INVESTIGATIONES LINGUISTICAE VOL. XXXIV, 2016

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

Wyznaczanie granicy zdania – człowiek vs maszyna[1]

Establishing Sentence Boundary – man vs machine

Michał Lipnicki

POZNAŃSKIE CENTRUM SUPERKOMPUTEROWO-SIECIOWE, PAN POZNAŃ
UL. JANA PAWŁA II 10, 61-139 POZNAŃ

mlipnicki@man.poznan.pl

Abstract

The paper aims at presenting the results of an experiment checking the possibility of an accurate reconstruction of text sentence boundary done by humans and computer programme with possible application in presenting the output of automatic speech recognition systems. The results are compare with the assesment of received sentences acceptability.

1. Wprowadzenie

Podział ciągłego tekstu – pozbawionego wszelkich graficznych operatorów metatekstowych (tj. znaków interpunkcyjnych oraz rozróżnienia na małe i wielkie litery), będącego rezultatem pracy systemów automatycznego rozpoznawania mowy (ASR) jest znanym, choć ciągle nieposiadającym zadowalającego rozwiązania problemem. Wyznaczenie w takim tekście fraz, a najlepiej poprawnych gramatycznie zdań nie jest wyłącznie zabiegiem estetycznym mającym na celu poprawienie jakości jego prezentacji, lecz również ułatwia czytanie oraz poprawia przetwarzanie zawartych w nim informacji. Oczywiście, zagadnienie to jest w pewnym stopniu powiązane z problemem ekstrakcji zdań z tekstów pisanych zawierających interpunkcję. Jednak o ile w przypadku takich systemów ich skuteczność często zależy od odpowiedniej interpretacji danego znaku przystankowego, czy też np. rozpoznawania skrótowców w tekście (por. Kiss, Strunk: 2006), o tyle w przypadku efektów rozpoznania mowy, należy przede wszystkim uwzględnić wskaźniki akustyczne (por m. in.: Huang,

[1] Badania wykonano w ramach projektu O ROB 0008 03 00 finansowanego przez NCBIR.

Zwaig, 2002; Baranowska, Francuzik, Karpiński, Kleśta: 2003) z możliwym wsparciem uzyskanych wyników przez analizy syntaktyczne i/lub statystyczne. Ponadto podział ciągłego tekstu jest również niezwykle istotny dla programów korzystających i dalej przetwarzających rezultaty pracy systemów ASR takich jak systemy wyszukiwania informacji (ang. *information retrieval;* por. Song, Anh, Kim: 2014), czy tłumaczenia automatyczne (por. Matusov, Mauser, Ney: 2006; Rao, Lane, Schultz: 2007).
Skupienie się na analizie akustycznej oczywiście wynika z tego, że z samej natury systemów ASR materiałem jest tu język mówiony (mniej lub bardziej spontaniczny). Ze względów na jego specyfikę – liczne nieciągłości, powtórzenia, elipsy itd. same tylko analizy statystyczne i składniowe, tj. nieuwzględniające parametrów akustycznych, mają dosyć ograniczone zastosowanie i nie dostarczają oczekiwanych rezultatów. Co więcej, możliwość poprawnej rekonstrukcji podziału na frazy jest w dużej mierze zależna właśnie od stopnia nieciągłości analizowanej wypowiedzi (por. Liu, Shriberg, Stolcke, Hillard, Ostendorf, Harper 2006). Pomimo tego analizy składniowe i statystyczne służą jako istotne wsparcie dla analizy akustycznej, pozwalające rozstrzygać wieloznaczności w proponowanym podziale i poprawiać adekwatność wskazań.

Mając to na uwadze, w niniejszej pracy przedstawimy wyniki badań empirycznych pokazujących, jakie są praktyczne możliwości rekonstrukcji podziału na zdania gotowych tekstów na podstawie testów wykonanych przez rodzimych użytkowników języka polskiego. Wyniki te były następnie porównywane z rezultatami pracy autorskiego parsera. Zdania będące rezultatem pracy programu zostały ponadto poddane ocenie pod kątem ich akceptowalności. W efekcie mogliśmy wyciągnąć dwa rodzaje wniosków: (a) dotyczących praktycznej możliwości rekonstrukcji interpunkcji zgodnej z intencjami nadawcy, oraz (b) sprawdzenie jak rezultaty (a) mają się do akceptowalności zdań zrekonstruowanych przez autorski parser działający w oparciu o informacje statystyczne oraz syntaktyczne.

1.1 Zasoby i statystyka

Informacje statystyczne wykorzystane w parserze zostały opracowane na podstawie danych z korpusu, w którego skład wchodziło 25 GB, tj. 5649839 plików, tekstów pisanych prasowych; łączna liczba tokenów = ok. 2 mld.; liczba typów (unikalnych jednostek) = ok. 9 mln.; liczba zdań = ok. 160 mln.[2] Na podstawie tych zasobów stworzyliśmy listę słów z przypisanymi wagami ich wystąpienia na początku oraz na końcu zdania obliczanymi wg wzorów:
(a) wystąpienie jednostki leksykalnej *x* na początku zdania:

[2] Wartości podawane są w przypliżeniu ze względu na fakt, że przy zasobach takich rozmiarów dokładne wyliczenia będą zawsze zależne od zbioru reguł wykorzystanego do ekstraksji odpowiednich informacji.

$$ws(x) = (fq(s(x)) \div fq(x)) \times 1000 \quad (1.1)$$

(b) wystąpienie jednostki leksykalnej *x* na końcu zdania:

$$we(x) = (fq(e(x)) \div fq(x)) \times 1000 \quad (1.2)$$

gdzie:
$fq(x)$ – całkowita frekwencja *x*; $fq(s(x))$ – frekwencja *x* na początku zdania; $fq(e(x))$ – frekwencja *x na* końcu zdania

Następnie dla pary wyrazów *xy* przypisywana jest waga dla wystąpienia granicy frazy między *x* i *y*. zgodna z poniższym wzorem:

$$gf(x, y) = (we(x) - ws(x)) + (ws(y) - we(y)) \quad (1.3)$$

We wzorze 1.3 postanowiliśmy uwzględnić nie tylko same wagi wystąpienia *x* na końcu zdania i *y* na początku, tylko sumę różnic wag ich wystąpienia na początku i końcu zdania. W ten sposób dla obydwu elementów otrzymujemy wartość odzwierciedlającą rzeczywistą tendencję ich występowania na danej pozycji w zdaniu, a nie wyłącznie częstość. W następnej kolejności tak obliczona waga *gf* została zmodyfikowana dla par jednostek silnie kolokujących. Wykorzystaliśmy tu sztucznie stworzoną listę kolokacji zawierającą ok. 80 tyś. dwuelementowych jednostek. Parom tym przypisano moc kolokacji obliczaną wg wzoru:

$$wk(x, y) = \frac{1}{\sqrt{((fq(x) + fq(y)) \div 2) \div fq(xy)}} \times 1000 \quad (1.4)$$

gdzie: $fq(xy)$ – frekwencja dla pary jednostek *x*, *y*.

Waga dla kropki między elementami znajdującymi się na liście kolokacji została pomniejszona o wagę *wk*:

$$gf'(x, y) = gf(x, y) - wk(x, y) \quad (1.5)$$

Powyższe wyliczenia były w trakcie pracy programu wspierane przez moduł syntaktyczny.

2. Parser i syntaktyka

Dla celów analiz przeanalizowaliśmy rezultaty prac trzech wersji opracowanego parsera, które różniły się między sobą stopniem rozbudowania komponentu syntaktycznego, podczas gdy moduł statystyczny pozostawał niezmieniony, w efekcie korzystaliśmy z:

1. programu wykorzystującego metody statystyczne i zawierającego jedynie prosty moduł syntaktyczny umożliwiający grupowanie

bigramów z przypisanymi wagami *gf/gf'* w grupy wyznaczone przez elementy otagowane jako czasowniki, przyjęte przez nas za jądro frazy; jednostki z najwyższą wagą *gf/gf'* w grupie oddzielano kropką.

2. wersji programu z rozbudowanym modułem syntaktycznym, który poszerzono o informacje na temat wiązania się jednostek leksykalnych w związki zgody i rządu (bez rekcji czasownikowej);
3. wersji programu z dalej rozbudowanym (w stosunku do (2)) modułem syntaktycznym korzystającym z ujednoznacznionych tagów gramatycznych[3]; na tym poziomie jednostkom leksykalnym przypisane zostały ujednoznacznione tagi gramatyczne zgodne z systemem zaproponowanym w: Woliński 2004; oraz dodano moduł analizujący rekcję czasownikową opracowany na podstawie informacji zawartych w 5 tomach *Słownika syntaktyczno-generatywnego czasowników polskich* (1980; 1984; 1988; 1990; 1992)

3. Wyniki

Zadanie, jakie postawiono przed trzema przedstawionymi wyżej programu, polegało na pofrazowaniu zbioru tekstów, które wcześniej zostały pozbawione wszelkich graficznych operatorów metatekstowych i następnie porównaniu rezultatów z modelem, tj. tekstami z zachowanymi oryginalnymi podziałami. Zbiór testowy zawierał łącznie 6 tekstów (326 zdań) z różnych dziedzin i rejestrów:

(a) tekst publicystyczny (TP) – model = 51 zdań;
(b) tekst popularnonaukowy (TN) – model = 64 zdania;
(c) tekst kulturalny (TK) – model = 48 zdań;
(d) 2 transkrypcje przemówień sejmowych – (S1) = 51 i (S2) = 15 zdań;
(e) transkrypcja wiadomości tv (TV) – model = 73 zdania.[4]

Teksty zostały tak dobrane, by zawierały zarówno język pisany (TP, TN, TK), jak i mówiony (S1, S2, TV). W celu uzyskania oglądu praktycznej możliwości zrekonstruowania oryginalnego podziału danego tekstu na zdania wyniki testu zostały porównane z rezultatami uzyskanymi w analogicznym badaniu, z tym że wykonanym na grupie pięciu ekspertów – rodzimych użytkownikom języka polskiego z wykształceniem filologicznym.

Znane są badania pokazujące, iż użytkownicy języka nie są zgodni co do sposobu podziału danego tekstu na zdania/frazy (por. m.in.: Beeferman et al. 1998; Stevenson, Gaisauzkas: 2000). W naszym badaniu każdy z ekspertów otrzymał 6 powyższych tekstów w identycznej formie, jaka została podana parserowi (bez podziału wielka/mała litera i znaków interpunkcyjnych) z zadaniem podzielenia tych tekstów na zdania zgodnie

[3] W tym celu korzystaliśmy z programu TaKIPI 1.8.

[4] W przypadku (d) oraz (e) model stanowiły ręcznie wykonane transkrypcje z podziałem dokonanym na podstawie wskaźników akustycznych.

z własną wiedzą językoznawczą oraz w niejednoznacznych przypadkach – intuicją językową. Oceny możliwości rekonstrukcji oryginalnego podziału na zdania dokonano przez porównanie otrzymanych tekstów ze stanowiącymi punkt odniesienia oryginałami. Stopień zgodności podziałów wykonanych przez program, jak i tych wykonanych przez ludzi z modelami obliczyliśmy wg miary *F1*. Uzasadnienie wykorzystania do tego celu właśnie takiej miary można znaleźć m.in. w artykule Stevensona i Gaizauskasa (Stevenson, Gaisauzkas: 2000). F1 jest liczona wg wzoru:

$$F = \frac{2PR}{P+R} \quad (3.1)$$

gdzie: dokładność (*precision*)*:* P = liczba zdań zgodnych z modelem podzielona przez całkowitą liczbę zdań utworzoną przez parser/eksperta; kompletność (*recall*)*:* R = liczba zdań zgodnych z modelem podzielona przez całkowitą liczbę zdań w modelu. Wyniki uzyskane przez parser przedstawiono w tabeli1, natomiast wyniki uzyskane przez ekspertów w tabeli2.

Tabela1: Zgodność wyników pracy parsera z modelem

Tekst	pr1	pr2	pr3
TP	-	7,61	9,25
TN	-	4,19	2,83
TK	6,94	4,72	6,83
S1	4,72	14,76	24,83
S2	-	-	6,25
TV	-	11,18	13,13
suma	2,65	8,29	11,62

Tabela2: Zgodność wyników pracy ekspertów z modelem

Tekst	H1	H2	H3	H4	H5	Średnia H
TP	60,21	86,86	63,46	66,66	70	69,44
TN	34	58,62	66,17	77,16	61,66	59,52
TK	34,88	44,23	19,4	42,01	30,92	34,29
S1	44,77	50	52,38	56,25	45,8	49,84
S2	35,71	50	41,17	41,17	27,58	39,13
TV	25	63,7	63,51	65,75	29,31	49,45
suma	39,58	59,36	52,2	60,54	46,54	51,64

W kolumnach pr1-3 tabeli1 znajdują się wyniki uzyskane przez kolejne wersje parsera, natomiast w kolumnach H1-5 tabeli2 wyniki uzyskane przez ekspertów. Ostatnia kolumna tabeli2 przedstawia uśrednione wyniki ekspertów dla danego tekstu, natomiast w ostatnich wierszach obydwu tabel przedstawiono wyniki dla sumy zdań ze wszystkich tekstów.

Niska zgodność uzyskana przez ekspertów, a zaprezentowana w tab.2 pokazuje, że rekonstrukcja oryginalnej, tj. zgodnej z intencjami nadawcy interpunkcji nie jest zadaniem prostym nawet dla osób z wykształceniem filologicznym. Wyniki dwóch pierwszych wersji programu wahają się na granicy losowej zbieżności, dlatego w dalszej części będziemy jedynie wykorzystywać rezultaty uzyskane przez trzecią, najbardziej rozbudowaną wersję programu. Najlepszy wynik uzyskany przez program 24,83 dla S1 oznacza, że poprawnie zrekonstruowanych zostało 19 na 75 zdań. Co ciekawe jest to tekst mówiony, a więc charakteryzujący się strukturą o większej nieciągłości niż np. najsłabiej zrekonstruowany TN będący artykułem popularnonaukowym. TN uzyskał zgodność rzędu 2,83, a więc 2 na 64 zrekonstruowane zdania. Wyjaśnienia takiego stanu rzeczy można upatrywać w tym, że w przypadku tekstów rodzaju TN pojawia się duża liczba słownictwa fachowego, z którą wykorzystywany tager TakiPi słabo sobie radzi. Najsłabszy średni rezultat rekonstrukcji struktury tekstu przez człowieka to 34,29 dla TK, a więc artykułu kulturalnego, charakteryzującego się dosyć artystycznym, literackim stylem. Z kolei najwyższy to 69,44 dla artykułu publicystycznego.

4. Ocena akceptowalności zdań

Fakt, że program nie uzyskał wysokiej zgodności w podziale z oryginałem nie znaczy oczywiście, że utoworzone przez niego zdania nie są akceptowalnymi zdaniami języka polskiego. Aby sprawdzić, na ile zdania utworzone przez program są akceptowalne dla rodzimych użytkowników języka polskiego, wybraliśmy 70% zdań z każdego zrekonstruowanego tekstu. Ponadto zasób poddawany testowi rozszerzyliśmy o zdania będące rzeczywistym rezultatem rozpoznania. Dodanie do zbioru zdań poddawanych ocenie zdań z dwóch tekstów będących wynikami rozpoznania o poprawności 61% oraz 73%, miało na celu sprawdzenie możliwości sensownego podziału tekstów z założenia pozbawionych ciągłości. Zagadnienie "akceptowalności", jako zgodne z założeniami Chomsky'ego, jest przeciwstawiane "gramatyczności" i mimo prób jego eksplikacji pozostaje pojęciem nieostrym (por: Featherson 2005: 701-702). Nie wdając się w zbytnie rozważania, podążamy tu utartą drogą, utożsamiając akceptowalność z poziomem wykonania (*performance*). Jej ocena ma na celu określić czy natywwni użytkownicy języka polskiego uznają dane zdanie za *dopuszczalne* w tym języku. W takim "klasycznym" ujęciu "gramatyczność" jest z kolei powiązana z poziomem kompetencji (*competence*), a jej ocena polega na stwierdzeniu, czy zdanie spełnia zasady gramatyki, czy też nie. Wobec tego gramatyczność jest tylko jednym z (niekoniecznych) czynników wpływających na ocenę akceptowalności (Chomsky 1965: 11). Tak przygotowany materiał daliśmy do oceny 12

natywnym użytkownikom języka polskiego. W rezultacie 12 ankietowanych oceniło 297 zdań według trzystopniowej skali: 1 – zdanie akceptowalne; 0 – trudno powiedzieć, -1 – zdanie nieakceptowalne. W sumie w przypadku 79 zdań ankietowani zgodnie orzekli, że są one akceptowalne i w przypadku trzech, że nie. Czyli procentowo 26% zdań zostało jednoznacznie ocenionych jako akceptowalne. 200 zdań zostało ocenionych jako tylko akceptowalne lub niepewne (bez -1), czyli 67%. 10 zdań zostało ocenionych jako nieakceptowalne lub niepewne (bez 1) – 3%. 141 zdań zostało tylko jeden raz ocenione jako nieakceptowalne. W pozostałych przypadkach brak było zgodności. Największy procentowy udział zdań jednoznacznie ocenionych jako akceptowalne był dla tekstów: publicystycznego – ok. 30%, popularno-naukowy – ok. 34%, oraz transkrypcji programów informacyjnych tv – ok. 39%. Zbieżność ocen obliczyliśmy wg wskaźnika kappa Fleissa (Fleiss 1971, 379), wg. wzoru:

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}} \quad (4.1)$$

Tabela3: Zbieżność ocen akceptowalności zdań

Tekst	kappa
TP	0,2
TN	0,28
TK	0,22
S1	0,31
S2	0,27
TV	0,3
Rozpoznanie 1	0,27
Rozpoznanie 2	0,53
całość	0,32

Stopień w jakim określona wartość kappa wyraża zgodność jest w dużej mierze kwestią w miarę subiektywnej decyzji (por. Carletta 1996: 252). W interpretacjach wyników najczęściej stosuje się skalę zaproponowaną w Landis & Koch (1977: 165) i przedstawioną w tabeli 5. Oprócz tego w literaturze można znaleźć interpretacje oparte na nieco uprosznonych skalach interpretacje (por. tabela4; tabela6). Skale interpretacji współczynnika wyglądają zatem następująco (za: Jarosz-Nowak 2007: 148):

Tabela4: *Interpretacja stopnia zbieżności wg Fleissa*.

kappa	zgodność
< 0,4	słaba
0,4-0,74	umiarkowana
0,75-1	perfekcyjna

Tabela5: Interpretacja stopnia zbieżności wg Landis, Koch.

kappa	zgodność
< 0,00	niewielka
0-0,2	słaba
0,21-0,4	dostateczna
0,41-0,6	średnia
0,61-0,8	znacząca
0,81-1	(prawie) całkowita

Tabela6: Interpretacja stopnia zbieżności wg *Cicchetti i innych.*

kappa	zgodność
< 0,40	słaba
0,4-0,59	umiarkowana
0,60-0,74	dobra
0,75-1	wyśmienita

Do naszych analiz wykorzystujemy skalę zaproponowaną w tabeli5. Zgodnie z nią w większości przypadków ankietowani uzyskali jedynie dostateczną zbieżność, co powoduje pewne trudności w jednoznacznej ocenie tego, na ile w wyniku działania parsera udało się uzyskać akceptowalne zdania języka polskiego. Natomiast w przypadku oceny zdań będących rezultatem działania systemu ASR uzyskano dwa rozbieżne wyniki.: 0,27 oraz 0,53, co jest związane ze specyfiką tekstów tego rodzaju. Przy dokładności rozpoznania rzędu 61% oraz 73%, odsetek niepowiązanych syntaktycznie ze sobą słów w tekście będzie dosyć wysoki. Stan taki jest mocno problematyczny dla parserów, których działanie wszak na takich właśnie regułach bazuje. Podobnie, trudno oczekiwać ciągłości na poziomie semantycznym, na którym bazują statystyki przypisujące wagi dla wystąpienia kropki między wyrazami. Mając na uwadze taki stan rzeczy, podział na zdania/frazy w tym przypadku pozostaje w dużej mierze losowy. Statystyki te wyglądają tylko nieco lepiej, jeśli poddamy je analizie bez uwzględnienia zdań z rozpoznania. W takiej

sytuacji 76 na 263 (29%) zdania zostały jednoznacznie ocenione jako akceptowalne, natomiast 169, czyli 69%, jako akceptowalne lub niepewne.

5. Podsumowanie

Pomimo że artykuł powstał w ramach pracy nad zastosowaniem analiz składniowych w podziale tekstu będącego rezultatem rozpoznania w systemach ASR, jego głównym celem było (1) zbadanie praktycznych możliwości dokonania sensownego podziału na zdania bez uwzględnienia informacji akustycznych oraz (2) stopień możliwego odstępstwa od oryginalnego/zamierzonego przez nadawcę podziału przy zachowaniu akceptowalności uzyskanych zdań. W odniesieniu do (1) rezultaty badań na grupie eksperckiej pokazały, że najwyższy średni stopień zgodności to 60,5%, podczas gdy najlepszy stopień zgodności uzyskany przez parser to 11,62. Warto zauważyć, iż najsłabszy stopień zgodności uzyskany przez eksperta to 39,6%, natomiast średni wynik dla całej grupy badanej 51,6%. Statystyki te świadczą o tym, że rekonstrukcja oryginalnej interpunkcji bez wskazówek akustycznych jest przedsięwzięciem trudnym i dającym dosyć kiepskie rezultaty, a w przypadku zastosowania analiz składniowych do podziału tekstów rozpoznanych przez systemy ASR konieczne jest wsparcie ze strony analiz akustycznych. Część (2) pokazała, że pomimo tego, iż podział tekstu dokonany przez parser uzyskał bardzo niski stopień zgodności podziału tekstu z modelem, to ocena uzyskanych w ten sposób zdań pod kątem akceptowalności była już o wiele bardziej pozytywna. W efekcie 169 z potencjalnych zdań (69%) zostało ocenionych jako nie nieakceptowalne (akceptowalne lub niepewne). W efekcie uzyskane wyniki pozwalają stwierdzić, że problemem w przypadku rekonstrukcji podziału na frazy pozostaje nie tyle poprawne zrekonstruowanie zdań jako takich, to jest w miarę dobre, tylko takie rekonstruowanie, które będzie zgadzać się z intencjami nadawcy wypowiedzi, która jest przez system rozpoznawana.

Bibliografia

Baranowska, E., Francuzik, K., Karpiński, M., Kleśta, J. 2003. Determining Phrase Boundaries in Written Texts for the Purpose of Polish Speech Synthesis. *Speech and Language Technology,* vol.7. pp. 71-78.

Beeferman, D., Berger, A., Lafferty. J. 1998. CYBERPUNC: A lightweight punctuation annotation system for speech. w: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 689-692.

Carletta, J. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computional Linguistics*. vol 22, no. 2. pp. 249-254.

Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press.

Dąbrowska, E. 2010. Naive v. expert intuitions: An experimental study of acceptability judgements. *The Linguistic Review*. vol 27. pp. 1-23.

Featherston, Sam. 2005. Universals and grammaticality: Wh-constraints in German and English. *Linguistics,* vol 43. pp. 667–711.

Fleiss, JL. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, vol. 76, no. 5 pp. 378–382.

Huang, J., Zweig, G. 2002. Maximum entropy model for punctuation annotation from speech. w: *Proceedings of ICSLP*. pp. 917-920.

Jarosz-Nowakm J. 2007. Modele oceny stopnia zgody pomiędzy dwoma ekspertami z wykorzystaniem współczynników kappa. *Matematyka stosowana*. vol.8. pp. 126-154.

Kiss, T., Strunk, J. 2006. Unsupervised Multilingual Sentence Boundary Detection. *Computional Linguistics*. vol. 32, no. 4. pp. 1-40.

Landis, JR., KochGG. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, vol. 33, no. 1. pp. 159-174.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Harper, M. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 5, pp. 1526-1540.

Matusov, E., Mauser, A., Ney, H. 2006, Automatic Sentence Segmentation and Punctuation Prediction for Spoken Language Translation. w: *Proceedings of the International Workshop on Spoken Language Translation*. pp. 158-165.

Polański, K. 1980-1992. *Słownik syntaktyczno-generatywny czasowników polskich*. tom 1-5. Wrocław, Warszawa, Kraków, Gdańsk, Łódź. Wydawnictwo Polskiej Akademii Nauk.

Rao, S., Lane, I., Schultz, T. 2007. Optimizing Sentence Segmentation for Speech Translation. w: *Preceedings of Interspeech, 2007*. pp. 2845-2848.

Song, Y., Ahn, H., Kim, H. 2014. Re-ranking ASR Outputs for Spoken Sentence Retrieval. w: *JIST Workshops & Posters*. pp. 6-11.

Stevenson, M., Gaizauskas, R. 2000. Experiments on Sentence Boundary Detection. w: Proceedings of the North American Chapter of the Association for Computational Linguistics annual meeting, pp. 24–30.

Woliński, M. 2004. System znaczników morfosyntaktycznych w korpusie IPI PAN. *Polonica XII*. pp. 39-54.