



Od ZX Spectrum do Jasnopisu

Włodzimierz Gruszczyński
IJP PAN

Kraków, 14 października 2016

Jak to się zaczęło?

- Miałem być inżynierem lub matematykiem...
- Dostałem się na polonistykę 😊
- Na I roku przeczytałem jeden ważny tekst:
Klasyfikacja gramatyczna leksemów polskich
Zygmunta Saloniego. Był wyjątkowo logiczny i ścisły.
- Dotarłem do autora i... zostałem jego uczniem.
- Był rok 1974.
- Jeszcze nigdy nie widziałem komputera. Nie miałem nawet kalkulatora...

Prehistoria

- Prace Jana Tokarskiego, a zwłaszcza *Fleksja polska, jej opis w świetle mechanizacji w urządzeniu przekładowym* (PJ 1961-64).
 - Opis fleksji w SJPDor. – tzw. notacja Tokarskiego.
 - Pierwsze próby wykorzystania komputerów do prac związanych z językiem polskim – koniec lat sześćdziesiątych XX w.*
 - Próby wykorzystania tzw. zmodyfikowanej notacji Tokarskiego w llnf. przez uczniów prof. Waligórskiego: J.S. Bienia, S. Szpakowicza w ich pracach magisterskich - to pierwsze programy odmieniające polskie wyrazy (na radzieckim komputerze GIER).
 - Eksperymenty wykazały, że informacja fleksyjna w SJPDor. była zrealizowana nie zawsze tak, jak zakładał to Tokarski i wymagała weryfikacji.
- *J.S.Bień, PJ 1983/8.

Planowane typy zastosowań

(za J.S. Bieniem)

- Do zastosowania „natychmiast”:
 - sporządzanie indeksów, które wymagają przekształcenia form fleksyjnych w postaci słownikowe (terminu *lematyzacja* jeszcze nie było).
- Do zastosowania wtedy, gdy „polska informatyka zacznie się zbliżać do poziomu światowego”:
 - Tworzenie programów sprawdzających ortografię przez „procesory tekstowe (ang. *word processors*), czyli mikrokomputery wypierające stopniowo klasyczne maszyny do pisania”.
- „Niezależnie od możliwych zastosowań praktycznych zrealizowanie komputerowego słownika morfologicznego umożliwi pogłębienie wiedzy o morfologii polskiej, a dysponowanie programami analizy i syntezy morfologicznej ułatwi badanie wyższych pięter języka polskiego”.

Seminarium *Formalny opis języka naturalnego**

- Inicjator i prowadzący: Zygmunt Saloni.
- Czas: 1977-1990 .
- Miejsce: Instytut Informatyki UW.
- Formuła: nieobowiązkowe cotygodniowe otwarte spotkania.
- Uczestnicy (od kilku do dwudziestu kilku osób na każdym spotkaniu):
 - informatycy, m.in. J.S. Bień, S. Szpakowicz, K. Szafran, M. Woliński, K. Laus-Mączyńska, B.Zawistowska, Z.Jurkiewicz.
 - językoznawcy, m.in. M. Świdziński, W. Gruszczyński, M. Bańko, D. Kopcińska, U.Andrejewicz, M.Wiśniewski, niekiedy też K.Kallas, M.Szupryczyńska.
- Jedno z pierwszych (pierwsze?) środowisk naukowych w Polsce, w którego skład wchodził zarówno lingwiści, jak i informatycy;
- Dominująca tematyka: dystrybucyjny opis języka (polskiego);
- Główne cele: ścisły opis fleksji i składni (zwłaszcza polskiej);
- Weryfikacja poprawności opisu – implementacja komputerowa.
- Tak powstał SALON.

*Wcześniej: seminarium magisterskie na ówczesnym Wydziale Filologii Polskiej i Słowiańskiej UW.

Weryfikacja informacji fleksyjnej w SJPDor.

- Powstanie ogromnej kartoteki „warszawsko-białostockiej”, która wchłonęła dwie swoje poprzedniczki*, zawierającej zweryfikowaną informację fleksyjną z SJPDor.
- Plany dwojakiego wykorzystania tej informacji:
 - w projektach informatycznych;
 - w leksykografii (o słownikach komputerowych wtedy się jeszcze nawet nie marzyło).
- *Kartoteki Z. Zagórskiego z UAM i M. Franjasz z IInf UW.

Morfologiczny poziom przetwarzania tekstów polskich (1979-80)

- Projekt J. S. Bienia mający doprowadzić do powstania programu syntetyzującego formy fleksyjne polskich leksemów.
- Przygotowywałem ręcznie instrukcje dla programisty, w formacie, który zbliżony był do zapisów w języku programowania PROLOG.
- Komputery widziałem na razie tylko z daleka ☹
- W projekcie zabrakło funduszy na komputerową weryfikację algorytmów ☹

Reguły tworzenia narzędnika (1979-80)

Narzędnik lp.

1. /indecl, -, -, -, FORMA HASŁOWA/
/15/ = FORMA HASŁOWA
ééé-----
2. /um, -, n, -, TEMAT/
/15/ = FORMA HASŁOWA
éééé-----
3. /-, -, f, -, TEMAT/
/15/ = TEMAT + a TYPY: radna /ADJ/, kobieta /f1/, radca/f2/

4. /ADJ, Mięk, -, -, TEMAT/
/15/ = TEMAT + im TYP: podskarbi

5. /ADJ, Stw, -, -, TEMAT/
/15/ = TEMAT + ym TYP: ~~myśliwy~~

6. /ADJ, Tw, -, -, TEMAT/
/15/ = TEMAT + ym TYP: myśliwy

7. /-, kg, -, -, TEMAT/
/15/ = TEMAT + iem TYP: ~~porządnic~~ róg, krzyk, oko

8. /-, -, -, -, TEMAT/
/15/ = TEMAT + em TYP: chłop, okno, kurczę, mieszczanin

Słownik gramatyczny języka polskiego

- Pierwsze pomysły, dyskusje, plany – koniec lat 70.
- Dwojaka inspiracja:
 - prace i dyskusje na seminarium FOJN,
 - dzieło A. A. Zaliznjaka: *Грамматический словарь русского языка* (1977).
- Prace prowadzone ekstensywnie przez nieformalną grupę, której przewodził Zygmunt Saloni.
- Brak jakiegokolwiek wsparcia finansowego i sprzętowego!

Zadanie dla doktoranta (WG): opisać ściśle fleksję rzeczowników

- Opis ma spełniać m.in. następujące warunki:
 - kompletność (ma obejmować wszystkie hasła rzeczownikowe z SJPDor.),
 - jednoznaczność, umożliwiającą napisanie programu oraz zastosowanie w słowniku (w szczególności – słowniku gramatycznym),
 - odniesienie do subkodu pisanego (diakrytem jest litera, a nie głoska),
 - ekonomiczność (posiłkowanie się danymi ilościowymi: najczęściej wykorzystywane wzory powinny być wzorami głównymi).

Tu pojawia się ZX Spectrum...

Opis za Wikipedią:

- jeden z pierwszych małych komputerów domowych wyprodukowanych przez angielską firmę Sinclair Research Ltd. w roku 1982,
- mikroprocesor Zilog Z80A pracujący z zegarem 3,54 MHz;
- pamięć operacyjna: RAM – 16 (lub 48 kB) i ROM – 16 kB;
- wyjście na odbiornik tv (TV-out), wbudowany modulator UHF pozwalający uzyskać obraz w rozdzielczości ekranu 256x192 piksel i 15 kolorach (wyłącznie tryb graficzny);
- gumowa klawiatura membranowa z 40 klawiszami realizującymi 193 polecenia wbudowanego języka BASIC;
- pamięć zewnętrzna na kasetach zapisywanych i odczytywanych przez magnetofon(!);
- Komputer w wersji podstawowej (16 kB RAM) sprzedawano w Wielkiej Brytanii w cenie 125 GBP (po pół roku cenę tę zmniejszono do 99,95 GBP), czyli ok. 150-180 USD (miesięczne wynagrodzenie asystenta na UW wnosilo ok. 12 USD).

ZX Spectrum i jego pamięć zewnętrzna - kasetą magnetofonową



Na tym sprzęcie napisałem doktorat!

- Pierwotnie tekst istniał tylko na kasetach magnetofonowych bez realnej perspektywy wydruku!
- Został wydrukowany (w 1986 r.) na 9-igłowej drukarce mozaikowej przeznaczonej do „wielkich” komputerów ODRA.
- Po obronie przeniesiony na PC XT i na nim złożony w systemie $T_E X$ (czas przetwarzania jednego rozdziału wynosił kilka godzin!).
- Ostateczny wydruk przeznaczony dla drukarni wykonano na drukarce laserowej (1989)!

Wydruk autoreferatu na drukarce igłowej

17
Autor: mjr Włodzimierz Gruszczyński
Promotor: doc. dr hab. Zysmund Saloni

Autoreferat

FLEKSJA RZECZOWNIKÓW WE WSPÓŁCZESNYM JEZYKU POLSKIM OPIS I PRÓBA JEGO ZASTOSOWANIA W SŁOWNIKU (autoreferat)

*W opracowaniu
nie ma opinii flek. sp.*

W pracy przedstawiam nowy opis fleksji polskich rzeczowników pospolitych (z wyjątkiem pluralis tantum, rzeczowników odmienianych się według wzorów przymiotnikowych i tzw. rzeczowników nieodmiennych). Opis ten z założenia służyć ma celom praktycznym, a w szczególności zastosowaniu w leksykografii jednojęzycznej (zwłaszcza w słownikach gramatycznych i ukierunkowanych gramatycznie) i dwujęzycznych jako fleksyjny komponent przyrzeczownikowej informacji gramatycznej. Jak się zdaje, dane wynikające z opracowania materiału mogą również znaleźć zastosowanie w różnego rodzaju opracowaniach poprawnościowych oraz – dzięki zachowaniu stosunkowo wysokiego stopnia dyscypliny formalnej – w automatycznym przetwarzaniu tekstów polskich (szczególnie jako morfologiczny komponent programów syntetyzujących).

Podstawą materiałową pracy stanowi kartoteka rzeczowników wyekscerpowanych z SJPdOr, zawierająca ok. 70 tys. fiszek. Kartoteka ta jest częścią tzw. "kartoteki białostockiej", obejmującej całość materiału leksykalnego zawartego w SJPdOr, a zestawionej przez uczestniczki seminarium maszynistycznych prowadzonych w FUW w Białymostku przez doc. dra hab. Zysmunda Salonia. Zasady opracowywania kartoteki omówione zostały we wstępie do pracy.

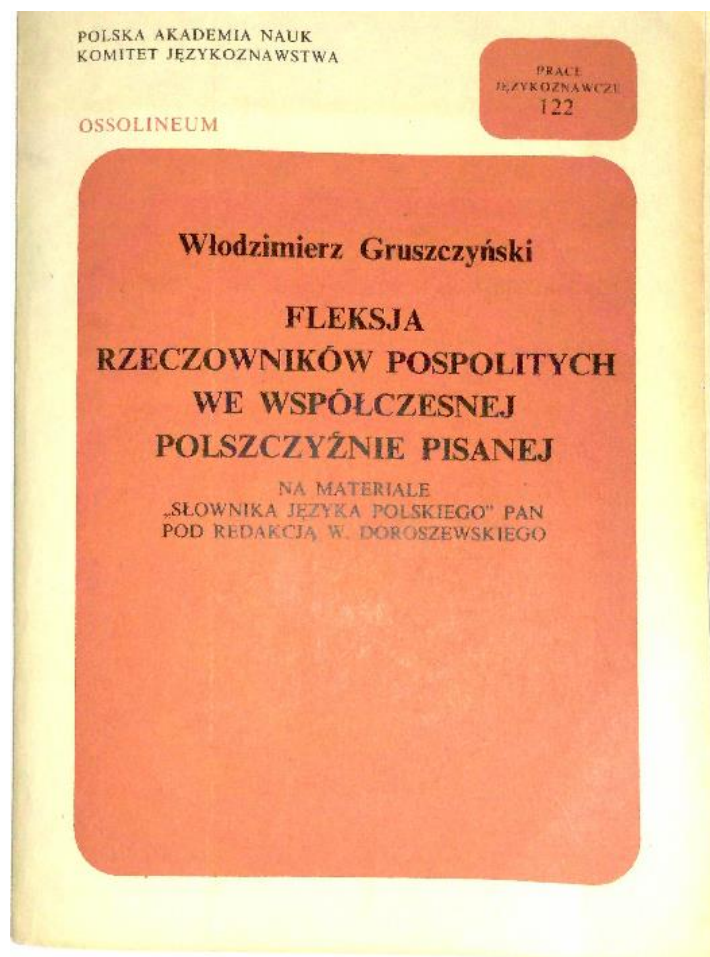
Praca składa się z krótkiego wstępu, trzech części i bibliografii przedmiotu. W części pierwszej podałem wyniki krytycznej analizy istniejących opisów polskiej fleksji nominalnej, zawartych zarówno w gramatykach, jak i w wybranych słownikach. Najdokładniej scharakteryzowałem opisy z gramatyki akademickiej i z SJPdOr. Dość szczegółowa analiza przyrzeczownikowej informacji gramatycznej w SJPdOr, potrzebna była, żeby wykazać celowość zasadniczej modyfikacji opisu Tokarskiego, takiej jednak, która utrzymywałaby jego ogólną koncepcję.

Druga część pracy ma charakter teoretyczny. Omawiam w niej problemy dyskusyjne związane z poszczególnymi katesoriami gramatycznymi (fleksyjnymi) rzeczownika, wprowadzam też aparat pojęciowy, którym posługuję się w części trzeciej – materiałowej. Tu obydwie zasadnicze części przedstawię łącznie.

Najistotniejsze pojęcia zostały w pracy szczegółowo zdefiniowane. Poniżej podaję tylko ich intuicyjny sens.

- SIEĆ PARADYGMATYCZNA to skończony zbiór, którego elementami są wszystkie zestawy wartości katesorii fleksyjnych charakterystyczne dla form leksemów określonego typu, zawierające puste miejsca dla kształtów graficznych form konkretnego leksemu. Jeden element sieci nazywam (za A. Zalzińskiem) klatką. Sieć paradygmaticzna rzeczownika z tradycyjnych opisów, uwzględniających tylko dwie katesorie gramatyczne: liczbę (dwie wartości) i przypadek (siedem wartości), miałaby czternaście klatek. W pracy powiększam asortyment katesorii fleksyjnych o dwie: deprecjatywność i uniformizm (obie przejęte z pewną modyfikacją z pracy Bienia i Salonia). W związku z tym liczba klatek sieci wynosi w moim opisie 17. Dość dobra intuicję pojęcia sieci

Wersja książkowa (1989)



Pierwsza (?) polska książka językoznawcza „z komputera”

Od redaktora

Rozprawa Włodzimierza Gruszczyńskiego należy do cyklu prac gramatyczno-leksyko-graficznych prowadzonych w ośrodkach warszawskim i białostockim. Podjęte przed kilkunastu laty prace te mają na celu poddanie możliwie dużych zbiorów polskiej leksyki stosunkowo dokładnej analizie gramatycznej. Ich materiałem jest przede wszystkim "Słownik języka polskiego" PAN pod redakcją Witolda Doroszewskiego, a wynikiem ma być propozycja opisu znacznych partii polskiej leksyki. Opis taki mógłby być wyzyskany w następnych syntezach gramatycznych i leksykograficznych współczesnej polszczyzny. Ze względu na swój charakter praca wychodzi naprzeciw podobnym oczekiwaniom, obejmuje bowiem analizę zbioru o olbrzymiej objętości. Liczący ponad 70 tysięcy jednostek zbiór rzeczowników pokrywa ok. 60% materiału zgromadzonego w SJPDor. Analiza takiego materiału jest rzadkością w wykonywanych na stopień pracach doktorskich. W danym wypadku nie wpłynęło to zresztą na pewno na obniżenie stopnia szczegółowości przeprowadzonej przez autora analizy.

Publikując rozprawę w kilka lat po jej obronie, autor, zgodnie z sugestiami recenzentów, zdetydował się nie poddawać jej zasadniczej przeróbce, lecz wprowadzić do niej poprawki i niewielkie uzupełnienia, pozostawiając nie zmienioną jej zasadniczą koncepcję. Charakter wprowadzonych do pracy, przede wszystkim pod wpływem uwag recenzentów zmian, omawia autor w zakończeniu, dodanym do wersji przeznaczonej do druku.

Specjalnie należy podkreślić, że praca Gruszczyńskiego została napisana i opracowana przez autora przy użyciu nowoczesnych narzędzi. Nie zasługiwałoby to nawet na specjalną wzmiankę, gdyby nie to, że w środowisku językoznawczym ma to charakter, można by rzec, pionierski. Rozpowszechnianie się tych narzędzi, także wśród przedstawicieli polskich nauk humanistycznych, jest tylko kwestią – krótszego czy dłuższego – czasu, dlatego sprawie tej warto poświęcić więcej uwagi i prześledzić drogę przebytą przez Gruszczyńskiego nieco bardziej szczegółowo.

Praca została więc, rozpatrując rzecz technicznie, od razu napisana na prymitywnym komputerze ZX SPECTRUM i była na nim poddawana wszelkim zabiegom redakcyjnym. Po wydrukowaniu na drukarce mozaikowej została przedstawiona jako rozprawa doktorska Radzie Wydziału Polonistyki Uniwersytetu Warszawskiego. W celu opracowania do druku

Próby współpracy z Pracownią Słownika C.K.Norwida (UW)

- W połowie lat 80. w pracowni pojawił się komputer...
- Odkondu się kilka seminariów FOJN z udziałem przedstawicieli pracowni Norwida.
- Istniała szansa na konstruktywną współpracę środowiska seminarium FOJN z pracownią języka Norwida.
- Zawiodła komunikacja międzyśrodowiskowa, brak wspólnego języka.
- Udało się zrealizować tylko jedno przedsięwzięcie: automatycznie zestawione konkordancje do *Vade-mecum*.
- Do faktycznego stworzenia komputerowego słownika języka Norwida doszło dopiero po 2010 r., a konsultantem w tych sprawach był (nieformalnie) mówiący te słowa – historia zatoczyła koło.

Gwarek

- Moduł polskiego słownika ortograficznego współpracujący z programem QuarkXPress (dla komputerów Macintosh).
 - Tworzony w latach 1989-91* na zlecenie firmy SAD, reprezentującej w Polsce firmę Apple.
 - Pierwotnie tworzony tylko przez programistę, później (za późno!) z pomocą lingwistów.
 - W tym wypadku informatycy niemający kontaktów ze środowiskiem lingwistycznym nie docenili skali problemu...
-
- *Dokładnych dat nie pamiętam, zwłaszcza daty rozpoczęcia projektu.

Schematyczny indeks a tergo polskich form wyrazowych

- Dzieło Jana Tokarskiego, opracowane przed rokiem 1973 w formie rękopisu (!).
- Dokończone, uzupełnione i udoskonalone przez Zygmunta Saloniego z pomocą jego uczniów i współpracowników w latach 1980-1993 (!).
- Wydane w formie książkowej dwukrotnie.
- Wykorzystywane wielokrotnie przez informatyków – m.in. jako podstawa pierwszej wersji analizatora morfologicznego *Morfeusz (SlaT)* napisanego przez Marcina Wolińskiego.

Korekty przesyłane z Uppsali

-rs	z V N	piers	
buś	m I N	labuś, rabuś, abbuś, chlebuś, skarbuś	
-buś	ż I IG	-busia	babuś, kolebuś, gębuś, lubuś?
-duś			klituś-bajduś
-cuś	m I N		pracuś, picuś, ojcuś, lancuś
duś	m I N		dziaduś, koroduś, cuduś, duduś
duś	VIa i	dusić	uduś (11)
-fuś	m I N		ryfuś
-guś	m I N		goguś, mizguś
-iuś	m I N		kiciuś, grzecznioś, psiapsiuś, dziadziuś, dzidziuś
-iuś	ż I IG	-iusia	paniuś
-kuś	m I N		lubkuś
kuś	VIa i	kusić	skuś (7)
iluś	Ka 2	ileś	
-luś			pluś(w)
-luś	m I N		laluś, trajluś
-luś	ż I IG	-lusia	laluś
jakiemuś	A 3	jakiś	niejakiemuś
(czyjemuś)	A 3	czyjś	
(czemuś)	Za D	coś	
-emuś	A 3	-yś	któremuś, jakowemuś
(komuś)	Za D	ktoś	
-muś	m I N		chamuś
-muś	ż I IG	-musia	damuś, mamuś
(muś)	VIIa i	-usiec	muś?
-muś	VIa i	-musić	zmuś (6)
-nuś	m I N		garnuś, wnuś, synuś
-nuś	ż I IG	-nusia	pannuś, żonuś, wnuś
-puś	m I N		kapuś, papuś
-ruś			truś(w)
-ruś	m I N		Jędruś, truś, struś
-ruś	ż I IG	-rusia	córuś, truś
-ruś	ż V NV		córuś
-ruś	VIa i	-rusić	bruś
-tuś	m I N		tatuś, kituś, centuś, kotuś
-tuś	ż I IG	-tusia	matuś, kotuś, kostuś?
-tuś	ż V NV		matuś
-tuś	VIa i	-tusić	ciotuś, krztuś (4)
-wuś	ż I IG	-wusia	kawuś?
-czuś	m I N		pączuś, ojczuś, wnuczuś
-czuś	ż I IG	-czusia	kaczuś, paczuś, szwaczuś
-rzuś	m I N		brzuś
-zuś	m I N		lizuś
-żuś	m I N		mężuś

-guś ż I IG -gusia Boguś

[wołacz: Ruś!]

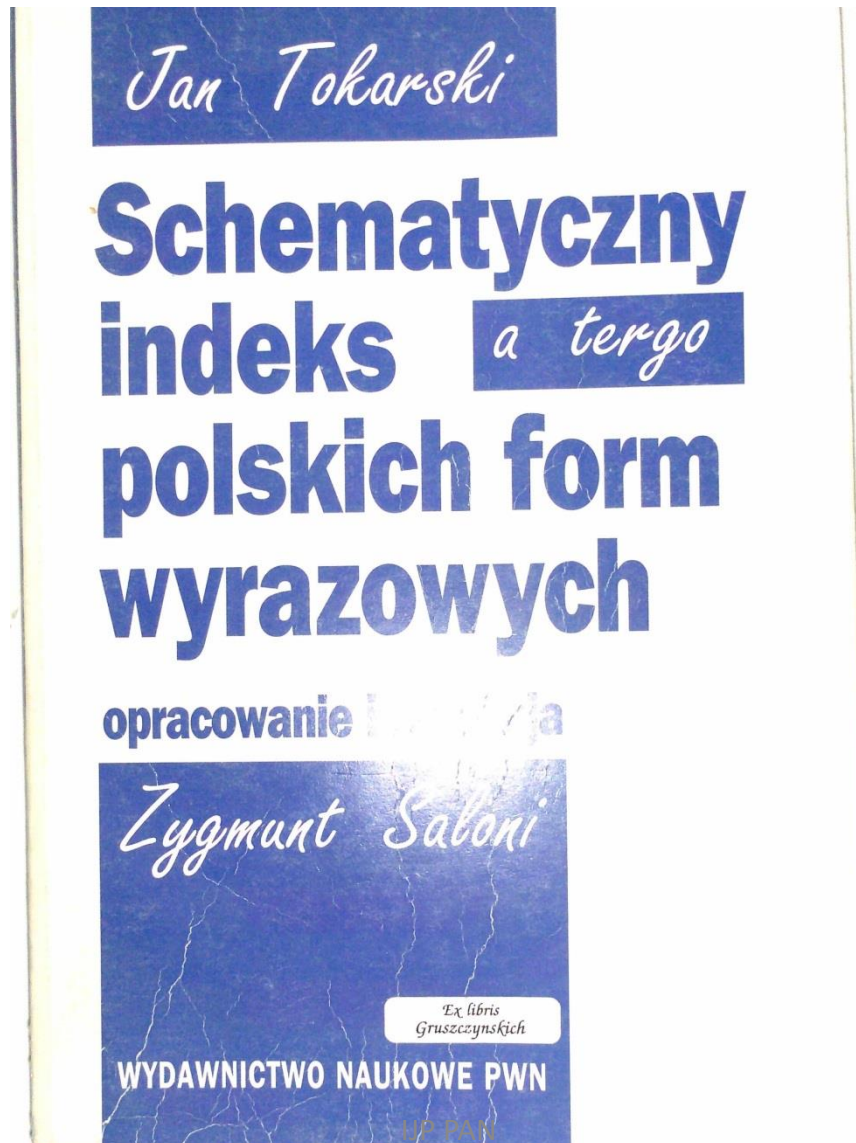
[wołacz!]

ruś ż V NV

FV (wołacz)

-czuś ż V NV Wnuczuś

I wydanie SlaT



Przykład zastosowania

The screenshot shows a web browser window with multiple tabs. The active tab is 'sgjp.pl/siat/'. The page has a dark blue header with the title 'Jan Tokarski i Zygmunt Saloni Schematyczny indeks a tergo polskich form wyrazowych' in orange text. Below the header, there is a green section containing text in Polish. The text discusses the work of Jan Tokarski and Zygmunt Saloni on a morphological index for Polish words, known as 'a tergo'. It mentions that the work was completed in 1973 but never published, and that it was a significant contribution to the study of Polish morphology. The text also mentions that the work was used by other researchers, including Stanisław Waligórski, and that it was a valuable tool for analyzing Polish texts.

Plik Edycja Widok Historia Zakładki Narzędzia Pomoc

Analizator morfol... Schematyczny in... maciej ogrodnic... Cyfrowi mówcy u... Seminarium „For... Weryfikacja korpu... *Komputerowa ... Macintosh – Wiki...

sgjp.pl/siat/

[in English](#)

Jan Tokarski i Zygmunt Saloni

Schematyczny indeks a tergo polskich form wyrazowych

[Z Przedmowy Zygmunta Saloniego:]

Niniejsza praca ma odzwierciedlać zjawiska nie konkretnego kanonu tekstów, lecz systemu językowego rozumianego jako potencja. Jej celem jest przyporządkowanie słowu, mogącemu wystąpić w polskim tekście, interpretacji gramatycznej formy wyrazowej, której wykładnikiem jest to słowo, oraz, w dalszej kolejności, formy podstawowej (hasłowej) odpowiedniego leksemu. Słowa tekstu polskiego reprezentowane są w niej jednak nie w całości, lecz przez swe części końcowe, różnej długości. Owe zakończenia słów są ułożone w odwróconym porządku alfabetycznym (zwanym *a tergo*).

Prezentowane obecnie dzieło Jan Tokarski opracował przed rokiem 1973, gdy prace redakcyjne nad *Słownikiem języka polskiego PAN* pod red. W. Doroszewskiego były już zakończone, lecz jeszcze nie ukazał się indeks *a tergo* do tego słownika. Chyba próbował zainteresować nim jakieś wydawnictwo, a na pewno jakąś instytucję naukową, która mogłaby patronować jego wydaniu. Pewnych danych na ten temat nie mam, mogę tylko przypuszczać z dużym prawdopodobieństwem, że starał się zyskać poparcie Witolda Doroszewskiego. Spotkało go niepowodzenie. [...]

Indeksem bezpośrednio po jego opracowaniu zainteresowali się tylko pracujący nad przetwarzaniem danych w języku naturalnym (a konkretnie — polskim) informatycy, którym autor udostępnił swą pracę w rękopisie. Byli oni przekonani o wartości dzieła i oryginalności jego koncepcji. Janusz Bień pisał, że indeks Tokarskiego będzie niezastąpioną pomocą w analizie fleksyjnej i rozpoznawaniu słów polskich. On to właśnie był w istocie inicjatorem dalszych losów dzieła. Drugi informatyk, Stanisław Waligórski, z indeksu czerpał bezpośrednio informacje użyteczne przy maszynowej analizie tekstów polskich. Sporządził on skróty wyciąg ze schematu Tokarskiego, ograniczający się do elementów wejściowych. Wyciąg ten pomagał mi w orientacji w rękopisie w fazie sporządzania i sprawdzania maszynopisu roboczego.

Sprawa schematycznego indeksu *a tergo* odżyła bowiem w roku 1980. Po wstępnych dyskusjach profesor Tokarski przekazał mi wówczas rękopis celem opracowania.

Pulpit Biblioteki Włodzimierz Komputer PL 23:12 2016-10-12

Morfeusz SIaT

Morfeusz SIaT
Pliki do pobrania

Morfeusz SIaT jest przestarzałą wersją analizatora morfologicznego Morfeusz, zob. [szczegóły](#)

Autorami i właścicielami praw autorskich programu Morfeusz SIaT są:

- w zakresie danych językowych: Zygmunt Saloni,
- w zakresie kodu programu: Marcin Woliński.

Programu *Morfeusz SIaT* można używać na zasadach zawartych w [umowie licencyjnej](#). W skrócie (poniższe punkty nie zastępują umowy):

- Autorzy udostępniają Morfeusza SIaT do bezpłatnego wykorzystania w celach niekomercyjnych (badania naukowe uznajemy za niekomercyjne).
- Nie udostępniamy kodu źródłowego, nie zezwalamy na modyfikację programu. *Morfeusz SIaT* nie stanowi oprogramowania swobodnego (ang. *free software*).
- W szczególności do dopuszczonego zakresu zastosowań nie należą próby wydobycia z *Morfeusza SIaT* danych lingwistycznych lub budowy na jego bazie innego programu o zbliżonej funkcji.
- Jeżeli komuś na tym zależy, umowę licencyjną możemy faktycznie podpisać. Jeżeli nie, zakładamy że osoba pobierająca program tym samym wyraża zgodę na stosowanie się do postanowień umowy.

wersja linuxowa 32-bit	morfeusz-linux-20080205.tar.gz
wersja linuxowa 64-bit	morfeusz-linux64-20080205.tar.bz2
wersja windowsowa	morfeusz-w32-20080205.zip
pakiet debianowy z wersją linuxową	morfeusz_20080205_i386.deb

Ku *Słownikowi gramatycznemu języka polskiego* – kolejne przybliżenia

- Współpraca Z.Saloniego i W.Gruszczyńskiego ze środowiskiem polskich scrabblistów. Konsultanci autora *Oficjalnego słownika polskiego scabblisty* na CD (PWN, 1998).
- Opracowanie wzorców odmiany leksemów polskich zastosowanych w analizatorze morfologicznym w Korpusie Języka Polskiego IPI PAN.
- Książka Z. Saloniego *Czasownik polski: odmiana, słownik*, Wiedza Powszechna 2001 (we współpracy z M. Wolińskim).
- Ręczne dopisanie informacji fleksyjnej do *Indeksu a tergo* do SJPDor.
- Wprowadzenie *Indeksu a tergo* do pamięci komputera (R. Wołosz).
- Opracowanie informacji fleksyjnej do *Słownika polsko-niderlandzkiego* (1997-2001, wyd. 2009) (W. Gruszczyński).
- *Informacja gramatyczna i tablice fleksyjne* do *Encyklopedycznego słownika języka polskiego od a do z*, Larousse Polska (Z. Saloni, W. Gruszczyński), 2006.

Fragment tabeli fleksyjnej ze *Słownika polsko-niderlandzkiego*

Rzeczowniki i zaimki / Substantiva en pronomina

1046

liczba pojedyncza

mianownik dopełniacz celownik biernik narzędnik miejscownik wołacz

singularis

nominatief genitief datief accusatief instrumentalis locatief vocatief

1	storczyk-	-a	-owi	-	-iem	-u	-u
2	stoł-ek	-ka	-kowi	-ek	-kiem	-ku	-ku
3	pie-niek	-ńka	-ńkowi	-niek	-ńkiem	-ńku	-ńku
4	ja-siek	-śka	-śkowi	-siek	-śkiem	-śku	-śku
5	pier-óg	-oga	-ogowi	-óg	-ogiem	-ogu	-ogu
6	ślizg-	-u	-owi	-	-iem	-u	-u
7	Quebe-c	-cu	-cowi	-c	-kiem	-cu	-cu
8	śnież-ek	-ku	-kowi	-ek	-kiem	-ku	-ku
9	st-óg	-ogu	-ogowi	-óg	-ogiem	-ogu	-ogu
10	kr-ąg	-ęgu	-ęgowi	-ąg	-ęgiem	-ęgu	-ęgu
11	rok	roku	rokowi	rok	rokiem	roku	roku

Informacja fleksyjna w Słowniku polsko-niderlandzkim

adidas – agitować

2

stemming (met) ♦ *ta kara jest nieadekwatna do czynu* de straf staat in geen verhouding tot de daad
adidas (m; 25) (euf., inf.) aids ♦ *złapać adidasa* aids oplopen
adidasz (m; 36) (sportt.) (zwykle pl.) sportschoen, tennischoen
adiunkt (m; 119) wetenschappelijk medewerker
adiustować (if; 1011), **za-** (pf) (boekw.) gereedmaken voor druk, persklaar maken
adiutant (m; 119) (mil.) adjudant
adm. (abbr) 1 (administracja) bestuur^h 2 (administracyjny/administracyjnie) administratief 3 (admiral) admiraal
administracja (f; 377) 1 (zarządzanie) bestuur^h, (nadzór) beheer^h ♦ *administracja majątkiem* beheer over een vermogen 2 (w przedsiębiorstwie) (zarząd) bestuur^h, (kierownictwo, dyrekcja) directie 3 (w państwie) (rząd; zarząd) bestuur^h, (władze) overheid ♦ *administracja państwowa* staatsbestuur; staatsapparaat
administracyjny (adj; 753) 1 (dot. zarządzania) bestuurlijk ♦ *koszty administracyjne* beheerskosten; *zdolności administracyjne* bestuurlijke capaciteiten 2 (dot. zarządu) ♦ *pracownik administracyjny* beleidsmedewerker 3 (dot. wykonywania władzy) administratief ♦ *podział administracyjny* administratieve indeling; *prawo administracyjne* (jur.) administratief recht
administrator (m; 121), **administratorka** (f; 297) administrateur, (zarządca) beheerder ♦ *administrator apostołski* (r.-k.) apostolisch administrator

aerodynamiczny (adj; 753) 1 (kształt) gestroomlijnd 2 (siła) aerodynamisch
aerozol (m; 217) 1 aerosol^h 2 (pojemnik) spuitbus ♦ *dezodorant w aerozolu* deodorant in spray
afekt (m; 52) (psych.) affect^h ♦ *zbrodnia w afekcie* crime passionnel
afektacja (f; 377) gemaaktheid, gekunsteldheid
afektowany (adj; 754) gemaakt, gekunsteld
afera (f; 321) 1 affaire ♦ *afera miłosna* liefdesgeschiedenis 2 (inf.) (awantura) rel ♦ *ale afera!* wat een toestand!
aferyzsta (m; 358), **aferyzstka** (f; 297) oplichter, zwendelaar
aferyzstka (f; 297) (vrouwelijke) oplichter, oplichtster, (vrouwelijke) zwendelaar, zwendelaarster
Afganistan (m; 44) (aard.) Afghanistan^h
Afganka (f; 297) Afghaanse (vrouw)
Afgańczyk (m; 15) Afghaan
afgański (adj; 742) Afghaans ♦ *chart afgański* (zoöl.) Afghaanse windhond
afisz (m; 161) affiche^h, aanplakbiljet^h ♦ *nie schodzić z afisza* repertoire houden (o sztuce, filmie itd.)
afiszować się (if; 1011) ((z) czymś¹) paraderen (met), (przechwalać się) te koop lopen* (met) ♦ *afiszować się swoim bogactwem* te koop lopen met zijn rijkdom; *afiszować się z kobietami* zich in het gezelschap van vrouwen vertonen
aforyzm (m; 44) aforisme^h
afrikaans (adj post; 837) (taalk.) ♦ *język afrikaans* de

Tabele fleksyjne ze słownika Larousse'a

R	Liczba pojedyncza							Liczba mnoga					
	M.	D.	C.	B.	N.	Msc.	W.	M.	D.	C.	B.	N.	Msc.
B01	delfin+ø	a	owi	a	em	ie	ie	y	ów	om	y	ami	ach
B02	kapitan+ø	a	owi	a	em	ie	ie	owie	ów	om	ów	ami	ach
B03	jolop+ø	a	owi	a	em	ie	ie	y	ów	om	ów	ami	ach
B04								warcab+y	ów	om	y	ami	ach
B05	intruz+ø	a	owi	a	em	ie	ie	i	ów	om	ów	ami	ach
B06	zakrystian+ø	a	owi	a	em	ie	ie	ie	ø	om	ø	ami	ach
B07	cygan+ø	a	owi	a	em	ie	ie	ie	ów	om	ów	ami	ach
B08	chłop+ø	a	u	a	em	ie	ie	i	ów	om	ów	ami	ach
B09	chłop+ø	a	u	a	em	ie	ie	y	ów	om	ów	ami	ach
B10	pan+ø	a	u	a	em	u	ie	owie	ów	om	ów	ami	ach
B12	skurwysyn+ø	a	owi	a	em	ie/u	u	i	ów	om	ów	ami	ach
B13	kwadrans+ø	a	owi	ø	em	ie	ie	e	ów	om	e	ami	ach
B14								finans+e	ów	om	e	ami	ach
B15	skarb+ø	u	owi	ø	em	ie	ie	y	ów	om	y	ami	ach
B16	AWF+ø	-u	-owi	ø	-em	-ie	-ie	-y	-ów	-om	-y	-ami	-ach
B17	dysonans+ø	u	owi	ø	em	ie	ie	e	ów	om	e	ami	ach
B18	raz+ø	u	owi	ø	em	ie	ie	y	y	om	y	ami	ach
B19	l+as	asu	asowi	as	asem	esie	esie	asy	asów	asom	asy	asami	asach
B20	listopad+ø	a	owi	ø	em	zie	zie	y	ów	om	y	ami	ach
B21	gad+ø	a	owi	a	em	zie	zie	y	ów	om	y	ami	ach
B22	lord+ø	a	owi	a	em	zie	zie	owie	ów	om	ów	ami	ach
B23	dziad+ø	a	owi	a	em	zie	zie	y	ów	om	ów	ami	ach
B24	chasyd+ø	a	owi	a	em	zie	zie	zi	ów	om	ów	ami	ach
B25	sąsi+ad	ada	adowi	ada	adem	edzie	edzie	edzi	adów	adom	adów	adami	adach
B26	zakład+ø	u	owi	ø	em	zie	zie	y	ów	om	y	ami	ach
B27	ZBOWiD+ø	-u	-owi	ø	-em	-zie	-zie	-y	-ów	-om	-y	-ami	-ach
B28	offsid+e	e'u	e'owi	e	e'em	zie	zie	e'v	e'ów	e'om	e'v	e'ami	e'ach

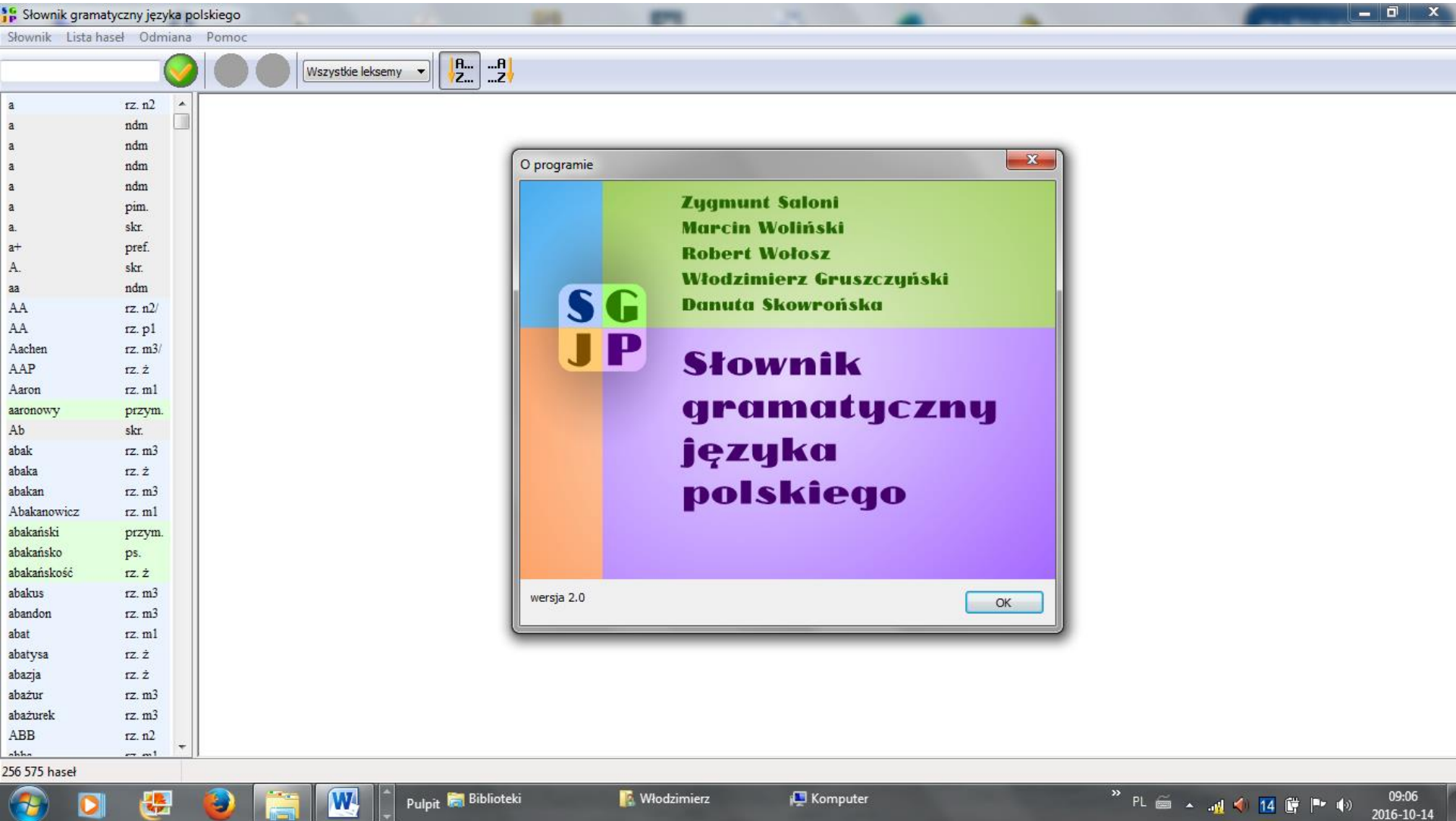
Słownik gramatyczny języka polskiego — wielki finał

- Projekt badawczy realizowany w latach 2003-2006 pod kierunkiem Z. Saloniego:
 - uformowanie zespołu autorskiego: Z. Saloni, W. Gruszczyński, R. Wołosz, M. Woliński;
 - podział obowiązków;
 - duża liczba współpracowników i wolontariuszy.
- Opis projektu i charakterystyka słownika:
 - Z. Saloni, W. Gruszczyński, M. Woliński, R. Wołosz, [Grammatical Dictionary of Polish. Presentation by the Authors.](#) „Studies in Polish Linguistics”. Vol. 4 (2008), 5–25.

Wydania SGJP

- 2007 – Wersja 1.0: CD + książka; program dostosowany do systemu Windows;
- 2012 – Wersja 2.0: CD + książka; poszerzony zespół autorski (D. Skowrońska), wiele zmian merytorycznych (m.in. szersze uwzględnienie danych korpusowych), uzupełnień; program został dostosowany do systemów Windows, Linux i Mac OS X/Intel;
- 2015 – wersja 3.0 online: www.sgjp.pl powstała w ramach grantu NCN Opus w latach 2011-2015.

SGJP 2.0



SGJP 2.0

Słownik gramatyczny języka polskiego

Słownik Lista haseł Odmiana Pomoc

funkcja

Wszystkie leksemy

funkcja

rzeczownik [SJP Dor.]

ż 0128

	1. p.	1. m.
M.	funkcja	funkcje
D.	funkcji	funkcji <i>hom.</i> funkcyj <i>char. arch.</i>
C.	funkcji	funkcjom
B.	funkcję	funkcje
N.	funkcją	funkcjami
Ms.	funkcji	funkcjach
W.	funkcjo	funkcje

256 575 haseł

Pulpit Biblioteki Włodzimierz Komputer

09:09 2016-10-14

SGJP 3.0 online

- Gruntowne zmiany w programie:
 - możliwości filtrowania wg różnych kryteriów;
 - możliwość wyświetlania wzorów;
- Gruntowne zmiany w opracowaniu lingwistycznym:
 - rezygnacja z kategorii uniformizmu;
 - zmiany w opisie rzeczowników;
 - wprowadzenie znacznej liczby nowych haseł (m.in. nazw własnych).

SGJP – zastosowania, perspektywy

- Najważniejsze zastosowania:
 - Analizator morfologiczny *Morfeusz* (SGJP) – program M. Wolińskiego.
 - Informacja fleksyjna w *Wielkim słowniku języka polskiego*, pod red. P. Źmigrodzkiego.
 - Narodowy Korpus Języka Polskiego.
- Perspektywy:
 - Systematyczne wprowadzanie nowych haseł (m.in. planowane wprowadzenie wszystkich polskich nazw miejscowych), co jest możliwe dzięki programowi Kuźnia,
 - Tworzenie „postarzonego” Morfeusza, czyli Korbeusza.

Elektroniczna rewolucja w *Słowniku języka polskiego XVII w. i 1. poł. XVIII w.*

Włodzimierz Gruszczyński
(Instytut Języka Polskiego PAN)

**Koncepcja cyfrowej wersji
*Słownika języka polskiego XVII
i 1. połowy XVIII wieku***

Festiwal Nauki, Warszawa, wrzesień 2005

Postulaty sprzed 11 lat...

Jak przyspieszyć i ułatwić pracę nad słownikiem?

- Ułatwić dostęp do materiałów źródłowych!
- Rozwiązać problem fiszek (Ale jak?!).
- Przenieść jak najwięcej źródeł na nośnik elektroniczny!
- Stworzyć KORPUS tekstów średniopolskich.
- Stworzyć narzędzia ułatwiające pisanie artykułów hasłowych.

Wątpliwości sprzed 11 lat...

Problem fiszek – podstawowego źródła materiału językowego na obecnym etapie prac

- Istnieje ok. 3,2 mln fiszek
- Większość fiszek pisana jest ręcznie
- Pewna część to powielaczowe kopie fiszek przepisywanych na metalowych matrycach.
- Czy przenosić je na nośnik elektroniczny?

17

Pesymizm nieuzasadniony...

Skanowanie fiszek jest nieracjonalne

- ☞ czas potrzebny na zeskanowanie jednej fiszki i zapisanie pliku na dysku porównywalny jest z czasem przepisania jej na komputerze;
- ☞ skanowanie fiszek doprowadziłoby do tego, że ten sam cytat zapisany na różnych fiszkach, byłby wprowadzony do pamięci komputera wielokrotnie;
- ☞ tekst na fiszkach nie nadaje się do automatycznego rozpoznawania (OCR)

20

Marzenia o korpusie

Konieczność stworzenia
korpusu tekstów źródłowych
XVII- i XVIII-wiecznych i
automatyczne wygenerowanie
ich indeksów

- Korpus w pewnym stopniu zastąpi kartotekę, a w znacznym ją uzupełni.
- Korpus przyspieszy pracę nad słownikiem, a w pewnym stopniu go zastąpi (!).

21

Pytania dziś już nieaktualne

Formy publikacji *SXVII*

- Wielotomowe wydanie na papierze czy edycja elektroniczna?
- Jeśli elektroniczna, to czy na CD-ROM, czy w Internecie?
- Jeśli w Internecie, to bezpłatna, czy płatna?

25

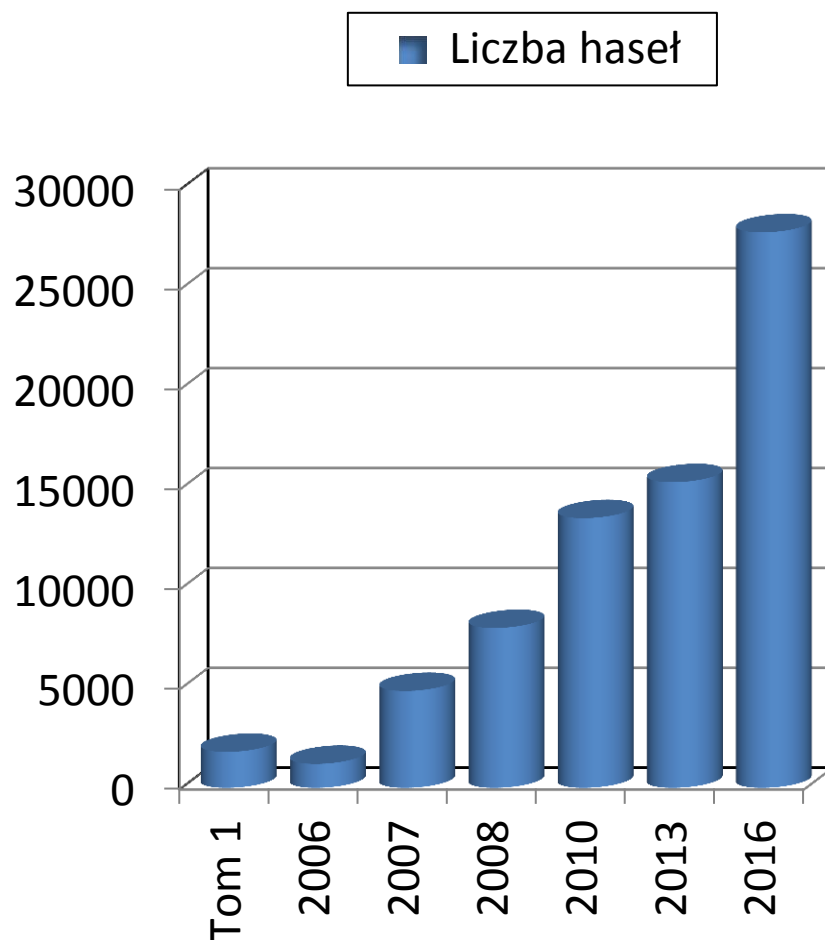
Rok 2005 – początek rewolucji

- Zawieszenie wydawania wersji drukowanej.
- Opracowanie założeń słownika w wersji elektronicznej.
- Zatrudnienie informatyka – Mateusza Żółtaka, który stworzył cały system informatyczny e-SXVII:
 - bazę danych,
 - formularza dla autorów haseł
 - program przekształcający dane z bazy na tekst artykułów hasłowych.

Trudne początki ☹️

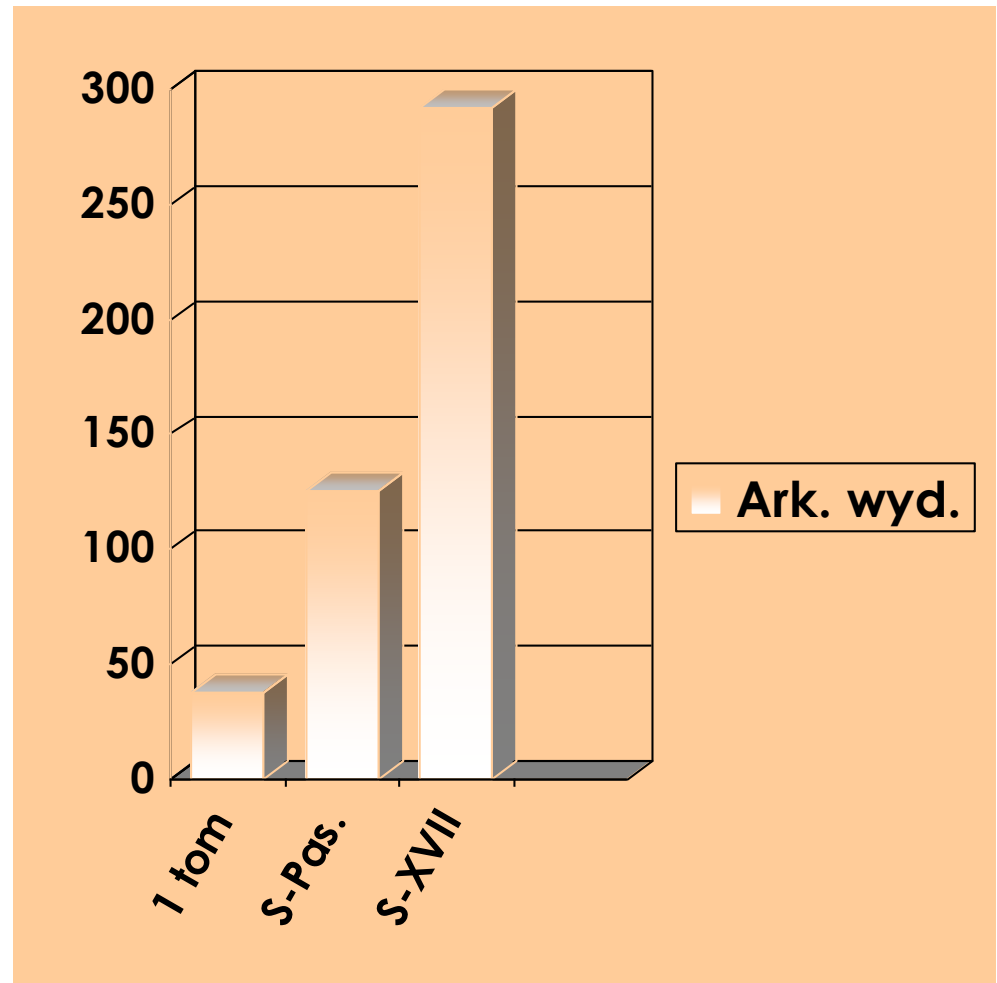
- Konieczność zmian wieloletnich przyzwyczajeń członków zespołu.
- Opór najstarszych członków zespołu i części środowiska językoznawczego.
- Czasowe zahamowanie prac.

Opłacało się 😊



Naprawdę się opłacało!

- I tom *S-XVII*: ok. 36-40 ark. wyd.
- *Słownik języka J. Ch. Paska*: ok. 125 ark. wyd.
- Elektroniczna wersja *S-XVII*: ok. 262 ark. wyd. (09.2010)



Narzędzia, wspomaganie...

Edycja hasła x Słownik XVII wieku

xvii-wiek.ijp-pan.krakow.pl/index.php

Formy hasłowe (1) Tylko odsyłacz rzecz. +

BODZIEC * m blp

>> Warianty fonetyczne (4) +

>> Słowniki

>> Formy gramatyczne 9 + Uwagi: +

>> Etymologia (0) +

>> Znaczenia (5) + + (niepewne)

>> + (0)

>> i (0)

>> Użycie metajęzykowe (0)

>> Podhasła (0) +

>> Odsyłacze (wypisać po przecinku)

Autor

BR-K

>> Wersje (1)

>> Informacje dodatkowe

Edytowana wersja

uwagi redakcyjne: WG: (3.12.2013) dodałem znaczenie 2., zmieniłem kolejność znaczeń 4 i 5. Dodałem przykłady użyc, formę fleksyjną (M.lm)

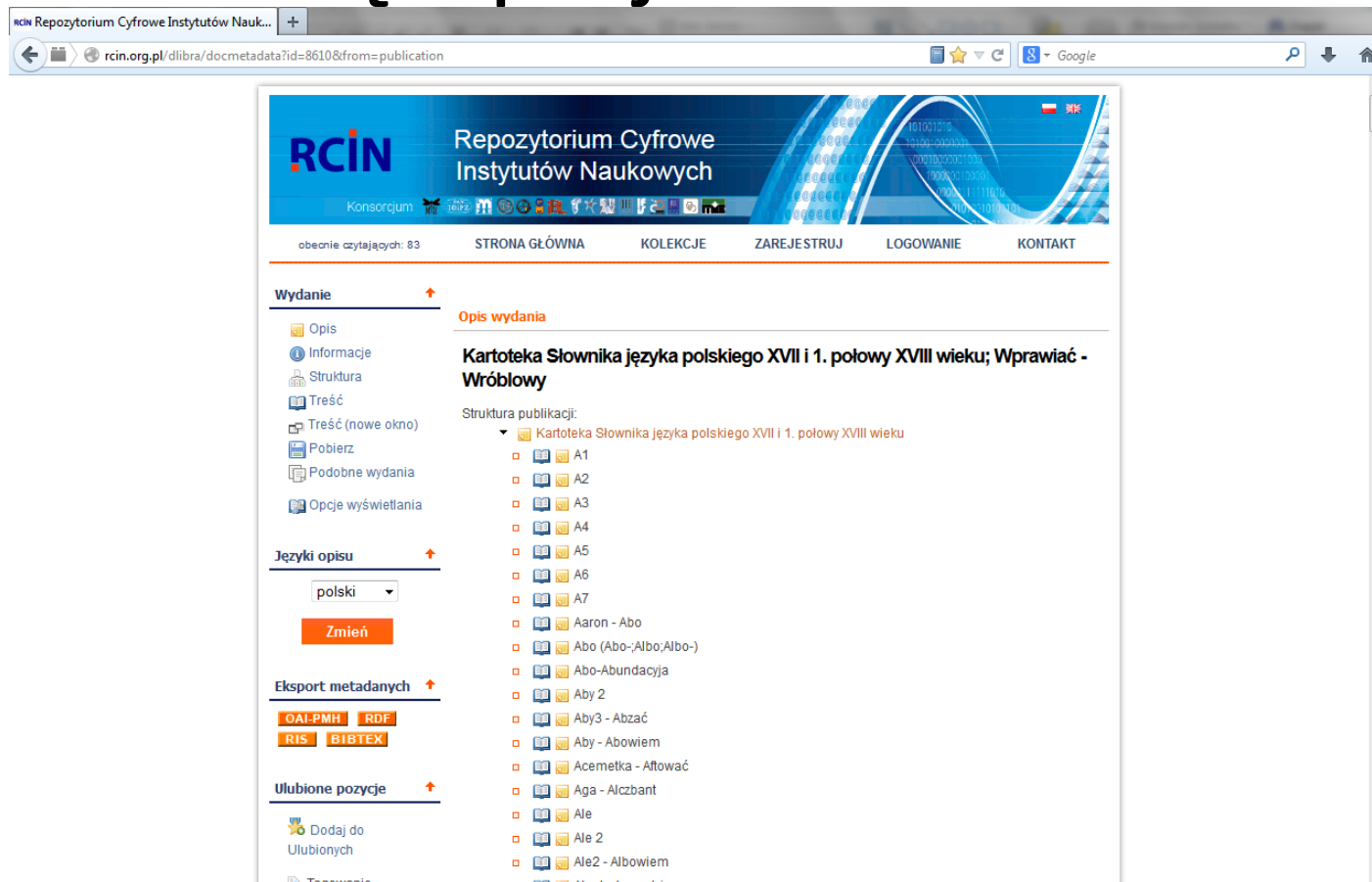
dziennik modyfikacji: włodek (04.12.2013 09:12:24) włodek (03.12.2013 23:12:33)

wybierz hasło nowe hasło usuń hasło podgląd zapisz i wybierz inne hasło zapisz

sprawdź poświadczenia indeks korpusu korpus źródła poczta wyloguj zgłoś błąd szerokość: 800

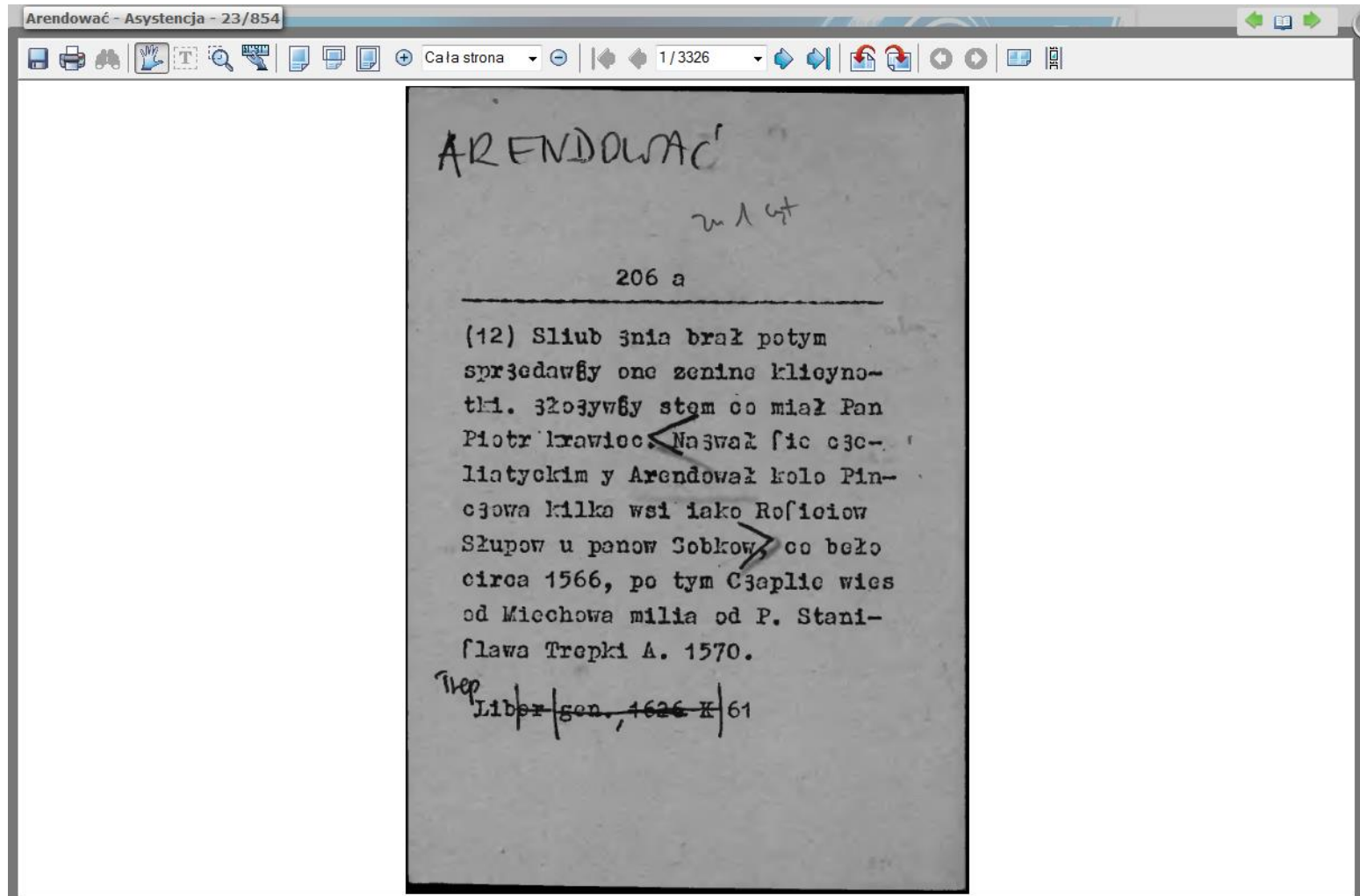
poprz. hasło BODARZ nast. hasło BODEJ

Fiszki w postaci elektronicznej dzięki projektowi RCIN



Kartoteka Pracowni zdigitalizowana w ramach projektu RCIN dostępna pod adresem:
<http://rcin.org.pl/dlibra/docmetadata?id=7222&dirds=1&tab=3>

Fiszki w postaci elektronicznej dzięki projektowi RCIN



Biblioteki cyfrowe

- „Wkład własny”:
 - cyfrowa edycja *Thesaurusa* Knapiusza opracowana przez członków Pracowni i wydana na CD przez BUW);
 - Cyfrowa Biblioteka Druków Ulotnych Polskich i Polski Dotyczących – obecnie na serwerze IJP PAN:
 - <http://cbdu.ijp-pan.krakow.pl/>
 - projekt zrealizowany w rekordowym czasie ośmiu miesięcy w interdyscyplinarnym zespole, w którego skład wchodził:
 - W. Gruszczyński (kierownik), K. Jasińska-Zdun (filolog klasyczny), a. Just (germanistka), E. Gruszczyńska (polonistka, translatolog), A. Kożuchowski (historyk), M. Ogrodniczuk (informatyk i językoznawca), J. Wilk (informatyk, programista), G. Oblas.

Korpus! (in statu nascendi)



- Projekt badawczy „Elektroniczny korpus tekstów polskich XVII i XVIII w. (do 1772 r.)”
- Finansowanie projektu: Narodowy Program Rozwoju Humanistyki na lata: 2013-2017
- Jednostka koordynująca: IJP PAN.
- Współpraca: IPI PAN.
- Kierownik: Włodzimierz Gruszczyński.
- Kryptonim: KorBa (KORpus BArokowy).

Najważniejsze założenia

- Chronologiczne rozszerzenie Narodowego Korpusu Języka Polskiego.
- Planowana objętość: 12 mln segmentów
 - mała w porównaniu z NKJP (300 mln w korpusie zrównoważonym),
 - relatywnie duża jak na korpus historyczny (np. tylko 1 mln w szwedzkich podkorpusach tekstów z okresu *Nysvenska*, tzn. 1520–1850).
- Znakowanie:
 - strukturalne, umożliwiające jak najdokładniejszą lokalizację cytatu,
 - socjolingwistyczne (informacje o autorze, wydawcy, tłumaczu),
 - stylistyczno-genologiczne,
 - tematyczne,
 - językowe (wtręty obcojęzyczne z dokładnością do języka),
 - morfosyntaktyczne (ręczne w podkorpusie 0,5-milionowym).

Korbowe narzędzia (dzięki współpracy z IPI PAN!)

- Wczytywacz,
- Anotatoria 2,
- Ameba,
- Korbeusz (postarzony Morfeusz),
- Poliqarp 2.

KorBa – perspektywa udostępnienia

- KorBa 0.1 będzie dostępna w internecie jeszcze w bieżącym roku.
- W wersji 0.1 możliwe będzie najprostsze przeszukiwanie (segmenty, ich części, ich sekwencje).
- Wersja 1.0 będzie udostępniona nie później niż w końcu 2017 r.

Skok w bok, czyli JASNOPIS

- JASNOPIS to wynik współpracy humanistów z informatykami w ramach projektu ***Mierzenie stopnia zrozumiałości polskich tekstów użytkowych (poza-literackich)*** realizowanego w latach 2012-2015 finansowanego przez NCN.
- Czym jest JASNOPIS?
- Narzędzie informatyczne służące do oceniania tekstów pod względem trudności językowej dostępne bezpłatnie w Internecie pod adresem jasnopis.pl.
- Nie wymaga żadnego dodatkowego oprogramowania ani wiedzy informatycznej lub lingwistycznej.
- Oprócz wyników łatwych do interpretacji podaje także informacje statystyczne przydatne bardziej zaawansowanym użytkownikom.
- Planowane jest udostępnienie tzw. wtyczek (plug-in) umożliwiających korzystanie z Jasnopisu w czasie pracy za pomocą programów MS Word i Open Office.
- Szczegółowe omówienie projektu i aplikacji dostępne w monografii *JASNOPIS, czyli mierzenie zrozumiałości polskich tekstów użytkowych*, red. W. Gruszczyński, M. Ogrodniczuk, Oficyna Wydawnicza ASPRA-JR, Warszawa 2015.

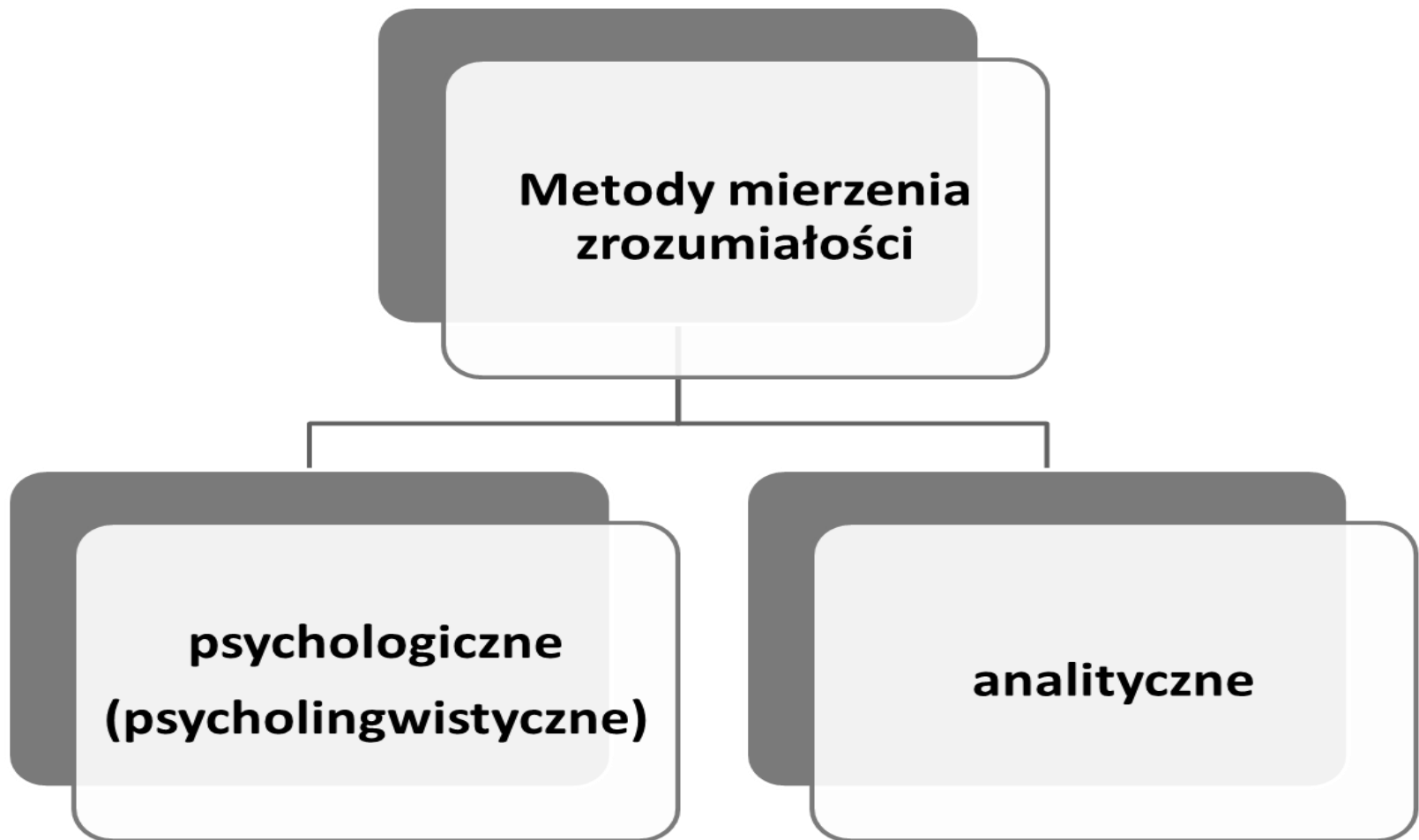
Skok w bok, czyli JASNOPIS

- Włodzimierz Gruszczyński – lingwistyka, kierownik
- Bartosz Broda (Samsung) – informatyk
- Edyta Charzyńska (UŚ) – psychologia i filozofia
- Łukasz Dębowski (IPI PAN) – matematyka, statystyka
- Milena Hadryan (UAM) – (legi)lingwistyka, skandynawistyka
- Bartłomiej Nitoń (IPI PAN) – informatyka
- Maciej Ogrodniczuk (IPI PAN) – lingwistyka i informatyka
- Grażyna Oblas (UW) – ekonomia, socjologia
- Laura Polkowska (UKSW) – lingwistyka
- studentki, studenci, koleżanki, koledzy, którzy podjęli trud przeprowadzania ankiet.

Cele projektu

- Ustalenie cech tekstu, które mają wpływ na jego zrozumiałość (ang. *readability*), przy założeniu, że badania ograniczają się w znacznym stopniu do językowych cech powierzchniowych:
 - gramatycznych (fleksja, składnia);
 - leksykalnych (dobór wyrazów).
- Stworzenie narzędzia informatycznego służącego do mierzenia stopnia zrozumiałości tekstu, a w szczególności:
 - do obliczania wskaźnika czytelności badanego tekstu i ustalania tych jego cech, które odpowiadają za zbyt duży stopień trudności (np. wskazanie trudnych wyrazów, zbyt skomplikowanych zdań, zbyt długich ciągów dopełniaczowych itp.);
 - umieszczanie wyniku na skali wypracowanej w ramach projektu m.in. za pomocą metod psycholingwistycznych.

METODY BADAWCZE



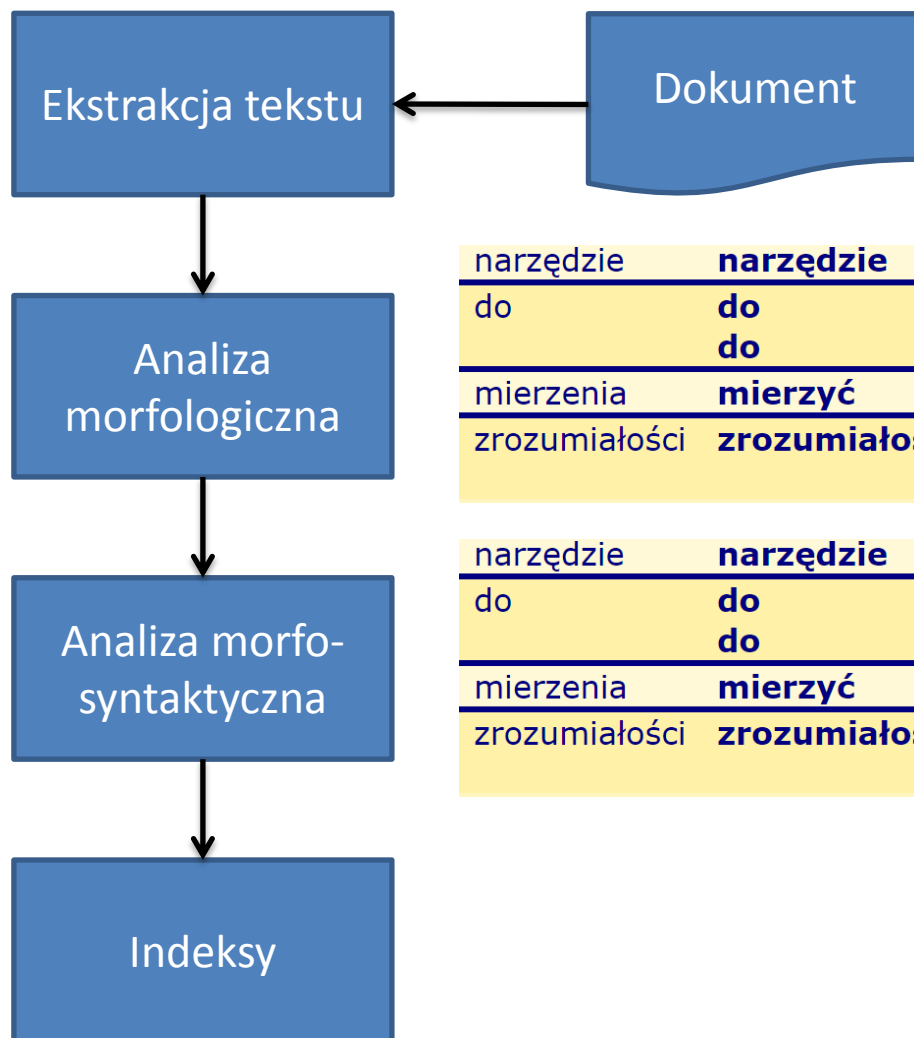
Wzór na trudność tekstu (Ł. Dębowski)

- **Klasa** = 12:25 – 4,12 * **Ridge**
- **Ridge** = 2:7 – 0:0034 *(średnia długość zdania)
 - 0:0027 *(procent słów trudnych)
 - + 0:0026 *(procent rzeczowników)
 - 0:0044 *(procent rzeczowników trudnych)
 - + 0:0037 *(procent czasowników)
 - + 0:0053 *(procent czasowników trudnych)
 - 0:00043 *(średnia długość akapitu)
 - 0:013 *(średnia długość łańcucha dopełniaczy)
 - 0:0033 *(procent dopełniaczy)
 - 0:0019 *(procent rzeczowników na -ość)
 - + ...



Jak JASNOPIS oblicza wynik(i)?

Analiza tekstu w języku polskim



<p> narzędzie do mierzenia zrozumiałości </p>

narzędzie	narzędzie	subst:sg:nom.acc.voc:n2
do	do	prep:gen
	do	subst:sg:pl:nom.gen.dat.acc.inst.loc.voc:n2
mierzenia	mierzyć	ger:sg:gen:n2:imperf:aff
zrozumiałości	zrozumiałość	subst:sg:gen.dat.loc.voc:f
		subst:pl:nom.gen.acc.voc:f

narzędzie	narzędzie	subst:sg:nom.acc.voc:n2
do	do	prep:gen
	do	subst:sg:pl:nom.gen.dat.acc.inst.loc.voc:n2
mierzenia	mierzyć	ger:sg:gen:n2:imperf:aff
zrozumiałości	zrozumiałość	subst:sg:gen.dat.loc.voc:f
		subst:pl:nom.gen.acc.voc:f

Klasa trudności: 4,0

FOG-Base: 11,6

FOG-Orth: 11,6

L-Pisarek-Base: 10,67

L-Pisarek-Orth: 10,67

Klasa trudności

$$\text{Klasa} = 12.25 - 4.12 * \text{Ridge}$$

Klasa	Propozycje etykiet	Wykształcenie odbiorcy
1	Teksty łatwe w odbiorze	Klasy 1-3 szkoły podstawowej
2	Teksty bardzo czytelne	Klasy 3-6 szkoły podstawowej
3	Teksty czytelne, zrozumiałe dla przeciętnego Polaka	Gimnazjum
4	Teksty nieco trudniejsze, zrozumiałe dla osób z wykształceniem średnim	Liceum
5	Teksty trudniejsze, zrozumiałe dla ludzi wykształconych	Studia licencjackie/inżynierskie
6	Tekst trudny w odbiorze przez przeciętnego Polaka	Studia magisterskie
7	Teksty bardzo skomplikowane, teksty fachowe	Doktorat

Automatyczny test Taylora

– autorski pomysł Bartosza Brody

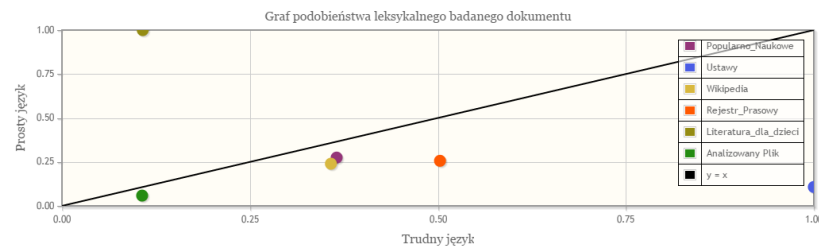


- Metoda Taylora – klasyczna metoda mierzenia czytelności poprzez uzupełnianie luk w tekście przez użytkowników języka
- Wytrenowanie modeli językowych na tekstach referencyjnych
- Uzupełnianie luk w tekście z wykorzystaniem modeli językowych
- Warianty:
 - Uzupełnianie co n-tego słowa poprzez model
 - Mierzenie odwrotności entropii (perplexity)

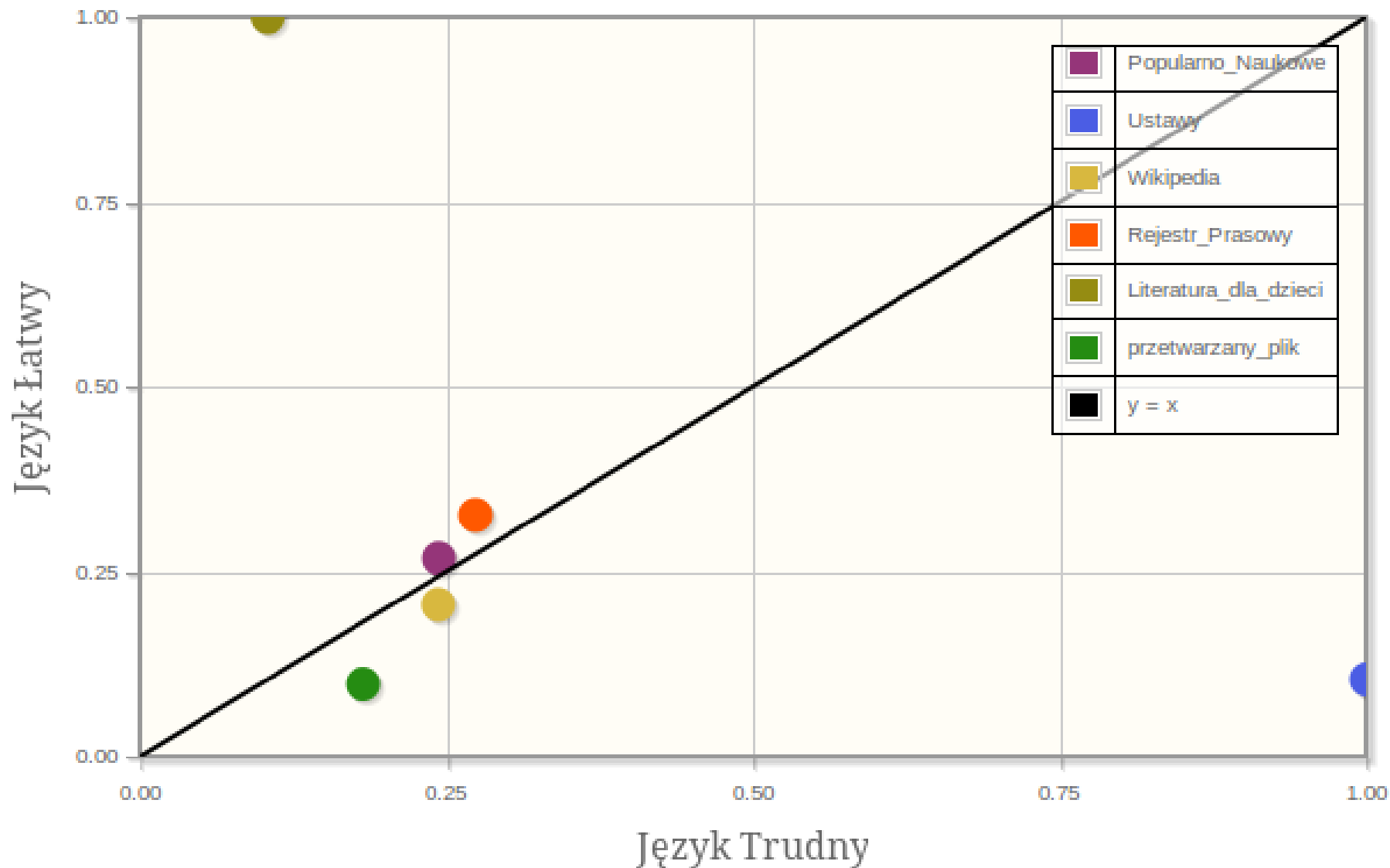
$$p(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_{w_i} c(w_{i-1}w_i)}$$

Grafy podobieństwa

- Podobieństwo pomiędzy korpusami referencyjnymi a tekstem użytkownika
- Worek słów
- Kosinus kąta pomiędzy wektorami jako miara podobieństwa
- Dwa modele porównywania tekstów
 - tf.idf
 - model binarny
- Porównanie na poziomie leksyki



Graf podobieństwa leksykalnego badanego dokumentu



Weryfikacja

- Korpusy:
 - literatura dziecięca (bajki),
 - Wikipedia,
 - artykuły prasowe (Rzeczpospolita),
 - ustawy,
 - teksty popularno-naukowe („Wiedza i Życie”)
- Ok. 40 tys. słów/korpus dla podobieństwa.
- Ok. 186 tys. słów/korpus dla automatycznego testu Taylora.
- Walidacja krzyżowa.



Weryfikacja

	Binarny	tf.idf
Literatura dla dzieci	100%	100%
Wikipedia	85,37%	85,37%
Ustawy	100%	100%
Artykuły prasowe	71,74%	73,91%
Popularno-naukowe	100%	100%

	Co 5 wyraz	Perplexity
Literatura dla dzieci	93,79%	97,18%
Wikipedia	80,56%	67,11%
Ustawy	86,29%	100%
Artykuły prasowe	71,66%	66,11%
Popularno-naukowe	73,77%	68,31%

Uwagi na marginesie Jasnopisu

- W projekcie wykorzystaliśmy ogromną część dorobku polskiej „lingwistyki komputerowej”, nie musieliśmy wywarzać otwartych drzwi.
- Było to możliwe dzięki temu, że większość aplikacji i wyników badań jest dostępna na podstawie bardzo liberalnych licencji.
- Współpraca naukowców z różnych środowisk, nie odbywała się bezproblemowo – najtrudniejsze było znalezienie wspólnego języka.

Obserwacje i wnioski końcowe

- Nieprawdopodobnie głębokie zmiany w warsztacie pracy lingwistów, jakie zaszły w ciągu życia jednego pokolenia (mojego pokolenia) dają ogromne możliwości, ale też pojawiły się w związku z nimi wyzwania i niebezpieczeństwa.
- Trudności we współpracy humanistów (zwłaszcza językoznawców) z informatykami spowodowane są zwykle brakiem wspólnego języka, a także różnicami w sposobie stawiania pytań i sposobie poszukiwania na te pytania odpowiedzi.
- Podział na informatyków (zajmujących się językiem naturalnym) i lingwistów (stosujących narzędzia informatyczne) staje się coraz bardziej anachroniczny, a rywalizacja między tymi środowiskami;
- W planowaniu i realizacji projektów z zakresu „lingwistyki cyfrowej” ważna jest współpraca różnych środowisk, dążenie do tego, by wyniki dały się połączyć we względnie spójną całość, a nie rywalizacja i robienie wszystkiego „od początku i po swojemu”.

Dziękuję za uwagę!