Rocznik Przekładoznawczy 3/4 2007 2008
Studia nad teorią, praktyką i dydaktyką przekładu

Jan Rybicki
Kraków

STYLOMETRIA KOMPUTEROWA W SŁUŻBIE TŁUMACZA
(NA PRZYKŁADZIE WŁASNYCH PRZEKŁADÓW)

Zarys treści: Artykuł przedstawia zastosowanie metod stylometrii komputerowej (analizy policzalnych elementów stylu) w stylistycznych badaniach porównawczych oryginału i przekładu literackiego. Wskazuje – na przykładzie własnych przekładów autora – jak pewne różnice stylistyczne w tekście oryginalnym mogą zostać zachowane lub przeciwnie – zaburzone, na podstawie świadomej lub nieświadomej decyzji tłumacza.

Wstęp

Stylometria to analiza policzalnych elementów stylu literackiego w celu ustalenia autorstwa, wykrywania plagiatu i badania różnic stylistycznych między autorami, indywidualnymi dziełami czy tekstami z różnych epok literackich. Najchętniej badane elementy stylu literackiego to długość wyrazów, zdań czy akapitów i bogactwo słownictwa (mierzone bardzo różnymi wzorami, od bardzo prostych, jak *type/token ratio*, po bardzo złożone, jak K Yule'a). Jednak najciekawsze jak dotąd wyniki otrzymywano przez analizę częstości względnych najczęściej występujących słów.

Pierwsze badania stylometryczne pojawiały się jeszcze w XIX w. Już w 1851 r. Augustus de Morgan usiłował potwierdzić autentyczność niektórych pism św. Pawła przez pomiar liczby liter w słowach w poszczególnych *Listach*. Ten sam pomysł zastosował kilka dekad później T. C. Mendenhall, badając dzieła m.in. J. S. Milla, Cervantesa, Cezara, Dickensa, Dumas i Szekspira. Wśród pionierów statystyki literackiej nie zabrakło Polaka – Wincenty

Lutosławski był obok Anglika Lewisa Campbella i Niemca Constantina Rittera pierwszym z bardzo wielu badaczy, którzy na podstawie względnej częstości słów usiłowali poprawić datowanie dzieł Platona. Mimo to do pełnego rozwoju stylometrii brakowało jeszcze właściwych narzędzi.

Po pierwsze potrzebne były nowoczesne metody statystyczne. Gdy te pojawiły się w wieku XX, było tylko kwestią czasu, by zostały zastosowane również do badań nad językiem i tekstem. Nic dziwnego, że pierwsza znacząca publikacja z tej dziedziny, *The Statistical Study of Literary Vocabulary* z 1944 roku, była dziełem jednego z ojców nauk statystycznych – George'a Udny Yule'a. Jednak prawdziwy rozwój tej nieco ekscentrycznej dyscypliny badawczej rozpoczął się wraz z powszechną dostępnością komputerów i służących badaniom nad tekstem programów komputerowych.

Pionierskie w tym względzie były prace Roberto Busa, włoskiego jezuity, nad słownictwem w pismach Tomasza z Akwinu. W roku 1951 ojciec Busa zaczął przerzucać swoje teksty na karty perforowane; pierwszy tom jego *Index Tomisticus* ukazał się drukiem w roku 1973, a ostatni – 7 lat później (Busa 1995).

Od lat 60. komputer na dobre rozgościł się w badaniach literackich. W latach 70. powstały pierwsze zespoły badawcze w Cornell, Holandii, Francji i Włoszech. Znów największym powodzeniem cieszyło się stylometryczne datowanie utworów, rozstrzyganie wątpliwego autorstwa czy wręcz tropienie plagiatów. Tematem studiów w tej dziedzinie były m.in. znów pisma św. Pawła (Morton 1966), anonimowe druki ulotne z czasów amerykańskiej wojny o niepodległość (Mosteller, Wallace 1964) i traktaty Arystotelesa (Kenny 1978). Powstały pierwsze podręczniki zastosowań komputerowych w badaniach literackich – *Guide to Computer Applications in the Humanities* Susan Hockey z 1980 r., choć opisuje zupełnie inne komputery i zupełnie inne programy, jest do tej pory podstawową pozycją w literaturze przedmiotu, dopiero od niedawna zastąpiony przez dostępny w Internecie *A Companion to Digital Humanities* (Schreibman i in. 2004).

Kolejny przełom dokonał się wraz z nadejściem ery komputera osobietego i Internetu. I sam komputer, i wielkie zasoby tekstu w formie elektronicznej są teraz na wyciągnięcie ręki; teraz każdy badacz może użyć komputera do przetwarzania wielkich porcji tekstu, w tym również obszernych powieści czy serii powieści.

Właśnie na powieści skupiły się badania prof. Johna Burrowsa z Uniwersytetu Newcastle w Australii, które doprowadziły do powstania nowego standardu w stylometrii komputerowej, pierwszej naprawdę powszechnie stosowanej metody badawczej w tej dziedzinie. W pracy *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*

z 1987 r. Burrows tak uzasadnia podjęcie badań literackich nad najczęstszymi, a nie kluczowymi słowami tekstu:

> Jest prawdą raczej niechętnie uznawaną, że badając dzieła literackie zachowujemy się tak, jak gdyby jednej trzeciej, dwóch piątych czy nawet połowy materiału po prostu w nich nie było. W przypadku Jane Austen owa jedna trzecia, dwie piąte czy połowa materiału to odpowiednio dwadzieścia, trzydzieści, pięćdziesiąt najczęściej przez nią używanych słów. W każdej z jej powieści są to w dodatku wciąż te same słowa: osiem zaimków osobowych, sześć czasowników posiłkowych , pięć przyimków, trzy spójniki, dwa przysłówki, przedimki określony i nieokreślony oraz cztery inne słowa („to", „that", „for" i „all"), z których każde pełni kilka różnych funkcji gramatycznych, zajmuje miejsce – w dodatku zwykle mniej więcej to samo miejsce – wśród trzydziestu najczęściej występujących słów w każdej z jej powieści. I nie tylko jej – w każdej w ogóle powieści napisanej w języku angielskim. (Burrows 1987: 5–6)

I dalej:

> Rozkłady częstości bardzo częstych słów dlatego są tak przydatne, że słowa nie funkcjonują w tekście jako indywidualne jednostki. Ponieważ swe pełne znaczenie uzyskują dopiero wchodząc w najróżniejsze związki między sobą, można widzieć w nich wyznaczniki tych związków i wszystkiego, co owe związki oznaczają w sensie semantycznym. I tak: tam, gdzie spotykamy więcej najczęstszych przyimków, mamy zazwyczaj do czynienia ze stylem bardziej opisowym czy refleksyjnym; jeżeli z kolei jest ich mało, oznacza to, że akcja rozgrywa się na znacznie skromniej udekorowanej scenie (12).

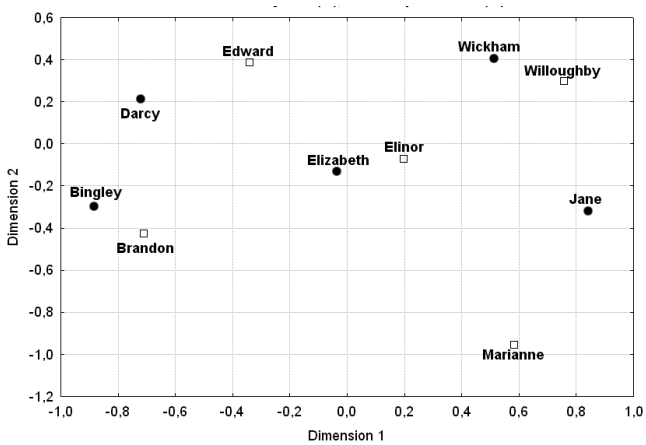
Jednak od takiego wyrywkowego badania na poszczególne słowa spośród tych, które zajmują najwyższe miejsca na listach rangowych znacznie ciekawsze jest spojrzenie całościowe, biorące pod uwagę nie jedną, lecz wiele cech stylu na raz, a możliwe właśnie dzięki statystyce – tę zaś oczywiście znacznie wygodniej liczyć z pomocą komputera. Odwracając znane już założenie, że zbadanie różnic w najbardziej „mechanicznych" elementach stylu może pozwolić na ustalenie autorstwa, Burrows badał różnice zachodzące w twórczości tego samego autora, a konkretnie podobieństwa między indywidualnymi językami, idiolektami podobnych bohaterów różnych powieści. W przypadku powieści Jane Austen jest to o tyle celowe, że w jej powieściach często pojawia się taki sam schemat zależności między głównymi postaciami. Są nimi: 1) nie bardzo bogata, za to inteligentna i niezależna panna, 2) jej mniej inteligentna siostra lub przyjaciółka, 3) z pozoru sympatyczny młody człowiek, który jed-

nak okaże się mniej sympatyczny i – oczywiście – 4) z pozoru niesympatyczny człowiek, który później okaże się porządny i szlachetny.

Znaczenie pracy Burrowsa polega nie tylko na tym, że jego hipoteza sprawdziła się: w powieściach Austen postacie z wyżej wymienionych kategorii rzeczywiście mówią dość podobnie. Australijski badacz był chyba pierwszym, który dość skomplikowaną metodę badawczą – sporządzanie macierzy korelacji między występowaniem kilkudziesięciu najczęściej pojawiających się słów w idiolektach postaci – doprowadził w końcu do formy (w miarę) czytelnej dla „zwykłego" humanisty: do dwuwymiarowych wykresów punktowych, a więc swoistych map odległości między idiolektami poszczególnych postaci.

Prześledźmy więc metodę Burrowsa na przykładzie dwóch powieści Austen, *Rozważna i romantyczna* oraz *Duma i uprzedzenie*. Po uzyskaniu wersji elektronicznej tekstu (w tamtych czasach stosowano jeszcze ręczne wprowadzanie tekstu lub skaner z oprogramowaniem do optycznego rozpoznawania tekstu, OCR) ustalał on np. 60 najczęściej występujących słów w dialogu. Potem obliczał częstości względne najczęściej występujących słów w idiolektach postaci i na tej podstawie ustalał współczynniki korelacji między idiolektami głównych postaci. Zebrane w tzw. macierz korelacji wyniki poddawał procedurze statystycznej zwanej analizą głównych składowych, w wyniku której otrzymywał wykresy, owe mapy odległości między idiolektami w utworach.

Dla wspomnianych dwóch powieści Austen wykres taki (wykres 1) przedstawiał się następująco[1]:



Wykres 1. Jane Austen
Rozważna i romantyczna **(□)**
Duma i uprzedzenie **(●)**

[1] Zaprezentowany tu wykres wykorzystuje dane Burrowsa, ale stosuje dokładniejszą metodę analizy głównych składowych, czyli skalowanie wielowymiarowe.

Jak widać, wykres dość dokładnie oddaje podobieństwa między idiolektami odpowiadających sobie postaciami z obu powieści: idiolekty „sióstr mądrych" – Elinor z *Rozważnej i romantycznej* oraz Elizabeth z *Dumy i uprzedzenia*, są do siebie bardzo podobne; ich nieco mniej rozważne siostry (Jane i Marianne) też znajdują się w tym samym „kącie" wykresu; podobne pary tworzą przyszli wybrańcy tak jednej (Edward i Darcy), jak drugiej siostry (Brandon i Bingley) oraz oba czarne charaktery (Willoughby i Wickham).

Choć prace Burrowsa nad stylem w powieściach Jane Austen dotyczyły podobnych postaci w dziełach pisanych przez tę samą autorkę i w tym samym języku, nic dziwnego, że prędzej czy później pojawiły się prace badające nie tyle podobieństwa, co zachowywanie pewnych podobieństw i różnic między tekstami powiązanymi w inny sposób: między oryginałem tekstu literackiego a jego przekładem. Prace te powoływały się najczęściej na te same i raczej skromne założenia teoretyczne: na tzw. hipotezę van Leuven-Zwart, według której „jeżeli oryginał i przekład różnią się między sobą na poziomie mikrostrukturralnym, będzie to miało wpływ na ich poziom makrostrukturalny" (van Leuven-Zwart 1989). Podobne założenia przyświecały Burrowsowi i współpracownikom, gdy zajęli się porównywaniem stylometrycznym oryginału trylogii Becketta (*Molloy*, *Malone Dies*, *Unnameable*) z jego własnym przekładem (McKenna i in. 2000) oraz rozpoczętym nieco wcześniej badaniom autora niniejszego doniesienia nad językiem trylogii Henryka Sienkiewicza w oryginale i dwóch angielskich przekładach Curtina i Kuniczaka (Rybicki 2006). Te ostatnie badania wykazały po pierwsze zgodne z tradycyjnymi odczytaniami Trylogii grupowanie idiolektów postaci pod kątem narodowości, klasy społecznej, płci i wieku oraz po drugie dość tajemnicze zachowywanie układu z wykresów oryginału w obu przekładach. Tym bardziej tajemnicze, że sporządzane wykresy oparte były przecież na zupełnie odrębnych danych – bowiem w listach najczęściej występujących słów w dwóch różnych językach nie można się dopatrzyć jakichkolwiek relacji odpowiedniości „jeden na jeden" czy nawet „dwa na jeden". O takie relacje trudno nie tylko w przypadku *nie*, *i*, *ja*, *the*, *to* czy *can*, ale nawet gdy chodzi o znacznie rzadsze słowa znaczące.

Materiał i metoda

W celu dalszego poznania stopnia i mechanizmów przechodzenia układów z oryginału do przekładu postanowiłem w podobny sposób zbadać dwie powieści psychologiczno-szpiegowskie Johna le Carrégo: *A Perfect Spy* (1986)

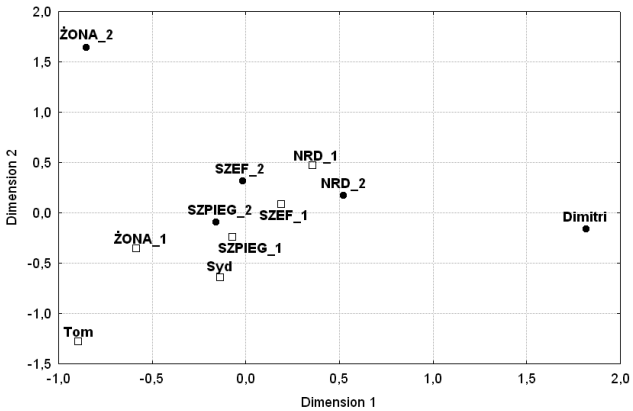
i *Absolute Friends* (2003), oraz moje przekłady tych dzieł, *Szpiega doskonałego* (koniec 2003) i *Przyjaźń absolutną* (początek 2004). Podobnie jak w przypadku opisanych we wstępie badań Burrowsa nad powieściami Jane Austen, obie książki łączy zestaw odpowiadających sobie postaci. Głównym bohaterem jest więc u le Carrégo angielski szpieg (odpowiednio Magnus Pym i Ted Mundy) – osobnik samotny, niezbyt rozgarnięty, ofiara angielskiego systemu oświaty i destrukcyjnej osobowości własnego ojca. W obu dziełach partneruje Anglikowi ułomny fizycznie, ale górujący nad nim intelektualnie podwójny agent z NRD (Axel i Sasha). Kolejną postacią jest szef Anglika, będący zarazem jego przyjacielem (Brotherhood i Amory).

Podczas pracy nad przekładem obu dzieł (w odróżnieniu od oryginałów, które powstały w odstępie 17 lat, ich polskie wersje zostały wydane w odstępie niespełna roku) tłumacz pozostawał pod nieodpartym wrażeniem silnych podobieństw nie tylko między odpowiadającymi sobie postaciami w obu dziełach, lecz może przede wszystkim właśnie między ich wypowiedziami – tym, co u Burrowsa określa się terminem idiolektu. Przedstawiona poniżej analiza stylometryczno-statystyczna jest więc próbą sprawdzenia słuszności owego nieodpartego wrażenia – tym bardziej wiarygodnego, że dokonanego przez tę samą osobę.

Na podstawie pozyskanych wersji elektronicznych obu oryginałów (Internet) i własnych przekładów (już po korekcie wydawniczej w Wydawnictwie Amber) uzyskano dwie listy 250 najczęściej występujących słów w dialogu: angielskich dla obu oryginałów i polskich dla obu przekładów. Rozszerzenie tych list w porównaniu z badaniami Burrowsa było możliwe dzięki wciąż rosnącej mocy obliczeniowej komputerów osobistych, a podyktowane doniesieniami innych autorów, że większa liczba badanych słów znacznie poprawia dokładność i wiarygodność badań (Hoover 2002). Następnie ustalono częstości występowania słów z tych list w każdym z oryginałów i w każdym z przekładów dla głównych postaci utworów. Częstości te podzielone przez liczbę słów (*tokens*) w partii każdej z postaci dały częstości względne. Następnie (w programie Statistica firmy Statsoft) uzyskano współczynniki korelacji między owymi częstościami względnymi dla poszczególnych postaci, które ujęto w macierz korelacji; ta zaś stała się punktem wyjścia do uzyskania dwuwymiarowych wykresów odległości między postaciami w każdym z oryginałów i w każdym z przekładów za pomocą skalowania wielowymiarowego (MDS), procedury podobnej do stosowanej przez Burrowsa analizy głównych składowych, ale posługującej się ulepszonymi algorytmami.

Wyniki

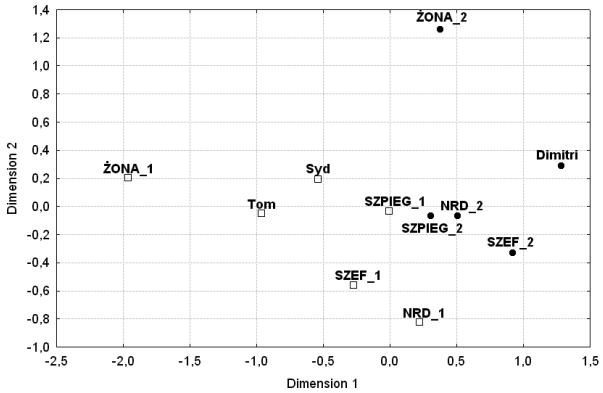
Wykres odległości idiolektów dopowiadających postaci w obu przekładach przedstawiał się następująco (wykres 2):



Wykres 2. Jan Rybicki
Idiolekty głównych postaci
w przekładach
***Szpieg doskonały* (□)**
***Przyjaźń absolutna* (●)**

Zgodnie z oczekiwaniami odpowiadające postacie występują parami: obaj angielscy szpiedzy, ich wschodnioniemieccy partnerzy oraz angielscy szefowie. Tylko obie postaci kobiece okazały się stylistycznie mniej podobne. Wykres uwzględniał poza tym postacie drugoplanowe – na uwagę zasługuje idiolekt jednej z nich, prowokatora Dimitriego z *Przyjaźni absolutnej*, którego odmienność stylu wypowiedzi widoczna była gołym okiem. Należy również zauważyć, że postaci z późniejszej powieści rozlokowane są na wykresie na większej powierzchni, co może sugerować większe zróżnicowanie stylistyczne *Przyjaźni absolutnej*.

Wykres dla oryginałów jest już jednak znacznie mniej jednoznaczny:



Wykres 3. John le Carre
Idiolekty głównych postaci w oryginale
***A Perfect Spy* (□)**
***Absolute Friends* (●)**

Łatwo zauważyć, że o ile oba idiolekty szpiega znajdują się blisko siebie, to wersja z *Absolute Friends* leży jeszcze bliżej swego partnera z NRD; znacznie większe niż w polskim przekładzie są natomiast różnice w pozostałych dwóch parach, a więc obu podwójnych agentów i szefów. Obie żony są w wersji angielskiej oddalone jeszcze bardziej niż w przekładzie; natomiast zachowana jest odrębność stylistyczna Dimitriego. Przekład tym jeszcze różni się od oryginału, że wyraźnie większe zróżnicowanie stylistyczne obserwujemy w *A Perfect Spy*.

Dyskusja

Trzeba przyznać, że zgodność wykresu dla polskiego przekładu z dokonaną przez tłumacza intuicyjną interpretacją podobieństw stylistycznych między postaciami jest zdumiewająca. Taki układ punktów danych zdaje się wskazywać, że pozornie mechanistyczna metoda bardzo dokładnie oddaje skomplikowane relacje podobieństwa i różnicy między stylami wypowiedzi postaci. Przynajmniej na poziomie analizy stylistycznej w obrębie jednego języka można zatem mówić o zgodności metody statystycznej z tradycyjną interpretacją literacką. Metoda najwyraźniej działa.

I właśnie dlatego kolejne obserwacje nie są już tak korzystne dla autora przekładu. Skoro metoda działa – a więc ukazuje odautorskie zróżnicowanie stylistyczne tekstu – nie ma powodu, by sądzić, że dzieje się tak tylko w jednym języku, i tak sugerują wyniki wspomnianych już studiów nad trylogiami Becketta i Sienkiewicza. Ale jeżeli tak, to wykres dla oryginału wskazuje, iż nie było w nim tych podobieństw, których dopatrzył się w nich tłumacz. Jeszcze inaczej: tłumacz najwyraźniej przecenił podobieństwa stylistyczne między idiolektami postaci i narzucił własną ich interpretację na znacznie mniej oczywiste relacje panujące w oryginale.

Pewnym usprawiedliwieniem tej sytuacji może być fakt, iż podczas gdy oryginały powstały w odstępie 17 lat, ich polskie wersje dzieliło w czasie najwyżej 10 miesięcy. O ile le Carré miał dość czasu, by zapomnieć o stylu idiolektów swych postaci, o tyle podobieństwa stylistyczne między obiema książkami musiałby być dla tłumacza znacznie bardziej oczywiste. Warto dodać, że zjawisko ewolucji stylu w czasie również było badane metodą Burrowsa; dłuższemu odstępowi chronologicznemu między porównywanymi dziełami prawie zawsze odpowiadała większa odległość na wykresie – niezależnie od tego, czy chodziło o twórczość jednego czy większej liczby autorów (Burrows 1994).

Jednak najbardziej kłopotliwym problemem nasuwającym się w wyniku przedstawionej powyżej analizy pozostaje pytanie o mechanizm zachowywania podobnych układów na odpowiadających sobie wykresach idiolektów – i nie tylko idiolektów, bo podobne zjawiska obserwowano również choćby w narracji (Rybicki 2006). Jak już wspomniano, nie ma przecież bezpośredniego związku między słowami, które plasują się na czele list rangowych częstości w obu językach. Jedynym, co łączy te dwie listy, jest właśnie to, że chodzi o słowa najczęstsze, nie zaś połączone jakimikolwiek relacjami odpowiedniości. Jak dotąd w literaturze przedmiotu brak pobieżnych choćby rozważań tej kwestii. Próby powiązania zjawiska z pewnymi założeniami językoznawstwa kognitywnego są na razie zbyt ogólnikowe i zbyt nieśmiałe, by mogły cokolwiek tłumaczyć (Connors 2006: 47). Choć znane powiedzenie Elżbiety Tabakowskiej o tym, że „poetyckość tekstu literackiego rodzi się z prozy końcówek i morfemów" (Tabakowska 1995: 175) można łatwo trawestować, by mówiło o prozie spójników i rodzajników, zadowalającego wytłumaczenia tego tajemniczego efektu jak dotąd nie widać.

Literatura

Burrows, J. F., 1987, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, Oxford.

Burrows, J. F., 1994, „Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative", [w:] *Research in Humanities Computing*, s. 1–33.

Busa, R., 1995, *Informatica e scienze umane*, Roma.

Connors, L., 2006, „Combining Cognitive Stylistics and Computational Stylistics", *Digital Humanities*, Paris.

Hockey, S., 1980, *A Guide to Computer Applications in the Humanities*, Baltimore.

Hoover, D. L., 2002, „New Directions in Statistical Stylistics and Authorship Attribution", [w:] *Proc. ALLC/ACH*, s. 57–60.

Kenny, A., 1978, *The Aristotelian Ethics*, Oxford.

Lutosławski, W., 1897, *The origin and growth of Plato's logic: with an account of Plato's style and of the chronology of his writings*, London.

McKenna, W. i in., 2000, „Beckett's Trilogy: Computational Stylistics and the Nature of Translation", [w:] *Revue informatique et statistique dans les sciences humaines*, s. 128.

Morton, A. Q., 1966, *Paul, the Man and Myth*, Boston.
Mosteller, F., Wallace, D., 1964, *Inference and Disputed Authorship: the Federalist Papers*, Reading.
Rybicki, J., 2006, „Can I Write like John le Carré?", *Digital Humanities*, Paris.
Rybicki, J., 2006, „Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations", [w:] *Literary and Linguistic Computing*, Oxford, s. 91–103.
Schreibman, S. i in., 2004, *A Companion to Digital Humanities*, Oxford.
Tabakowska, E., 1995, „Przekład a językoznawstwo kognitywne", [w:] *Mała encyklopedia przekładoznawstwa*, U. Dąmbska-Prokop (red.), Częstochowa.

Computational stylistics in the translator's work (on the basis of the author's own translations)

(summary)

This paper presents a stylometric analysis of two "most literary spy novels" by John le Carré, A Perfect Spy (1986) and Absolute Friends (2003). Written 17 years apart, they were translated by the author of this paper into Polish less than months one from the other in 2003 and 2004. From the very start, it was evident for the translator that the two novels would be an interesting subject of study due to their being built according to a very similar model, especially where characterization is concerned. Both feature a slightly foolish British agent (le Carré's famous trademark), his highly intellectual yet physically handicapped East German nemesis, the British agent's boss/friend, etc. Since these two very similar works shared their Polish translator – who continued to experience a very strong feeling of déjà vu while working on the two novels, this case seemed perfect for a study of stylistic relationships between original and translation. The main effect observed in this study was that of the three above-mentioned couples of corresponding characters, two are very expectedly similar, while one (the two East-German double agents) is not. Their similarity is "regained" in the translation – an interesting corroboration of the translator's "intuitive" suspicion during his work on the Polish version. These results show that, at least in this – very special – case, the accuracy of studies performed by Multidimensional Scaling of correlation matrices of relative frequencies of the most frequent words is quite considerable when applied to translation. This is true despite the disquieting fact that, like previous statistical authorship attribution techniques, this correspondence lacks any compelling theoretical justification. The tentative explanations proposed so far by van Leuven-Zwart's postulate of microstructural chan-

ges influencing the text's macrostructure, 1995) or by McKenna, Burrows and Antonia are certainly not enough. Since overlapping semantic fields of the most frequent words of texts and divergent linguistic systems make one-on-one correspondences impossible, a more general underlying mechanism must be found. At the same time, empirical studies hinting at the existence of such a mechanism have still been very few. This is why more are needed to explain the compelling yet somewhat mysterious successes of Burrows's "old" method.