Uniwersytet im. Adama Mickiewicza
w Poznaniu
Zakład Lingwistyki Informatycznej i Sztucznej Inteligencji
PL-61-614 Poznań, ul. Umultowska 87, tel. +48-61-8295380, fax +48-61-8295315
Kierownik Zakładu: prof. zw. dr hab. Zygmunt Vetulani
http://vetulani.home.amu.edu.pl
vetulani@amu.edu.pl

Poznań, 15.02.2019.

Recenzja rozprawy doktorskiej mgr. Mateusza Kopcia
"Summarization of Polish Press Articles Using Coreference"

Do oceny przedłożona została rozprawa w języku angielskim w postaci tekstu, na który składa się siedem rozdziałów merytorycznych, bibliografia (153 pozycje) oraz jeden dodatek (Appendix A).

Przedmiot rozprawy.
Rozprawa dotyczy automatycznego streszczania tekstów w języku polskim, a tezą postawioną przez Doktoranta jest stwierdzenie, że „automatycznie wykryta koreferencja pozwoli poprawić wyniki automatycznego systemu streszczającego polskie artykuły prasowe, tak aby działał ze skutecznością przekraczającą najlepsze dostępne rozwiązania". Tak postawione zagadnienie wymaga nie tylko rozwiązania problemu teoretycznego, polegającego na ustaleniu roli koreferencji w streszczaniu tekstów, lecz także umiejętności wykorzystania wyników teoretycznych (dotyczących zjawiska koreferencji) dla uzyskania systemu przewyższającego wszystkie dostępne systemy, co wymaga pracochłonnej i wnikliwej analizy istniejącego stanu technologii w tym zakresie.

Nowatorskość podejścia
Problem automatycznego streszczania tekstu stanowi ważne wyzwanie dla lingwistyki komputerowej i inżynierii języka polskiego. Jego podjęcie przez Doktoranta było cenną inicjatywą ze względu na niewielką liczbę prac poświęconych w tym zakresie językowi polskiemu. Należy zauważyć, że znacząca część istniejących przyczynków zawarta jest w nielicznych pracach magisterskich wykazujących niekiedy istotne braki warsztatowe (jak mała liczba danych, brak walidacji wyników). Stosunkowo nikłe zainteresowanie problemami automatyzacji streszczeń ze strony środowiska polskiego powoduje, że istniejące próby na ogół nie wychodzą poza powszechne stosowanie metodologii i praktyk wypracowanych w dziedzinie ekstrakcji informacji z tekstu, a więc korzystających z klasycznych metod przetwarzania języka naturalnego (NLP). Powszechnie ignorowany jest przy tym fakt, że streszczanie winno być traktowane jako zagadnienie z dziedziny *przetwarzania wiedzy* raczej niż *przetwarzania tekstu*, a więc jako zagadnienie wymagające technik *głębokiego rozumienia (deep understanding)*. I tu wypada odnotować, że badania Doktoranta idą w tym kierunku, gdyż relacja koreferencji wyrażeń jest relacją definiowalną semantycznie i zawartą *implicite* w pojęciu głębokiego rozumienia tekstu. Jest to niewątpliwie nowatorski element rozprawy, przynajmniej w stosunku do prac prowadzonych do tej pory dla języka polskiego.

Treść rozprawy

Wśród siedmiu merytorycznych rozdziałów rozprawy wyróżniam trzy części: wstępną, warsztatową i główną. Część wstępna to rozdziały „Introduction" oraz „Task definition" (a także część rozdziału 3 pt. „State of the art" prezentująca rys historyczny dziedziny badań). Opis aktualnego stanu badań, a także Rozdział 4 „Polish Summaries Corpus" zaliczam do części warsztatowej, a następne trzy („Coreference-based content selection", „Coreference-based summary revision", „Implementation") do głównej.

Rozdziały 1 i 2, jak należało oczekiwać, stanowią wprowadzenie do rozprawy. Najciekawszym fragmentem tego wprowadzenia jest podrozdział „Motivation". Zaczyna się on od zdania zawierającego nieformalną definicję pojęcia „streszczenia" (*summary*) z Oxford Advanced Learner's Dictionary". Niewątpliwie pojęcie *streszczenia* jest najważniejszym słowem kluczowym dysertacji. Autor omawia pozostałe pojęcia w Rozdziale 2 „Task definition", jednak na ogół poprzestając na omówieniach, niekiedy zdawkowych („*Text* and *document* are understood as synonyms"), niedbałych („We may say that two expressions are *coreferent* when they describe the same (or similar) referent"), a nawet wewnętrznie sprzecznych („The definition of a *referent* is, therefore, crucial for the understanding of coreference, yet it is not a term with a well-established (sic!) meaning"). Mimo tych wad Rozdział 2 generalnie spełnia rolę wprowadzającą do części zasadniczej rozprawy. W szczególności istotna jest sekcja 2.3.2 „Why coreference?" gdzie, między innmi, Autor zwraca uwagę na słabą stronę technik sumaryzacyjnych, bazujących na ekstrakcji pełnych zdań z tekstu streszczanego (*extract summaries*), co jest istotnym argumentem za uwzględnieniem koreferencji, i co stanowi wiodącą ideę rozprawy.

Rozdział 3 zatytułowany „State of the art" odnosi się do historii i stanu obecnego badań w zakresie streszczania automatycznego tekstów, ze szczególnym uwzględnieniem języka polskiego. Wyróżnić tu możemy dwie części: pierwszą (3.1 Automatic summarisation), poświęconą ogólnie rozumianemu zagadnieniu streszczania automatycznego i drugą (3.2 Coreference-based summarisation), poświęconą przyczynkom uwzględniającym zjawiska koreferencji w celu poprawy jakości algorytmów sumaryzacyjnych. W obu częściach Autor dokonuje przeglądu historycznego najważniejszych wyników, z tym, że w części pierwszej przegląd ten obejmuje okres od początku dyscypliny, tj. od lat 1950-tych, a w części drugiej od roku 1989, jako, że za pierwsze powiązane prace uznaje on prace Naumera związane z „anaforą zerową" w kontekście generowania tekstu. (Nb. doktorant nie odnotował, że w tym samym roku zagadnienia anafory zerowej, elipsy i koreferencji w języku polskim w kontekście tłumaczenia maszynowego były omawiane w monografii „Linguistic Problems in the Theory of Man-Machine Communication in Natural Language /Z. Vetulani/).

O ile Rozdział 3 zawiera wiele cennych informacji, to brakuje w nim konsekwencji. W części pierwszej (3.1) cytowane prace obejmują okres od roku 1958 do 2011 (jeśli nie liczyć jednej publikacji z 2017), a w części drugiej, obejmującej streszczenia wykorzystujące zjawisko koreferencji i tematy pokrewne (*related work*) – od 1989 do 2016. Dla kompletności obrazu dodajmy, że w tej części winny pojawić się także informacjie o pracach Ogrodniczuka i Kopcia, niewątpliwie *pokrewne* tematowi rozprawy (i częściowo do niej bezpośrednio włączone).

Z punktu widzenia celów projektu doktorskiego (patrz wyżej, sekcja „Przedmiot rozprawy") bardzo ważny i rzetelnie opracowany jest rozdział 3.1.2 „Automatic summarisation for Polish". Autor opisuje prace nad systemami sumaryzacyjnymi dla języka polskiego powstałymi w okresie pomiędzy 2002 i 2018, których autorami byli między innymi Głowińska i Głowiński, Gajęcki i Branny, Dudczak, Świetlicka, Pakulska, Pawluczuk. Przegląd tych projektów Autor uważa za reprezentatywny dla prac w zakresie sumaryzacji dla języka polskiego (z wyłączeniem prac własnych) i, co za tym idzie, za właściwy dla przeprowadzenia zpowiedzianego na wstępie porównania wykazującego wyższą niż pozostałych skuteczność systemu NICOLAS zaprojektowanego i zaimplementowanego przez Doktoranta w ramach

badań własnych i prowadzonych we współpracy z innymi osobami. W podsumowaniu przeglądu, Doktorant wyróżnia dwa spośród sumaryzatorów, a mianowicie system opracowany przez Świetlicką (w rozprawie pod nazwą *Świetlicka summariser*) oraz system LAKON Dudczaka, z których ten pierwszy w porównaniach bezpośrednich wypada najlepiej, jeżeli pominąć porównania z trywialnym sumaryzatorem BASELINE. (Algorytm BASELINE polega na uznaniu za streszczenie krótki, początkowy fragment tekstu.) BASELINE okazuje się być bardzo skuteczny dla tych klas tekstów, dla których obowiązuje zasada „to co najważniejsze umieść na początku tekstu". (Ta praktyka, jak zauważa Autor, jest często obserwowana w doniesieniach prasowych typu „news". Tłumaczy ona odnotowaną przez Doktoranta wysoką ocenę BASELINE przy założeniu, że korpus streszczeń wykorzystywany przy stosowaniu ROUGE jako miary oceny sumaryzatorów to korpus PSC („*The POLISH SUMMARIES CORPUS contains manual single-document summaries of press articles.*", str. 68)).

Za część zasadniczą rozprawy uważam rozdziały kolejne.

Rozdział 4 „Polish Summaries Corpus" jest bardzo istotną, merytoryczną częścią rozprawy, która zawiera materiał źródłowy stworzony na potrzeby konstruowania i ewaluacji systemów automatycznego streszczania pojedynczych artykułów prasowych pisanych w języku polskim. Metodologia tworzenia i wymogi stawiane temu korpusowi zostały starannie opisne i uzasadnione. Najważniejsze oczekiwane cechy zostały zebrane w selcji 4.1.2. „Corpus desiderata". Wśród założonych i zrealizowanych wymogów są, między innymi, następujące:
- postulat dostatecznie dużej liczby tekstów („not less then a few hundered"),
- uwzględnienie różnych stopni streszczania („different ratios of compression") (5%, 10%, 20%),
- zachowanie równowagi pomiędzy rodzajami streszczeń („extractive summaries" vs „abstractive summaries"),
- zapewnienie reprezentatywności tematycznej streszczanych tekstów,
- odpowiedni dobór wieloosobowego (11) zespołu osób streszczających celem uniknięcia indywidualnego obciążenia przez kompetencję językową osoby streszczającej („single-annotator bias"),
- pełna publiczna dostępność korpusu.

Niezależnie od wykorzystania korpusu PSC w samym projekcie doktorskim, będzie on niewątpliwie cennym zasobem tekstowym przydatnym do badań pokrewnych w stosunku do tych opisywanych w rozprawie.

Rozdział 5 „Coreference-based content selection", wraz z dwoma kolejnymi, stanowi zasadniczą część rozprawy. Główne części Rozdziału 5 to :

a) część wstępna (5.1 „Preliminary study"), referująca badania poprzedzające prace nad tworzeniem systemów streszczających przeprowadzone przed rokiem 2015, mające na celu stwierdzenie korelacji pomiędzy decyzjami osób wykonujących streszczenia ekstrakcyjne metodami tradycyjnymi a występowaniem koreferencji wewnątrz wyekstrahowanych fragmentów, szczegółowo przedstawione w (wielokrotnie cytowanej w rozprawie) monografii Ogrodniczuk i inni (2015), której Doktorant był wspólautorem,
b) opis prac nad systemem EMILY (5.2 „Greedy content selection using coreference information"), będący wersją („revised version") opublikowanego w roku 2015 samodzielnego artukułu Autora („Coreference-based Content Selection for Automatic Summarization of Polish News"). Opis systemu EMILY przedstawiony przez jego autora jest wyczerpujący i klarowny. Istotnym elementem tej części Rozdziału 5 są rozważania dotyczące stosowanych powszechnie miar jakości streszczeń zaliczanych do rodziny ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*). Ideą tych miar jest ocena jakości systemu streszczającego przez przyrównanie:

(i) liczby wystąpień (occurrence) interesującego nas obiektu tekstowego (tutaj n-gramu) obserwowanego zarówno w streszczeniach wzorcowych sporządzonych przez ekspertów (ludzi) (*gold summaries*), jak też w ich odpowiednikach sporządzonych przez system, do
(ii) liczby wystąpień tego obiektu tekstowego (n-gramu) w streszczeniach sporządzonych przez ludzi (*gold summaries*).
„Przyrównanie" to ma formę ilorazu. Autor proponuje modyfikację polegającą na tym, żeby zamiast brania pod uwagę wszystkich *gold summaries*, brać tylko najlepsze.
Jest to propozycja zastąpienia miary ROUGE

$$ROUGE_n = \frac{\sum_{S \in RS} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in RS} \sum_{gram_n \in S} Count(gram_n)}$$

przez miarę ROUGE-M

$$ROUGE\text{-}M_n = \max_{S \in RS} \frac{\sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{gram_n \in S} Count(gram_n)}$$

gdzie ROUGE-M bierze pod uwagę tylko te streszczenia automatyczne, które najmniej różnią się od sporządzonych przez człowieka. Miary typu ROUGE-M lepiej oddają oczekiwania, które ma człowiek wobec maszyn, od których to maszyn wymagamy działań zbliżonych do ludzkich. (W rozdziale tym oraz kolejnych stosuje się do oceny eksperymentów streszczania rozmite miary z rodziny ROUGE, w tym z miary ROUGE-M.)

c) opis systemu NICOLAS (5.5 „Mention-level classification and cluster-level aggregation"). Zważywszy, że ewaluacja autorska systemu EMILY w zestawieniu z innymi systemami streszczającymi (BASELINE, systemy Dudczaka i Świetlickiej) nie wypadła w opinii Doktoranta zadawalająco (*„The results are not encouraging, as EMILY manages to significantly outperform only OPENTEXTSUMMARIZER"*, str. 105) zaproponował on system NICOLAS będący wynikiem wnikliwych badań streszczeń generowanych przez EMILY. O ile podstawowa zasada wykorzystania zjawiska konotacji jako podstawy tworzenia streszczeń została utrzymana, o tyle zmodyfikowany algorytm EMILY bierze ponadto pod uwagę szereg istotnych aspektów. Zasadniczym pomysłem było wprowadzenie dwóch kategorii dla rozróżnienia pomiędzy przydatnymi do streszczania elementami tekstu („*good mentions*"), a elementami nieprzydatnymi („*bad mentions*"). W kolejnych sekcjach Autor opisuje statystyczny klasyfikator rozpoznający elementy przydatne, uczący się na podstawie korpusu PSC z naniesionymi anotacjami wskazującymi „dobre" i „złe" elementy tekstu streszczenia, a także sam algorytm systemu NICOLAS. Ewaluacja systemu przeprowadzona według tej samej metodologii co w przypadku EMILY wykazuje, że system wyraźnie przewyższa dotychczasowe. Szczegółowe eksperymenty pokazują przewagę nad systemami BASELINE, ŚWIETLICKA, TEXTRANK i POINTERGENERATOR (dwa ostatne także omówione w rozprawie) przy zastosowaniu różnych miar rodziny ROUGE+ROUGE-M dla 3 wariantów n-gramów (n=1,2,3) i streszczeń 20-to procentowych.

Rozdział 6 („Coreference-based summary revision") nie ma istotnego znaczenia dla oceny całości rozprawy. Zawiera on dość luźnie rozważania dotyczące możliwości polepszenia jakości streszczeń przez rekonstrukcję „brakujących" elementów tekstu („*introducing zero subjects automatically*") i rozszerzenie systemu NICOLAS do NICOLAS+ZERO przeprowadzającego rekonstrukcję tekstową podmiotu domyślnego. W tym przypadku Autor ogranicza się do eksperymentowania z korekcją streszczeń przez rekonstrukcję eliptycznego (zerowego) podmiotu. Zagadnienie zerowej powierzchniowej realizacji jest dobrze znane i rozpoznane szczególnie dla języków o bogatej fleksji rzeczownikowej (w szczególności

języków słowiańskich). Zerowa realizacja powierzchniowa jest pokrewna zjawisku elipsy, od tej ostatniej odróżnia się tym, że ma charakter systematyczny (jest quasi-obowiązkowa). Tak jest w przypadku podmiotu, który jest na ogół zakodowany w sposób morfologiczny i jego odtworzenie nie wymaga wyjścia poza zdanie. W sumaryzacji ekstraktywnej jego przywrócenie lub nieprzywrocenie nie ma na ogół większego znaczenia z punktu widzenia informatywności streszczenia, a jedyna korzyść to zwiększenie redundacji, przy tym za wadę może być uważane zwiększenie objętości tekstu. (Problem często spotykany w praktyce np. przy konieczność sporządzenia abstraktu z użyciem określonej, drastycznie małej liczby znaków. Doświadcza go niewątpliwie każdy z potencjalnych czytelników recenzowanej rozprawy). W sytuacji gdy z takich czy innych powodów „przywrócenie" zerowego podmiotu uważać chcemy za wartość dodaną algorytmu, należy pamiętać, że tej poprawy nie interpretuje pozytywnie miara ROUGE-M. Wyjaśnia to komentarz Autora do Tabeli 6.6 porównującej działanie systemów streszczających BASELINE, ŚWIETLICKA, NICILAS i NICOLAS+ZERO (patrz str. 142 „Data clearly shows that a zero subject injection doesn't improve ROUGE scores of the summarizer. This is understandable when we use an n-gram based evaluation metric: each revision of original text sentence comes with a high risk of decreasing the number of matching n-grams between gold and system summaries").

Znacznie istotniejszy byłby problem niepodnoszony przez Autora w kontekście polepszania sumaryzatora NICOLAS, a mianowicie problem intencjonalnie stosowanej elipsy. Problem jest istotny, gdyż rekonstrukcja elipsy wymaga często analizy (i zrozumienia) obszernego kontekstu, na ogół gubionego przy powierzchownym streszczaniu. Niewątpliwie rozwiązanie problemu wymaga głębokiego rozumienia tekstu, co niekiedy może być nieosiągalne. (O związkach elipsy, anafory, koreferencji pisałem ja (Vetulani 1989), a także wielu innych).

Rozdział 7 („Implementation"). Rozdział ten jest opisem implementacji stystemu NICOLAS z automatyczną rekonstrukcją podmiotu domyślnego (NICOLAS+ZERO). Zawiera on krótki opis aplikacji zawierający szczegóły techniczne interesujące w pierwszym rzędzie aktualnych lub potencjalnych projektantów systemów streszczających.

Rozdział 8 („Conclusion"). Najbardziej interesującą częścią tego, zamykającego rozprawę, rozdziału jest sekcja 8.2 „Feature work", gdzie Autor przedstawia swoją wizję dalszych prac, zwracając uwagę na szereg ważnych, i godnych zajęcia się nimi, aspektów, jak np. uwzględnienie rozmaitych kategorii streszczeń (*multi-documment summarisation* czy *update summarisation*), a także poprawienie instrumentarium (np. przez wykorzystywanie sieci neuronowych o złożonej strukturze, czy powiększenie zasobu danych treningowych).

Doceniajac w pełni osiągnięte wyniki i zrealizowanie zamierzonego celu, jakim było poprawienie parametrów streszczeń w stosunku do najlepszych do tej pory wyników osiaganych przez inne systemy, uważam, że w planie najbliższych prac winna się znaleźć pogłębiona analiza uzyskiwanych przez systemy NICOLAS i NICOLAS+ZERO streszczeń wyznaczająca kolejne kierunki prac własnych. Przykładowo, warto byłoby ustalić czy, i w jakim stopniu, zaimplementowane algorytmy są wrażliwe na cechy stylistyczne streszczanych tekstów. Materiałem do eksperymentów mogłyby być w tym przypadku przekłady tego samego tekstu, na przykład literackiego, wykonane przez różnych tłumaczy. Interesujące byłoby także zbadanie, czy (i ewentualnie w jakim stopniu) wpływ na wynik mieć będzie jakość narzędzi NLP wykorzystywanych przez algorytmy.

Dobre wyniki uzyskane zostały dzięki wykorzystaniu zjawisk koreferencji typowych dla tekstów prasowych czy literackich, a więc zjawisk o charakterze semantycznym. Był to skuteczny krok w dobrym kierunku. Tym bardziej dziwi mnie, że Autor nie stawia przysłowiowej kropki nad „i", przewidując, że w perspektywie dalszej należałoby podjąć badania idące w kierunku wykorzystania technologii *głębokiego rozumienia tekstu (deep understanding)* (nb. istniejących już dla języka polskiego). O ile bowiem trudno jest wymagać uzyskania streszczenia wysokiej jakości od osoby, która tekstu nie rozumie, o tyle

nierealistyczne jest oczekiwanie uzyskania odpowiednio dobrego rezultatu od systemu wykorzystującego technologie NLP niskiego poziomu do streszczania czy tłumaczena maszynowego. Czyniąc te uwagi, mam świadomość istnienia sytuacji, gdzie *pełne rozumienie głębokie* jest nieosiągalne, a streszczanie pozostaje zadaniem sensownym i wykonalnym.

Bibliografia. Na uznanie zasługuje obszerna bibliografia obejmująca 153 pozycje dobrze dokumentujące opis stanu wiedzy i wykonanych prac. Pomocne dla czytelnika są odsyłacze z pozycji bibliograficznych do tekstu.

W konkluzji do przeglądu treści rozprawy stwierdzam, że rozprawa obejmuje imponujący dorobek Doktoranta, ale także zespołu współautorów, w zakresie streszczania tekstów pisanych, a dokładniej artykułów prasowych w języku polskim, uzyskany na przestrzeni ostatnich kilku lat.

Uwagi szczegółowe

Język rozprawy

Decyzję wyboru języka angielskiego jako języka rozprawy uważam za niefortunną. Co prawda decyzja ta pozornie ułatwiła napisanie rozprawy zastosowaną tu techniką kompilacji tekstu z publikacji własnych i współautorskich, lecz – zapewne wbrew intencjom Autora – w obecnej postaci nie doprowadzi to do zwiększenia jej zasięgu oddziaływania, gdyż przyjęta technika częściowego kolażu własnych tekstów czyni dysertację niepublikowalną w aktualnej postaci mimo imponującej zawartości merytorycznej. W tej sytuacji będę doradzał Autorowi gruntowne jej przeredagowanie do formy monografii. (Por. też uwagi redakcyjne poniżej).

Terminologia

Czytając spis treści, można odnieść wrażenie, że Doktorant szczególną uwagę przypisuje definiowaniu i wprowadzaniu nowych pojęć. W miarę lektury rozprawy to wrażenie mija. Przykładem może być ważne w dysertacji słowo *mention,* pojawiające się w rozdziale 2.2 „Coreference". Drugie zdanie tego rozdziału zapowiada definiowanie podstawowych pojęć („*Therefore, we need to present some definitions regarding the concept of coreference*"). Niestety zapowiedź nie została poprawnie zrealizowana. Przykładowo następujący po tej zapowiedzi fragment (str. 26) nie spełnia oczekiwań (poniżej):

> *Coreference is usually defined as a relation between two or more expressions (text spans) in a single text. Coreference resolution is the process of finding the coreference relations in a given text. Such expressions connected (or potentially connected) by coreference relations are denoted as mentions or markables. (...) We may say that two expressions are coreferent when they describe the same (or similar) referent. The definition of a referent is, therefore, crucial for the understanding of coreference, yet it is not a term with a well-established meaning.*

Mianowicie:

- pierwsze zdanie nie jest definicją pojęcia „koreferencja", lecz jedynie zwraca uwagę na jedną z przypisywanych mu „zazwyczaj" („*usually*") własności („bycia relacją").
- nie wiadomo do czego odnieść wyrażenie „such expression" (brak antecedentu anafory), co czyni, że pojęć *mention* i tym samym *markable* nie można uznać za zdefiniowane.
- dodatkowo w zdaniu pretendującym do definicji tych pojęć nieprawdłowo użyta została forma *(to be) denoted*
- definicję pojęcia *coreferent* (i implikowaną przez nią definicję pojęcia *coreference*) możnaby zaakceptować, gdyby nie to, ż odwołuje się ona do pojęcia *referent,* o którym Doktorant wypowiada się, że nie ma on „dobrze określonego znaczenia" („*it is not a term with a well-established meaning*" (sic!))

Co prawda nie ma formalnych reguł, poza dobrym obyczajem i zdrowym rozsądkiem, wprowadzania terminologii, pod warunkiem, że znaczenie wprowadzanych terminów jest dobrze wprowadzone za pomocą definicji, co w tym przypadku nie ma miejsca. W tym przypadku nie przyjdą czytelnikowi z pomocą potoczne tłumaczenia słów *mention* i *markable* (odpowiednio *wzmianka, nadmienienie* oraz *znakowany).*
Zilustrowane tu problemy z terminologią i w szczególności z definiowaniem pojęć mógłyby być znacznie mniej dotkliwe dla czytelnika polskiego (będącego głównym adresatem rozpawy) gdyby stosowane w tekście terminy miały przypisane im odpowiedniki terminologii polskiej.

Inedeks rzeczowy i indeks nazwisk

Brakiem formalnym utrudniającym dokładne studiowanie rozprawy jest brak indeksu rzeczowego oraz indeksu nazwisk (standardowych w rozprawach i monografiach naukowych). Autor najprawdopodobniej nie wyobraził sobie sytuacji, gdzie czyta się tekst w innym formacie niż elektroniczny, a jednak takie sytuacje istnieją (znaleźli się w tej sytuacji np. recenzenci niniejszej rozprawy). Wykonanie indeksu jest czynnością pracochłonną, lecz znacznie polepsza odbiór dzieła. Indeks rzeczowy byłby też okazją do powiązania terminologii polskiej z terminologią angielską. Dominacja terminologii angielskiej w publikacjach naukowych nie zmniejsza obowiazku kultywowania przez świadomych naukowców terminologii rodzimej, choćby na potrzeby dydaktyki i popularyzacji nauki.

Uwagi redakcyjne

Główny problem redakcyjny wynika z przyjętej przez Autora metody autokompilacyjnej, która polega na kompilacji fragmentów własnych publikacji (na ogół sygnalizowanych w tekście odesłaniem bibliograficznym, lecz bez wyraźnego zaznaczeni fragmentów przenoszonych dosłownie). Stosowanie metody przeklejania całych (własnych lub współautorskich) fragmentów tekstu z artykułów jest technicznie wygodne dla Autora (i co do zasad nie uważam tego postępowania za naganne), lecz często ma niekorzystny wpływ na całość dzieła. Przykładem może być zdanie występujące w sekcji 5.1.4 „Conclusion" brzmiące „*Further research is required to investigate how to incorporate this coreference information into sentence selection procedure.*" To zdanie, a w szczególności słowo „further" było na swoim miejscu we wnioskach do artykułu ongiś opublikowanego, z którego pochodzi, ale nie w rozprawie, gdyż w dalszym ciągu tekstu rozprawy czytamy o zapowiedzianych „przyszłych badaniach", które już zostały przeprowadzone.

Przykładem nie przemyśanego dostatecznie przeklejania z różnych artykułów własnych całych fragmentów tekstu jest – z natury rzeczy kompilacyjny – Rozdział 3 „State of the art", gdzie tekst czyni wrażenie chaotycznego. To wrażenie wzmacnia tytuł rozdziału sprawiający wrażenie, że autor utożsamia „rys historyczny" z opisem „aktualnego stanu badań".

Współautorstwo

Rozprawa zawiera wyniki w dużej części opublikowane przez Doktoranta, przy tym w niektórych przypadkach przytoczone zostały *in extenso* obszerne fragmenty odnośnych publikacji. Artykuły źródłowe zostały na ogół w tekście wskazane, zwykle we fragmencie wprowadzającym rozdziału. Przykład tej techniki redakcyjnej stanowi rozdział 4 (Polish Summaries Corpus) gdzie doktorant deklaruje, że „the corpus was originally introduced in [Ogrodniczuk and Kopeć, 2014], part of this chapter was also published in an article about PSC by Ogrodniczuk et al. [2015]." Porównanie tekstu rozdziału z pierwszym z tych artykułów pokazuje, że został on niemal dosłownie przeniesiony do tekstu rozprawy. Jednocześnie Czytelnikowi trudno jest ustalić na podstawie tekstu rozprawy autorstwo poszczególnych sformułowań, a ogólniej, precyzyjnie oszacować wkład własny doktoranta w referowane wyniki.

W pracach współautorskich wykorzystywanych przez Autora brak jest co prawda procentowego określenia udziału w uzyskanych wynikach, czy też analitycznego wskazania tego udziału (co jest coraz częściej praktykowane jako dobry zwyczaj), ale nie ulega

wątpliwości, że udział ten jest znaczny (w szczególności w pracach referowanych w rozdziałach 5, 6 i 7), a Doktorant będzie miał okazję wyjaśnić ewentulane wątpliwości podczas obrony.

Konkluzja

Biorąc pod uwagę sumę poczynionych obserwacji, a przede wszystkim fakt osiągnięcia zadeklarowanych przez Doktoranta celów, stwierdzam, że przedłożona rozprawa spełnia warunki stawiane przez Ustawę o Stopniach i Tytułach Naukowych przed rozprawami doktorskimi i wnioskuję o jej przyjęcie oraz o dopuszczenie jej autora do dalszych etapów przewodu doktorskiego.

Prof. zw. dr hab. Zygmunt Vetulani