza w Poznaniu

O wiele więcej niż Google Translate, czyli komputerowe przetwarzanie języka naturalnego (NLP) w translatoryce i translatologii

Powieść graficzna autorstwa Catherine Anyago oraz Davida Zane'a Mairovitza to intersemiotyczny przekład *Jądra ciemności* Josepha Conrada. Duet ten podjął się niemal niemożliwego wyzwania, jakim jest przełożenie Conradowej mglistości za pomocą obrazu. Mimo konieczności redukcji treści, jaką pociąga za sobą zmiana medium, w myśl zasady, że mniej znaczy więcej, autorzy odważnie sięgnęli po abstrakcyjne grafiki zajmujące całą stronę:

> Szare tło, na którym pośrodku jawi się czarna kula, okrągłe skupisko czarnych plam, narzucające myśl o tytułowym jądrze. […] To może być słońce […]. Ale to jądro, ta czarna kula na następnej stronie pojawia się, wyraźnie zmniejszona, na prostokątnym białym polu. W kolejnym kadrze, w tym samym prostokącie widzimy już dwie kropki umiejscowione symetrycznie, jedna u dołu, druga u góry, by w kadrze trzecim odkryć, że kropki te są oczkami na kostkach domina, ustawionymi na pierwszym planie na tyle blisko, że przesłaniają widok. W szczelinie między nimi dostrzec można tylko kawałek morza i latarnię morską. Kostki domina ustawione pionowo i poziomo wyglądają na pierwszy rzut oka jak domy, a czarne kropki jak okna. W kolejnych kadrach czarna kropka przyjmuje inną tożsamość: widoczna nad krajobrazem miejskim obrazuje słońce (Jarniewicz 2018: 226-227).

Artyści w marginalnym przedmiocie dostrzegli potencjał wizualnej metafory – zbliżona lub oddalona kostka domina gra oczkami, sugerując raz słońce, raz okienka, raz sama siebie. „Poprzez zmieniającą się tożsamość przedmiotu – przypominając o niewiarygodności doświadczenia naocznego, o ułomności percepcji, o konieczności korygowania naszych rozpoznań" (Jarniewicz 2018: 227) kostka ma ilustrować słowami opisaną w oryginale mglistość. Mglistość obrazu, narracji, formy.

Mamy więc geniusz tłumacza (intersemiotycznego) i nie mniej uderzający polot krytyka tego tłumaczenia pozwalający, w dobie tłumaczeń automatycznych, triumfować potędze człowieka na polu przekładu. Ale przecież pomiędzy niedoścignionym umysłem ludzkim a wątpliwym sukcesem translatora Google w rodzaju „Ala ma kota" oddanym jako „Ala owns a cat" czy prostym klasyfikatorem przypisującym nacechowanie emocjonalne spod znaku „pozytywny" i „negatywny", jak gdyby nie istniała cała skala ocen pośrednich, pomiędzy sztywnym, maszynowym wykonaniem zadania a twórczą ludzką pracą istnieje dziś pewien pomost.

Okazuje się, że niejeden geniusz translatorski lub translatologiczny może swoje niezaprzeczalnie ludzkie dzieło wspomagać komputerowymi metodami przetwarzania tekstu z korzyścią dla efektu i bez szwanku dla honoru. I że najczęściej tak właśnie robi, choć nie zawsze świadomy, iż w jego domu oprócz własnej, wciąż jeszcze niezawodnej inteligencji zamieszkała również ta sztuczna.

1. Przekład automatyczny nie tylko w translatorze

Kiedy komputer próbuje zagościć w dziedzinie przekładu, niemal natychmiast zaczyna i kończy w szufladzie przekładu maszynowego (komputerowego, automatycznego). I choć moim celem jest pokazać, że komputerowe przetwarzanie języka pojawia się i przydaje w zupełnie innych obszarach procesu i analizy przekładu, to jednak przekładu maszynowego nie sposób całkiem z tej opowieści wykluczyć.

Historia tłumaczenia automatycznego sięga końcówki lat czterdziestych ubiegłego wieku, kiedy to Warren Weaver zaproponował pierwsze koncepcje przekładu wykonanego z pomocą komputera, a także wprowadził pojęcie „computer translation". Mohamed Amine Cheragui zaproponował wyróżnienie pięciu stadiów rozwoju tej myśli na przestrzeni ostatnich dekad: faza początkowa, parsowanie i porzucenie złudzeń, nowe pomysły i nadzieje, szturm Japończyków oraz sieć i nowy zarys tłumacza (Cheragui 161-163). Zmieniały się głównie architektury systemów, podejścia (od pojedynczych słów przez całe frazy, po uwzględnianie kontekstu), a także sposoby ewaluacji wyników (por. Cheragui).

Cheragui zwraca też uwagę na jeszcze jeden istotny podział – funkcjonalny, bowiem przekład maszynowy (MT) to ogólna kategoria obejmująca zarówno narzędzia, jak i całe procesy zorientowane na zupełnie różne cele. I tak mamy przekład automatyczny generowany z myślą o:

1) służbach,
2) redaktorach,
3) tłumaczach,
4) autorach.

To zróżnicowanie wywodzi się z jednej strony z asymetrii potrzeb w zakresie jakości – moderator treści czy agent służb wymaga szybkiego dostępu do semantycznego kondesatu przy dużej dozie tolerancji wobec struktur gramatycznych czy stylistyki otrzymanego tłumaczenia, podobnie jest w przypadku redaktorów, którzy stanowią drugi człon translatorskiego duetu – maszyna dostarczyć ma przekładu „na brudno", nad którego redakcją pracować będzie człowiek. W zestawieniu Cheragui autorzy natomiast mają korzystać z MT w celach autopromocji, tutaj zatem waga przykładana do jakości otrzymanego przekładu będzie z pewnością większa, skoro od niej zależeć może potencjalny zagraniczny splendor.

Z drugiej strony natomiast różnorodność typów wynika z pojemności samej kategorii – pod pojęciem przekładu maszynowego oprócz popularnych translatorów, stanowiących osobne aplikacje webowe lub mobilne, z przystępnym interfejsem użytkownika i obsługiwanych świadomie do wykonania zadania tłumaczenia podanego tekstu, znajdują się również narzędzia pomniejsze, towarzyszące. I te skierowane są właśnie do samych tłumaczy, których pracy – znów – nie mają zastępować, lecz ją wspomagać. A należą do nich między innymi wszelakie pamięci tłumaczeniowe (translation memory) czyli tzw. narzędzia typu CAT (computer assisted translation)[1] – edytory zawierające glosariusze tworzone na podstawie wykonanych przez jednego (lokalnie) lub wielu (jeśli współdzielimy wyniki w chmurze) użytkowników. Narzędzia te są szczególnie przydatne w tłumaczeniach specjalistycznych, gdzie istotna jest konsekwencja terminologiczna. Ich pułapką oczywiście może być powtarzanie utartych, lecz błędnych zwrotów, które komputer będzie z uporem podsuwał jako statystycznie słuszne. Współczesne słowniki on-line to również w dużej mierze translatory, które nie tylko wyszukują hasła w statycznej bazie leksykonu, lecz dynamicznie

1 Więcej na temat narzędzi CAT zob. Krüger, natomiast refleksje na temat wpływu pamięci tłumaczeniowych na świat translatorski można znaleźć w: Pym.

generują przekład z korpusów równoległych (patrz Callison-Burch, Talbot, Osborne; Koehn), nieustannie reagując na „język ulicy" czy raczej Internetu. Takim słownikiem są na przykład babla.com czy context.reverso, prezentujące tłumaczenia na podstawie opublikowanych w sieci dwóch wersji językowych dokumentu zawierającego poszukiwaną przez nas frazę, oraz przeglądarka Google, która po wpisaniu słowa obcojęzycznego przeszukuje między innymi Wikipedię, by na jej podstawie przedstawić polskie tłumaczenie, jeśli znajdzie odpowiadające hasło.

Tłumacze nierzadko wykorzystują również intralingwalne narzędzia pomocnicze, takie jak słowniki synonimów, rymów, narzędzia sprawdzające pisownię czy gramatykę wbudowane w edytory tekstowe. Kolejnym istotnym elementem codziennego warsztatu tłumacza jest też interfejs człowiek-maszyna, czyli narzędzia pozwalające wczytywać, przesyłać czy dekodować dane, takie jak moduły rozpoznawania mowy w mobilnych notatnikach czy programy typu OCR do zamiany obrazu (skanu dokumentu) na edytowalny tekst – jedne wykorzystują tak zwane metody klasyczne, regułowe i statystyczne, inne zaprzęgają do pracy głębokie sieci neuronowe, ale wszystkie wymienione tu narzędzia opierają się na algorytmach przetwarzających język.

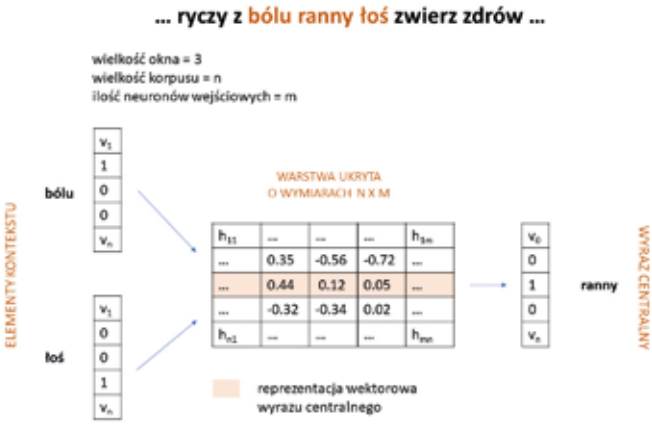
2. Jak komputer przetwarza język naturalny?

W jaki sposób komputer przeprowadza operacje na języku naturalnym[2]? Otóż musi zacząć od przekładu – zamienić słowa na liczby, które będą stanowiły zbiór domknięty ze względu na konkretne operacje, czyli taki, na którego elementach można dokonywać tych operacji w takich sposób, by ich wynik nadal mieścił się w tym zbiorze (te operacje to przede wszystkim miara cosinusowa, norma, odległość euklidesowa). A to znaczy przede wszystkim tyle, że komputer nie może przydzielać kolejnym słowom wyłącznie arbitralnych zestawów liczb (tak jak człowiek od zarania dziejów przydzielał obiektom arbitralne ciągi fonetyczne!), tylko musi postępować zgodnie z jakąś przyjętą logiką, która pozwoli połapać się, co jest czym (w jakiej relacji wobec czego).

Obecnie najczęściej wykorzystywanym pomysłem na kodowanie słów są tzw. osadzenia słów (*word embeddings*), czyli wektory od długości najczęściej 100 lub 300 elementów, w zależności od ustalonej liczby neuronów wejściowych.

2 Świadomie unikam tu pytania o to, jak komputer „rozumie" język, ze względu na fakt, że kwestia rozumienia wywołuje gorące, lingwistyczno-filozoficzne debaty w kontekście sztucznej inteligencji i przytoczenie możliwych stanowisk oraz prezentacja własnego w tym zakresie to materiał na całą osobną publikację.

Najpopularniejsze algorytmy tworzenia osadzeń słów to Word2Vec (Mikolov et al.) lub Glove (Pennington, Socher, Manning). Wektory te powstają w wyniku uczenia jednowarstwowej sieci neuronowej w procesie, który na wejściu przyjmuje możliwie duży korpus tekstu (np. zrzut Wikipedii w danym języku) i próbkuje go na wektory opisujące częstość występowania danego słowa w korpusie (w formie one-hot, czyli 1 zaznacza obecność, 0 brak), by na wyjściu zależnie od wykorzystanego algorytmu uczenia skipgram lub continuous-bag-of-words zwrócić odpowiednio przewidziany kontekst słowa (o zadanej wielkości okna) lub, otrzymawszy kontekst, przewidzieć jego słowo centralne (również w postaci wektorów o elementach ze zbioru {0,1}).



Ryc. 1. Uproszczony schemat ideowy działania algorytmu Word2Vec w konfiguracji CBOW, czyli przewidującej wyraz centralny na podstawie zadanego kontekstu (opracowanie własne)

Proces uczenia polega z grubsza na wstecznym przypisywaniu wag (liczb) w warstwie ukrytej, które tworzą ogromną macierz o wymiarach równych ilości próbek w korpusie (wiersze) oraz ilości neuronów wejściowych (kolumny), gdzie ta ostatnia decyduje później o długości wektora, bowiem wiersz macierzy odpowiadający pozycji wejściowego słowa to właśnie wspomniane osadzenie. W ten sposób skonstruowana reprezentacja wektorowa sprawia, że zachodzą między tymi wektorami wymierne zależności, a z ich kombinacji można pozyskać informacje między innymi o semantyce, pozycji (kontekście) czy nacechowaniu emocjonalnym danego elementu. Takie podejście okazało się przełomowe nie tylko w zadaniach przetwarzania tekstu jednojęzycznego,

ale również w tłumaczeniu maszynowym, ponieważ tak skonstruowana wektoryzacja okazuje się doskonale rzutować słowa i relacje jednego języka na inny. Wektory osadzeń to więc niezwykle przydatne *tertium comparationis* czy wręcz wymarzony przez Waltera Benjamina przedjęzyk, wspólny wszystkim i nieodzowny w słynnym zadaniu tłumacza:

> Otóż tak jak skorupy naczynia – po to, by można je było złączyć – muszą nie tyle być identyczne, ile odpowiadać sobie w najdrobniejszych szczegółach, tak też przekład – zamiast upodabniać się do sensu oryginału, musi raczej ukształtować się we własnym języku według właściwego oryginałowi sposobu wskazywania, […] tak by – podobnie jak skorupy rozpoznajemy jako ułomki jakiegoś naczynia – oryginał i przekład można było rozpoznać jako ułomki pewnego większego języka (Benjamin 37).

We współczesnym przekładzie maszynowym zarówno oryginał, jak i przekład reprezentowane są za pomocą wektorów liczb o takiej samej strukturze, w której zanikają wszelkie kwantyfikatory narodowe czy kulturowe, mamy więc rzeczywiście do czynienia z owym „pewnym, większym językiem", językiem liczb i wymiarów, a to otwiera niezbadane dotychczas możliwości analizy czy symulacji[3].

3. Algorytmy przetwarzania języka naturalnego w translatologii – krytyka i badania przekładoznawcze

Jedną z prawdziwych perełek pośród imponujących zbiorów architektonicznie imponującej biblioteki Beinecke na Uniwersytecie Yale w New Haven jest tzw. manuskrypt Wojnicza: ni to podręcznik, ni magiczna księga, najprawdopodobniej stworzona w XV lub XVI wieku w Europie, zawierająca rozdziały botaniczne, kosmogoniczne i anatomiczne lub nawet medyczne (por. Voynich Manuscript, Beinecke Library). Księga ta zyskała swój przydomek od nazwiska polsko-amerykańskiego antykwariusza, który odkupił ją w 1912 roku. Autor tekstu pozostaje nieznany, podobnie jak treść manuskryptu, jego pochodzenie, gatunek czy nawet język, w jakim został napisany. Ten ostatni fakt, jak łatwo sobie wyobrazić, od lat frapuje naukowców i to – jak się okazuje – nie tylko specjalistów w dziedzinie językoznawstwa. Niezwykła to również gratka z ducha translatologiczna, bo przekład sam w sobie jest przecież zadaniem arcytrudnym czy – jak chciał

3 O kreatywności sieci neuronowych na polu literackim zob. Okulska 2019.

Benjamin – w zasadzie niemożliwym, ale choćby najbardziej skazany na porażkę przekład zacząć trzeba od identyfikacji oryginału – a jak decydować o strategii czy procesie przekładu, jeśli nie wiadomo nawet, z jakiego języka tłumaczymy? Tym bardziej, że istnieje zgodna hipoteza o tym, iż tekst manuskryptu został poddany szyfrowaniu, a więc sam w sobie jest już przekładem i to przekładem intersemiotycznym (podejrzewano substytucję alfabetyczną oraz wizualizację za pomocą nieznanego alfabetu).

Artykuły donoszące cyklicznie o rzekomym złamaniu kodu i rozszyfrowaniu manuskryptu można by kolekcjonować na osobnym regale, choć rozbieżność doniesień wymagałaby przynajmniej odrębnych półek. Niektórzy badacze byli wręcz zdania, że manuskrypt jest mistyfikacją i nie został napisany w żadnym istniejącym języku naturalnym. W styczniu 2018 roku pojawiły się kolejne entuzjastyczne nagłówki, ale tym różniły się od poprzednich rewelacji, że zapowiadały udział sztucznej inteligencji. Duet badaczy z zakresu informatyki technicznej, Bradley Hauer oraz Grzegorz Kondrak z University of Alberta w Kanadzie postanowili użyć klasyfikatorów, które w efekcie miałyby wskazać język lub zawężoną grupę języków, które pod kątem wzorców strukturalnych najbardziej odpowiadają „wyrazom" zawartym w manuskrypcie. Koncepcja zakładała wykorzystanie jednego tekstu istniejącego w możliwie dużej liczbie oficjalnych tłumaczeń (a więc o potwierdzonej ekwiwalencji[4]) – i tu z pomocą przyszła Deklaracja Praw Człowieka, dostępna w 380 wersjach językowych. Zadaniem algorytmów było rozpoznanie i nauczenie się wzorców morfosyntaktycznych charakterystycznych dla danego języka. Ta jakże fascynująca koncepcja identyfikacji unikalnych strukturalnych sygnatur języków narodowych opierała się głównie na zadaniu szyfrowania i deszyfracji, z wykorzystaniem substytucji i anagramizacji (Hauer, Kondrak), czyli alfabetycznego zaszumiania (i odszumiania) danych, czego przykład można zilustrować następująco:

```
(a) organized compositions through improvisational music into genres
(b) fyovicstu dfnrfecpcfie pbyfzob cnryfgcevpcfivm nzecd cipf otiyte
(c) otvfusyci cpifenfercfd bopbfzy fgyiemcpfcvrcnv nczed fpic etotyi
(d) adegiknor ciimnooopsst ghhortu aaiiilmnooprstv cimsu inot eegnrs
(e) adegiknor compositions through aaiiilmnooprstv music into greens
```

Ryc. 2 Przykładowe etapy szyfrowania i deszyfracji tekstu (z użyciem substytucji i anagramizacji). Źródło: Hauer, Kondrak 80

4 Nie sposób tu nie pomyśleć o Theo Hermansie i jego cennych refleksjach na temat pojęcia ekwiwalencji w kontekście tekstów urzędowych. Por. Hermans.

Kiedy osiągnięty został odpowiednio niski błąd klasyfikacji, badacze zastosowali algorytmy do manuskryptu, by stwierdzić, że napisano go w języku starohebrajskim, zanim został poddany szyfrowaniu. Co więcej, pokusili się nawet o przełożenie pierwszego zdania na angielski, które miało brzmieć „She made recommendations to the priest, man of the house and me and people" (Vincent)[5], a więc robiło niezwykłe wrażenie ze względu na jednoczesny potencjał semantyczny i nieco tajemniczą składnię.

Niestety, po pierwszej fali zachwytu znów nadeszło zwątpienie i jeszcze silniejsza fala krytyki, zarzucająca badaczom zmarnowanie potencjału koncepcji – po pierwsze ich statystyczne metody okazały się bardzo proste (wcale nie ma tu mowy o oczywistych, zdawałoby się, sieciach neuronowych czy głębokim uczeniu), po drugie sygnatur szukano w normach statystycznych języków nowożytnych (skąd więc starohebrajski?) (por. Vincent) i po trzecie, przekład wstępu wykonany za pomocą Google Translate przelał czarę goryczy – to my tu studiujemy języki od lat, a kod łamie bezduszna maszyna? A figę! I, jakby na złość, zupełnie w środowisku badaczy tajemniczego starodruku odcięto się od rezultatów dociekliwych informatyków.

Zabrakło doskonalenia narzędzia od strony technicznej, ale również współpracy z ewentualnymi hebraistami, zabrakło czynnika ludzkiego, który zechciałby zaufać mocy obliczeniowej maszyny. Choć ambiwalentna i bez jednoznacznego sukcesu, historia ta powinna jednak służyć za przykład na to, że w dyskursie towarzyszącym przekładowi, w jego niezbędnej analizie, komputerowe metody przetwarzania języka mogą okazać się niezwykle przydatne choćby po to, by zaprezentować nową perspektywę, by jedne hipotezy wykluczyć, a inne zasugerować.

3.1. To nie w jego/jej stylu – stylometria w służbie przekładu

Statystyczne metody pod postacią stylometrii znalazły zresztą szerokie zastosowanie w literaturoznawstwie historycznym i niejednokrotnie pomogły rozstrzygnąć kwestie autorskich atrybucji na podstawie właśnie autorskich sygnatur nieświadomie zaszytych w tekście. Na gruncie naszych rodzimych badań nieocenionym źródłem zarówno własnych narzędzi (pakiet *stylo* dla języka R)

5 Zdanie to można przełożyć jako: „Wydała zalecenia księdzu, panu domu, mnie oraz ludziom". Jako zdanie otwierające oczywiście budzi ono wiele pytań i niedomowień – kim jest ona, dlaczego pierwszoosobowe „ja" jest wyodrębnione w tej enumeracji obok pozostałych ludzi oraz czego owe zalecenia dotyczyły, skoro zostały skierowane do takiego zestawu adresatów.

(Eder, Rybicki 107-121), jak i eksperymentów prezentujących intrygujące interpretacje jest grupa krakowskich językoznawców-anglistów z Uniwersytetu Jagiellońskiego pod przewodnictwem Jana Rybickiego. Rybicki oprócz literackiej mapy stylometrycznej prezentował ciekawe wyniki dotyczące stylistycznych profili tłumaczy i klasteryzacji dzieł pod kątem stylu narzucanego w tłumaczeniu (Rybicki 2014: 106-128; Rybicki 2017). Fantastycznie pokazał również, że choćby najskrupulatniejsza redakcja mająca na celu scalenie tekstu, by złagodzić czytelniczy dysonans, nie zamaskuje prawdy ukrytej w liczbach: kiedy po śmierci Anny Kołyszko tłumaczenie *Nights and Days* Virginii Woolf przejęła Magda Heydel, klasyfikator opracowany przez Rybińskiego i jego zespół dobitnie wyznaczał linię szwu przebiegającą w miejscu rzeczywistej zmiany tłumaczek (Heydel, Rybicki 708-717). Przydatność statystycznej obróbki tekstu w analizie różnych poziomów ekwiwalencji przekładowej poetyki sugerowałam również niegdyś nieco przewrotnie na podstawie drobnego narzędzia, jakim jest kalkulator stopnia „zaciemnienia" tekstu, czyli tzw. fog index, przypisujący tekstowi poziom wykształcenia autora (i adekwatnie poziom wykształcenia czytelnika wymagany do zrozumienia takiego tekstu) (por. Okulska 2015: 152-165).

Zresztą bardzo na czasie, bo po modzie na bliską lekturę, ujmującą tekst albo jego pojedyncze elementy w kadrze makro, która interesuję się tylko zbliżeniem i pomija wszystko, co się w tym kadrze nie mieści, nastała era zachwytu nad czytaniem odległym. Kolejne kroki w tył w celu złapania szerszej perspektywy obracają w palcach już nie tylko tekst, ale cały kon-tekst, autora, biografię, czasy, dyskursy, sąsiedztwo, potencjał: the big picture. Z pomocą przychodzi tu z pewnością humanistyka cyfrowa dostarczająca ogromnych zbiorów danych, a więc big data, z których można wydobywać wiedzę (*text mininig*), czyli obliczać zależności, trendy i prawdopodobieństwa.

3.2. Wydobywanie tematyki

Klasyfikatory uczone z nadzorem lub bez niego są dziś tylko bardziej złożone, precyzyjne, a nawet kreatywne, pytanie brzmi więc nie: jak chcemy odpowiedzieć na dane pytanie, a raczej: o co zapytać – dyskurs przekładowy, krytyka i analiza samych tłumaczeń, ale także ich serii, kontekstów, relacji międzyludzkich proszą się o interpretację danych obliczeniowo przekraczającą zdolności człowieka. Na przykład cyfrowe zasoby Żydowskiego Instytutu Historycznego oferują ciekawe moduły wizualizacji grafów reprezentujących połączenia ludzi, miejsc i tematów na podstawie informacji zawartych w zdigitalizowanych dokumentach – a więc dane, które mogą się okazać niezwykle cenne w budowaniu historyczno-społecznego kontekstu danego przekładu czy życiorysu tłumacza.

Warto badać też sam dyskurs – naukowcy (Goldstone, Underwood 359-384) przeanalizowali ponad 22 tysiące tekstów z zakresu literaturoznawstwa autorstwa ponad 13 tysięcy badaczy, które zostały opublikowane na przestrzeni całego XX wieku, by przyjrzeć się dynamice dyskursów reprezentowanych przez grupy pojęć (np. temat nr 82: przemoc, siła, krew, strach itd.). Na wykresach można prześledzić, w jakich latach wzrastała popularność danych tematów i z omawianiem jakich dzieł łączył się dany temat. Interpretacja tych danych wymaga zresztą analizy wielokryterialnej, bo rosnący trend mógł jednocześnie oznaczać nagromadzenie danego tematu w literaturze aktualnie publikowanej lub wzrost zainteresowania dziełami dawniejszymi, które dany temat poruszały, lub też że starsze dzieła zostały poddane interpretacyjnej rewizji i nauczono się czytać je przez pryzmat modnych tematów.

3.3. Kto dyktuje kanon?

Niezwykle ważkim tematem jest również kanon literatury w przekładzie. Franco Moretti, popularyzator wspomnianego pojęcia *distant reading,* wraz z Matthew Jockersem parę lat temu założył na Uniwersytecie Stanforda Literary Lab, który jest nieocenionym źródłem doniesień na temat statystycznych badań przeprowadzanych na korpusach literackich i literaturoznawczych. W „Pamfletach" (Porter, Moretti) znajdują się między innymi refleksje na temat istoty i możliwych definicji kanonu oraz roli cyfrowych archiwów w badaniach nad kanonem. Kontynuując myśli Franka Morettiego oraz J.D. Portera, można powiedzieć, że archiwum to korpus, a kanon to odpowiednio sparametryzowana norma korpusu.

Algorytmy przetwarzające język, w tym wypadku ujęty w ramy tekstu literackiego, pozwalają bowiem na różnorodne statystyczne obliczenia, które mogą powiedzieć coś nowego o podobieństwie i różnicach dzieł czy trendach odbiorczych. Popularność czytelniczych portali społecznościowych, takich jak analizowane przez laborantów goodreads.com albo nasze rodzime lubimyczytac.pl[6], gdzie nieprofesjonalni czytelnicy mogą wystawiać oceny książkom (w formie zarówno dziesięciogwiazdkowej skali, jak i komentarzy opisowych – niektóre z nich to pełnoprawne gatunkowo recenzje), sprawiła, że pojęcie kanonu uległo zachwianiu. Odtąd oprócz kanonu tworzonego przez elitarną grupę profesjonalnych krytyków (choć i tu z powodu idiosynkrazji zawsze królował pluralizm) mamy namacalnie do czynienia z kanonem oddolnym, czytelniczym, liczonym

6 Analiza ocen czytelniczych opublikowanych na tym portalu wchodzi w skład bazy nieprofesjonalnej krytyki literackiej, nad badaniem której prowadzę obecnie prace.

nie tylko poczytnością, ale i wyrażoną explicite oceną tej lektury – słupki sprzedaży mówiłyby przecież tylko o rozprzestrzenianiu się danego nazwiska wśród odbiorców w myśl zasady „nieważne jak mówią, byleby mówili".

Jak to kanoniczne otwarcie ma się do przekładu? Już dawno temu Jerzy Jarniewicz przypomniał, że tłumacze to bardzo ważni twórcy kanonu literatury w przekładzie, i wyróżnił dwa ich rodzaje – tłumaczy legislatorów i tłumaczy ambasadorów. Ci pierwsi mieli kierować się własną intuicją wobec potencjalnej wartości dodanej, jaką wniesie do języka i kultury rodzimej dane dzieło obce, niezależnie od jego pozycji w jakimkolwiek kanonie. Ci drudzy natomiast, jak pisze badacz:

> nie próbują ustanawiać własnych hierarchii wartości. Nie piszą swojej historii literatury angielskiej, francuskiej, niemieckiej czy rosyjskiej, nie proponują przewartościowań czy ryzykownych samodzielnych ocen. Przeciwnie, przyglądają się temu co w kulturze, z której tłumaczą, uchodzi za najlepsze, najciekawsze, śledzą to, co tę kulturę współkształtuje (Jarniewicz 2012: 14).

Przy tak sformułowanym zadaniu pojawia się pytanie o miarę oceny tego, co najlepsze, najciekawsze – czy decyduje o tym garstka profesjonalistów publikująca swoje recenzje i omówienia w branżowej prasie, czy społeczność czytelnicza? Dostęp do ogromnych zasobów danych i algorytmów odnajdujących w tym gąszczu istotne informacje pozwalają stworzyć obraz tego, „co w kulturze uchodzi za najlepsze" z punktu widzenia użytkowników tej kultury, laików, odbiorców, do których teoretycznie adresowana jest literatura. Decyzja wydawnicza o wyborze utworu do przekładu oparta na rankingu czytelniczym byłaby bliższa idei przekładu funkcjonalnego, gdzie oprócz dzieła tłumaczy się również recepcję – można uznać, że immanentnym elementem dzieła, tak jak gatunek, staje się również jego pozycja (w tym wypadku wysoka, ze względu na sukces czytelniczy). I ta pozycja właśnie mogłaby podlegać próbie przekładu – na etapie wyboru samego dzieła, strategii translatorskiej, projektowania okładki, tytułu czy innych elementów „obliczonych" na sukces w kulturze docelowej[7]. Ciekawym aspektem analizy jest również ewentualna niemożliwość przekładu – tym razem już nie tylko samego dzieła, lecz również jego pozycji.

7 Tutaj ciekawym przykładem może być tłumaczenie kryminału *Bezcenny* Zygmunta Miłoszewskiego, który we współpracy z tłumaczką Antonią Lloyd-Jones przepisywał całe fragmenty, tłumacząc swoją wizję artystyczną na oczekiwania anglosaskich czytelników. Daleko idące ustępstwa podyktowane były taką właśnie próbą przełożenia nie tylko samego dzieła, ale jego pozycji (sukcesu) w kulturze rodzimej.

• • •

Komputerowe metody przetwarzania języka przychodzą z pomocą zarówno w procesie przekładu, ze względu na szereg narzędzi wspomagających, takich jak adaptacyjne słowniki kontekstowe, pamięci translatorskie czy przyjazne interfejsy człowiek-maszyna, ale także w całym translatologicznym dyskursie, bo pozwalają zdobyć i wydobyć wiedzę na temat kontekstu i zależności tekstów, tematów, osób i innych. Współczesne możliwości pozwalają na akwizycję gigantycznych zasobów danych – dzięki szeroko zakrojonym projektom cyfryzacji dziedzictwa kulturowego oraz popularności platform oferujących przestrzeń wypowiedzi. W ten sposób mamy dostęp i do dzieł, i do dyskursu krytycznego i ta dostępność zachęca, by myśleć szeroko, patrzeć daleko za horyzont – thing big, big data, chciałoby się rzec. Wiedza to statystyka, ale statystyka poddana interpretacji. Więc bez obaw – algorytmy serwują sugestie i wizualizują tendencje, ale za ostateczną interpretację i tak odpowiada człowiek – doświadczony profesjonalista w swojej dziedzinie, który wie, jak pozyskane informacje połączyć w spójną, ciekawą opowieść. Czynnik ludzki w przekładzie nadal ma się świetnie.

| Bibliografia

Benjamin, Walter. „Zadanie tłumacza". Przeł. Adam Lipszyc. *Literatura na Świecie* 5-6 (2011). S. 27-41.

Callison-Burch Chris, Talbot David, Osborne Miles. "Statistical machine translation with word-and sentence-aligned parallel corpora". *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (2004). S. 175-182.

Cheragui, Mohamed Amine. "Theoretical Overview of Machine translation". *Proceedings ICWIT* (2012). S. 161-163.

Eder Maciej, Rybicki Jan, Kestemont Mike. "Stylometry with R: a package for computational text analysis". *R Journal* 8(1) (2016). S. 107-121. Web. 15.05.2019. <https://tinyurl.com/ybjvxc7g>

Goldstone Andrew, Underwood Ted. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us". *New Literary Studies* 45.3 (2014). S. 359-384.

Hauer Bradley, Kondrak Grzegorz. "Decoding anagrammed texts written in an unknown language and script". *Transactions of the Association for Computational Linguistics* 4 (2016). S. 75-86.

Hermans, Theo. *Narada języków*. Red. M. Heydel, K. Szymańska. Przeł. Magda Heydel et al. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego, 2015.

Heydel Magda, Rybicki Jan. "The Stylistics and Stylometry of Collabrative Translation: Woolf's Night and Day in Polish". *Literary and Linguistic Computing* 28 (2013). S. 708-717.

Jarniewicz, Jerzy. *Gościnność słowa*. Kraków: Wydawnictwo Znak, 2012.

Jarniewicz, Jerzy. *Tłumacz między innymi*. Wrocław: Ossolineum 2018.

Koehn, Philipp. "Europarl: A parallel corpus for statistical machine translation". *MT summit* 5 (2005). S. 79-86

Krüger, Ralph. *From Situated Translation to CAT Tool Usability*, EST Congress 2016. Translation Studies: Moving Boundaries. Web. 20.08.2019. <https://tinyurl.com/ycp3w8qm>.

Mikolov Tomas, et al. "Distributed representations of words and phrases and their compositionality". *Advances in neural information processing systems* (2013). S. 3111-3119.

Moretti, Frank. "Patterns and Interpretation". *Pamphlet* 15 (2017). Web. 20.09.2019. <https://tinyurl.com/y834za4h>

Okulska, Inez. „Poetyka przekładu". *OderÜbersetzen* 5 (2015). S. 152-165.

Okulska, Inez. „Sieci neuronowe typu GAN i GPT-2, słowa zużyte i kreatywność, czyli literacki second-hand". *Forum Poetyki* 18 (2019). Web. 12.03.2020. <https://tinyurl.com/y7y5mq3w>

Pennington Jeffrey, Socher Richard, Manning Christopher D. "Glove: Global vectors for word representation". *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014). S 1532-1543.

Porter, J.D. "Popularity/Prestige". *Pamphlet* 17 (2018). Web. 20.09.2019. <https://tinyurl.com/y7uvxofl>

Pym, Anthony. "What technology does to translation". *Translation and Interpreting* 3 (2011). S. 1-9.

Rybicki, Jan. „Pierwszy rzut oka na stylometryczną mapę literatury polskiej". *Teksty Drugie* 2 (2014). S. 106-128.

Rybicki, Jan. „Drugi rzut oka na stylometryczną mapę literatury polskiej". *Forum Poetyki* 10 (2017). Web. 15.09.2019. <https://tinyurl.com/y99rqngz>

Vincent, James. "AI didn't decode the cryptic Voynich manuscript – it just added to the mystery". *The Verge* 1 Feb (2018). Web. 17.08.2019. <https://tinyurl.com/ybn24fo8>

Voynich Manuscript. Opis zasobu na stronie biblioteki Beinecke Library. Web. 24.09.2019. <https://tinyurl.com/semce6e>

| Abstrakt

Inez Okulska
O wiele więcej niż Google Translate, czyli komputerowe przetwarzanie języka naturalnego (NLP) w translatoryce i translatologii

Przewrotna jest rola postępu – im więcej technologicznego rozwoju, tym większy udział człowieka – w koncepcji, formułowaniu zadań, interpretacji wyników, nadzorze i korekcie. Hierarchia jest zachowana, człowiek wciąż nieodzowny, ale to nie znaczy, że w pewnych obszarach maszynowy potencjał rzeczywiście nie przewyższa ludzkiego i że nie warto z tej przewagi skorzystać. Przetwarzanie języka naturalnego (NLP) to dziedzina niemłoda, ale w ostatnich latach dzięki rozkwitowi metod uczenia głębokiego (deep learning), mody na maszynowe wnioskowanie (data/knowledge mining) czy nowym sprzętowym interfejsom (m.in. zaawansowane rozpoznawanie obrazu) komputerowa analiza tekstu przeżywa istny renesans. W odniesieniu do translacji przyjęło się mówić i pisać głównie o coraz doskonalszych lub właśnie zupełnie niemożliwych algorytmach dla kolejnych par języków czy coraz większej precyzji samego tłumaczenia. Niniejszy artykuł przedstawia natomiast nieco szersze spektrum procesu tłumaczenia i przygląda się elementom przekładowi towarzyszącym (jak choćby krytyka), w których wykorzystanie metod NLP może przynieść nowe, ciekawe wyniki. Wyniki, których ze względu na ograniczoną moc obliczeniową człowiek nie jest w stanie osiągnąć. Omówione zostały takie aspekty jak wektorowa reprezentacja języka, stylometria i jej zastosowania czy analiza wielkich zbiorów danych – wszystko to na potrzeby szeroko rozumianych translacji i translatologii.

Słowa kluczowe: przekład literacki, teoria przekładu, przetwarzanie języka naturalnego, lingwistyka komputerowa, przekład maszynowy, uczenie maszynowe, big data

| Abstract

Inez Okulska
Far Beyond Google Translate: Natural Language Processing (NLP) in Translation and Translatology

The more technological development, the greater the participation of the human – in formulating tasks and problems, supervising and improving automated processes

and interpreting their outcomes. The hierarchy is preserved, humans are still indispensable, but it does not mean that in certain areas of machinery the potential does not really exceed that of the human and that this advantage is not worth exploiting. Natural language processing (NLP) is not a young field, but in recent years, thanks to the thrive of deep learning methods, data and knowledge mining or new human-machine interfaces, computer text analysis is experiencing a real renaissance. As far as translation is concerned, it is mostly algorithms for machine translation that are being discussed. This article, on the other hand, presents a slightly broader spectrum of the translation process and looks at the accompanying elements (such as criticism) in which the use of NLP methods may bring new and interesting results. Results which, due to limited computing power, humans are unable to achieve. The discussion in the paper covers such aspects as the vector representation of language, stylometry and its application, or the analysis of large data sets – all for the purposes of translation and translatology.

Keywords: literary translation, theory of translation, natural language processing, computational linguistics, machine translation, machine learning, big data

| Nota o autorze

Inez Okulska – doktor nauk humanistycznych w zakresie literaturoznawstwa. Po przejściu barwnej ścieżki humanistycznej (na którą składały się m.in. lingwistyka, kulturoznawstwo, filozofia) zakończonej podoktorskim stażem na Harvard University (Wydział Komparatystyki Literackiej), ukończyła studia magisterskie na Automatyce i Robotyce Politechniki Warszawskiej. Metody sztucznej inteligencji, a w szczególności metody przetwarzania języka naturalnego, którymi obecnie zajmuje się naukowo w Państwowym Instytucie Badawczym NASK, doskonale łączą te, odległe zdawałoby się, dziedziny. Zwłaszcza że najchętniej aplikuje je właśnie do analizy materiału literackiego.
E-mail: inez.okulska@gmail.com