dr inż. Mirosław Gajer
Akademia Górniczo-Hutnicza w Krakowie
Katedra Automatyki

INTERAKTYWNY SYSTEM TRANSLACJI AUTOMATYCZNEJ Z JĘZYKA SZWEDZKIEGO NA POLSKI

Translacja automatyczna jest dziedziną zastosowań informatyki, której celem jest opracowanie metod tworzenia programów komputerowych zdolnych do automatycznego generowania przekładów tekstów z jednego języka naturalnego na drugi. Niestety opracowanie uniwersalnego programu translacji automatycznej, który tłumaczyłby dowolne teksty, w sposób całkowicie automatyczny, jest wciąż problemem otwartym. Zdając sobie sprawę z możliwości i ograniczeń systemów translacji automatycznej autor stworzył program interaktywny, który dzięki pomocy człowieka tworzącego wypowiedzi w języku źródłowym przekładu pozwala na uzyskanie poprawnych gramatycznie i semantycznie przekładów. W artykule omówiono zaproponowane przez autora podejście na przykładzie systemu tłumaczącego z języka szwedzkiego na polski. Rozważany system został oparty na technice translacji automatycznej opartej na wzorcach, która została dodatkowo zmodyfikowana przez autora pod kątem mechanizmów gramatyki języka polskiego.

THE INTERACTIVE SWEDISH-TO-POLISH MACHINE TRANSLATION SYSTEM

Machine translation is a field of applied computer science the aim of which is to develop the methods that would allow us to write computer programs which were be able to translate in an automatic way texts written in one natural language into the other. However, writing a machine translation program that would be able to translate general texts in a fully automatic way, and which would deliver the perfect translation of these texts is still an open problem. Taking into account all the possibilities and restrictions of machine translation systems this author has created an interactive machine translation program, which can generate syntactically and semantically correct translation of input source language texts, while being aided by a human. In the paper the Swedish-to-Polish interactive machine translation system is described. The system is based on a pattern-based machine translation technique that was adjusted by this author especially for the grammatical structure of the Polish language.

1. WSTĘP

Translacja automatyczna jest dziedziną informatyki zajmującą się wypracowaniem metod pisania programów komputerowych, które byłyby w stanie dokonywać w sposób automatyczny przekładów pomiędzy językami naturalnymi (np. pomiędzy polskim i francuskim) [1].

Początków translacji automatycznej można doszukiwać się aż w XVII w., kiedy to wielcy matematycy i uczeni tej epoki podejmowali próby stworzenia za pomocą symboli matematycznych pewnych uniwersalnych dla całej ludzkości języków. Z kolei w 1933 r. w Moskwie uczony polskiego pochodzenia Piotr Trojański opatentował urządzenie nazwane przez niego uniwersalną maszyną przekładową. Maszyna ta miała być pewnym rodzajem mechanicznego słownika, ale niestety z braku środków i większego zainteresowania w świecie nauki nigdy nie doczekała się praktycznej realizacji. W 1946 r. zaraz po zbudowaniu pierwszego komputera zaczęto zastanawiać się na możliwościami innych niż numeryczne zastosowań maszyn cyfrowych. Z propozycją zastosowania cyfrowych komputerów do automatyzacji przekładu jako pierwszy wystąpił w 1947 r. Amerykanin Warren Weaver. W 1952 r. w Stanach Zjednoczonych na MIT została powołana pierwsza grupa badawcza, której głównym zadaniem było dokonanie analizy możliwości zastosowania komputerów do mechanizacji przekładu. Prace tej grupy zaowocowały w 1954 r. pierwszym publicznym pokazem działającego systemu translacji automatycznej. Podczas tego pokazu komputer przetłumaczył 49 prostych rosyjskich zdań na język angielski. Po tym wydarzeniu w świecie nauki zapanował spory entuzjazm i w związku z pierwszymi sukcesami wieszczono rychłe uporanie się z zagadnieniem automatyzacji przekładu i całkowite wyeliminowanie zawodu tłumacza w ciągu najbliższych 20 lat. Niestety szybko okazało się jak bardzo się mylono i równie szybko uczeni uświadomili sobie, jak niezwykle złożonym zagadnieniem jest automatyzacja przekładu oraz jak liczne przeszkody o charakterze fundamentalnym należy pokonać, aby osiągnąć zamierzony sukces. Wydanie w 1966 r. przez Amerykańską Akademię nauk tzw. raportu ALPAC praktycznie położyło kres badaniom nad translacją automatyczną na dobre kilkanaście lat. Renesans nastąpił dopiero pod koniec lat 70-tych, kiedy to na rynku pojawiły się komputery o dużej wydajności obliczeniowej oraz o odpowiednio pojemnych pamięciach, które pozwoliły na dokonanie implementacji pierwszych systemów translacji automatycznej ukierunkowanych na praktyczne zastosowania związane z tłumaczeniem dokumentacji technicznej, tekstów medycznych, komunikatów meteorologicznych itp.

Obecnie badania nad translacja automatyczną nabierają szczególnej wagi, w związku ze zjawiskiem globalizacji światowej gospodarki, rozrastaniem się wielonarodowych tworów politycznych (Unia Europejska), a także z ekspansją Internetu (automatyczne tłumaczenie stron www). Oczekuje się, że w przyszłości zapotrzebowanie na wszelkiego rodzaju tłumaczenia może bardzo gwałtownie wzrosnąć i tłumacze nie będą mogli podołać nadmiarowi pracy. Wszystko to sprawia, że obecnie w translacji automatycznej zdecydowaną przewagę uzyskał nurt badań praktycznych nad badaniami teoretycznymi. Zatem badacze działający w obszarze translacji automatycznej zajmują się raczej pisaniem konkretnych programów komputerowych tłumaczących pomiędzy wybranymi językami, niż snuciem jałowych w swej istocie, ściśle teoretycznych rozważań nad tym co będzie, a co nie będzie możliwe w przeszłości (zresztą prawdopodobnie wszystko na ten temat napisano już w latach 60.).

2. TRANSLACJA OPARTA NA PRZYKŁADACH

W latach 90. powstało kilka nowych metod translacji automatycznej, stanowiących alternatywę dla uprzednio stosowanych metod opartych głównie na gramatykach i językach formalnych. Szczególnie obiecująco zapowiada się metoda automatycznego tłumaczenia oparta na przykładach EBMT (ang. Example-Based Machine Translation). Istotę tej metody stanowi to, iż nie są tłumaczone pojedyncze wyrazy, lecz frazy składające się z kilku wyrazów. W tym wypadku system translacji automatycznej współpracuje z elektronicznym leksykonem fraz lub korzysta z bilingwicznego korpusu tekstów, w którym w sposób automatyczny wyszukuje przekłady fraz języka źródłowego. A zatem działanie systemu polega po prostu na dokonywaniu podmiany fraz tekstu źródłowego ich odpowiednikami w języku docelowym przekładu. Jakość tłumaczenia systemów EBMT jest o wiele wyższa niż w przypadku innych stosowanych dotychczas metod, a uzyskane przekłady tekstów brzmią o wiele bardziej naturalnie i bardziej przypominają dzieło człowieka niż bezrozumnej maszyny [2]. Niestety poważną wadą systemów EBMT jest konieczność wyposażenia ich w bazę danych o potężnych rozmiarach, która byłaby w stanie zawrzeć odpowiednio dużą liczbę przykładów translacyjnych. Również zdobycie bilingwicznych korpusów tekstów o odpowiednio dużej objętości, w których każdemu zdaniu w języku źródłowym odpowiada dokładnie jedno zdanie w języku docelowym nie jest sprawą łatwą. Ponadto metoda EBMT okazuje się zawodna w przypadku języków z bogatą fleksją lub aglutynacją, ponieważ w tym wypadku rozmiary baz danych rosną dramatycznie.

Mając na uwadze powyższe ograniczenia autor zaproponował pewną modyfikację metody EBMT, którą nazwał translacją opartą na wzorcach PBMT (ang. Pattern-Based Machine Translation).

3. WZORCE TRANSLACYJNE

Metoda translacji PBMT oparta jest na następujących założeniach. Wypowiedzi formułowane w języku naturalnym budowane są ze zdań. Zdania złożone zawsze składają się z pewnej liczby zdań prostych. Każde zdanie może być zbudowane z frazy stanowiącej jego podmiot (np. „gruby chłopiec"), frazy będącej jego orzeczeniem (np. „zjadł"), frazy pełniącej funkcje dopełnienia bliższego (np. „tłustą kiełbasę") oraz pewnej liczby fraz pełniących funkcje dopełnień dalszych (np. „swojemu bratu") bądź okoliczników odpowiadających na pytania: jak?, gdzie?, kiedy?, po co?, na co?, dlaczego?, w jakim celu?, jakim sposobem? (np. „szybko", „w poniedziałek", „w kuchni", „na śniadanie") itp. Rozważane przez autora wzorce translacyjne są właśnie takim frazami języka źródłowego, pełniącymi w zdaniach określone funkcje. Ponadto wzorce translacyjne zawierają przekłady tych fraz na język docelowy, uwzględniające wszelkie możliwe do utworzenia ich formy fleksyjne. Istnieją trzy typy wzorców translacyjnych: wzorce typu rzeczownikowego (podlegają deklinacji), wzorce typu czasownikowego (podlegają koniugacji) oraz wzorce afleksyjne.

Przykładowe rekordy z bazy danych wzorców typu rzeczownikowego przedstawiono na rys. 1. W wierszu „source" umieszczone są kolejne frazy rzeczownikowe należące do języka źródłowego (w tym wypadku Szwedzkiego). Użytkownik programu dokonuje za pomocą specjalnego menu wyboru spośród fraz rzeczownikowych, tych które pełnią w tworzonym przez niego zdaniu funkcję podmiotu lub dopełnienia bliższego. Kolejny wiersz tabeli – zatytułowany „comments" – zawiera pewne wyjaśnienia, komentarze dotyczące frazy zawartej w wierszu „source". Na przykład na rys. 1. szwedzka fraza „en

glas" pojawia się dwukrotnie, ale pola „comments" tabeli są za każdym razem różne. W pierwszym wypadku pojawia się tam słowo „dricksglas", a w drugim „snapsglas", które objaśniają znaczenie frazy „ett glas". To właśnie użytkownik programu kierując się informacją zawartą w polu „comments" musi wybrać tę jednostkę leksykalną, o którą mu akurat chodzi w tworzonym przez niego zdaniu.

source	ett glas	ett glas	goda vänner
comments	dricksglas	snapsglas	
case1	szklanka	kieliszek	dobrzy przyjaciele
case2	szklanki	kieliszka	dobrych przyjaciół
case3	szklance	kieliszkowi	dobrym przyjaciołom
case4	szklankę	kieliszek	dobrych przyjaciół
case5	szklanką	kieliszkiem	dobrymi przyjaciółmi
case6	szklance	kieliszku	dobrych przyjaciołach
numgen	12	11	21

Rys. 1. Przykładowe wzorce translacyjne typu rzeczownikowego

Pole „comments" może niekiedy być puste. Taka sytuacja ma miejsce w przypadku frazy „goda vänner", której semantyka jest jednoznaczna i w związku z tym nie są potrzebne żadne dodatkowe objaśnienia. W kolejnych wierszach od „case1" do „case6" znajdują się przekłady szwedzkich fraz źródłowych na język polski. Uwzględnione zostały kolejno formy fleksyjne mianownika, dopełniacza, celownika, biernika, narzędnika i miejscownika. Nie uwzględniono tutaj formy fleksyjnej wołacza, który jest przypadkiem niezależnym (nie wchodzi w związek zgody z orzeczeniem zdania) i w związku z tym tłumaczony jest za pomocą wzorców afleksyjnych.
W ostatnim wierszu zamieszczono wartości specjalnego atrybutu „numgen", który to atrybut zawiera pełną informację o liczbie i rodzaju gramatycznym fraz polskich stanowiących przekład frazy szwedzkiej. Na przykład na rys. 1 wystąpiła fraza „dobrzy przyjaciela" i ponieważ fraz ta występuje w liczbie mnogiej i jest rodzaju męskiego, atrybutowi „numgen" została przypisana wartość 21. Pierwsza cyfra atrybutu „numgen" oznacza zawsze liczbę i przyjmuje wartość 1 dla liczby pojedynczej i 2 dla liczby mnogiej. Z kolei druga cyfra oznacza rodzaj gramatyczny (1 – męski, 2 – żeński i 3 – nijaki). Zatem atrybut „numgen" może zawsze przybierać tylko jedną z 6 możliwych wartość: 11, 12, 13, 21, 22 i 23.
Kolejnym typem wzorca translacyjnego są wzorce zawierające frazy czasownikowe, które pełnią w tłumaczonych zdaniach funkcję orzeczenia. Pole „source" zawiera frazę czasownikową w języku źródłowym, czyli w tym przypadku szwedzkim. Funkcja spełniana przez pole „comments" jest identyczna, jak w przypadku omówionych uprzednio wzorców rzeczownikowych, czyli jej zadaniem jest rozwianie wszelkich wątpliwości użytkownika odnośnie semantyki frazy źródłowej.
Kolejne wiersze od „numgen11" do „numgen23" zawierają wszelkie możliwe do utworzenia formy fleksyjne polskiego przekładu szwedzkiej frazy źródłowej. Z kolei występowanie pola związanego z atrybutem „case" związane jest ze związkiem rządu jaki zachodzi w języku polskim pomiędzy orzeczeniem a dopełnieniem zdania. Po prostu wystąpienie danej frazy czasownikowej wymusza użycia jej dopełnienia w odpowied-

nim przypadku gramatycznym i właśnie pole „case" zawiera informację o przypadku w jakim należy użyć dopełnienia tłumaczonego zdania.

source	har drukkit	åker till	åkte hem
comments			
numgen11	wypił	jedzie do	pojechał do domu
numgen12	wypiła	jedzie do	pojechała do domu
numgen13	wypiło	jedzie do	pojechało do domu
numgen21	wypili	jadą do	pojechali do domu
numgen22	wypiły	jadą do	pojechały do domu
numgen23	wypiły	jadą do	pojechały do domu
case	4	2	1

Rys. 2. Przykładowe wzorce translacyjne typu czasownikowego

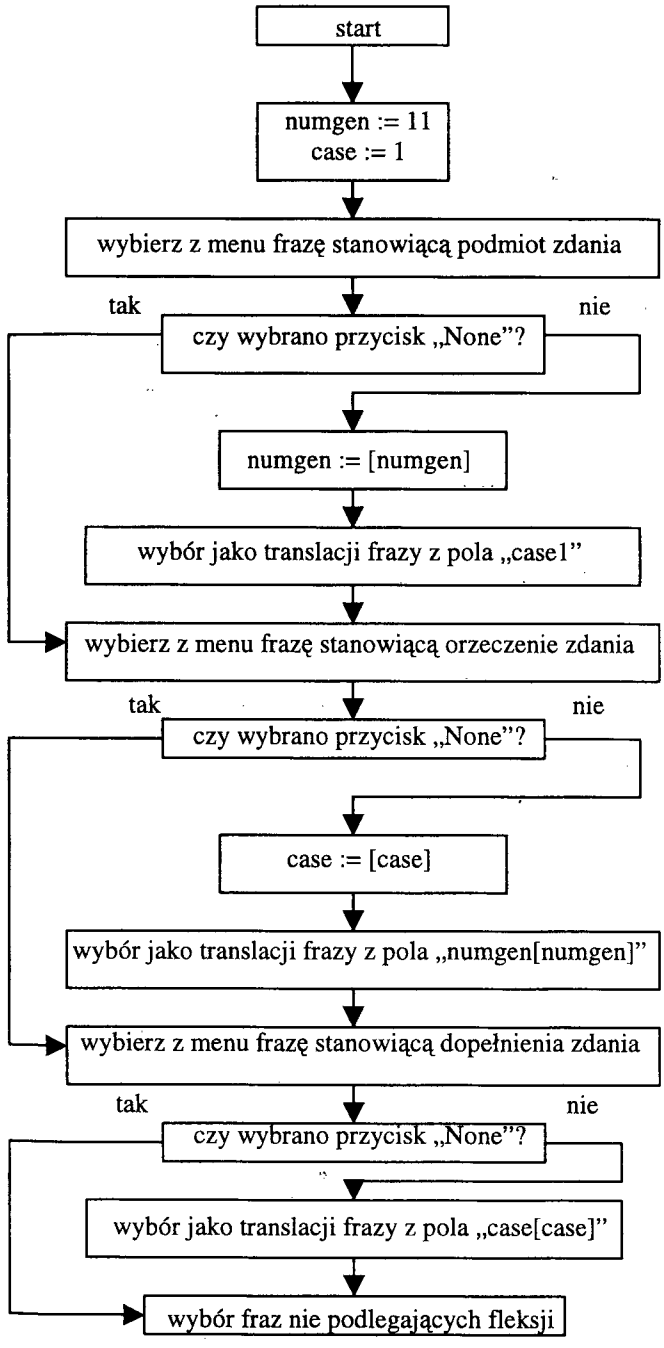
Ostatnim typem wzorca translacyjnego są wzorce nie podlegające fleksji. Przykład takich wzorców został zamieszczony na rys. 3. Baza danych zawierająca wzorce afleksyjne składa się jedynie z trzech pól rekordów: „source", „comments" i „target". Funkcje pełnione przez dwa pierwsze z nich są identyczne jak w opisanych uprzednio przypadkach, natomiast pole „target" zawiera tłumaczenie frazy wejściowej zawartej w polu „source" na język docelowy, czyli w rozważanym wypadku polski.

source	den här veckan	i arbete	i norra Sverige
comments			
target	w tym tygodniu	w pracy	w północnej Szwecji

Rys. 3. Przykładowe wzorce translacyjne typu afleksyjnego

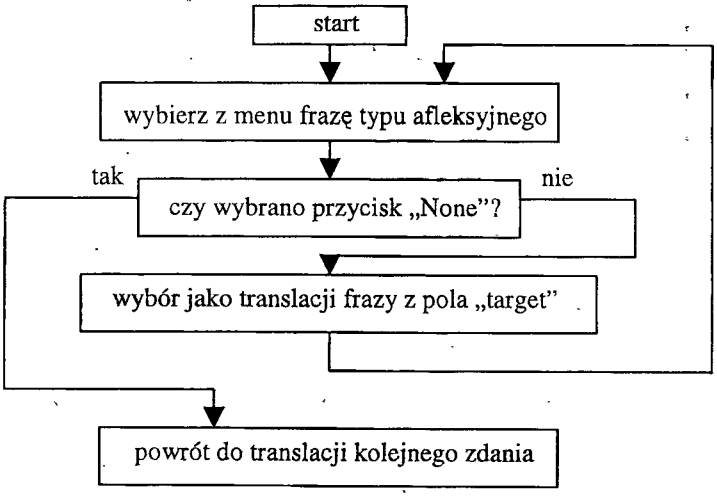
4. IMPLEMENTACJA SYSTEMU

Rozważany interaktywny system translacji automatycznej został zaimplementowany w środowisku Visual Studio 6.0 dla platformy Windows. Zasadę działania programu translacji automatycznej ilustruje zamieszczony na rys. 4 algorytm. Jego zasada działania jest następująca. Zawsze przed przystąpieniem do translacji nowego zdania atrybutowi „numgen" zostaje przypisana wartość 11, a atrybutowi „case" zostaje nadana wartość 1 (podmiot zdania występuje zawsze w mianowniku). Następnie program prosi użytkownika o dokonanie wyboru frazy stanowiącej podmiot tłumaczonego zdania. Wyboru tego użytkownik dokonuje za pomocą specjalnego menu, gdzie w odpowiednim okienku wyświetlane kolejne frazy rzeczownikowe. Ewentualnie użytkownik może wybrać przycisk „None", co będzie oznaczało, że rezygnuje z wyboru frazy stanowiącej podmiot zdania. Po wybraniu frazy pełniącej funkcje podmiotu atrybutowi „numgen" zostaje przypisana nowa wartość, która zostaje odczytana z bazy wzorców translacyjnych. Z kolei jako przekład frazy w języku źródłowym zostaje wybrana pozycja z pola „case1", ponieważ uprzednio atrybutowi „case" została właśnie przypisana wartość 1.



Rys. 4. Algorytm interaktywnego programu translacji automatycznej.

Postępując podobnie użytkownik wybiera z menu frazę stanowiącą orzeczenie tłumaczonego zdania. W tym przypadku jako przekład wybierana jest fraza języka polskiego zapisana w polu „numgen[numgen]" o numerze takim, jaki został uprzednio nadany atrybutowi „numgen" podczas translacji frazy stanowiącej podmiot zdania. Następnie atrybutowi „case" zostaje przypisana nowa wartość, dzięki czemu dopełnienie zdania będzie mogło zostać użyte we właściwym przypadku (związek rządu). W kroku kolejnym dokonywana jest translacja frazy stanowiącej dopełnienie zdania. Jako przekład frazy źródłowej wybierana jest polska fraza zawarta w polu „case[case]" o numerze takim, jaki został przypisany atrybutowi „case" podczas tłumaczenia frazy stanowiącej orzeczenie zdania. Na koniec użytkownik ma możliwość dokonania wyboru dowolnej liczby fraz typu afleksyjnego. Algorytm postępowania podczas wyboru tych fraz został przedstawiony na rys. 5.



Rys. 5. Algorytm translacji fraz nie podlegających fleksji

5. ZAKOŃCZENIE

W artykule opisano system translacji automatycznej z języka szwedzkiego na polski. Zaproponowane podejście do translacji automatycznej jest podejściem nowatorskim, z co najmniej dwóch powodów.

Po pierwsze zastosowano w nim opracowaną przez autora metodę wzorców translacyjnych. Metoda ta wywodzi się ze znanych uprzednio podejść opartych na tłumaczeniu poprzez analogię i tzw. przykłady translacyjne, ale w odróżnieniu od nich pozwala na uwzględnienie niezwykle istotnych zjawisk językowych związanych z fleksją języka polskiego. Bez tej modyfikacji, bezpośrednie zastosowanie metody tłumaczenia opartej na przykładach dla języka polskiego nie dałoby zadowalających rezultatów, ze względu na olbrzymią mnogość form wyrazowych w języku polskim.

Po drugie omówiony w artykule system translacji automatycznej jest systemem interaktywnym. Jest to również nowość w stosunku do spotykanych dotychczas rozwiązań, w

których na wejście systemu translacji automatycznej podawany jest pewien gotowy tekst zapisany w języku źródłowym, natomiast na wyjściu otrzymuje się przekład tego tekstu na język docelowy. Niestety jakość tego przekładu jest w zdecydowanej większości przypadków niezadowalająca. Dzieje się tak dlatego, że analizator składniowy programu translacji automatycznej nie radzi sobie po prostu ze zdaniami o zawiłej budowie gramatycznej, których rozbioru na części składowe można często dokonać na kilka alternatywnych sposobów. Ponadto trudnością praktycznie nie do pokonania dla klasycznych systemów translacji automatycznej jest wieloznaczność wyrazów, które w języku przekładu mogą mieć nawet kilkanaście ekwiwalentów o zupełnie różnych znaczeniach (np. przetłumaczenie polskiego słowa „pokój" na język angielski słowem „room", zamiast „peace" poskutkuje prawdopodobnie zupełnym brakiem zrozumienia przekładu przez angielskojęzycznego odbiorcę).

W omówionym w artykule interaktywnym systemie translacji automatycznej sprawy mają się jednakże inaczej. Tutaj sam użytkownik systemu wprowadzając do niego tekst źródłowy decyduje o tym, jaki fragment zdania pełni w nim funkcję podmiotu, orzeczenia, dopełnienia bądź okolicznika. Podejście takie uwalnia komputer od niezwykle trudnego zadania analizy tekstu zapisanego w języku naturalnym (istnieje niezwykle małe prawdopodobieństwo, aby komputer był w stanie przeprowadzić taką analizę poprawnie). Ponadto użytkownik, korzystając z pola „comments" sam wybiera, z pośród kilku możliwych, właściwe znaczenie tłumaczonego wyrazu bądź frazy (jest to rzecz w zasadzie niewykonalna dla komputera [3]).

Na zakończenie autor pragnąłby jeszcze zwrócić uwagę na fakt, iż zaprezentowany w artykule system translacji automatycznej z języka szwedzkiego na polski jest prawdopodobnie (zgodnie z posiadaną przez autora wiedzą) pierwszym tego typu systemem, jaki został w ogóle zbudowany, a zatem ma on całkowicie pionierski charakter. Dalsze prace autora koncentrować się będą na implementacji omówionej w artykule metody translacji automatycznej dla innych ważnych oficjalnych języków używanych w Unii Europejskiej. Ponadto system zostanie zaimplementowany w postaci serwera www i z opisanego w artykule programu już wkrótce będzie mógł skorzystać każdy za pośrednictwem Internetu.

W dalszej perspektywie planowana jest również implementacja systemu dokonującego przekładu w kierunku przeciwnym, tzn. z języka polskiego na język szwedzki. Przewiduje się także integrację obu programów translacji automatycznej z programem typu „chat". Powstanie zatem system komunikacyjny, który pozwoli na bezpośrednie prowadzenie dialogu dwóm osobom, z których jedna posługuje się językiem polskim, a druga szwedzkim.

6. LITERATURA

[1] Arnold D. Balkan L., Meijer S., Humphreys R., Sadler L.: *Machine translation: An introductory guide*, NCC Blackwell, London, 1994

[2] Whitelock P., Kilby K.: *Linguistic and computational techniques in machine translation system design*, UCL Press, London, 1995

[3] Melby A.: *Machine translation and philosophy of language*, Machine Translation Review, No. 9, April 1999, pp. 6-17