INVESTIGATIONES LINGUISTICAE VOL. XXXI, 2014

AL. NIEPODLEGŁOŚCI 4, 60-874, POZNAŃ – POLAND

Błędy programu do obróbki korpusu, podczas badań korpusowych słownictwa biznesowego i prawnego w języku wietnamskim, na przykładzie programu AntConc

Errors of corpus research software while researching business and legal corpus of Vietnamese language, the example of AntConc software

INSTYTUT JĘZYKOZNAWSTWA, UNIWERSYTET IM. ADAMA MICKIEWICZA
AL. NIEPODLEGŁOŚCI 4, 61-874 POZNAŃ

krolczyk.jakub@gmail.com

Abstract

On the one hand corpus research and corpus linguistics are relatively new fields of science but on the other hand, according to some people, there are one of fastest developing methods of linguistic research. To perform a corpus research, it is necessary to have a text corpus and a proper kind of software. The range of software kinds is wide and its easy to find free of charge on or license based software. Nevertheless, what the choice is, it is possible to encounter problems or the software will have low efficiency. Low efficiency of AntConc can be seen while researching a corpus compiled from an isolating language. After processing the corpus, consisting of 18 text items in the Vietnamese language (that is 290 pages of typescript) dedicated to the field of management and law, the software outputted incorrect results. Starting with counting the number of words in a corpus and ending with concordance plotting. There are two ways to deal with this problem. The method involves "teaching" AntConc how to read the Vietnamese language, in other words it is necessary to input a list of all words in the Vietnamese language. The second method is more time consuming because it involves replacing the spaces between syllables to a sign that will not be recognized by the software as a space. Using one of these methods could potentialy end in raising AntConc efficiency.

Abstrakt

Badania korpusowe, jak i językoznawstwo korpusowe są dość młodymi dziedzinami nauki, są też według niektórych najszybciej rozwijającymi się metodą badawczą językoznawstwa. W badaniach korpusowych wykorzystuje

się korpusy tekstów i specjalne oprogramowania komputerowe. Oprogramowanie to może być darmowe albo płatne, niestety, nie ważne, na jakie oprogramowanie się zdecydujemy, mogą pojawić się błędy lub program może mieć małą skuteczność. Niska skuteczność programu AntConc jest widoczna podczas badania korpusów języków izolujących. Po wprowadzeniu do programu AntConc korpusu, składającego się z 18 pozycji w języku wietnamskim(tj. 290 stron maszynopisu), poświęconych zagadnieniom zarządzania i prawa, program przedstawiał błędne wyniki. Począwszy od policzenia ilości słów i wytypowaniu jakie pojawiają się najczęściej do tworzenia list konkordancji. Istnieje kilka sposobów na zaradzeniu takiej sytuacji, pierwszą metodą jest „nauczenie" programu AntConc czytania języka wietnamskiego innymi słowy wprowadzenie listy słów które występują w języku wietnamskim. Inną metodą, znacznie trudniejszą i wymagającą dużego nakładu pracy, jest zamiana spacji między sylabami na inny znak który nie był by czytany jako odstęp między słowami przez program AntConc. Jeżeli by zastosować jedną z wyżej wymienionych metod, program ten miał by bardzo wysoką sprawność gdyż język wietnamski nie posiada końcówek fleksyjnych i jest typowym językiem SVO.

Badania korpusowe jak i językoznawstwo korpusowe są dość młodymi dziedzinami nauki. Są też, według niektórych opinii uznawane za najszybciej rozwijające się metody badawcze. W badaniach korpusowych wykorzystuje się korpus tekstów i specjalne oprogramowanie komputerowe. Oprogramowanie to może być darmowe albo płatne, niestety, nie ważne na jakie oprogramowanie się zdecydujemy mogą pojawić się błędy w przetwarzaniu korpusu. W językach należących do rodziny indoeuropejskiej, praca programu przebiega dość sprawnie i z małą ilością błędów. Głównym tego powodem jest struktura pojedynczych słów, które mogą się składać z wielu sylab, jednakże między poszczególnymi sylabami nie występują przerwy. Ten zabieg powoduje, że programy do obróbki korpusów „widzą" sekwencje sylab jako jedno słowo gdyż algorytm programu jest zaprogramowany na rozpoznawanie spacji (domyślnie, można zaprogramować wiele znaków jako odstęp między słowami) jako koniec jednego wyrazu i początek drugiego.

Sprawy mają się zupełnie inaczej w przypadku języków austroazjatyckich i chińsko-tybetańskich. W tych językach poszczególne sylaby mają określone znaczenie, a kombinacja tych sylab powoduje zmianę znaczenia ciągu. Biorąc za przykład język wietnamski, słowo może składać się z maksymalnie czterech sylab.

Celem niniejszej pracy jest wykazanie, jakie błędy występują podczas stosowania programu do badania korpusów, gdy korpus jest w języku wietnamskim. Drugim celem jest zaprezentowanie rozwiązania tego problemu, rozwiązanie to jest proste i eleganckie ale wymaga dużego nakładu pracy ze strony badacza.

Stwierdziłem że warto zająć się tym problemem, gdyż osobiście się na niego natknąłem w moich badaniach. W dalszym toku rozważań i wyjaśnień, będę się dzielić swoimi spostrzeżeniami na temat pojawiających się błędów a mówiąc dokładniej zafałszowań wyników badań. Zanim przejdę do analizy przypadku i próby znalezienia remedium na powyższy problem, należy powiedzieć słów kilka na temat tłumaczenia specjalistycznego i wagi badań korpusowych, od których zależy wybór odpowiedniego słownictwa do tworzenia słowników, leksykonów i glosariuszy.

1 Badania korpusowe

Pierwszym pytaniem, które się nasuwa jest, czym jest korpus, albo szerzej spoglądając, czym jest lingwistyka korpusowa? Termin ten najlepiej i przystępnie jest wyjaśniony w pracy autorstwa Sambora Gruczy. Według jego pracy, lingwistyka korpusowa jest „doborem i elektronicznym przetwarzaniem określonych zbiorów tekstów, określanych jako korpusy tekstowe".[1]

Przejdźmy do objaśnienie pojęcia „korpus". Tak jak w przypadku terminu „kultura", tak i w przypadku korpusu nie istnieje jedna dobra definicja. Każda definicja proponowana przez badaczy wnosi coś do ogólnego obrazu. Poniżej znajduje się kilka z popularniejszych definicji korpusu tekstowego:

> *[korpus to] teksty, dane itp. zgromadzone ze względu na swoją reprezentatywność, stanowiące podstawę do analizy naukowej.*[2]

> *Corpus–a collection of materials that has been made for a particular purpose, such as a set of textbooks which are being analyzed and compared or a sample of sentences or utterances which are being analyzed for their linguistic features.*[3]

> *A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.*[4]

> *A large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria.*[5]

---

1 Grucza S., 2007, *O konieczności tworzenia korpusów tekstów specjalistycznych*, [w:] S. Grucza et al. [red.], *W kręgu teorii i praktyki lingwistycznej*, Warszawa: WUW, s. 103-122.

2 Słownik Języka Polskiego PWN [wersja online]: <http://sjp.pwn.pl>, data dostępu: 23 kwiecień 2015.

3 Richards J.C., 1996, *Longman Dictionary of Language Teaching & Applied Lingusistics*, Harlow: Longman.

4 Sinclair J., 1996, EAGLES: *Preliminary recommendations on Corpus Typology,* EAG-TCWG-CTYP/P. Pisa: ILC-CNR.

5 Bowker L, Pearson J., 2002, *Working with Specialized Language. A practical guide*

> *The term 'corpus' may be defined as a body or collection of linguistic data, especially the one considered complete and representative, from a particular language or languages, in the form of recorded utterances or written texts, which is available for theoretical or/and applied linguistic investigation.*[6]

Jeżeli przyjmiemy definicję słownikową i Bowkera za kluczowe dla dalszych dywagacji, to zarysują nam się dwie cechy korpusów. Pierwszą cechą jest to, że korpus jest zapisywany w formie elektronicznej, a drugą, że są to zgromadzone dane. Z tych dwóch cech powstają następujące pytania: Dlaczego akurat forma elektroniczna? W jakim celu są one gromadzone?

Na pierwsze pytanie odpowiedź jest stosunkowo łatwa i nie trzeba się w nią zbytnio wgłębiać. Dwudziesty pierwszy wiek wymaga od nauki cyfryzacji, dzięki temu przepływ informacji jest bardziej płynny a zgromadzone zbiory znajdują się w środowisku bardziej stabilnym niż np.: w bibliotekach. Poza tym obróbka materiału w formie elektronicznej jest łatwiejsza niż w formie papierowej. Komputery wraz ze specjalnym oprogramowaniem są w stanie wykonać miliony obliczeń w ciągu sekundy, które metodą tradycyjną zajęłyby lata. To samo się tyczy badań korpusowych z dobrym oprogramowaniem pojedyncza osoba jest w stanie wykonać pracę kilkudziesięciu osób w czasie krótszym niż kiedyś było to wykonalne.

Co do pytania o cel gromadzenia danych, można śmiało odpowiedzieć, że gromadzi się dane korpusowe, by realizować wszelakie cele badawcze w rożnych dziedzinach językowych. Dziedzinami, które są najczęściej wymieniane są glottodydaktyka, lingwistyka, terminologia, leksykografia i translatologia.[7] Z punktu widzenia tematyki poruszanej w tej pracy, najbardziej interesuje mnie dziedzina translatologii, w której tworzy się bazy tekstów porównywalnych, czyli tekstów w języku docelowym należących do tego samego gatunku, co teksty wyjściowe. Daje to możliwość tłumaczowi wyszukania wyrażeń, które są niezbędne do wykonania tłumaczenia. Poza tym badania korpusowe pozwalają na tworzenie konkordancji równoległych, a także wykorzystywanie ich w tzw. pamięciach tłumaczeniowych, i innych programach wspomagających tłumaczenie.

Na sam koniec, by lepiej pojąć lingwistykę korpusowa należy jeszcze nadmienić, od czego zależy dobór korpusu. Istnieje wiele kryteriów, które dają rożnego rodzaju wyniki. Teoretycznie każdy zbiór tekstów jest korpusem, ale taki zbiór jest zbyt szeroki by przeprowadzić na nim miarodajne wyniki. Korpus należy dobrać odpowiednio do zakresu badań,

---

to using corpora, London: Routledge.

6 Burkhanov I., 1998, *Lexicography. A Dictionary of Basic Terminology*, Rzeszów: Wydawnictwo Wyższej Szkoły Pedagogicznej.

7 Łukasik M., 2007, *Narzędzia lingwistyki korpusowej w warsztacie terminologa, terminografa i tłumacza tekstów specjalistycznych*, Katedra Języków Specjalistycznych Uniwersytetu Warszawskiego, Warszawa.

jakie chce się przeprowadzić, np. chcąc badać język zarządzania, należy zebrać dwa rodzaje tekstów, pierwsza grupa to korpus teksów związanych tylko i wyłącznie z dziedziną zarządzania. Drugi korpus powinien się składać z tekstów referencyjnych. Dzięki takiemu zabiegowi jesteśmy w stanie łatwiej zdeterminować, które słowo jest należy do grupy specjalistycznej, a które do potocznej. Ważna uwaga – nie należy generalizować pojęcia zarządzania do biznesu, gdyż w sferę biznesu zaliczają się takie dziedziny, jak negocjacje, marketing itd. Dokonując takiej generalizacji niepotrzebnie poszerzamy zakres słownictwa szukanego i wyniki mogą być niezgodne z rzeczywistością. Poniżej zamieszczam listę kryteriów, względem których należy się kierować przy doborze korpusu.

Kryteria pod względem których dobiera się korpus[8]

1. **zakres**: korpusy referencyjne, obejmujące wszystkie rodzaje tekstów – (ang. *reference corpora*) vs. korpusy specjalne, obejmujące wybrane rodzaje tekstów, np. specjalistyczne – (ang. *special/specialized* corpora);
2. **forma tekstów**: korpusy tekstów mówionych (transkrybowanych) – (ang. *spoken corpora*) vs. korpusy tekstów pisanych – (ang. *written corpora*);
3. **rodzaj lingwalności**: korpusy jednojęzyczne (monolingwalne) – (ang. *monoligual corpora*) vs. korpusy wielojęzyczne (multilingwalne) – (ang. *multilingual corpora*), które można dalej podzielić na korpusy złożone wyłącznie z tekstów porównywalnych, niebędących własnymi translatami – (ang. *comparable corpora*) oraz złożone z oryginałów i ich translatów w poszczególnych językach;
4. **stopień otwartości**: korpusy statyczne (zamknięte), czyli nieposzerzane o nowe teksty – (ang. *closed/static*) vs. korpusy dynamiczne (monitorujące), czyli nieustannie uzupełniane oraz weryfikowane – (ang. *open/monitor corpora*);
5. **odniesienie temporalne**: korpusy synchroniczne – (ang. *synchronic corpora*) vs. korpusy diachroniczne – (ang. *diachronic corpora*);
6. **stopień kompletności**: korpusy zawierające pełne teksty–(ang. *full-text corpora*) vs. korpusy z tekstami o określonej długości, np. zawierające wyłącznie streszczenia, wstępy itd. artykułów naukowych – (ang. *sample corpora*);
7. **rodzaj opracowania danych**: korpusy anotowane – opatrzone specjalnymi znacznikami (tagami), które mogą definiować dany wyraz w zakresie informacji gramatycznej (np. części mowy: *Part-Of-Speech tagging*), syntagmatycznej czy semantycznej – (ang. *annotated/tagged corpora*) vs. korpusy nieanotowane – (ang. *raw*

8 Łukasik M., 2007, *Narzędzia lingwistyki korpusowej w warsztacie terminologa, terminografa i tłumacza tekstów specjalistycznych*, Katedra Języków Specjalistycznych Uniwersytetu Warszawskiego, Warszawa.

text/data corpora). [9]

2 Język wietnamski

Język wietnamski jest szczególnym przypadkiem wśród języków izolujących, gdyż zapisywany jest alfabetem łacińskim. Alfabet ten nosi nazwę *Quốc ngữ*; powstał on w XIX w. i został stworzony przez portugalskiego misjonarza. Oficjalnie został przyjęty przez państwo wietnamskie w 1934 roku.[10] Składa się z 29 znaków, z czego 7 to znaki diakrytyczne. Poza tym, w alfabecie jest 5 znaków oznaczających tony. Znaki łączą się w 10 dyftongów i 1 tryftong. By najlepiej zobrazować jak wygląda język wietnamski, posłużę się prostym przykładem, który jest znany wszystkim adeptom nauki wietnamskiego:

Przykład po wietnamsku: *Tôi là sinh viên ở trường đại học.*
Tłumaczenie: Jestem studentem na uniwersytecie.

Dla niewprawionego oka przykład wietnamski składa się z 8 słów, ale naprawdę zdanie to składa się tylko z 5 słów.

Przykład po wietnamsku: *Tôi / là / sinh viên / ở / trường đại học.*

W języku wietnamskim, jako że jest językiem sylabicznym, sylaba w niektórych przypadkach stanowi samodzielny wyrazy aczkolwiek w większości przypadków sylaba występuje jako morfem w wyrazach złożonych.

học – uczyć się / *học sinh* - uczeń
trời – niebo / *mặt trời* – słońce

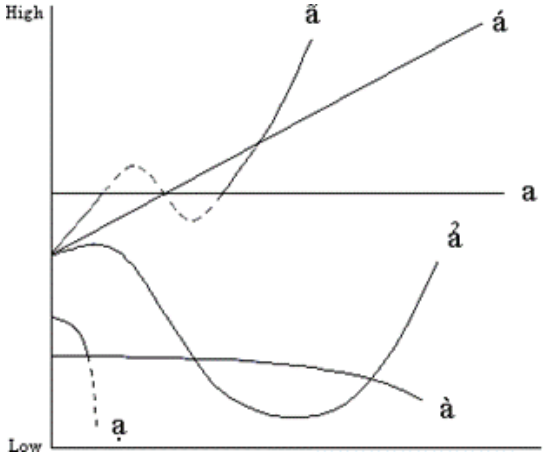
Z fonetycznego punktu widzenia, sylaba jest autonomiczną jednostką, a każdą jednostkę zapisuje się oddzielnie, w niektórych przypadkach sylaba jest słowem i w takiej sytuacji mowimy o słowie prostym jednosylabowym.[11] Szyk zdania pełni bardzo ważną funkcję znaczeniową, gdyż funkcję gramatyczną wyrazu nie wyznacza jego postać morfologiczna, lecz jego pozycja w zdaniu. Ważną cechą języka wietnamskiego jest stały szyk wyrazów w zdaniu: jest to typowy język SVO (Podmiot – Orzeczenie - Dopełnienie). Cecha ta jest także główną i niezmienna zasadą tegoż języka. Istnieją oczywiście odstępstwa od tej zasady; jest to grupa czasowników z tak zwanym dopełnieniem pustym.

Tôi	xem	phim
Ja	ogladam	film

9 Łukasik M., 2007, *Narzędzia lingwistyki korpusowej w warsztacie terminologa, terminografa i tłumacza tekstów specjalistycznych*, Katedra Języków Specjalistycznych Uniwersytetu Warszawskiego, Warszawa.
10 Halik T., 2009, *Język wietnamski*, Wydawnictwo Akademickie Dialog, Warszawa.
11 Halik T., 2009, *Język wietnamski*, Wydawnictwo Akademickie Dialog, Warszawa.

S	V	O

Czasowniki nie ulegają odmianie przez osoby i czas, co skutkuje tym, że należy używać innych słów do określania czasu, takich jak przysłówki, rzeczowniki w funkcji przysłówkowej lub partykuły czasowe. Zaimki osobowe posiadają bardzo wysoka rangę, gdyż dzięki nim określa się relacje społeczne. Znajomość zaimków osobowych jest jednym z czynników, które czynią nas taktownymi. W zależności, czy rozmawiają rówieśnicy, ojciec z synem, czy szef z podwładnym, należy użyć odpowiedniego zaimka; pomyłka może być uznana za obrazę. Ostatnim aspektem języka wietnamskiego jest jego tonalność. Każdy fonem ma 6 odpowiedników tonalnych skutkujących zmianą znaczenia. W zależności od tonu zmienia się wysokość i długość samogłoski. Poniżej zamieszczam wykres wszystkich 6 tonów, które występują w języku wietnamskim.



Wyk. 1 Wykres tonów języka wietnamskiego[12]

3 Analiza przypadku

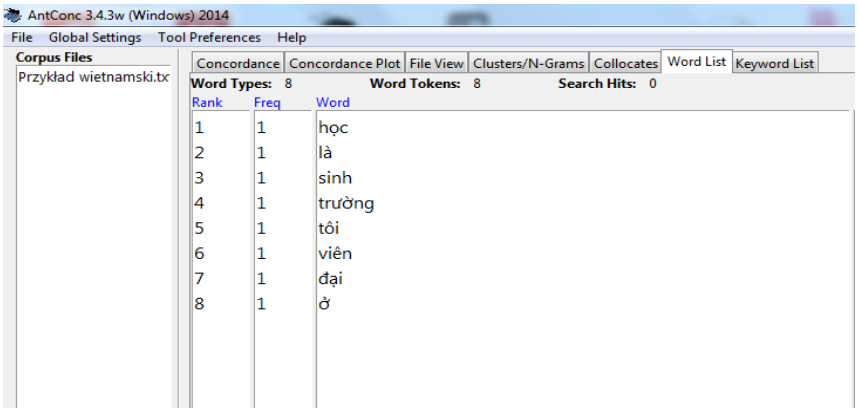
Po tym krótkim wstępie przejdę do analizy przypadku. Podczas badań korpusowych, które przeprowadzałem, natknąłem się na szereg słabych punktów w programie AntConc. Jak już wcześniej wspomniałem, program ten działa dobrze, gdy w swojej bazie ma korpusy zapisywane w językach indoeuropejskich. Problemy zaczynają się przy językach izolujących, takich jak chiński, koreański czy wietnamski.

Korpus, który skompilowałem do swych badań, składa się z 18 pozycji (tj. 290 stron maszynopisu), z czego 4 to pozycje książkowe, a pozostałe to artykuły naukowe i umowy handlowe. Pozycje są poświęcone prawnym aspektom zarządzania i zarządzaniu *sensu stricte*. Większość

12 Wykres tonów języka wietnamskiego, https://outonlimb.files.wordpress.com/2009/10/picture7.gif

z nich to tłumaczenia książek np.: Jima Colinsa, Dona Faila lub Petera F. Drucera. Poza książkami są też studia przypadku różnych amerykańskich firm, przetłumaczone na wietnamski. W korpusie umieściłem też umowę handlową, gdyż jak dobrze wiadomo sfera zarządzania jest bardzo powiązana ze sferą prawną.

Przejdźmy zatem do prezentacji błędów, na jakie można się natknąć pracując w AntConc'u na korpusie w języku wietnamskim. Biorąc za przykład zdanie, które pojawiło się w poprzednim rozdziale: *Tôi là sinh viên ở trường đại học*. Po zapisaniu tego zdania w podstawowym edytorze tekstu i zamieszczeniu pliku w programie, otrzymamy poniższy wynik.



Obr. 1 Korpus z przykładowym zdaniem po wietnamsku.

Jak widać na załączonym powyżej zrzucie ekranu, AntConc podzielił wprowadzone zdanie na 8 części składowych, co oczywiście jest błędem. Powinien on był wyszczególnić tylko 5 elementów.
Oczywiście słowa wyszczególnione przez program posiadają znaczenie. Jednakże w odpowiednich kombinacjach znaczenie może diametralnie się zmienić, np.:

sinh – urodzić się (czasownik)
viên – 1. klasyfikator do rzeczy o regularnych kształtach.
2. klasyfikator do osób piastujących wysokie stanowisko.
sinh viên – student (rzeczownik)

Jak łatwo można dostrzec, charakterystyka języka wietnamskiego i sposób pracy programu AntConc uniemożliwiają poprawna analizę korpusu. Wszakże prezentowany przykład jest tylko prostym zdaniem. Dlatego też następne przykłady będą prezentowane na pełnym korpusie, który składa się z 18 pozycji z czego 4 to pozycje książkowe, daje to 290 stron maszynopisu. Po wprowadzeniu pełnego korpusu i dokonaniu wstępnej analizy, otrzymujemy taki wynik.

Pierwsza dziesiątka najczęściej pojawiających się słów w korpusie	Tłumaczenie
một	jedne
của	partykuła oznaczająca przynależność
và	razem, i, również (spójnik)
là	być
có	mieć
những	przysłówek oznaczający liczbę mnogą
các	przysłówek oznaczający liczbę mnogą
người	człowiek
trong	wewnątrz, w (przyimek)
công	1. paw, 2. wysiłek, 3. publiczny

Tab. 1 Pierwsza dziesiątka najczęściej pojawiających się słów w korpusie

W pierwszej dziesiątce jest parę nieścisłości. Według obliczeń programu najczęściej pojawiającym się słowem w całym korpusie jest słowo „*một*" oznaczające jeden(liczba). Jest to dość nietypowy wynik, a jego nietypowość jest utwierdzana dodatkowo przez wykresy częstotliwości występowania danego słowa w poszczególnych tekstach korpusu. W niektórych tekstach słowo jedne występuje tak często, że wykres przybiera czarny kolor.

Obr. 2 Lista konkordancji dla słowa „một" w formie graficznej

W takim razie, pytanie które się rodzi, to: dlaczego niektóre teksty korpusu składają się głównie ze słowa jeden? Słowo „*một*", tak jak większość słów w języku wietnamskim, jest samodzielnym wyrazem jak i morfemem większego złożonego wyrazu. Poza swoim pierwotnym liczebnikowym znaczeniem, słowo „*một*" przyjmuje funkcję przedimka, czyli jest stawiany przed rzeczownikiem i wskazuje jego określoność bądź nieokreśloność.

Wyraz z „một" jako przedimkiem	Tłumaczenie
một lĩnh vực	pole
một doanh nghiệp	biznes
một thanh niên	młody człowiek

Tab. 2 Wyraz z „một" jako przedimkiem

Ten sposób stosowania niektórych wyrazów powoduje, że liczba ich w tekście rośnie, ale nie przekłada się zbytnio na jakość tłumaczenia np.: gdyby przy wyrazie „*một lĩnh vực*" oznaczającego pole, usunięto wyraz „*một*", to znaczenie nie uległoby zmianie.

Ostatnim rażącym uchybieniem programu AntConc jest zła gradacja wyrazów pod względem ich występowania, która wynika z niemożności rozpoznawania co jest wyrazem, a co sylabą. Jest to niestety najpoważniejszy problem, jaki napotyka się w AntConc przy badaniu języka wietnamskiego. Poniżej zamieszczam tabele, w której skonfrontowałem wyniki domyślne podawane przez program z wynikami wspomaganymi przez użytkownika. Do tej próby wybrałem następujące wyrazy: *quyết định* – decyzja lub decydować, *kinh doanh* - przedsiębiorstwo, *kết quả* - rezultat, *đơn vị* – jednostka(organizacja lub wojsko).

Słowo bazowe	Częstotliwość występowania w korpusie	Kombinacja sylab	Częstotliwość występowania w korpusie	Procentowy udział występowania kombinacji sylab w częstotliwości słowa bazowego	Ilość kombinacji sylab ze słowa bazowego
quyết	866	quyết định	571	65,94%	12
kinh	1478	kinh doanh	910	61,57%	30
kết	896	kết quả	370	41,29%	21
đơn	353	đơn vị	88	24,93%	10

Tab. 3 Procentowy udzial wystepowania kombinazji sylab w czestotliwosci wystepowania słowa bazowego.

Pierwsza sylaba tych wyrazów została nazwana przez mnie „słowem bazowym". Słowa bazowe zostały wytypowane przez program i mieszczą się w pierwszej dwusetce jest to dość wysoko gdyż cały korpus według programu składa się z 10178 typów słów. By łatwiej opisać to zjawisko, skupimy się na jednym wyrazie np.: *quyết* – decydować. Słowo to powtórzyło się w korpusie 866 razy, aczkolwiek słowo to ma aż 12 wariantów, co oznacza, że po dołączeniu do niego kolejnej sylaby, jego znaczenie ulegnie zmianie. Jeden z tych wariantów, który jest najbardziej interesujący to *quyết định* - decyzja lub decydować. Jak widać znaczenie jest podobne do słowa bazowego, ale ten wariant stosuje się w pismach sądowych, pozwoleniach i decyzjach, więc można przyjąć, że jest to wyrażenie specjalistyczne. W badanym korpusie wyraz ten występuje 571 razy oznacza to, że jest on częstszy niż samo słowo bazowe, które dzieli pozostała liczbę trafień z pozostałymi wariantami. Można obliczyć z łatwością, że średnio na pozostałe wariant przypada po 26 trafień.

Reasumując, gdyby program Antconc rozróżniał wyrazy od pojedynczych sylab i odpowiednio je klasyfikował według częstotliwości, to wyszukiwanie najczęściej występujących wyrazów stałoby się niezwykle proste i efektywne. Nie powstawałyby sytuacje przedstawione w tablicy 1. Słowa w niej przedstawione automatycznie spadałyby na niższe pozycje, ukazując w czołówce słowa, które mogą posłużyć do tworzenia glosariuszy.

4 Remedium

Na początku artykułu wspomniałem o dwóch metodach radzenia sobie z problem przedstawionym w niniejszej pracy. Pierwsza metoda polega na wgraniu w program jak największej liczby słów języka wietnamskiego, tak aby program wiedział jakie słowo występuje w formie pojedynczej sylaby a jakie w formie wielosylabowej. Drugi sposób, polega na uprzednim przygotowaniu korpusu do obróbki przez zamianę spacji między sylabami na inny znak.

Nawiązując do pierwszej metody, ma ona swoje dobre strony, jak i złe. Pozytywnym aspektem tej metody jest to, że program zacznie rozpoznawać poszczególne słowa języka wietnamskiego, a następnie będzie poprawnie liczyć częstotliwość ich występowania. Z drugiej strony, by wykorzystać tą metodę jest potrzebna obszerna lista słów, rzędu kilkudziesięciu tysięcy lub narodowy korpus. Niestety nie wszystkie języki takowe listy posiadają. Wietnamscy językoznawcy są w trakcie opracowywania Korpusu Narodowego i może to potrwać jeszcze jakiś czas. Na chwilę obecną jedynym rozwiązaniem jest gromadzenie słowników w formie elektronicznej albo zgrywanie zawartości zamieszczonej w sieci.

Druga wymieniona metoda jest niezwykle pracochłonna gdyż polega ona na ręcznej zamianie spacji między sylabami, które należą do tej samej jednostki wyrazowej na inny znak. W przypadku AntConc, najlepiej działa metoda z ręczną likwidacja spacji między sylabami. Nie jest to rozwiązanie najbardziej eleganckie, ale skuteczne. Za przykład wezmę już pojawiające się zdanie: *Tôi là sinh viên ở trường đại học*. Po usunięciu spacji zdanie będzie wyglądać następująco: *Tôi là sinhviên ở trườngđạihọc*. Następnie po umieszczeniu korpusu z powyższym zdaniem otrzymujemy poprawny wynik.

Obr. 3 Zdanie z połączonymi sylabami

Jak widać, AntConc podzielił poprawnie korpus, aczkolwiek rozwiązanie to nie jest eleganckie. Ideałem byłoby zmodyfikowanie tekstu tak, by między sylabami znajdowały się myślniki, a następnie zmodyfikowanie tak programu AntConc, by rozpoznawał myślniki jako część wyrazu. W dobie cyfryzacji dokonanie takich zmian w tekście i w programie nie powinno stanowić dużego wyzwania, a efekt ciężkiej pracy przekładałby się na skuteczność badań korpusowych.

5 Podsumowanie

Badania korpusowe są jedną z ważniejszych metod badawczych. Tempo ich rozwoju jest zaskakujące, bo jeszcze kilkanaście lat temu, pracownie językoznawcze były pełne kartotek roboczych z papierowymi fiszkami. W chwili obecnej jest już potrzebny tylko komputer i odpowiednie oprogramowanie. Niestety, nie zawsze oprogramowanie jest w 100% skuteczne. Pewne grupy języków stanowią jeszcze wyzwanie dla językoznawców, ale wyzwania są po to, by je przezwyciężać.

Do tej grupy języków należy wietnamski, język izolujący, tonalny i sylabiczny. Jego charakterystyka zradza problemy, które uniemożliwiają poprawne jego badanie za pomocą programu do przetwarzania korpusów. Głównym powodem takiej trudności jest sposób zapisu słów, są one zapisywane za pomocą pojedynczych sylab lub ich ciągu. Program AntConc nie rozpoznaje, jaka cześć jest autonomicznym słowem, a co jest ciągiem sylab tworzących jedno słowo.

Na zaistniały mankament można zaradzić na kilka sposoby: i) należy dokonać zmian w algorytmie programu i korpusie. Między ciągami sylab tworzących jedno słowo należy umieścić myślniki, a algorytm tak zmodyfikować by rozpoznawał taki ciąg jako całość, ii) ręczna lub automatyczna eliminacja spacji między ciągami sylab. Jednakże ta metoda skutkuje wprowadzeniem chaosu, gdy badający nie zna języka, iii) zaimplementowanie Korpusu Narodowego języka wietnamskiego do programu tak, aby program „umiał" odróżnić ciągi od pojedynczych jednostek.

Z praktycznego punku widzenia i przy założeniu, że badający zna język wietnamski na poziomie dobrym, najprostszą metoda jest eliminacja spacji między ciągami sylab. Takie działanie przekłada się na otrzymanie dobrych wyników, a następnie w przypadku prezentacji, można ponownie wprowadzić spacje, by tekst był przejrzysty i zrozumiały dla odbiorców.

Program AntConc jest wspaniałym narzędziem pracy dla początkującego językoznawcy, gdyż jest darmowy i daje dużo możliwości analizy korpusu niestety, staje się nieskuteczny przy językach izolujących. Stosując jedną z wymienionych wyżej metod, można znacznie zwiększyć jego sprawność a co za tym idzie, zacząć badać nowe obszary w lingwistyce i translatologii. Należy też pamiętać, że szybkość rozwijania się kontaktów biznesowych na świecie, narzuca wymóg na nowe techniki tłumaczeń specjalistycznych w różnych, od angielskiego, językach. Dlatego powinno się usprawniać narzędzia pracy językoznawców by rezultaty otrzymywane w ich dochodzeniach były adekwatne do stanu rzeczywistego.

Bibliografia

AntConc, wersja 3.4.3w., autor programu: Laurence Anthony, dostępny nieodpłatnie na stronie domowej autora: <http://www.antlab.sci.waseda.ac.jp/>.

Bowker L, Pearson J., 2002, *Working with Specialized Language. A practical guide to Using corpora*, London: Routledge.

Burkhanov I., 1998, *Lexicography. A Dictionary of Basic Terminology*, Rzeszów: Wydawnictwo Wyższej Szkoły Pedagogicznej.

Collins, J., 2001, *Từ tốt đến vĩ đại*, Nhà xuât, T. T. N. Tuyến, Hanoi.

Doanh, N.K., 2011, *100 điều Doanh nhân cần biết*, Vitanco, Hanoi.

Drucker, P.F., 2011, *Tinh Hoa Quan Tro*, Nha xuat Ban tre, Ho Chi Minh City.

Grucza S., 2007, *O konieczności tworzenia korpusów tekstów specjalistycznych*, [w:]S. Grucza et al. [red.], *W kręgu teorii i praktyki lingwistycznej*, Warszawa: WUW, s. 103-122.

Halik T., 2009, *Język wietnamski*, Wydawnictwo Akademickie Dialog, Warszawa.

Hạ nh, M., 2005, *Thu hút khách hàng-vấn đề của doanh nghiệp*, Vitanco,Hanoi.

Hoàng Anh, 2008, *Bí mật của những tổ chức vĩ đại - Phần 1*, Vitanco, Hanoi. Lam, N.H., 2005, *Môi trường Bên Ngoài: Cơ hội, Đe dọa, Cạnh tranh*, Nha xuat Ban tre, Ho Chi Minh City.

Łukasik M.,2007, *Narzędzia lingwistyki korpusowej w warsztacie terminologa, terminografa i tłumacza tekstów specjalistycznych*, Katedra Języków Specjalistycznych Uniwersytetu Warszawskiego, Warszawa.

Mandino, O. & Tâm, T.H., 2009, *Người Bán Hàng Vĩ Đại Nhất Thế Giới*, Nha xuat Ban tre, Ho Chi Minh City.

Nguyễn Đinh Hoà, 2007, *Vietnamese-English Dictonary*, Tuttle Publishing, Singapore.

Nguyễ n Thị Trà, 2010, *Mối quan hệ nhượng quyền - nhận quyền cho bạn biết tình hình kinh doanh nhượng quyền của mình*.

Nye, J., 2006, *Quyền lực mềm, quyền lực cứng và việc lãnh đạo*.

Phi, L.T., 2005, Để kinh doanh hiệu quả hơn: hội thảo chuỗi thị trường,

Richards J.C., 1996, *Longman Dictionary of Language Teaching & Applied Lingusistics*, Harlow: Longman.

RIES, A., 2009, Định Hướng Tập Trung Tương Lai Của Cty Các Bạn Phụ Thuộc Vào Điều Này.

Sinclair J., 1996, *EAGLES: Preliminary recommendations on Corpus Typology*. EAG-TCWG-CTYP/P. Pisa: ILC-CNR.

Słownik Języka Polskiego PWN [wersja online]: <http://sjp.pwn.pl>, data dostępu: 23 kwiecień 2015.

THÂN, H.T., 2007, *Tâm lý quản lý*, Hanoi.

Trà, N.T., 2010, *Howard Schultz - Ông chủ thương hiệu Starbucks*, Hanoi

Trầ n Phư ơ ng Minh, 2008, *Tìm kiếm nhân viên bán hàng như thế nào?* Entrepreneur, Hanoi.

Trang, M., 2005, *Làm thế nào để tìm và bán hàng cho thị trường mục tiêu của bạn?*, Hanoi.

Viễ n Quang, 2002, *Làm thế nào Thực hiện doanh nghiệp lớn và thành đạt Trong kinh doanh theo mạng*, Hanoi.