Zeszyty Naukowe Wydziału Informatyki
Wyższej Szkoły Informatyki Stosowanej i Zarządzania
„Informatyka Stosowana”
Nr 1/2014

WSPOMAGANIE LOKALIZACJI JĘZYKOWEJ DOKUMENTÓW

Milena Kowalska, Jan W. Owsiński

Wyższa Szkoła informatyki Stosowanej i Zarządzania,
01-447 Warszawa, ul. Newelska 6

Streszczenie

Artykuł, oparty na pracy dyplomowej pierwszej autorki, przedstawia zagadnienie językowej lokalizacji dokumentów z punktu widzenia możliwości automatyzacji kluczowych funkcji, związanych z lokalizacją. Przedstawiono zarys zagadnienia i powstające w związku z nim konkretne kwestie techniczne i merytoryczne. Skupiono się na wybranych elementach, w tym przede wszystkim na rozpoznawaniu języka dokumentu. Przedstawiono obszerne wyniki badań, opartych na opracowanych i zastosowanych aplikacjach, w stosunku do konkretnych zbiorów dokumentów. Zaproponowano możliwe rozwinięcie wykonanych prac.

Słowa kluczowe: lokalizacja, analiza tekstu, języki, klasyfikacja

1. Wprowadzenie

Artykuł, oparty na pracy dyplomowej Kowalska (2014), poświęcony jest zagadnieniom, jakie powstają w trakcie tzw. lokalizacji językowej, a więc w trakcie realizacji procedur, dotyczących tłumaczenia. Celem jest ocena efektywności wykorzystania różnych technik w ramach jednego z etapów lokalizacji językowej, a mianowicie rozpoznawania języka. Ponieważ znaczna część czynności, jakie są wykonywane w profesjonalnych biurach lokalizacji, jest wspomagana i ułatwiana przy pomocy odpowiednio skonstruowanych, a następnie wykorzystywanych i stale uaktualnianych narzędzi informatycznych, praca niniejsza może stanowić przyczynek do dalszego ułatwienia, lub poprawy, jakości odpowiednich czynności.

Po opisaniu kontekstu rozpatrywanego zagadnienia, przedstawiono techniki, jakie mogą znaleźć zastosowanie do jego rozwiazywania, a następnie przedstawiono wyniki ich użycia do konkretnego zbioru dokumentów. Wyniki te są interesujące same w sobie, zarówno w warstwie merytorycznej, jak i technicznej, a także pozwalają na nakreślenie potencjalnych dalszych prac.

2. Lokalizacja językowa a „Translation Memory”

Dzięki postępom w rozwoju telekomunikacji wymiana danych z osobami na całym świecie przestała nie tylko być niemożliwa, ale stała się niezmiernie łatwa, niwelując barierę odległości geograficznej. Jedyny zatem praktyczny problem, jaki można w tym kontekście napotkać, to bariera językowa.

Choć globalizacja ujednolica świat, język jest jednym z trwałych aspektów, wyróżniających większość państw i społeczeństw. Języki ewoluują, ale ostatecznie, żaden z krajów nie chce poddać swojej „suwerenności” językowej. Istniejące odrębne kultury i języki spowodowały stworzenie odmiennych rozwiązań, zarówno w zakresie technologii i produktów innowacyjnych, jak i prostych, codziennych urządzeń i procesów. Już w latach 60-tych XX wieku powstawały organizacje o zasięgu globalnym, które próbowały ujednolicać normy, dotyczące technologii informacyjnych oraz komunikacyjnych. Dzięki nim zachowane zostały modele technologiczne, pozwalające, mimo istniejącego zróżnicowania, tworzyć produkty i usługi spełniające standardy zapewniające odpowiednią ich zgodność.

Niejako dualnym zagadnieniem do standaryzacji jest tzw. lokalizacja. Dla potrzeb lokalnego rynku trzeba dostosować programy, dokumentacje itp. informacje pod względem językowym. W roku 1990 powstało międzynarodowe stowarzyszenie LISA (ang. *Localization Industry Standards Association*), zajmujące się opracowaniem standardów, niezbędnych do dostosowania produktów do rynków lokalnych, dla osób prywatnych oraz firm. Za sprawą LISA powstała, między innymi, TM (Translation Memory), czyli baza segmentów tekstowych.

Wykorzystanie TMów w znacznym stopniu wspomaga proces tłumaczenia. Powstają programy, które automatycznie wyszukują już przetłumaczone segmenty. Istnieją systemy, wyszukujące tylko treści identyczne oraz takie, które także szukają segmentów podobnych. W zależności od zastosowanych algorytmów, fragmenty, które odpowiadają szukanej treści, albo zawierają najwięcej wspólnych słów, są uznawane za najlepiej dopasowane i prezentowane tłumaczowi jako „kandydaci” do tekstu tłumaczenia. W projektach lokalizacyjnych, w których teksty składają się ze zdań istotnie zależnych od siebie, a ich budowa jest prosta, i bywa, że treść powtarza się, przy skorzystaniu z TM tłumacze oszczędzają sporo czasu. Kwestią zasadniczą jest stworzenie TMu dla określonej tematyki, co prowadzi do znacznie łatwiejszego znajdowania tłumaczeń fragmentów tekstów. Dodatkową zaletą takich systemów jest dołączanie glosariuszy, czyli słowników terminologicznych, zawierających terminy stosowane w danej dziedzinie i ich odpowiedniki w innych językach.

Istotnym aspektem, zarówno bazy, jak i odpowiedniego systemu, jest długość segmentów tekstów. Wybór długości zależy od charakteru tłumaczonych tekstów. Segmenty mogą zostać stworzone na podstawie całego zdania, czy też całego akapitu. Im dłuższa treść, tym trudniej znaleźć tłumaczenie w pamięci, dlatego najczęstsze jest tworzenie segmentów na podstawie zdania. Automatyczne dzielenie treści za pomocą kropki bywa jednak kłopotliwe, ponieważ niekiedy znak ten znajduje się w środku zdania, np. w skrótach. Należy zatem dołączyć do zasad dzielenia wyjątki, bądź też opracować inne reguły.

Różnice w podziale mogą wpływać na tzw. „natłumaczenie" – wprowadzenie do nowych plików tłumaczeń znalezionych dla identycznych segmentów, przechowywanych w TMie. Źle podzielone zdanie, ze względu na inny szyk, przyjęty w różnych językach, może wprowadzać do TMu błędne tłumaczenia. Inne problemy, dotyczące podziału zdania i jego natłumaczenia, wynikają z formatów, w jakich otrzymujemy pliki. Przykładem jest otrzymanie plików w formacie, powiedzmy, doc, i za kilka dni otrzymanie tej samej treści, lub jej dalszego ciągu, w formacie xml. Format doc, z licznymi stylami, powoduje tworzenie tagów w treści i jest on dzielony w ustalony sposób. Natomiast format xml pozbawiony jest rozmaitych czcionek oraz narzuca poprzez swoją budowę określony podział. W takich sytuacjach od razu natłumaczenie nie jest możliwe, mimo iż cała treść znajduje się w bazie.

3. Praca biura lokalizacji językowej

Biura lokalizacji językowej są pośrednikami między klientem a tłumaczem. Dzięki nim klient szybciej otrzymuje tłumaczenia oraz może znacznie obniżyć koszty tłumaczenia. Natomiast tłumacz ma znacznie łatwiejszą pracę, w pełni może skupić się na tłumaczeniu i, korzystając z nowych możliwości technologicznych, szybciej przetłumaczyć przekazany tekst.

Lokalizacji językowej są poddawane, między innymi:

- Oprogramowania systemowe,
- Serwisy internetowe,
- Systemy geograficzne,
- Aplikacje, zwłaszcza np. graficzne, tekstowe,
- Pisma urzędowe,
- Dokumentacje techniczne,
- Filmy, reklamy,
- Gry komputerowe.

Poza samą automatyzacją, zmniejszenie czasu i kosztów wynika z użycia TM. Jeśli liczba zgromadzonych tłumaczeń jest duża, to objętość potrzebnego (nowego) tłumaczenia znacznie się zmniejsza. Ponadto, podczas analizy tekstu można również określić ilość segmentów identycznych. Zdarza się, że wyszukiwane segmenty podobne, pomocne dla tłumaczy, stanowią 90% nowego tłumaczenia, a różnice polegają zaledwie na jednym dodanym słowie lub znaku interpunkcyjnym. Te informacje mogą podczas negocjacji wpłynąć na warunki finansowe odpowiedniej umowy.

Oprócz zarządzania bazami tłumaczeń, sporządzania informacji o istniejących tłumaczeniach dla nowego pliku i znajdowania najlepszych rozwiązań w kwestii natłumaczenia plików, biura lokalizacji językowej poświęcają czas na sporządzenie poprawnych formatów plików do tłumaczenia. Tłumacze nie muszą posiadać zaawansowanej wiedzy informatycznej, stąd sporządzone pliki i systemy, które je obsługują, muszą być logiczne, proste i zrozumiałe.

Zmęczenie tłumacza lub przypadkowe naciśnięcie klawiszy mogą powodować błędy. Niekiedy i automatyzacja może przynosić niespodziewane rezultaty. Toteż specjalista, otrzymujący pliki z tłumaczenia, musi być przygotowany na każdą sytuację. Tworzone są różne programy do walidacji plików oraz programy do „zapewnienia jakości", czyli Quality Assurance (QA), które określają niezbędne działania, podejmowane dla spełnienia wymagań co do jakości końcowych plików.

4. Sporządzanie plików dla tłumaczy

Pliki mogą mieć różne formaty i mogą być dostarczane na kilka sposobów. Jednym z najprostszych jest ich przesłanie drogą elektroniczną. Dostarczone przez pocztę elektroniczną lub za pośrednictwem serwera ftp, wracają tą samą drogą do klienta. Innym sposobem jest przesłanie plików przez serwis. Korzystając ze specjalnego oprogramowania, klient może przesłać do tłumaczenia wielostronicowe dokumentacje, zawierające różnego rodzaju części. Dzięki temu, kontrahent ma wgląd do pliku wynikowego oraz wpływ na części do przetłumaczenia i do zachowania w oryginalnym języku.

Powszechnie otrzymywane pliki mają rozszerzenia typu: .xml, .html, .doc, .ppt, .txt, .xls itp. W celu ich możliwie jednolitej obróbki powstały narzędzia CAT tools, pozwalające na wyodrębnienie z dokumentów treści do tłumaczenia. Narzędzia te mogą też tworzyć pliki, które za pomocą specjalnego interfejsu graficznego są łączą się z bazą danych, przechowującą wcześniej tłumaczone treści. Ułatwia to wyświetlanie segmentów do tłumaczenia w przejrzysty sposób, a także daje możliwość generowania dodatkowych podpowiedzi.

Jeśli pliki są nowe, w niestandardowym formacie, należy najpierw poznać ich budowę. Dla dużego projektu może to być czasochłonne, ale pomaga uniknąć błędów i wykluczyć niektóre części, które nie należą do tłumaczenia. Tego etapu nie da się uniknąć, jeśli pliki dostarczane są rzadko, a wiedza na temat ich budowy nie jest wystarczająca, by skonstruować standardowe narzędzie przyspieszające pracę.

XML to format, który służy do przenoszenia informacji. Pozbawiony stylów i innych właściwości, jest łatwy do obróbki i konwersji. W branży tłumaczeniowej, jego pochodne typy dobrze spełniają rolę plików pośredniczących w tłumaczeniu. Przesyłane do tłumaczy, formaty oparte na xml, umożliwiają skoncentrowanie się na treści, bez zastanawiania się nad wyglądem dokumentu, co w znacznym stopniu ułatwia pracę tłumaczom. Specyficzna budowa, oparta na formacie xml, pozwala również na szybszą i łatwiejszą analizę treści po tłumaczeniu.

Jednym z formatów, opartych na XMLu jest XLIFF (XML *Localisation Interchange File Format*). Zaczyna się on od deklaracji XML, po której następuje deklaracja dokumentu XLIFF. W każdym pliku znajdziemy informacje na temat języka źródłowego oraz języka, na który plik jest tłumaczony. Zawarte są w nim też elementy i atrybuty do gromadzenia z różnych oryginalnych formatów, a także odpowiadające im tłumaczenia.

Inne pliki, pojawiające się przy tłumaczeniu, to .rtf, tj. *Rich Text Format* Microsoftu. Większość rozwiniętych edytorów tekstowych może czytać i zapisywać

informacje w niektórych wersjach RTF. Specyfikacje nowych wersji są zmieniane i publikowane wraz z głównymi wersjami Microsoft Word i pakietu Office.

Oprócz tych kilku wymienionych formatów, przesyłanych do tłumaczy, istnieją również inne. Ich głównym zadaniem, w sensie realizacji procedur lokalizacji, jest przechowywanie informacji o tekście źródłowym, jego tłumaczeniach oraz o własnościach oryginalnego formatu pliku, z którego została wyodrębniona treść. Dzięki temu możemy zapamiętać tłumaczenia oraz w szybki sposób przywrócić dokumentom stary wygląd dla nowej treści.

5. Problemy występujące przy lokalizacji

Po otrzymaniu plików od tłumaczy, przed ich przekształceniem w pliki docelowe, należy potwierdzić ich poprawność. Standardowe systemy QA, stworzone pod określone formaty plików, zostają wzbogacone o sprawdzenia, specyficzne dla danego projektu. Należy przy tym pamiętać, że niektóre problemy będą pojawiać się w każdym projekcie. Główną przyczyną mogą być błędy w pliku źródłowym.

Podejść do tłumaczenia jest kilka, można je określać jako różne strategie przenoszenia tekstu i jego znaczenia między językami, a więc i kulturami. Jednym z nich jest możliwie wierne odwzorowanie tekstu, innym – przetłumaczenie tekstu tak, aby stał się bardziej zrozumiały dla końcowego czytelnika[1]. Ponadto, w tekście oryginalnym zdarzają się błędy językowe, np. pominięcie znaków interpunkcyjnych, lub błędy gramatyczne czy ortograficzne, które niekiedy mogą być celowym zabiegiem, ale często wynikają z pomyłki. Tłumacząc, trzeba rozstrzygać kwestię faktycznego zamiaru autora. Proste, dosłowne przetłumaczenie pomiędzy dwoma językami nie istnieje, ze względu na różnice znaczeniowe dla słów i wyrażeń, nawet teoretycznie oddających tę samą rzeczywistość. Dlatego też tłumaczenie nie jest procesem łatwym i mechanicznym. Oprócz zdolności językowych oraz technik wyszukiwania, pozwalających precyzyjnie oddać sens oryginalnej treści, należy mieć sporą znajomość tłumaczonej tematyki, języka (“dialektu”) właściwego dla tej tematyki w oryginale i w języku tłumaczenia oraz wrażliwość kulturową.

Dla sprawdzenia poprawności tłumaczeń, tworzy się pliki, do których tłumacze bądź recenzenci klienta mają wgląd przed ostatecznym zatwierdzeniem. W przypadku gier, lub innych aplikacji, sprawdzanie polega na zainstalowaniu produktu i przetestowaniu wszystkich możliwych wariantów i elementów przez natywnych językowców. Jeśli w owym procesie zostaną zgłoszone niejasności, projekt wraca do poprawy. Taka taktyka daje dobre rezultaty, ale też może znacznie podnieść koszty. Z tych przyczyn nie jest stosowana we wszystkich przypadkach.

Podczas tłumaczeń nietrudno o pomyłki. Typowymi błędami, jak wspomniano, mogą być literówki, które obecnie, dzięki możliwości automatycznego sprawdzania pisowni, mogą zostać szybko wykryte. Inne pomyłki mogą występować

[1] Ta opozycja jest szczególnie istotna dla tekstów literackich, ale w obszarze lokalizacji ma również niebagatelne znaczenie.

w tłumaczeniach fragmentów zawierających daty, wielkości fizyczne, czy modele urządzeń. Nietrudno o pomyłkę, kiedy cyfr jest dużo. Stosowane są również tłumaczenia automatyczne, zwłaszcza przy tłumaczeniu krótkich fragmentów, zawierających np. daty. W celu uniknięcia problemów, jakie mogą w takich sytuacjach wystąpić, porównuje się występowanie i liczność wystąpień cyfr.

Tłumaczenia bywają sporządzane przez różne firmy, korzystające z usług wielu biur tłumaczeniowych. Aby zapewnić jednolitość treści, można zachować w tym celu tłumaczenia z plików od klienta. Wobec braku dokładnej wiedzy na temat pochodzenia tłumaczeń, przypisywane jest plikom tego rodzaju "zmniejszone zaufanie". Tak więc, jeśli znajdujemy fragment, dla którego mamy tłumaczenie z plików klienckich, nie zostanie ono od razu wstawione jako właściwe tłumaczenie, ale pozostanie w podpowiedziach dla tłumacza. Tłumacz zweryfikuje jego poprawność dla danego tekstu.

Innym stosowanym kryterium jest poprawność budowy plików. Dla plików o określonej budowie, na przykład w standardzie xml, istnieją programy do walidacji. Sprawdzają one otwarcia i zamknięcia tagów oraz ich liczbę. Podczas tłumaczenia, tłumacz nieumyślnie może skasować fragmenty kodu, powodując problemy przy konwersji plików do docelowego formatu. Gdy operujemy plikami w formacie .doc prostsze jest porównanie optyczne plików i zwrócenie uwagi czy istnieją treści pogrubione, czy odnośniki do stron zostały zlokalizowane odpowiednio itp.

Wszystkie narzędzia i podejścia do tłumaczenia, w tym i w kontekście lokalizacji, mają swoje plusy i minusy. Bazy są stosowane zarówno dla przyśpieszenia, jak i dla zachowania jednolitości tłumaczenia. Pewne błędy jednakże wynikają właśnie z korzystania z nich. Jednym z trudniejszych problemów jest mieszanie się tłumaczeń. Zdarzają się projekty, w których w jednym pliku klient pragnie mieć zawarte wszystkie tłumaczenia na różne języki dla określonej treści. Podczas kopiowania tłumaczeń może, na przykład, wystąpić pomyłka w rubrykach. Problem mieszania się tłumaczeń może także pojawić się podczas natłumaczenia plików niepoprawnym językiem. Wykrycie faktu, że języki są bardzo do siebie podobne, może być utrudnione, szczególnie jeśli w treści występują fragmenty, których nie należy tłumaczyć, lecz zostawić w postaci oryginalnej. Pomyłka może zostać wychwycona podczas sprawdzenia plików przez lingwistów, lub przez tłumaczy. Nie wszystkie jednak projekty przewidują wykonanie kroku testowego.

Pliki mogą być różnie kodowane. Przy konwersji między formatami lub korzystaniu z różnych systemów, litery odpowiednie dla danego języka mogą ulec zmianie. Sprawdzanie poprawności kodowanych znaków, po części wyklucza występowanie niepoprawnych języków, a po części pokazuje błędy w odczycie.

Przy realizacji procedur zapewniania jakości należy pamiętać by proces sprawdzania plików nie stał się niepotrzebnie długi. Dzięki TMom zmniejszamy czas trwania tłumaczeń i, co za tym idzie, także koszt. Kontrola jakości powinna być dostosowana do projektu i jego specyfikacji oraz przemyślana, tak by nie stała się źródłem potrzeby wprowadzania dodatkowego nadzoru.

6. Tekst pisany i identyfikacja języka

6.1. Kodowanie znaków

W różnych miejscach na ziemi, w różnych kulturach oraz w różnym czasie, powstawały odmienne systemy umownych znaków i symboli. Stąd rozróżnienie metod zapisu, takich jak, między innymi:

a) pismo logograficzne – złożone z logografów, czyli ideogramów odpowiadających na ogół wyrazom lub morfemom (np. pismo piktograficzne)
b) sylabariusz – system pisma, w którym pojedyncze znaki reprezentują całe sylaby (np. chińskie pismo kobiet nüshu)
c) pismo alfabetyczne
d) pismo alfabetyczno-sylabiczne – rodzaj pisma, w którym jedna litera reprezentuje zasadniczo całą sylabę wraz z domyślną samogłoską (np. hindi).

Współczesne języki europejskie używają systemu alfabetycznego. Choć większość języków ukształtowanych na obszarze Europy ma wspólne korzenie, to można dostrzec znaczące odmienności zapisu, np. między alfabetem łacińskim, cyrylicą, czy alfabetem greckim. Dochodzimy do wniosku, iż tylko na podstawie wyglądu można już w przybliżony sposób sklasyfikować pochodzenie informacji.

Na przestrzeni kilku dekad powstało wiele systemów kodowania tekstów. Jednakże nie każda metoda kodowania daje się zastosować do każdego języka. Możemy wziąć pod uwagę chociażby pismo chińskie, które zawiera tysiące znaków, oraz pismo łacińskie, które potrzebuje znacznie mniejszej pamięci. Nie było możliwe stworzenie kodowania, które zawierałoby wszystkie potrzebne znaki, występujące w tekstach pisanych, głównie ze względu na ograniczenia pamięci.

Alfabet łaciński jest najtrudniejszym z alfabetów do detekcji – rozróżniania – języka. Liczba języków zapisanych w ten sposób jest znacznie większa od 25. Możemy jednak zwrócić uwagę na pewne osobliwości poszczególnych języków. I tak, w języku angielskim używa się tylko 26 podstawowych liter alfabetu łacińskiego: a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z. W języku polskim z użycia wyszły litery: q, v, x, i dlatego polski alfabet, tj. alfabet używany do zapisywania tekstów całkowicie pochodzących z języka polskiego (bez wyrazów obcych i obcych nazw własnych) zawiera tylko 24 litery, ale oprócz tego stosowanych jest 8 innych liter: ź, ż, ś, ó, ń, ę, ć, ą, które nie występują w innych alfabetach.

W przeszłości, kiedy liczył się każdy bit, nawet dla języka angielskiego tworzono oddzielne kodowania. Z tego powodu Unia Europejska wymagała kilku różnych kodowań, aby istniała możliwość zapisu treści w komputerze we wszystkich językach urzędowych.

Zdarza się, że różnorodność kodowań przysparza kłopoty. Dany komputer może potrzebować obsługi różnych rodzajów kodowania w tej samej chwili, np. do przeglądania plików na dysku, przeglądania stron internetowych, czy korzystania ze

skrzynek mailowych. Dwa różne kodowania mogłyby używać tych samych liczb dla dwóch różnych znaków. W momencie, gdy dane są przekazywane pomiędzy różnymi rodzajami kodowania, lub platformami, może dojść do deformacji treści, w wyniku której informacja jest źle odczytana lub nie do odczytania.

6.2. Słowa i stop lista

Wyrazy są najmniejszymi i względnie samodzielnymi jednostkami językowymi, wykorzystywanymi do zapisu wyrażeń. Jednakże, na przykład w języku polskim występują różne formy słów, wynikające z fleksji. Ponadto język polski ma mnóstwo możliwości słowotwórczych (zdrobnienia, zgrubienia, zdrobnienia zdrobnień, przedrostki lub przyrostki itp.). Słownictwo może mieć różne formy. Inaczej porozumiewamy się ze względu na różne gwary, czy na różnice pokoleniowe, miejsce pracy, w którym używamy zwrotów technicznych, nie do końca zrozumiałych dla osób o innym wykształceniu, czy też choćby przy tworzeniu treści tajnej, aby być zrozumianym tylko przez określone osoby. Każdy używa innego słownictwa, zależnie od miejsca urodzenia, zamieszkania, środowiska, ale także możliwości intelektualnych i twórczych. Potwierdzeniem oryginalności językowej każdego z nas są studia na temat identyfikacji autorów (Calix i in., 2008; Gerritsen, 2003; Narayanan i in., 2012). Problem ten jest bardzo istotny z punktu widzenia np. prawa, czy dziennikarstwa, ale jest ważny także dla pozyskiwania informacji i lingwistyki komputerowej.

Na świecie funkcjonuje mnóstwo języków i ich odmian. Identyfikacja języka może odbywać się na podstawie wykrywania słów, właściwych dla danych języków. Jeśli do identyfikacji języka chcielibyśmy posłużyć się całymi słownikami, musielibyśmy posiadać niewyobrażalnie wielką bazę. Rodzi się zatem pytanie: co uznamy za słowa należące do danego języka i czy nie warto byłoby („optymalnie") ograniczyć zasobów naszej hipotetycznej bazy dla przyśpieszenia obliczeń? Szczególnie, jeśli zakres naszych badań jest zawężony i dotyczy informacji technicznych, możemy zapomnieć o zdrobnieniach, czy też o nomenklaturze medycznej, lub zoologicznej.

Czy faktycznie, do celów identyfikacji języka, możemy znacząco zawęzić ilość informacji? Takie postawienie kwestii prowadzi do pytania o zawartość merytoryczną tekstów w porównaniu z hipotetycznymi słownikami, ale także i do pytania o własności statystyczne tekstów. Najczęstszymi słowami, znajdującymi się w dokumentach, są słowa o małym znaczeniu. Zwykle są to rodzajniki, przyimki, spójniki, zaimki zwrotne, ale również liczebniki, czy też niekiedy czasowniki. W systemach wyszukiwania informacji tekstowej, na podstawie częstości występowania słów w różnych tekstach wyznacza się wagi tych słów dla konkretnych tekstów, co ma prowadzić do charakteryzacji („indeksowania") poszczególnych dokumentów.

Zgodnie ze znanym – zresztą nie tylko z lingwistyki – empirycznym prawem, znanym jako prawo Zipfa, częstość występowania słów może być opisana prostą malejącą zależnością, przy czym słowa o największych częstościach, mają z reguły niewielkie lub żadne znaczenie. Dopiero od pewnego miejsca w utworzonym w ten

sposób rankingu pojawiają się słowa znaczące dla danego tekstu. Pierwsze słowa tego rankingu, o znikomej wadze w sensie znaczenia, tworzą najczęściej tzw. stop-listę, listę słów, które bez straty dla znaczenia treści, można z niej usunąć. (Nie znaczy to jednak bynajmniej, by słowa bez znaczenia nie mogły się pojawiać daleko w ewentualnym rankingu częstości, np. właśnie „bynajmniej"). Usuwanie słów ze stop list jest przydatne przy klasyfikacji dokumentów lub wyszukiwania informacji.

Stop listy mogą być wyodrębnione ze statystyki częstości słów określonych dokumentów w założonym przedziale, lub też mogą być określone przez konkretne słowniki. Strona *http://www.Ranks.nl* zawiera stop listy 19 języków, używane w ramach silnika do analizy tematyki stron WWW i artykułów. Użyta tam stop lista dla języka angielskiego zawiera 174 słowa. Dodatkowo, dla języka angielskiego przedstawiono też pełniejszą stop listę, z której korzysta MySql. Zawiera ona aż 543 słowa, ale nie jest to najdłuższa listą, z jakiej można skorzystać w celach praktycznych. Podano tam także jeszcze dwa inne przykłady stop list, najdłuższą – liczącą 667 słów oraz bardzo krótką, składającą się tylko z 32 wyrazów.

6.3. Metoda oparta na modelu słownikowym

Podczas nauki języka, człowiek uogólnia zasłyszane elementy. Cały proces nauki nie jest procesem skończonym, lecz otwartym, ale już na podstawie szczątkowej wiedzy jesteśmy w stanie określić język rozmówcy, a nawet cel użytych słów/wyrazów. Nie jest przy tym konieczne rozumienie przekazywanych treści.

Modelem słownikowym można, w sensie potocznym, nazywać po prostu zbiór wszystkich wyrazów używanych w danym języku. Metoda słownikowa w identyfikacji języka polega na stworzeniu modeli językowych, opartych na grupie wyrazów z danego języka (na podzbiorze słownika całościowego). Następnie nowy, analizowany tekst jest reprezentowany przez wektor słów z liczbą n wejść, która odpowiada wszystkim znalezionym słowom z modelu. Otrzymujemy relację między liczbą n a wskaźnikiem obecności lub nieobecności słów z danego korpusu. Metoda słownikowa, identyfikując język, sprawdza występowanie poszukiwanych wyrazów we wszystkich dostępnych słownikach językowych. Rozpoznanie języka następuje poprzez wskazanie modelu słownikowego o największej liczbie trafionych słów.

Do tego celu możemy użyć różnych modeli słownikowych:

a) słownik statyczny – który jest zdefiniowany odgórnie, np. słownik języka polskiego PWN
b) słownik quasi-statyczny – słownik zbudowany na podstawie analizy danych
c) słownik dynamiczny – słownik adoptowany i zmieniający się po dostarczeniu nowych informacji.

Aby określić język nie potrzeba, w rzeczywistości, dużej ilości informacji. Czasem sama stop lista może okazać się zbiorem „wiedzy" wystarczającym do tego celu. Przyjrzyjmy się, mianowicie, następującemu przykładowi:

Język niemiecki: *Ich* spreche *nicht* gut deutsch
Język holenderski: *Ik* spreek *niet* goed Nederlands

Język francuski: *Je ne* parle *pas* bien *le* français
Język polski: *Nie* mówię dobrze *po* polsku

Podkreślone w powyższych tekstach słowa są typowe dla stop list. Liczba słów pasujących do danego języka, czyli „liczba wejść", równa jest w tym wypadku cztery dla języka francuskiego i po dwa dla pozostałych języków (Tabela 1).

Tabela 1. Wykrycie słów w słownikach dla czterech słów i wybranych języków

Słowa:	*je*	*ne*	*pas*	*le*
niemiecki	+	+	-	-
holenderski	+	-	+	-
francuski	+	+	+	+
polski	+	-	+	-

Metoda słownikowa w swojej postaci klasycznej jest bardzo restrykcyjna. Jeśli zostanie popełniona tzw. literówka, lub, ze względu na kodowanie, program nie będzie mógł właściwie zinterpretować znaków, szanse na poprawne przypisanie języka znacznie spadają. Dodatkowo, trzeba zwrócić uwagę na wielkość słownika oraz na poprawność i wagę słów.

6.4. Model n-gramowy

„Jeden sposób na odczytanie zaszyfrowanej wiadomości, gdy wiemy w jakim języku została napisana, polega na znalezieniu innego tekstu w tym samym języku, na tyle długiego, by zajął mniej więcej jedną stronę, i obliczeniu, ile razy występuje w nim każda litera. (...) Następnie bierzemy tekst zaszyfrowany i również znajdujemy najczęściej występującą w nim literę, zastępując go najczęściej występującą literą innego tekstu (...)"

z rękopisu arabskiego uczonego z IX wieku, Al-Kindiego

W większości stosowanych systemów identyfikacja języka jest oparta na matematycznym modelu, który został opracowany jako fragment podejścia do łamania kodów. W każdym języku częstość występowania określonej litery jest inna. Podobnie jest ze zróżnicowaniem częstości występowania grup kolejnych dwóch, trzech, lub więcej liter. Zauważmy, że większość słów ze stop listy to wyrazy o małej liczbie liter, a ponadto języki, które podlegają odmianie, posiadają specyficzne, powtarzające się, kilkuliterowe końcówki, lub inne złączenia.

Ponadto, jak już wspomniano, wiele języków posiada pewne litery, lub warianty liter które są charakterystyczne wyłącznie dla danego języka. Według Beesleya (1988) istnienie – lub nie istnienie – liter charakterystycznych jest często wystarczające do identyfikacji konkretnego języka.

Obecnie, wśród najbardziej popularnych metod identyfikacji języka, ważne miejsce zajmują metody bazujące na n-gramach, czyli sekwencjach liter o długości

n[2]. Modele n-gramowe są stosowane w statystyce języków i kodów, rozpoznawaniu mowy, lingwistyce komputerowej, kompresji danych i w wielu innych dziedzinach. Metoda z wykorzystaniem modelu n-gramowego jest nieco podobna do rozpoznawania za pomocą bazy słów (o czym wspomniano, przytaczając przykład możliwości zastosowania stop listy do identyfikacji języka). Posiada także kilka dodatkowych atutów: nie trzeba wykrywać znaków ani też martwić się o znaki interpunkcyjne, o ile tylko statystyka i pliki przygotowane są w odniesieniu do tego samego kodowania, poza tym, metoda ta może być użyta dla każdego języka oraz akceptuje tekst napisany z błędem.

Celem stworzenia modelu językowego opartego na n-gramach jest przewidywanie prawdopodobieństw naturalnie występujących sekwencji liter, oparte na odpowiednich statystykach. Metoda n-gramowa wymaga zgromadzenia, zwłaszcza dla większych wartości n, bardzo dużego zasobu danych, ponieważ dla większych n liczba tekstów do analizy, w której będą uwzględnione wszystkie n-gramowe wystąpienia, powinna być adekwatnie większa. Im większa ilość tekstów w zasobie, tym wyższa jakość modelu (tj. tym bardziej uzasadnione przybliżenie prawdopodobieństwa przy pomocy częstości wystąpień).

I tak, dla wyrażenia *Test hamulców* otrzymujemy 11 następujących trigramów (3-gramów) o liczności dla każdego równej jeden i częstości równej 0.09(09):

Tes / est / st / t h / ha / ham / amu / mul / ulc / lcó / ców

Tworzenie n-gramów ze spacjami jest przydatne, bo pozwala na badanie – jednocześnie – także n-1-gramów, a nawet n-2-gramów, co może być ważne, jeśli – a bywa tak często – istotną rolę w języku pełnią początki i/lub zakończenia wyrazów. I tak, np., w języku angielskim występuje wiele słów zaczynających się na „Th", zaś w niemieckim występuje znacznie więcej wyrazów rozpoczynających się od litery „Z" niż w angielskim.

Prostota i możliwość skalowalności to główne zalety modeli opartych na n-gramach. Zmieniając „n", możemy otrzymać modele, które nie wymagają obszernych danych treningowych, ale nie dają dużej mocy predykcyjnej, i możemy otrzymać modele wymagające dużego zbioru danych, ale za to oferujące duże możliwości predykcyjne.

Grefestett (1995) prezentuje wyniki, ilustrujące skuteczność identyfikacji języka za pomocą techniki „małych słów", które są zbiorem częstych określników, spójników i przyimków (kwalifikujących się w dużej mierze do stop-listy). Wyniki te zostały zestawione z metodą trigram, tj. opartą na częstości występowania sekwencji trzech znaków. Wyniki dla tekstów złożonych z małej liczby słów są lepsze w przypadku zastosowania metody trigram. Już przy treści, przeznaczonej do

[2] Ogólnie, model n-gramowy może także dotyczyć słów, nie tylko liter, czy znaków. Zasada bowiem dotyczy możliwości oceny częstości wystąpień, a nie samych obiektów, których częstości są zliczane, a następnie testowane statystycznie.

rozpoznania, zbudowanej z jednego lub dwóch słów, metoda ta potrafi osiągnąć przewidywalność od prawie 60% do ponad 90%, podczas gdy dla tych samych warunków metoda słownikowa daje dwa razy gorsze rezultaty. Z drugiej strony, metoda słownikowa znacznie szybciej osiąga stabilizację wyników przy rosnących długościach sekwencji i pewniej działa dla dłuższych zdań (Tabela 2). Obecnie nadal prowadzone są badania nad skutecznością metod, zarówno statystycznych, Shuyo (2012), jak i słownikowych, Zampieri (2013).

Tabela 2. Średnie udziały procentowe poprawnej identyfikacji języka, w funkcji długości zdania, dla metod opartych na trigramie i technice małych słów, jako metody słownikowej, dla języków: norweskiego i hiszpańskiego.

Norweski								
Długość zdania	1 lub 2	3 do 5	6 do 10	11 do 15	16 do 20	21 do 30	31 do 50	> 50
trigram	70.8	91.3	98.1	99.5	99.7	99.9	99.9	100
technika małych słów	87.5	97.4	99.2	99.8	99.9	100	100	100
Hiszpański								
Długość zdania	1 lub 2	3 do 5	6 do 10	11 do 15	16 do 20	21 do 30	31 do 50	> 50
trigram	73.8	86.9	97.3	99	99.8	99.9	99.9	100
technika małych słów	8.1	81.5	98.8	99.7	100	100	100	100

6.5. Sporządzanie zbioru danych oraz klasyfikacja

Mając bazę n-gramów lub słów oraz ich statystykę, możemy zacząć identyfikację języka treści. Polega ona na zaklasyfikowaniu odpowiedniego tekstu do jednego z założonych języków, lub grupy języków. Należy przy tym uwzględnić kilka szczególnych sytuacji dla tekstów występujących w plikach po tłumaczeniu:

- W bazie może znajdować się identyczna dana, taka jak, np. „Ich”, występująca w więcej niż jednym języku, w tym przypadku – w języku polskim i niemieckim. Takich danych może być więcej.
- Dodatkowo, w treści tekstu możemy natrafić na sytuację, w której nie będziemy w stanie jasno określić przynależności do żadnego języka.
- Teksty do identyfikacji są krótkie i wnoszą mało informacji.
- Niektórych treści nie tłumaczy się i pozostają w języku źródłowym.
- Tekst może być w innym języku niż klasy, które założyliśmy na początku.

Celem klasyfikacji jest przyporządkowanie nowych danych, dla których wartość atrybutu decyzyjnego (klasy) nie jest znana, do odpowiedniej klasy. Jako całość postępowania, jest to metoda eksploracji danych z nadzorem. W celu ustalenia języka, dla nowego tekstu należy określić jego relacje do poszczególnych, znanych już, kategorii (tutaj: języków, grup języków), ustalonych dzięki zbiorom

treningowym. Po otrzymaniu klasyfikatora, tj. zależności, wskazującej klasę dla podanych na wejściu danych, w drugim etapie stosujemy go do klasyfikacji nowych danych. Podczas testowania weryfikujemy, w oparciu o zbiór testowy, dokładność naszego klasyfikatora. Jeśli jest ona akceptowalna, model możemy użyć do klasyfikacji.

Istnieje szereg metod i algorytmów klasyfikacji, por. np. Mitchell (1997), Kuncheva (2004), które przyporządkowują nowe elementy do najbardziej pasujących klas.

W celu identyfikacji języka przetestowano różne rodzaje danych np. duże dokumenty, zawierające wiele słów, czy też krótkie zwroty lub zapytania, używane w wyszukiwarkach tekstowych. Oprócz długości treści do identyfikacji należy zwrócić uwagę na jakość i charakter danych. Podczas gdy dokumenty naukowe nie zawierają, zazwyczaj, licznych błędów stylistycznych itp., to dla wiadomości na tweeterze, Han i Baldwin (2010), potrzebna jest uprzednia normalizacja danych. Przeprowadzono też badania dla dokumentów, w których może występować kilka języków, Jim et al. (1999).

Różnorodność problemów zaowocowała dużą liczbą bardziej szczegółowych badań. Dunning (1994), korzystając z badań Cavnar'a i Trenkle'a (1994), zaproponował system, który używa łańcuchów Markowa z regułą decyzyjną Bayesa. W 1996 r., przy użyciu entropii względnej, wygenerowano prawdopodobieństwo rozkładu dla danych treningowych i testowych, a następnie zmierzono odległość między prawdopodobieństwami przy użyciu odległości Kullbacka-Leiblera (Sibun i Reynar, 1996). McCallum i Nigam (1998) badali dwie różne metody oparte na założeniach klasyfikatora naiwnego Bayesa. Poutsma (2001), natomiast, zaproponował system oparty na metodzie Monte Carlo. Badano również klasyfikator oparty na metodyce SVM (Support Vector Machine), Campbell i in. (2004), a także wiele innych rozwiązań.

Oprócz poszukiwań nowych rozwiązań, pracowano również nad ulepszaniem starych technik. Dokonywano rozmaitych porównań starych oraz nowych metod. Brano pod uwagę długości tekstów, liczbę i różnorodność języków, kodowania znaków itp. (Hughes i in., 2006, czy www.slideshare.net). Poutsma (2001) przedstawił zależność liczby znaków i prognozy w sensie określenia prawidłowego języka dla 6 różnych metod (3 słownikowych i 3 opartych na n-gramach) oraz zależność liczby znaków od logarytmu czasu prognozy. Wynik badań jest następujący: dla krótkich tekstów najlepsze wyniki dają modele oparte na n-gramach i ich występowaniu lub częstościach, metody słownikowe natomiast dają lepsze rezultaty ze względu na mniej kosztowne obliczenia i lepszą dokładność przy dłuższych zdaniach.

Ponieważ opisy popularnych technik – naiwnego klasyfikatora Bayesa oraz k-NN (k najbliższych sąsiadów), które rozważano w prezentowanej pracy, zostały przedstawione w artykule Ryczaja i Owsińskiego (2015), nie będziemy ich tutaj przytaczali.

7. Eksperymenty

7.1. Zastosowanie identyfikacji w biurze lokalizacji językowej

W celu udoskonalenia procesów, związanych z lokalizacją językową, oraz ujednolicenia rozwiązań, stosuje się różne modele, które pomagają w ustaleniu przyczyn problemów i ich skutków. Na jakość dostarczanych plików wpływają różne czynniki i niektórych nie możemy zmienić, czynniki związane z wykwalifikowaniem, skupieniem, a także emocjonalnym i fizycznym zmęczeniem pracownika. Ponadto swój wpływ ma również przebieg procesu, wewnętrzne ustalenia, ruch plików pomiędzy różnymi systemami, błędy oraz szereg innych aspektów.

Mieszanie się języków nie jest częstą sytuacją, dlatego można to traktować jako mające „średnio" niski wpływ na jakość. Stąd, w sytuacjach trudnych do realizacji nie powinniśmy być może w ogóle zajmować się tym problemem. Jest jednak całkiem inaczej, jeśli do druku może pójść nakład kilku tysięcy, lub więcej, niezrozumiałych instrukcji. Klient może zażądać wysokiej kary, uwzględnionej w umowie. Przyjęte rozwiązanie musi być możliwie proste. Kwestia wpływu na jakość pozostaje najczęściej drugorzędna, jako wynikająca wprost ze specyfiki projektu.

Przy ocenie usprawnienia sprawdzamy efektywność dostępnych rozwiązań. Podczas weryfikacji przetłumaczonych plików jednym z etapów jest wykrywanie błędów wynikających z „wykrzaczonych" znaków. Etap ten nazywamy „corruption check". Korzystając z metod „corruption check" niekiedy udaje się nam wykryć źle umieszczony w treści język, czyli, oprócz wykrywania błędów powstałych na skutek zmiany kodowań, możemy wykrywać błędny język. Odpowiedni skrypt, zaimplementowany dla całego etapu weryfikacji, może spowodować zaniechanie poszukiwań alternatywnego sposobu wykrywania błędnych tłumaczeń. Niezależnie od tego, dodatkowy algorytm mógłby wydłużyć znacznie czas sprawdzania plików.

Poszczególne projekty posiadają własne bazy. Treści zawarte w bazie łączy wspólna tematyka oraz często ci sami tłumacze, ewentualnie biura tłumaczeniowe. Wykorzystanie do identyfikacji języka własnych zasobów, ze względu na specyficzność treści o określonej tematyce, może okazać się dobrym pomysłem.

Identyfikacja języka jest potrzebna w wielu systemach, dlatego na stronach internetowych łatwo znaleźć biblioteki dla różnych języków programowania. Niektóre biblioteki są wyposażone od razu we wdrożone modele językowe. Należy jednak pamiętać, że im większa liczba języków tym większa różnorodność odpowiedzi.

Większość instrukcji technicznych, czy dokumentacji, wobec globalizacji i ujednolicania, jest początkowo pisanych po angielsku. Zdarzają się też tłumaczenia z języków mniej powszechnie znanych, takich jak: portugalski, niemiecki, rosyjski, czy innych. Liczby języków docelowych mogą zawierać się w przedziale od jednego do kilkunastu, czy też kilkudziesięciu. Każda baza zawiera dane dla określonych tematów, tak aby móc dopasować kontekst treści. Dla jednych źródeł segmenty są krótkie, dla innych zawierają same tytuły bądź wręcz przeciwnie, zawierają głównie

zdania złożone, zaś dla jeszcze innych powstają z całych akapitów w dokumentach.

7.2. Plan przebiegu badania

Dane do badania zostały przygotowane na podstawie 14 różnych baz zawierających tłumaczenia z: gwarancji, biuletynów obsługi technicznej, materiałów treningowych, dokumentów zawierających informacje o symptomach i diagnostyce procedur, dokumentacji, instrukcji technicznych oraz innych. Głównym ich tematem jest motoryzacja. Wszystkie dokumenty, które posłużyły do sporządzenia zestawu danych, były tłumaczone z języka angielskiego lub niemieckiego.

7.2.1. Badanie za pomocą „corruption check"

Celem pierwszej części badania było zwrócenie uwagi na łatwość wyróżnienia niektórych języków na podstawie używanych znaków. W większości, uwzględnione języki (brytyjski, amerykański, polski, grecki, wietnamski, norweski, duński, francuski, hiszpański, szwedzki, portugalski), używają pisma łacińskiego.

Następnie przebadano efektywność dostępnego rozwiązania, tj. metodyki „corruption check". Do badania użyto tłumaczeń na 12 języków: czeski, francuski, portugalski, polski, węgierski, hiszpański, wietnamski, włoski, niemiecki, szwedzki, duński, norweski, a dla każdego - 1000 segmentów. W każdym języku akceptowalne są wszystkie litery używane w angielskim. Różnice polegają na dodatkowych literach, wymienionych w Tabelach 4 i 5. Weryfikacja szczególnych znaków miała pokazać szanse znalezienia błędnych tłumaczeń przy pomocy tej techniki „corruption check". Przeliczono ilość znaków szczególnych w całej treści w relacji do ilości wszystkich znaków oraz ilość segmentów, w których występują te znaki.

7.2.2 Przygotowanie danych dla metody słownikowej i metody modeli n-gram

Głównym celem pracy stało się znalezienie lepszej metody do identyfikacji języków pisanych pismem łacińskim. W tym celu przebadano dwie metody. Pierwsza była oparta na wykrywaniu słów w segmentach i weryfikacji na ich podstawie przynależności do języka, druga opierała się na znajdowaniu n-gramów i klasyfikacji segmentów za pomocą naiwnego klasyfikatora Bayesa. Zestaw danych został skonstruowany na podstawie języków, dla których można było otrzymać ponad 150 000 segmentów, czyli w analizowanym przypadku - dla duńskiego, hiszpańskiego, francuskiego, norweskiego, polskiego, portugalskiego i szwedzkiego.

Ze względu na różne zaszumienia, których automatycznie nie można w całości usunąć, dla modelu n-gram wyszukano także na stronach wikipedii oraz wikibook treści w celu stworzenia dodatkowego modelu. Stworzenie dodatkowego modelu nie jest własnym pomysłem, skorzystano w tym zakresie ze sposobu opisanego przy niektórych bibliotekach służących do identyfikacji języka. Dla każdego języka ilość znaków nowych danych była większa niż milion.

Aby zwiększyć szanse identyfikacji, automatycznie zmieniono treść segmentów na wszystkich plikach, przy użyciu wyrażeń regularnych, co wpłynęło na ujednolicenie formy zbiorów. Przeprowadzono także inne operacje „czyszczące" zbiory danych. Dotyczyło to, w szczególności, tagów, placeholderów, itp.

Znaki interpunkcyjne i cyfry zostały podczas przetwarzania zamienione na spacje. W szczególności, cyfry w językach pisanych pismem łacińskim nie wnoszą żadnej dodatkowej informacji, a wręcz przeciwnie, mogą powodować zaszumienie. Następnie wszystkie dane zostały ujednolicone dla uniknięcia ponadwymiarowych spacji, które powstały w trakcie opisanych tutaj zmian.

Oprócz procesów słowotwórczych oraz adaptacyjnych w stosunku do określonego środowiska, coraz częściej do tekstów wkradają się zapożyczenia, skróty, rozszerzenia skrótów w języku oryginału lub cytaty. To może prowadzić do dużego szumu i otrzymania niepoprawnego wyniku. Niektóre najbardziej oczywiste nazwy i skróty zostały usunięte.

Większość zmian opisanych powyżej dotyczy danych z baz klienta. Tylko zmiany dotyczące adresów mailowych i stron, znaków interpunkcyjnych oraz cyfr zostały również przeprowadzone dla zbiorów otrzymanych ze stron internetowych.

Po usunięciu powyższych treści, niektóre segmenty w ogóle przestają istnieć. Pojawia się też znaczna liczba segmentów identycznych, które początkowo różniły się wartościami wielkości fizycznych, nazwami własnymi, placeholderami, lub tagami. Dane zostały zmniejszone o treści identyczne i liczba segmentów zmniejszyła się o ok. 30%, co przy ponad 150 tys. segmentów jest dużą wielkością.

Tabela 3. Zmiany ilości segmentów

Etap	Duński	Hiszpański	Francuski	Norweski	Polski	Portugalski	Szwedzki
1	177 847	187 896	179 725	150 347	260 256	268 170	268 200
2	177 679	187 733	178 707	149 062	260 027	267 926	267 990
3	117 896	122 321	150 865	129 579	160 261	164 646	168 505

Tabela 3 przedstawia liczby segmentów w trzech etapach obróbki danych. Pierwszy etap informuje o całkowitej liczbie segmentów z bazy, w drugim etapie liczba zmniejszyła się o segmenty, w których skasowano tagi, placeholdery, nazwy własne itp. Ostatni etap to liczba segmentów po usunięciu powtarzających się treści.

Długość segmentów ma znaczenie. W celu otrzymania odpowiednich danych dla zbioru treningowego i testowego przebadano długość segmentów. Zauważono przy tym pewne ciekawe właściwości. Mianowicie, do około 20 wyrazów widać progresywne uzyskiwanie znacznej większości danych. Natomiast po przekroczeniu 40 wyrazów liczba segmentów wzrasta już bardzo niewiele. Porównując liczbę wyrazów w zależności od ilości segmentów, np. dla języka duńskiego i hiszpańskiego, przy podobnej liczbie dostępnych segmentów widać, że język duński posiada znacznie więcej segmentów o mniejszej liczbie wyrazów niż język hiszpański. Może to świadczyć o specyficznej budowie zdań dla określonych języków.

7.2.3. Plan przebiegu identyfikacji segmentów

Po wyczyszczeniu danych oraz ich wstępnej analizie można było przystąpić do opracowania kroków badania.

I tak, najlepszy zbiór danych dla modelu słownikowego opartego na 300 wyrazach został porównany ze skutecznością identyfikacji modelu słownikowego dla

150 i 20 słów. Do określenia języka został stworzony skrypt w języku python. Zakładał on „dodatkowy punkt" dla klasy, w której kolejne słowo z segmentu znajduje się w jego modelu.

Dla krótkich segmentów identyfikacja języka jest utrudniona, z powodu małej ilości danych. Częsty jest wówczas brak dopasowania treści do jakiegokolwiek modelu językowego. Przedstawiono najczęstsze możliwe pomyłki dla segmentów zwierających mniej niż 6 wyrazów.

Metoda następnie została nieco ulepszona poprzez zwrócenie uwagi na częstość słowa w odpowiednim modelu językowym. Pomogło to w jasny sposób określić język oraz zwiększyć efektywność predykcji.

Następnie badano metodę opartą na modelach n-gramowych. Do badań wykorzystano skrypt z www.cavar.me dostępny jako open source, w języku python, napisany pod licencją GNU. Na potrzeby projektu został on trochę zmieniony. W pierwotnej wersji skrypt korzysta z modelu trigram i wykorzystuje naiwny klasyfikator Bayesa. Badania zostały przeprowadzone dla tych samych danych, co badania dla metody słownikowej. Dodatkowo badania przeprowadzono ze zmianą zbiorów treningowych na zbiory pochodzące ze stron internetowych. Dla porównania, skrypt został dostosowany dla metody opartej na modelach bigram i unigram również dla różnych zbiorów treningowych.

7.3. Różnorodność zapisu treści

W odniesieniu do wspomnianego już stwierdzenia o możliwości identyfikacji języka na podstawie charakterystycznych znaków, poniżej zaprezentowano zdanie w jedenastu różnych językach po tłumaczeniu zdania angielskiego. Kontekst tłumaczeń jest różny, dlatego tłumaczenie słowa „timing" nie jest takie samo.

Brytyjski:	Software installation timing
Amerykański:	Software installation timing
Polski:	Termin instalacji oprogramowania
Grecki:	Χρονοδιάγραμμα εγκατάστασης λογισμικού
Wietnamski:	Cài đặt phần mềm
Norweski:	Tidsplan for programvareinstallasjon
Duński:	Tidsplan for softwareinstallation
Francuski :	Calendrier d'installation du logiciel
Hiszpański :	Fechas de la instalación del software
Szwedzki:	Tidsplan för installation av programvaran
Portugalski:	Momento da instalação do software

Niektóre języki znacząco wyróżniają się budową. W powyższym zestawieniu są to języki grecki oraz wietnamski (ten drugi - mimo iż zapisany znakami pisma łacińskiego).

Odróżnienie w tym przykładzie języka angielskiego od amerykańskiego jest niemożliwe, a bardzo podobny efekt obserwujemy dla duńskiego i norweskiego.

W ogólności, przytoczone zdanie jest dobrym przykładem na ograniczoną możliwość identyfikacji języka na podstawie znaków szczególnych, jeśli tekst poddany analizie nie jest wystarczająco długi, w sensie zawierania minimalnego zestawu odpowiednich znaków.

W języku wietnamskim, oprócz dodatkowych liter używanych w alfabecie, wyróżnia się też znaczną część znaków tonalnych, których tutaj nie wymieniamy.

7.4. Weryfikacja efektywności corruption check w identyfikacji języka

W celu przebadania szansy wykrycia nieprawidłowego języka ze zbioru danych wyodrębniono 1 000 segmentów, z liczbą znaków około 4 700 dla dwunastu języków: czeskiego (CS), francuskiego (FR), portugalskiego (PT), polskiego (PL), węgierskiego (HU), hiszpańskiego (ES), wietnamskiego (VI), włoskiego (IT), niemieckiego (DE), szwedzkiego (SV), duńskiego (DK), norweskiego (NO). Tabela 5 przedstawia udział zdań, w których znajdują się znaki z Tabeli 4. Dodatkowo, zanotowano liczbę wystąpień znaków w zdaniach, co doprowadziło do wyznaczenia średniej ilości szukanych znaków w stosunku do ilości wszystkich użytych znaków.

Tabela 4. Litery szczególne występujące w niektórych językach pisanych pismem łacińskim wykrywane przez „corruption check”

CZESKI	FRANCUSKI	PORTUGALSKI	POLSKI	WĘGIERSKI	HISZPAŃSKI
Á	à	Á	Ą	á	á
Č	â	À	Ć	é	é
Ď	ç	Â	Ę	í	í
CZESKI	FRANCUSKI	PORTUGALSKI	POLSKI	WĘGIERSKI	HISZPAŃSKI
É	é	Ã	Ł	ó	ñ
Ě	è	Ç	Ń	ö	ó
Í	ê	É	Ó	ő	ú
Ň	ë	Ê	Ś	ú	ü
Ó	î	Í	Ź	ü	
Ř	ï	Ô	Ż	ű	
Ť	ô	Ó			
Ú	œ	Ú			
Ů	ù	Ü			
Ý					
Ž					
WIETNAMSKI	WŁOSKI	NIEMIECKI	SZWEDZKI	DUŃSKI	NORWESKI
Ă	à	ä	Å	æ	æ
Â	è	ö	Ä	ø	ø
đ	ì	ß	Ö	å	å
ê	ò	ü			
ô	ù				
ơ					
ư					

Dla języka wietnamskiego widzimy, iż około 1/3 wszystkich znaków to znaki spoza „standardowego" alfabetu łacińskiego, czyli, że już co trzeci znak może dać nam informację o użytym języku. Przewidywalność odnalezienia krótkiego zdania za pomocą tych znaków wynosi praktycznie 100% (tutaj: 99.9%).

Najtrudniej jest określić występowanie języka włoskiego, bo w rozpatrywanym przykładzie tylko w 14.1% zdań występują znaki specjalne, a następnie języka niemieckiego, gdzie występują one w niecałych 50% zdań. W szwedzkim, mimo iż wyszukiwane były tylko 3 litery, częstość ich występowania jest większa niż w portugalskim czy francuskim, dla których zostało wyodrębnionych po 12 liter.

Niektóre litery diakrytyzowane występują w analogicznej postaci graficznej i kodowej w kilku językach. Oznacza to, że mając model oparty na 12 językach i stosując „corruption check" mamy faktyczne szanse na wyłapanie nieprawidłowości tłumaczeniowych, które są jeszcze mniejsze niż przedstawione wyniki z Tabeli 5.

Tabela 5. Procentowe udziały wystąpień liter „szczególnych"

JĘZYK	procent zdań	procent znaków
VI	99.9	32.55
HU	92.5	10.30
CS	86.7	8.86
PL	83.0	5.16
SV	81.3	4.09
PT	77.4	3.55
FR	69.8	2.90
DK	66.4	2.62
NO	57.6	2.00
ES	54.3	1.81
DE	46.9	1.48
IT	14.1	0.33

W celu określenia szans poprawnej identyfikacji na podstawie corruption check dla 12 języków, zbiory zostały przebadane, a wyniki pokazano w Tabeli 6.

Alfabet duński jest identyczny z alfabetem norweskim i języki te są bardzo podobne. Posługując się znakami szczególnymi nie możemy odróżnić ich od siebie, czyli szansa znalezienia błędnego tłumaczenia norweskiego w duńskim pliku lub odwrotnie jest zerowa. Dla wielu przebadanych języków, trudno jest jednoznacznie określić wystąpienie tego właściwego. Udało się to najlepiej dla wietnamskiego (96.4%) i polskiego (78.5%), zaś najgorzej dla francuskiego (3.4%), hiszpańskiego (1.7%), i wreszcie włoskiego, duńskiego i norweskiego w 0%.

Jeśli wyłączymy język norweski to język duński jednoznacznie określić będzie można w 57.8%, natomiast, jeśli wyłączymy język duński, to język norweski jednoznacznie zostanie określony w 43.5%. Problem istotny dotyczy włoskiego, dla

którego w 86.2% segmentów nie zostały znalezione żadne znaki szczególne.

W segmentach mogą występować inne znaki wynikające z błędów – problemów z kodowaniami, czy też z pozostawionymi nazwami własnymi lub segmentami, które miały pozostać nie tłumaczone itp. Jak widać z wyników, kilka takich segmentów zostało wychwyconych. Przykładem jest język portugalski, w którym 2% segmentów zidentyfikowano jako wietnamskie, lub szwedzki, który w 0.1% zawiera znaki znajdujące się w alfabecie francuskim, portugalskim lub wietnamskim, dla norweskiego i duńskiego ten udział wynosi 0.4% segmentów.

Tabela 6. Rozpoznawanie języka za pomocą liter „szczególnych" (dla każdego języka podano, w %, kolejno, udziały: nierozpoznanych segmentów, segmentów rozpoznanych jako zapisane we właściwym języku, oraz segmentów rozpoznanych jako odpowiadające innym językom)

PT		HU		FR		DE	
-	22.6	-	6.5	-	30.3	-	55.1
pt	26.3	hu	58.4	cz fr pt hu es vi	29.2	pt hu es de	18.2
cz pt hu es vi	20.3	cz pt hu es vi	21.5	fr pt vi	18.7	de	10.8
pt vi	14.4	pt hu es	6.1	fr vi	7.7	de sv	9.4
fr pt vi	4.9	cz fr pt hu es vi	3.9	fr vi it	5.5	hu de sv	4.7
cz fr pt hu es vi	3.8	cz pt pl hu es vi	2.8	fr pt vi it	5.2	hu de	1.8
cz pt pl hu es vi	3	pt hu es de	0.8	fr	3.4		
fr pt	2.2						
vi	2						
fr pt vi it	0.5						

SV		NO		DA		ES	
-	19.7	-	42.4	-	37	-	45.8
de sv	32.5	da no	43.5	da no	57.8	cz pt hu es vi	27.7
sv	27.2	sv da no	13.4	sv da no	4.5	cz pt pl hu es vi	21.7
hu de sv	11	cz fr pt hu es vi	0.4	cz fr pt hu es vi	0.4	cz fr pt hu es vi	2.9
sv da no	9.5	cz fr pt hu es vi da no	0.2	cz fr pt hu es vi da no	0.3	es	1.7
fr pt vi	0.1	cz fr pt hu es vi sv da no	0.1				

VI		IT		CS		PL	
-	2	-	86.2	-	14.2	-	18.3
vi	96.4	fr pt vi it	6.3	cz	60.3	pl	78.5
cz pt hu es vi	0.6	fr vi it	6.3	cz pt hu es vi	18.3	cz pt pl hu es vi	3.2
fr pt vi	0.5	vi it	1.1	cz vi	4.7		
pt vi	0.5	cz fr pt hu es vi	0.4	cz fr pt hu es vi	2.1		

Przetestowana metoda jest przeznaczona stricte do wyszukiwania błędów wynikających ze zmiany kodowań znaków. Daje niekiedy szansę na wyszukanie pomyłkowego natłumaczenia, ale jego prawdopodobieństwo jest niskie. Występowanie pomyłek jest marginalne, a dodatkowo, jeśli szansa ich wykrycia jest niewielka, to problem może w ogóle pozostać niezauważony.

7.4. Identyfikacja języka

Wszystkie dane zostały otrzymane z dokumentów o zbliżonej tematyce. Początkowo powstały trzy zbiory, z których każdy zawierał zbiór treningowy i testowy. Na podstawie zbiorów danych treningowych stworzono modele słownikowe. W celu stworzenia słowników do identyfikacji segmentów użyto 300 najczęściej powtarzających się wyrazów, zidentyfikowanych w zbiorach danych treningowych. Porównując ilość wyrazów słownikowych z liczbą 300 najczęściej występujących dochodzimy do wniosku, że stanowią one około 1% wszystkich wyodrębnionych wyrazów słownikowych. Mimo tego, ich sumaryczna częstość dla większości języków wynosi ponad 50% wszystkich słów.

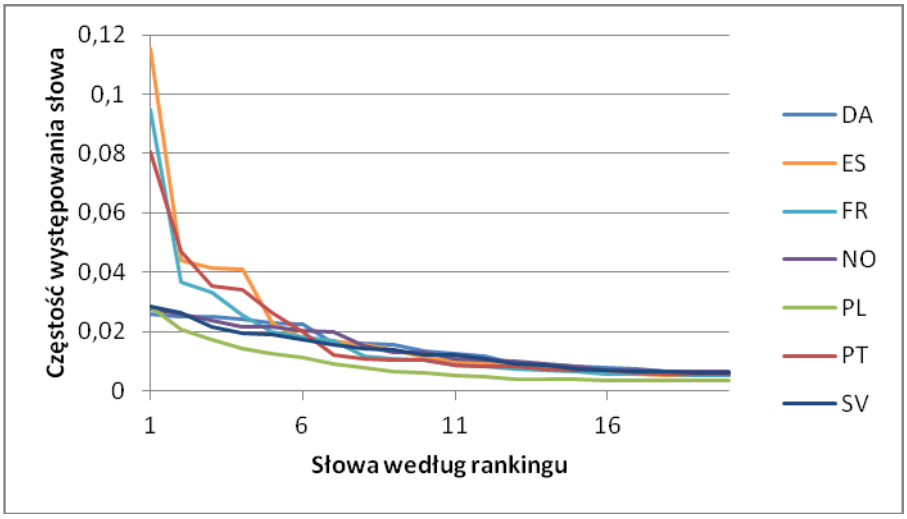
Dla każdego języka stop lista ma, naturalnie, inną długość. Podobnie, jest dość oczywiste, że do 15-20 pierwszych najczęstszych wyrazów stop-lista nie miesza się z wyrazami znaczącymi. Z uwagi na tę samą tematykę zbiorów, słowniki zostały wzbogacone o nieco większą ilość wyrazów, by tym samym zwiększyć efektywność rozpoznawania przynależności zdań.

Wyniki (por. Rys. 1) sugerują podobieństwo w obrębie dwóch grup języków: (1) hiszpański, portugalski, francuski, oraz (2) duński, norweski, szwedzki. Dla pierwszej grupy podobieństwo widać w rozkładzie częstości pierwszych wyrazów z rankingu. Grupa ta charakteryzuje się znacznie wyższą częstością dla pierwszych wyrazów, które są przedimkami, czy przyimkami. Ponadto języki te charakteryzują się możliwością większej predykcji. Dla drugiej grupy najwyższą częstość przyjmują przyimki lub spójniki. Wyniki dla języka polskiego nieco odstają od pozostałych grup.

Choć te obserwacje mogą się wydawać trywialne wobec znajomości pokrewieństwa poszczególnych grup języków, jednak dla prowadzonej analizy są bardzo ważne, ponieważ potwierdzają poprawność realizowanej procedury.

7.4.1 Metoda słownikowa

Wykrywanie języków jest uzależnione od długości segmentów. Podzielono zatem zbiór testowy na 5 grup w zależności od ilości słów w treści, a mianowicie: bardzo krótkich segmentów (do 5 wyrazów), krótkich segmentów (6 do 10 wyrazów), średnich segmentów, (11 do 20 wyrazów), długich segmentów (21 do 35 wyrazów), oraz bardzo długich segmentów, dłuższych niż 36 wyrazów.



Rys. 1. Wykresy częstości występowania dla 20 najczęstszych słów w badanych językach

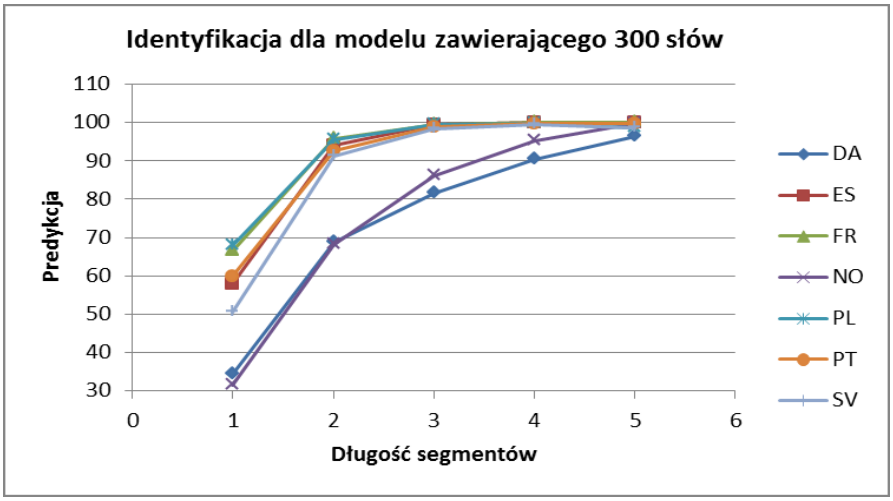
Zastosowanie metody słownikowej przynosi bardzo dobre rezultaty dla segmentów o liczbie wyrazów powyżej 5 (Tabela 7). Dla duńskiego i norweskiego poziom identyfikacji jest bardzo podobny, znacznie niższy niż dla innych języków.

Tabela 7. Wyniki procentowe poprawnej identyfikacji według długości segmentu przy zastosowaniu słownika o 300 najczęstszych wyrazach

Wyrazów w segmentach:	DK	ES	FR	NO	PL	PT	SE
0-5	34.18	58.04	66.31	31.34	67.75	59.81	50.44
6-10	68.83	94.06	95.69	68.26	95.51	92.53	91.29
11-20	81.53	99.38	99.51	85.97	99.33	98.77	98.23
21-35	90.21	99.88	99.89	95.01	99.83	99.75	99.42
> 36	96.26	100.00	99.93	99.67	98.61	99.65	98.59

Mimo tego, że słowa zawarte w modelu dla języka hiszpańskiego stanowiły ponad 68% wszystkich wyrazów ze zbioru treningowego, podczas, gdy dla języka polskiego było to tylko 45%, to okazało się, że względem zbioru testowego najlepszą przewidywalność ma język polski.

W celu sprawdzenia wpływu wielkości słownika na wyniki, zbadano dodatkowo zbiór słownikowy dwa razy mniejszy (150 wyrazów – Tabela 8) i zbiór złożony z praktycznie samej stop listy (Tabela 9). Bez znajomości danego języka, trudno jednak określić skład stop listy. Dlatego jako „stop lista" zostało w sposób automatyczny wybranych 20 najczęstszych słów dla każdego z języków.



Rys. 2. Poprawna identyfikacja z podziałem na grupy według długości segmentu przy zastosowaniu słownika o 300 najczęstszych wyrazach

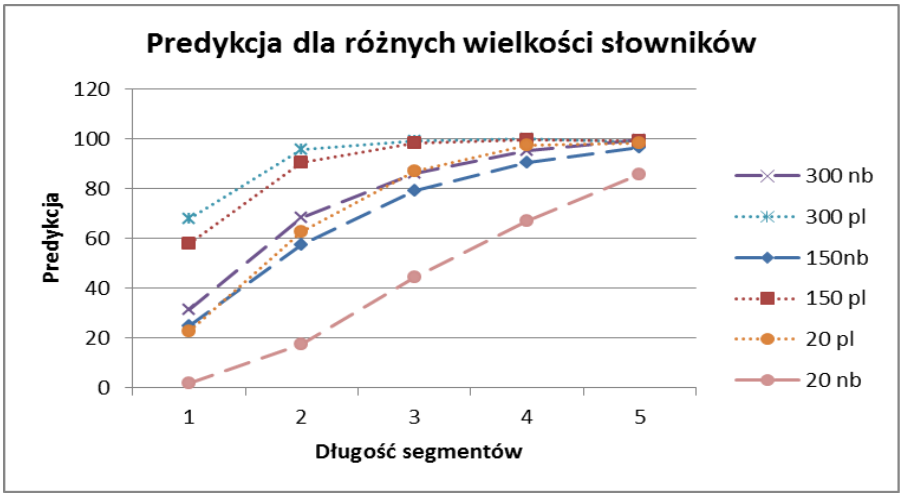
Tabela 8. Wyniki procentowe poprawnej identyfikacji z podziałem według długości segmentu przy zastosowaniu słownika o 150 najczęstszych wyrazach

Wyrazów w segmentach:	DK	ES	FR	NO	PL	PT	SE
0-5	27.32	53.52	57.03	24.77	57.68	52.09	43.95
6-10	61.26	91.49	92.52	57.46	90.63	89.04	86.35
11-20	77.87	99.05	99.07	79.17	98.20	97.82	96.67
21-35	90.95	99.86	99.79	90.55	99.64	99.71	99.03
> 36	94.56	100.00	99.80	96.37	99.08	99.65	98.59

Tabela 9. Wyniki (w %) poprawnej identyfikacji z podziałem według długości segmentu przy zastosowaniu słownika o 20 najczęstszych wyrazach (hipotetycznej „stop listy")

Wyrazów w segmentach:	DK	ES	FR	NO	PL	PT	SE
0-5	8.93	31.56	26.43	1.62	22.52	27.62	22.66
6-10	27.03	74.83	71.56	17.54	62.46	70.57	64.28
11-20	48.60	94.99	93.14	44.22	87.09	90.61	87.48
21-35	62.07	99.54	98.38	66.91	97.24	97.95	94.54
> 36	66.67	99.84	99.67	85.48	98.38	99.13	97.46

Wraz ze zmniejszeniem ilości wyrazów w modelu słownikowym zmniejsza się, naturalnie, trafność identyfikacji. Zmniejsza się ona także dla zdań z małą liczbą słów. Dla modelu zawierającego 150 słów otrzymujemy całkiem dobre rezultaty dla duńskiego i norweskiego i bardzo dobre rezultaty dla pozostałych języków, już przy segmentach składających się z więcej niż 6 słów. Przewidywalność na podstawie stop listy jest oczywiście znacznie gorsza, szczególnie dla norweskiego i duńskiego.



Rys. 3. Poprawna identyfikacja dla języka polskiego i norweskiego w zależności od wersji modelu słownikowego

Rys. 3 prezentuje przewidywalność jednego z najlepiej zidentyfikowanych języków, języka polskiego, w porównaniu z jednym z najgorzej zidentyfikowanych języków, mianowicie norweskim. Wyniki dotyczą trzech modeli słownikowych. Różnica jest znacząca. Przy modelach językowych o 300 bądź 150 słów dla norweskiego osiągnięto przewidywalność na poziomie, osiągniętym dla języka polskiego korzystającego z modelu o 20 słowach. Różnice między metodami opartymi na 300 i 150 słowach są nieduże, w zakresie zaledwie kilku procent dla krótkich segmentów.

Aby nieco poprawić wyniki identyfikacji, zmodyfikowano klasyfikację. I tak, w pierwszej metodzie dla każdego modelu wyznaczono ilość słów znalezionych w modelu słownikowym na podstawie wszystkich wyrazów zawartych w identyfikowanym segmencie. Dla kolejnej metody oprócz ilości słów, pod uwagę wzięta została częstość występowania znalezionych wyrazów w słowniku, otrzymana na podstawie zbioru treningowego. Suma łącznej ilości słów oraz sumy częstości tworzyła ostateczny wynik. Identyfikacja w tym wypadku polegała na określeniu najwyższego rezultatu. Wyniki zamieszczone zostały w Tabeli 10.

Identyfikacja uległa znacznej poprawie. Zniknęło określanie kilku języków dla badanego segmentu. Dzięki temu przy krótkich zdaniach można osiągnąć lepsze rezultaty. Znacznie poprawiła się identyfikacja dla języka duńskiego i norweskiego.

Przebadano także aspekt ilościowy segmentów, w których nie zostały znalezione żadne słowa pochodzące z któregokolwiek modelu językowego. Stosowanie identycznych wielkości słowników tylko z różnymi rodzajami klasyfikacji nie zmieniało ilości nieokreślonych segmentów. Liczba segmentów niezidentyfikowanych była większa dla krótszych zdań i dla metod opartych na mniejszych słownikach, co widać w Tabeli 11.

Tabela 10. Wyniki procentowe poprawnej identyfikacji z podziałem według długości segmentu przy zastosowaniu słownika o 300, 150 i 20 najczęstszych wyrazach i ulepszonej metodzie klasyfikacji segmentów.

Model z 300 słów	DK	ES	FR	NO	PL	PT	SE
0-5	47.71	77.13	74.46	54.09	69.77	73.39	57.78
6-10	84.53	98.79	97.62	85.00	96.77	96.05	95.09
11-20	93.20	99.74	99.73	93.40	99.61	99.37	99.06
21-35	96.55	99.89	99.93	97.51	99.86	99.83	99.64
> 36	98.30	100.00	100.00	99.67	98.85	99.77	98.87
Model z 150 słów	DK	ES	FR	NO	PL	PT	SE
0-5	41.13	73.66	65.45	48.56	59.78	67.51	51.95
6-10	79.93	98.42	95.77	79.82	92.81	94.05	91.74
11-20	91.37	99.70	99.52	89.77	98.79	98.83	98.33
21-35	96.74	99.93	99.88	95.34	99.75	99.78	99.42
> 36	97.96	100.00	99.93	98.35	99.31	99.77	99.15
Model z 20 słów	DK	ES	FR	NO	PL	PT	SE
0-5	21.07	65.28	33.84	28.45	24.24	49.60	26.91
6-10	56.77	97.72	74.72	57.21	65.95	83.31	71.22
11-20	75.86	99.66	93.90	70.66	88.76	94.37	91.25
21-35	81.27	99.91	98.52	81.30	97.36	98.64	95.71
> 36	80.95	99.92	99.74	90.43	98.61	99.48	97.74

Tabela 11. Wyniki procentowe dla segmentów nie zawierających słów z jakiegokolwiek języka dla modeli o długości 300 i 20 słów.

Model z 300 słów	DK	ES	FR	NO	PL	PT	SE
0-5	27.38	14.63	14.88	26.18	23.97	12.20	27.80
6-10	0.16	0.03	0.14	0.26	0.96	0.06	0.35
11-20	0.00	0.02	0.05	0.01	0.03	0.01	0.01
21-35	0.00	0.00	0.00	0.04	0.00	0.00	0.00
> 36	0.34	0.00	0.00	0.00	0.00	0.06	0.00
Model z 20 słów	DK	ES	FR	NO	PL	PT	SE
0-5	49.76	29.20	36.80	51.52	69.20	27.61	56.40
6-10	3.18	0.88	2.51	4.30	19.61	1.26	5.96
11-20	0.17	0.03	0.20	0.21	2.86	0.08	0.33
21-35	0.14	0.04	0.00	0.12	0.25	0.02	0.19
> 36	0.34	0.00	0.07	0.00	0.46	0.06	0.28

Podstawowym minusem metody słownikowej jest niska przewidywalność dla krótkich segmentów, związana z trudnością ze znalezieniem wyrazów odpowiadających wyrazom ze słowników. Ponadto treści dla stworzenia modelu słownikowego powinny zawierać podobny temat i powinny być pisane podobnym stylem, by zwiększać prawdopodobieństwo identyfikacji. Należy też dysponować znacząco dużym zbiorem tekstów dla danej tematyki, aby móc efektywnie stworzyć model.

Największe błędy wynikają z podobieństwa języków, jak to widać z wyników dla duńskiego i norweskiego. Jednakże przy segmentach zawierających więcej niż 6 słów identyfikacja osiąga ponad 85% dla języka duńskiego i norweskiego oraz ponad 90% dla pozostałych języków, co jest bardzo dobrym rezultatem.

7.4.2. Metoda n-gramowa

Dla identyfikacji za pomocą skryptów Damira Cavara na początek należy stworzyć modele językowe oparte na trigramach oraz na odpowiadającym im częstościach. Na podstawie modeli oraz wybranej metody klasyfikacji (np. naiwnego klasyfikatora Bayesa) zostaje dokonana identyfikacja.

Sporządzenie zbiorów danych może okazać się utrudnione, kiedy nie mamy dostępnych danych do utworzenia zbiorów treningowych. Zbiory te muszą być odpowiednio duże. Niektóre programy korzystają ze stron internetowych i sporządzają potrzebne statystyki, aby ułatwić proces rozpoznania tekstu użytkownikowi. Różnica w stosunku do modelu słownikowego opartego na większej ilości wyrazów niż tylko uwzględniającej stop listy jest taka, iż dane nie muszą dotyczyć tego samego tematu.

Na stronach Wikipedii oraz Wikibooks wyszukano artykuły ręcznie, aby sprawdzić, na jakie problemy lub błędy można się natknąć. Strony te zawierają sporo odnośników oraz powtarzających się treści dotyczących np. nawigacji. Niektóre strony zawierają mało przydatne informacje, mogące wprowadzać dużo szumu. Ponadto, dla niektórych języków można otrzymać dość dużą próbkę raczej szybko i bez trudu, a w innych przypadkach jest to utrudnione. Różnorodność treści jest dużo większa, niż przy danych pochodzących z baz, ze względu na liczbę osób tworzących strony, posługujących się różnymi stylami i dialektami. Dane zostały przygotowane i przebadane na tych samych zbiorach testowych, które były wykorzystane przy metodzie słownikowej.

Tabela 12. Wyniki procentowe poprawnej identyfikacji według długości segmentu dla metody n-gram opartej na zbiorach ze stron internetowych

Wyrazów w segmentach:	DK	ES	FR	NO	PL	PT	SE
0-5	55.58	63.01	68.34	65.27	91.45	53.3	48.77
6-10	67.52	77.52	83.23	75.27	98.35	66.97	63.13
11-20	77.66	85.25	91.15	80.82	99.56	76.59	73.32
21-35	85.14	90.95	95.36	87.54	99.61	85.76	78.65
> 36	84.35	91.62	97.71	90.76	98.61	93.33	82.2

Porównując wyniki z Tabeli 12 z wynikami dla metody słownikowej opartej na modelu 300 słów widać spore pogorszenie przewidywalności dla szwedzkiego, portugalskiego i hiszpańskiego.

Tabela 13. Wyniki procentowe poprawnej identyfikacji według długości segmentu dla metody n-gram opartej na zbiorach z bazy tłumaczeniowej

Wyrazów w segmentach:	DK	ES	FR	NO	PL	PT	SE
0-5	79.20	73.22	87.26	75.11	95.42	71.51	74.15
6-10	92.23	91.41	96.95	85.39	99.56	87.75	87.68
11-20	95.70	96.42	99.31	91.45	99.92	94.60	94.31
21-35	97.44	98.65	99.75	96.50	99.83	97.85	97.37
> 36	94.22	98.94	99.87	99.01	98.85	99.25	99.15

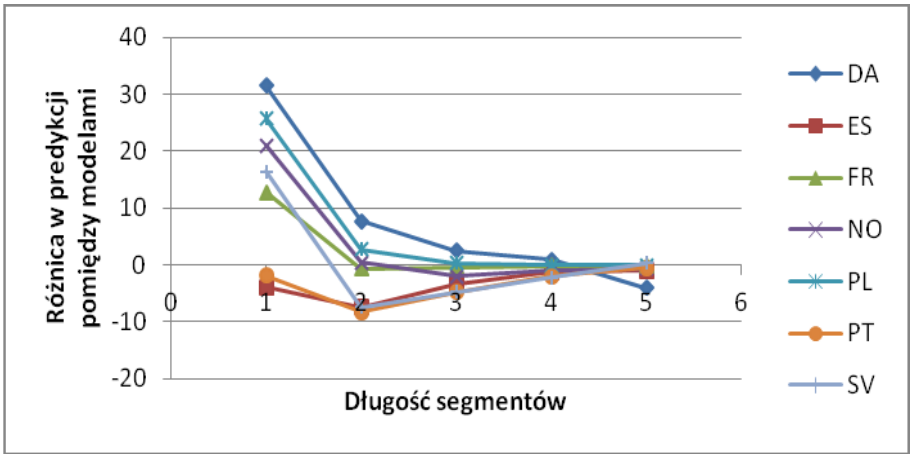
Badanie powtórzono dla statystyk sporządzonych na zbiorach treningowych otrzymanych z baz klienta (Tabela 13). Wyniki są w każdym przypadku lepsze niż dla modeli, które były otrzymane na zbiorach danych pochodzących ze stron internetowych. Poprawa przewidywalności sięga nawet ponad 20% (Tabela 14).

Tabela 14. Różnica poprawnej identyfikacji według długości segmentu między modelami n-gram opartymi na zbiorach z bazy tłumaczeniowej i ze stron internetowych

Wyrazów w segmentach:	DK	ES	FR	NO	PL	PT	SE
0-5	23.62	10.21	18.92	9.84	3.97	18.21	25.38
6-10	24.71	13.89	13.72	10.12	1.21	20.78	24.55
11-20	18.04	11.17	8.16	10.63	0.36	18.01	20.99
21-35	12.3	7.7	4.39	8.96	0.22	12.09	18.72
> 36	9.87	7.32	2.16	8.25	0.24	5.92	16.95

Zaskakująco dobre są wyniki dla języka polskiego dla każdej długości segmentów względem identyfikacji innych języków: ponad 90%. Dane treningowe dla języka polskiego obejmują prawie 9 milionów znaków dla tekstów z bazy tłumaczeniowej. Dla modelu trigram opartego na treściach ze stron internetowych, jak wspomniano, dane wynoszą niespełna jeden milion znaków. Powstaje model trigram, zawierający ponad 13.5 tysiąca trigramów, a dla drugiego modelu ponad 8 tysięcy trigramów. Dla pozostałych języków jest podobnie. Jeśli prawie dwukrotna różnica wielkości bazy modelu miała niewielki wpływ na przewidywalność języka polskiego, można się spodziewać, że podobnie powinno być dla pozostałych języków.

Metoda trigramowa daje znacznie lepsze wyniki dla niektórych języków przy identyfikacji bardzo krótkich segmentów, ale osiąga gorsze rezultaty, niż metoda słownikowa oparta na 300 słowach, dla portugalskiego i hiszpańskiego we wszystkich długościach segmentów (Rys. 4). Na plus dla metody trigramowej należy zaliczyć identyfikację, nawet dla bardzo krótkich segmentów, i to dla wszystkich języków, osiągającą ponad 70%.



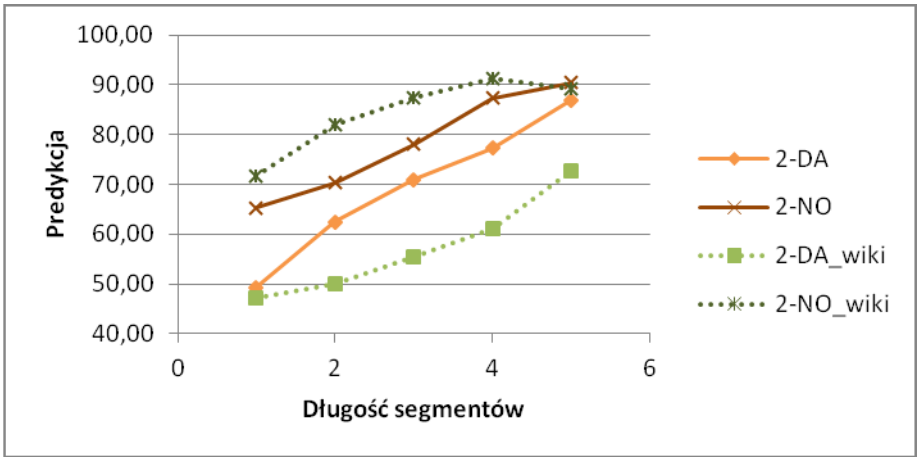
Rys 4. Różnica w identyfikacji metodą trigram opartą na bazach klienta i metody słownikowej z 300 słów

Najpowszechniej używany w bibliotekach jest trigram. Przeprowadzono jednak także badanie dla metody opartej na bigramie oraz unigramie. Przy zmniejszaniu *n* pogarsza się predykcja, zarówno dla modeli stworzonych na podstawie stron internetowych, jak i dla modeli stworzonych na podstawie danych z baz klienta.

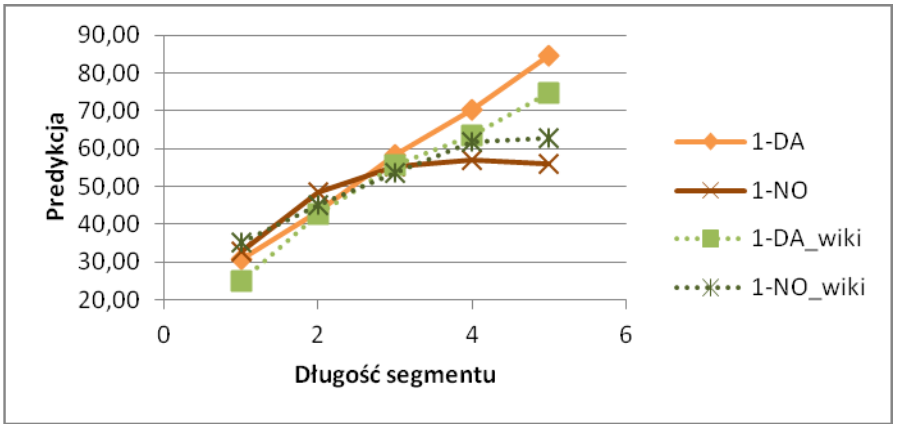
Potwierdza się w uzyskanych wynikach silna zależność duńskiego i norweskiego, znacznie większa niż dla pozostałych języków (Rys. 5 i 6). Wraz ze wzrostem długości segmentów dla każdego języka rośnie poprawna identyfikacja. Jednak dla duńskiego i norweskiego kształt krzywej zmian jest odwrotnie monotoniczny.

Statystyka językowa zależy od dokumentów, na których jest oparta. Różnice statystyk, odpowiadających sporządzonym modelom, wpływają znacząco na wyniki. Wyższa przewidywalność jest osiągana dla takiego badania, które bazuje na modelach opartych na podobnych danych, co dane do testów. Dla niektórych treści należy przeprowadzić operacje, prowadzące do ujednolicenia. Dlatego przebadano również modele, w których nie uwzględniono różnic w wielkości liter. Wyniki niewiele się zmieniły. Różnice wahają się w granicach kilku procent, co okazuje się nieznacznym polepszeniem lub pogorszeniem w odniesieniu do ostatecznych wyników.

Skoro predykcja maleje wraz ze zmniejszeniem n, przeprowadzono jeszcze jedno doświadczenie i zwiększono liczbę liter w modelu o jedną.



Rys. 5. Predykcja dla norweskiego i duńskiego przy zastosowaniu modelu bigram na różnych danych treningowych



Rys. 6. Predykcja dla norweskiego i duńskiego przy zastosowaniu modelu unigram na różnych danych treningowych

Tabela 15. Wyniki procentowe poprawnej identyfikacji z podziałem według długości segmentu dla metody 4-gram opartej na zbiorach z baz klienta.

Wyrazów w segmentach:	DK	ES	FR	NO	PL	PT	SE
0-5	84.38	80.08	91.67	83.07	95.82	79.64	85.98
6-10	94.80	93.26	98.78	92.51	99.59	93.70	95.93
11-20	97.65	97.60	99.79	97.64	99.93	97.93	98.98
21-35	98.97	99.26	99.99	98.83	99.94	99.38	99.45
> 36	98.98	99.92	99.93	100.0	99.08	99.71	99.72

Wyniki są bardzo dobre nawet dla krótkich segmentów (Tabela 15). Nadal są one najlepsze dla polskiego. Bardzo dobrze identyfikowany jest również francuski (ponad 90% bez względu na długość segmentów).

Wraz ze wzrostem liczby *n*, czyli liczby kolejnych sekwencji liter w modelach n-gram, rośnie też czas trwania obliczeń, poniekąd przez znaczny wzrost objętości danych dla modeli językowych.

8. Podsumowanie

8.1. Najlepsze rozwiązania

Identyfikacja języka dynamicznie rozwijała się w latach 90-tych XX wieku. Obecnie można niedużym kosztem osiągnąć bardzo dobre rezultaty. Dla biura tłumaczeniowego, w którym większość segmentów opiera się na krótkich zdaniach, lepsze okazują się do rozpatrywanego tutaj celu modele oparte na n-gramach. Metoda n-gramowa daje bowiem lepsze wyniki dla krótszych segmentów i tylko nieco gorsze wyniki dla dłuższych treści niż metoda słownikowa. Jej wyniki są zrównoważone dla każdej długości segmentów i nieco tylko rosnące dla dłuższych treści.

W biurze lokalizacji językowej nie powinno być problemu z danymi do stworzenia modelu językowego. Według przeprowadzonych badań, lepsze wyniki dają modele oparte na treściach o podobnej tematyce. Wadą rozpatrywanych modeli jest trudność automatycznego wykluczenia danych, prowadzących do zaszumienia: adresów mailowych, stron, nazw firm, nazw modeli, wielkości fizycznych itp.

Projektów w biurach tłumaczeniowych może być bardzo dużo. Przy tym, każdy projekt oparty jest na innej stawce i muszą być spełnione inne warunki co do jakości. Jedyną wspólną cechą dla metody powinna być prostota w implementacji i łatwość w korzystaniu, która nie wydłuża w znaczący sposób procesu tłumaczenia.

Jeżeli projekt wymaga wyższej jakości, można zastanowić się nad innymi rozwiązaniami np. połączeniem metody słownikowej, która przejmowałaby kontrolę nad treściami dłuższymi z metodą n-gramową dla segmentów bardzo krótkich. Ewentualnie może zamiast sporządzać modele, oparte na n-gramch lub na słownikach, można stworzyć za pomocą reguł asocjacyjnych silne związki identyfikacyjne.

Wyniki dla krótkich segmentów sięgają w powyższych badaniach do 80% przy użyciu dobrych modeli. Przy plikach z tłumaczenia można by określić pewien próg różnicy między oczekiwanym językiem podanym w treści pliku, a najlepszym dopasowaniem, aby liczba segmentów do sprawdzenia została nieco obniżona. Problem może się pojawić przy językach bardzo podobnych, jak duński i norweski.

8.2. Cele na przyszłość

W biurach tłumaczeniowych niektóre projekty zawierają różne odmiany tego samego języka (np. tłumaczenia na: francuski europejski i kanadyjski, lub hiszpański europejski i meksykański). Ponieważ osoby posługujące się tymi

językami potrafią zrozumieć siebie w znacznym stopniu, w większości systemów te różnice są zaniedbywane. W biurze zajmującym się lokalizacją powinna istnieć możliwość wyłapywania takich sytuacji, dlatego należałoby przetestować metodę opierającą się na już powstałych pomysłach lub próbować znaleźć inne rozwiązania.

W badaniach tu opisanych skupiono się na identyfikacji języków pisanych pismem łacińskim. Metody słownikowej nie możemy stosować do pisma logograficznego. Mimo, iż języki pisane pismem łacińskim są bardziej powszechne i tłumaczenia na nie w Europie mają większe znaczenie, należałoby przetestować metody dla języków, nie korzystających z systemu alfabetycznego, lub z alfabetu łacińskiego. W tłumaczeniach, np. na język arabski lub chiński, część nazw jest zapisywanych pismem łacińskim. Jednym ze sposobów identyfikacji jest wówczas usunięcie jednego z pism, które stanowi mniej niż 20% znaków. Ciekawym studium może być badanie udziałów znaków w bardzo krótkich segmentach.

Bibliografia

Baldwin T., Liu M. (2010) *Language Identification*: *The Long and the Short of the Matter*, University of Melbourne.

Beesley K. R. (1988) *Language identifier: A Computer Program for Automatic Natural-Language Identification of On-line Text*, a.l.p. Systems, October 1988.

Calix K., Connors M., Levy D., Manzar, H., McCabe, G. and We, S. (2008) *Stylometry for E-mail Author Identification and Authentication* Proc. of CSIS Research Day, Pace University, May 2008.

Campbell W.M., Singer E., Torres-Carrasquillo P.A., Reynolds D.A. (2004) *Language Recognition with Support Vector Machines,* MIT Lincoln Laboratory.

Cavnar W. B., Trenkle J. M. (1994) *N-Gram_Based Text Categorization*, Environmental Research Institute of Michigan.

Cowie J., Ludovik Y., Zacharski R. (1999) *Language Recognition for Mono- and Multilingual Documents*, Proceedings of the Vextal Conference, 209-214. Venice, November 22-24, 209-214.

Dunning T. (1994) *Statistical Language Identification,* Computing Research Laboratory, New Mexico State University.

Gerritsen C.M. (2003) *Authorship Attribution Using Lexical Attraction.* MIT, Dept. of Electrical Engineering and Computer Science, M. Eng. and BS Thesis.

Grefestett G. (1995) *Comparing two language identification schemes*, JADT, December 1995.

Han Bo, Baldwin T. (2010) *Lexical Normalisation of Short Text Messages: Makn Sens a twitter,* NICTA Victoria Research Laboratory, The University of Melbourne.

http://www.cavar.me/damir/LID/pyfiles/ strona zawierająca skrypt do identyfikacji języka na podstawie tri-gramu, stworzony przez Damir Cavar'a

http://www.ranks.nl/resources/stopwords.html strona holenderskiej firmy *1ste Keuze B.V*.

http://www.slideshare.net/shuyo/short-text-language-detection-with-infinitygram-12949447

Hughes B., Baldwin T., Bird S., Nicholson J., MacKinlay A. (2006) *Reconsidering Language Identification for Written Language Resources*, University of Melbourne.

Kowalska M. (2014) Wspomaganie procesu lokalizacji językowej oprogramowania – detekcja języka. Praca dyplomowa magisterska, WSISiZ.

Kuncheva L. I. (2004) *Combining Pattern Classifiers, Methods and Algorithms*, John Wiley & Sons, Inc., Hoboken, New Jersey.

McCallum A., Nigam K. (1998) *A Comparison of Event Models for Naive Bayes Text*

Classification, Pittsburgh. W: *AAAI/ICML-98 Workshop on Learning for Text Categorization*, pp. 41-48. Technical Report WS-98-05. AAAI Press. 1998.

Mitchell T. (1997) *Machine Learning*, McGraw Hill.

Narayanan A., Paskov H., Zhenqiang Gong N., Bethencourt J., Stefanov E., Chul E., Shin R., Song D. (2012) *On the Feasibility of Internet-Scale Author Identification*. Symposium on Security and Privacy. IEEE Computer Society. DOI 10.1109/SP.2012.46

Owsiński J. W., Zadrożny, S. (2003) *Metody i systemy wyszukiwania informacji tekstowej,* Wykłądy, http://wit.edu.pl

Owsiński J. W. (2014) *Wprowadzenie do wyszukiwania informacji tekstowych: modele, techniki, zasadnicze zagadnienia*. WSISiZ, Warszawa.

Poutsma A. (2001) *Applying Monte Carlo Techniques to Language Identification,* SmartHeaven, Amsterdam.

Ryczaj W., Owsiński J. W. (2015) Im więcej, tym lepiej? O pewnej analizie z dziedziny wyszukiwania informacji tekstowej. *Zeszyty Naukowe Wydziału Informatycznych Technik Zarządzania WSISiZ „Współczesne Problemy Zarządzania"*, 1/2015, 81-124.

Shuyo N. (2012) *Short Text Language Detection with Infinity-Gram,* NARA Institute of Science and Technology, 14 May 2012.

Sibun P., Reynar J. C. (1996) *Language Identification: Examining the Issues,* The Institute for the Learning Science.

Zampieri M. (2013) *Using Bag-of-words to Distinguish Similar Languages: How Efficient are They?* Saarland University.

SUPPORTING LANGUAGE LOCALIZATION OF DOCUMENTS

Abstract: The present work refers to the processes concerning the language localization offices. The appropriately constructed, used and constantly updated tools support and facilitate the quality of translations. The aim of the work is to focus on the effectiveness of the various techniques for language recognition, and, as a result, on further improvement of the quality of translations. In the study, the relationships between identity recognition methods and language models are presented.