Od skanów do Unicode

Janusz S. Bień

Katedra Lingwistyki Formalnej
Wydział Neofilologii
Uniwersytet Warszawski

28 września 2012 r.
Dni Technologii Językowej
IPI PAN, Warszawa

Zespół

Pracownicy

- Janusz S. Bień
- Joanna A. Bilińska (formalnie od 1.10.2012)

Zespół

Pracownicy

- Janusz S. Bień
- Joanna A. Bilińska (formalnie od 1.10.2012)

Współpraca

Zespół

Pracownicy

- Janusz S. Bień
- Joanna A. Bilińska (formalnie od 1.10.2012)

Współpraca

- Narzędzia dygitalizacji tekstów …
- IMPACT — IMproving ACcess to Text
- …

Projekty

Narzędzia dygitalizacji tekstów na potrzeby badań filologicznych

Grant MNiSzW, 13.05.2009 - 12.05.2012

`https://bitbucket.org/jsbien/ndt`

- Janusz S. Bień (kierownik projektu),
- Joanna Bilińska, Krzysztof Szafran, Jakub Wilk,
- Grzegorz Chimosz, Tomasz Olejniczak, Michał Rudolf, Piotr Sikora.

IMPACT — IMproving ACcess to Text

7. PR, (1.01.2008) 1.02.2010 — 31.12.2011 (30.06.2012)

`http://www.digitisation.eu/tools/language-resources/historical-lexicon-polish/`

- Janusz S. Bień (kierownik zespołu),
- Krzysztof Szafran, Monika Kresa

Narzędzia i metody

DjVu

Format i techniki
do reprezentacji dokumentów
przy pomocy warstwy graficznej i tekstowej
z metadanymi i adnotacjami

Narzędzia i metody

DjVu

Format i techniki
do reprezentacji dokumentów
przy pomocy warstwy graficznej i tekstowej
z metadanymi i adnotacjami
oraz do efektywnego udostępniania
tak reprezentowanych dokumentów w Internecie.
[JSB]

Narzędzia i metody

DjVu

Format i techniki
do reprezentacji dokumentów
przy pomocy warstwy graficznej i tekstowej
z metadanymi i adnotacjami
oraz do efektywnego udostępniania
tak reprezentowanych dokumentów w Internecie.
[JSB]

- Warstwa tekstowa — czysty tekst w Unicode
- Warstwa graficzna — wyrafinowane metody kompresji
  - tło (*background*)
  - zadruk (*foreground*) — słowniki kształtów

Narzędzia i metody

The Library 2.012 worldwide virtual conference

October 4, 2012, 12-13 CET

Janusz S. Bień

Scanned publications in digital libraries:
new Open Source DjVu tools

`http://bc.klf.uw.edu.pl/298/`

`http://www.library20.com/forum/topics/scanned-publications-in-digital-libraries-new-open-source-djvu`

Narzędzia i metody

Tworzenie dokumentów DjVu — Jakub Wilk

- pdf2djvu
  Debian+Ubuntu popcon installed/votes: $\sim$ 45 000/1000
- didjvu
  Debian+Ubuntu popcon installed/votes: $\sim$ 200/20
- ocrodjvu
  Debian+Ubuntu popcon installed/votes: $\sim$ 2 000/100
- djvusmooth
  Debian+Ubuntu popcon installed/votes: $\sim$ 1 500/80

Narzędzia i metody

Korpusy DjVu

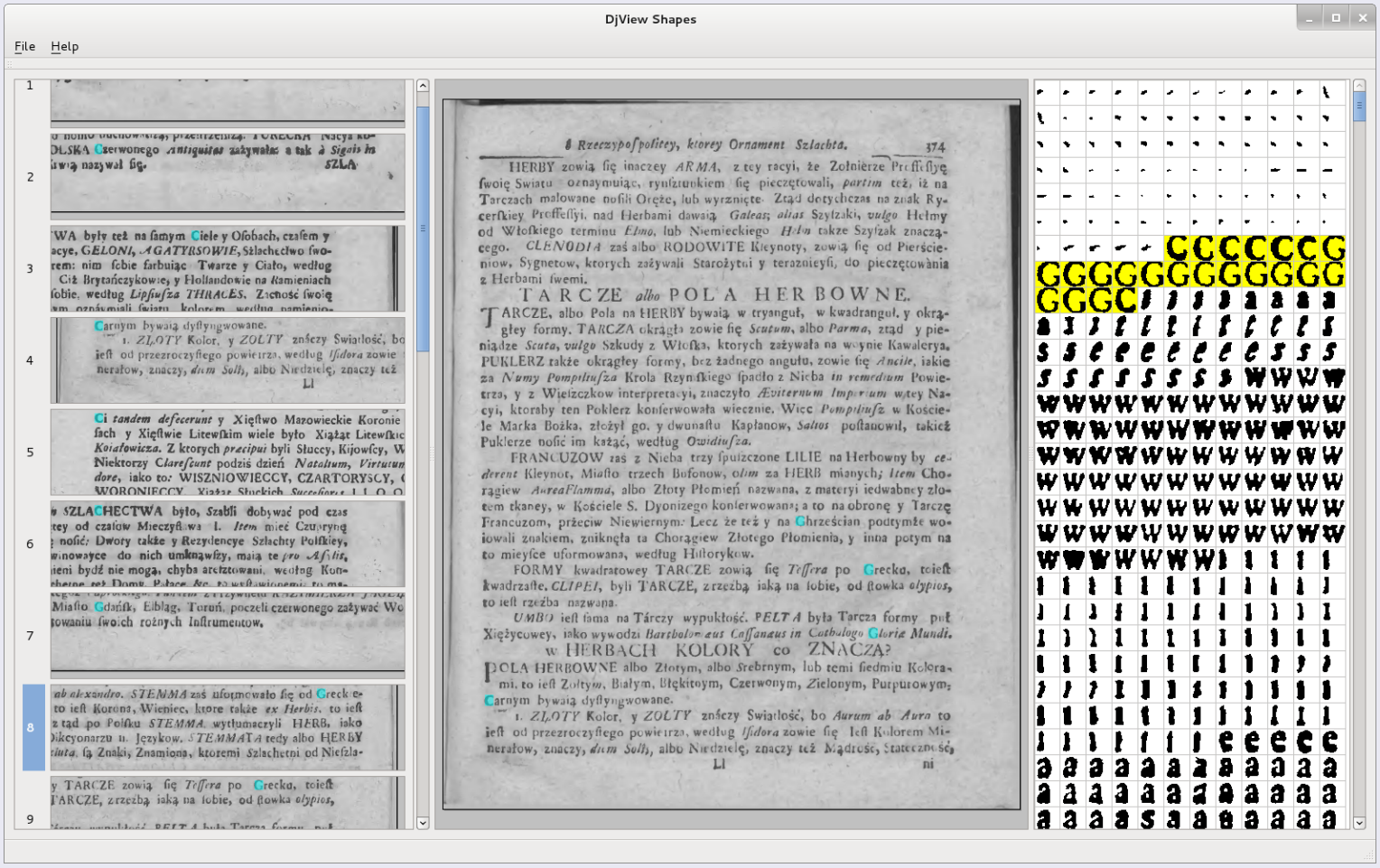
- Poliqarp for DjVu — serwer
- marasca — klient WWW
- Djview for Poliqarp — klient zdalny

Analiza kształtów zadruku

- Lokalna przeglądarka kształtów wspólnych
- Narzędzia typu klient-serwer
  - eksporter
  - etykieciarka

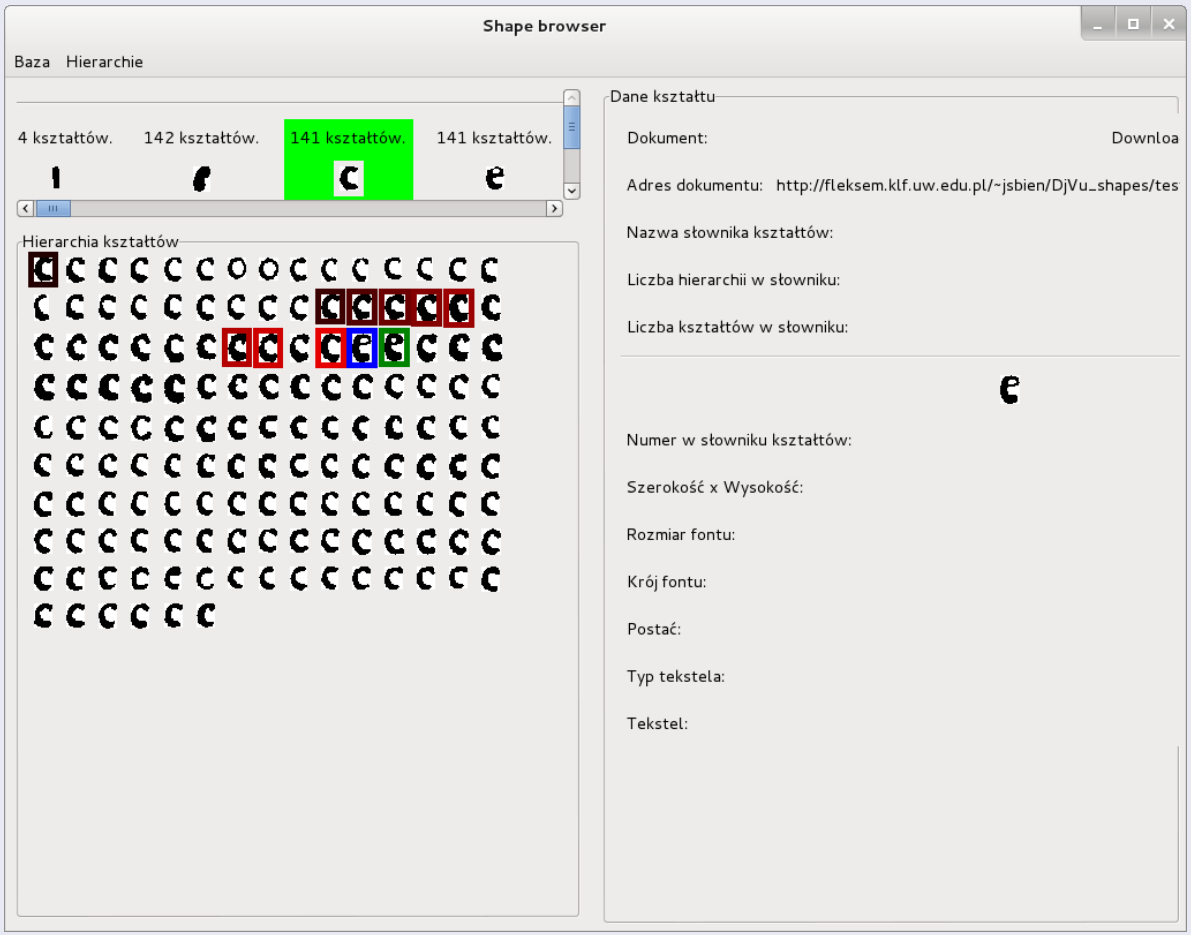
Lokalna przeglądarka kształtów

Wykaz wystąpień kształtów podobnych do litery *C*



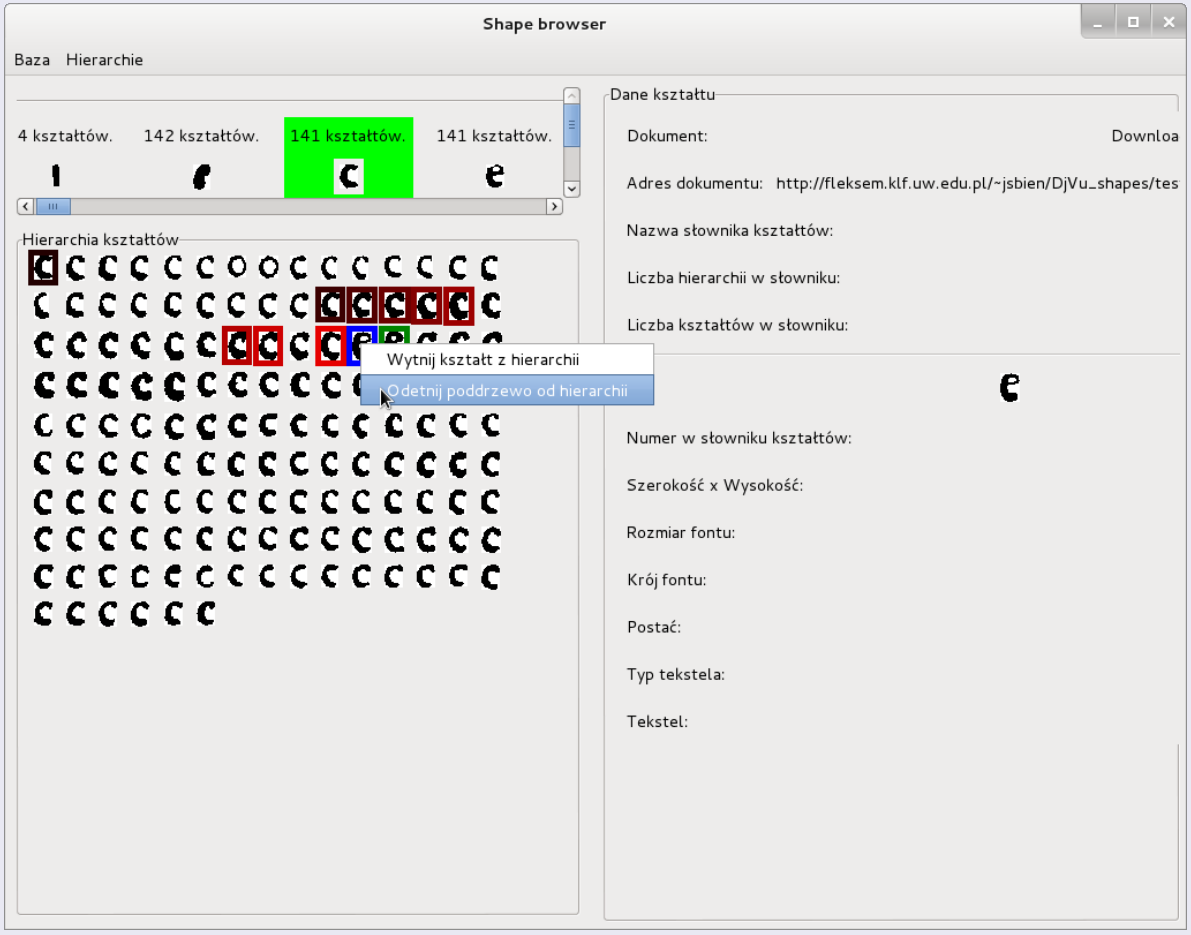
Analiza kształtów

Hierarchia kształtów litery *c* z „zabłąkaną” literą *e*



Analiza kształtów

Eliminacja „zabłąkanej” litery *e*



Przykłady korpusów DjVu

`http://poliqarp.wbl.klf.uw.edu.pl`

Słownik polszczyzny XVI wieku (wychodzi od 1966 r.)

Ukazało sie 35 tomów (*A* do *ROWNY*), razem około 19 000 stron

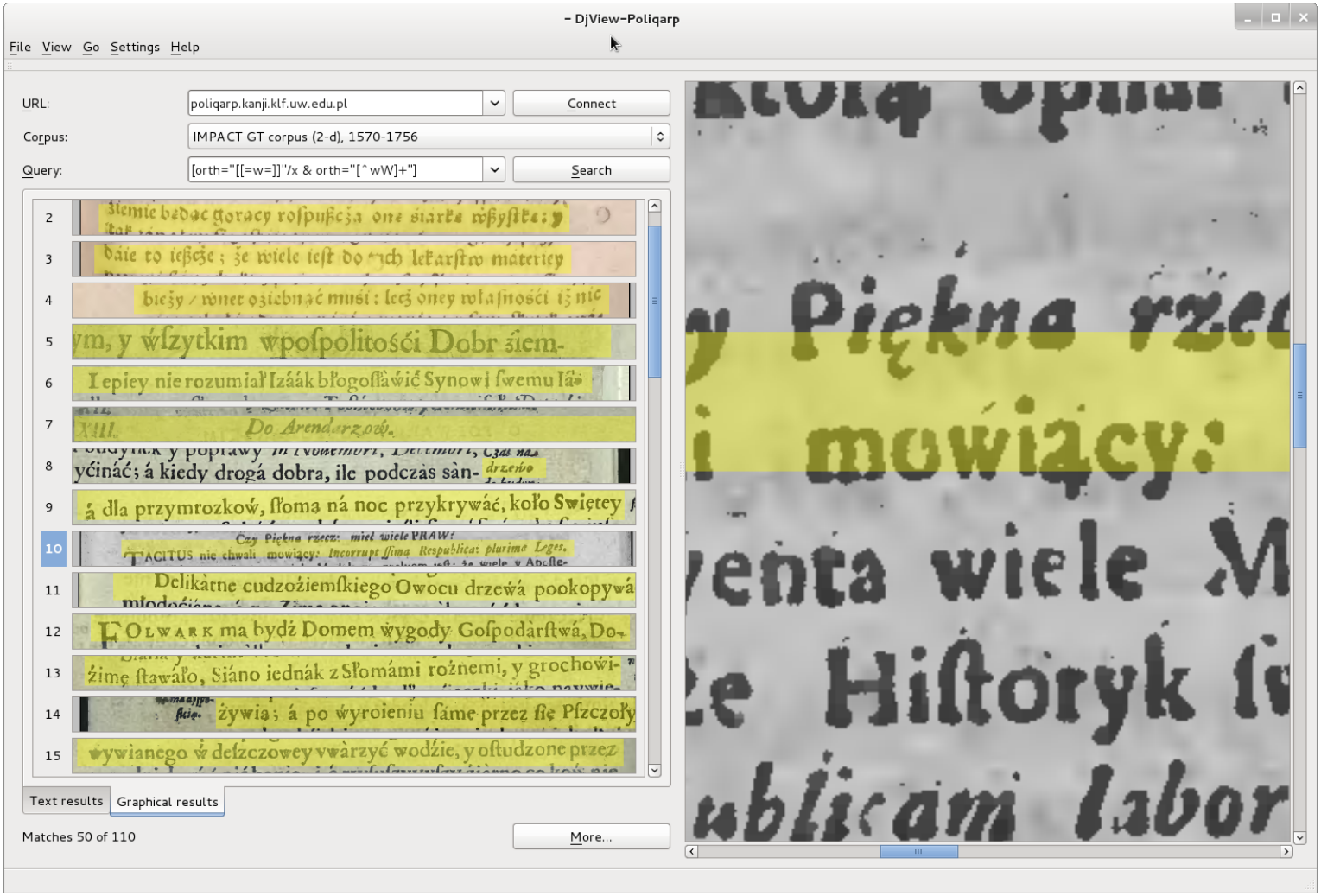
Korpus tekstów wzorcowych projektu IMPACT

Prawie 5 000 stron tekstów z lat 1570–1756

Słownik Lindego

Pierwsze wydanie 1807–1814, drugie wydanie 1854-1861,
razem około 4 500 stron.

Przykład kwerendy — korpus IMPACT

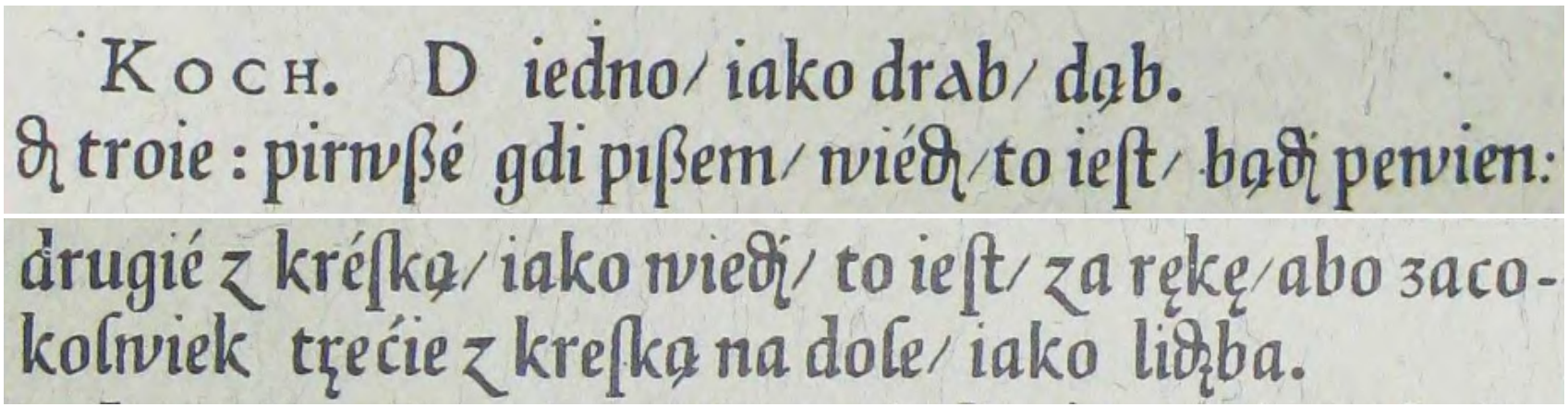


Przykłady tekstów — Słownik polszczyzny XVI wieku

[*3 r.*]; [*przyklady na konsonantes*] ą Cuąv ſynowie zá-
ſtáṗeli ſie. Wuoą moi/ i znáiomy moi. *JanNKarGórn* G4,
G3v; D iedno/ iako drab/ dąb. ą troie: pirwſzé gdi piſzém/
wiéą/ to ieſt/ bąą pewien: drugié z kréſką/ iako wieą/
to ieſt/ za rękę/ abo zacokolwiek, tṗećie z kreſką na dole/

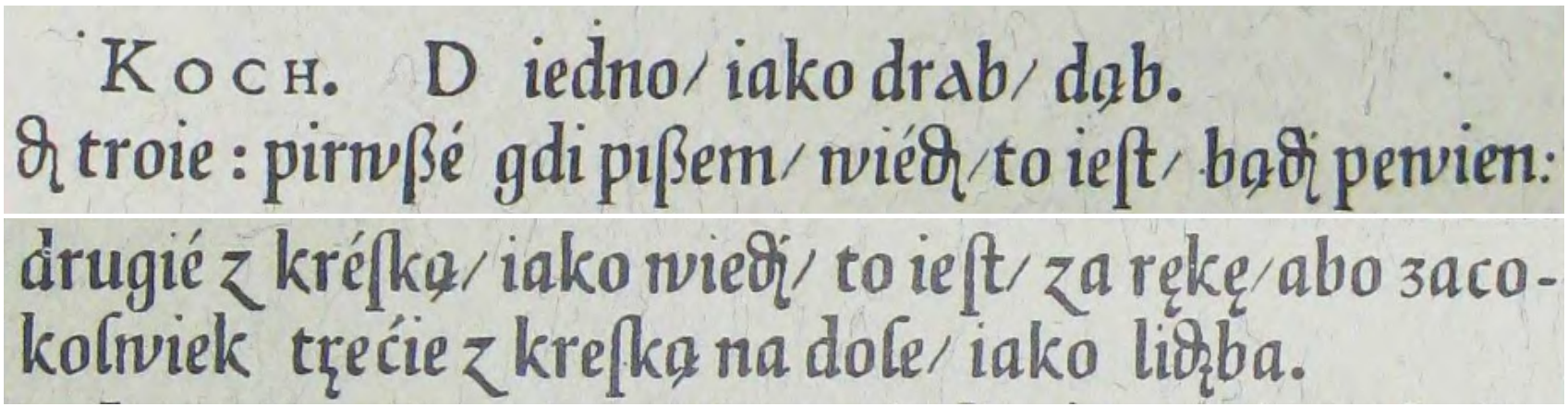
Przykłady tekstów — Słownik polszczyzny XVI wieku

[3 r.]; [*przyklady na konsonantes*] ą Cudzy ſynowie zá-
ſtáṗeli śie. Wuoą moi/ i znáiomy moi. *JanNKarGórn* G4,
G3v; D iedno/ iako drab/ dąb. ą troie: pirwſzé gdi piſzém/
wiéą/ to ieſt/ bąą pewien: drugié z kréſką/ iako wieą/
to ieſt/ za rękę/ abo zacokolwiek, tṗećie z kreſką na dole/


Koch. D iedno/ iako drab/ dąb.
ą troie: pirwßé gdi pißem/ wiéą/ to ieſt/ bąą pewien:
drugié z kréſką/ iako wieą/ to ieſt/ za rękę/ abo zaco-
kolwiek tṗećie z kreſką na dole/ iako liḋba.

Przykłady tekstów — Słownik polszczyzny XVI wieku

[3 r.]; [*przyklady na konsonantes*] ą Cudzy ſynowie zá-
ſtáṗeli śie. Wuoą moi/ i znáiomy moi. *JanNKarGórn* G4,
G3v; D iedno/ iako drab/ dąb. ą troie: pirwſzé gdi piſzém/
wiéą/ to ieſt/ bąą pewien: drugié z kréſką/ iako wieą/
to ieſt/ za rękę/ abo zacokolwiek, tṗećie z kreſką na dole/


Koch. D iedno/ iako drab/ dąb.
ą troie: pirwßé gdi pißem/ wiéą/ to ieſt/ bąą pewien:
drugié z kréſką/ iako wieą/ to ieſt/ za rękę/ abo zaco-
kolwiek tṗećie z kreſką na dole/ iako liḋba.

ligatura *dz*, *r* z ogonkiem, …

Przykłady tekstów — korpus IMPACT

Przykłady tekstów — korpus IMPACT

Przykłady tekstów — korpus IMPACT

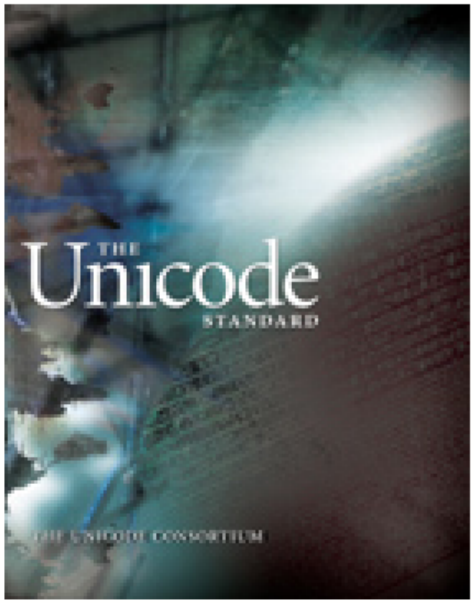
Przykłady tekstów — korpus IMPACT

Co to jest Unicode?

http://www.unicode.org/standard/WhatIsUnicode.html

What is Unicode?

Unicode provides a unique number for every character,
no matter what the platform,
no matter what the program,
no matter what the language.



Co to jest znak (character)?

Terminologia

(abstract) character = znak (piśmienny)

Co to jest znak (character)?

Terminologia

(abstract) character = znak (piśmienny)
(encoded) character = (piśmienny) znak kodowy

Co to jest znak (character)?

Terminologia

(abstract) character = znak (piśmienny)
(encoded) character = (piśmienny) znak kodowy
code point = współrzędna kodowa

Co to jest znak (character)?

Terminologia

(abstract) character = znak (piśmienny)
(encoded) character = (piśmienny) znak kodowy
code point = współrzędna kodowa

Unicode 6.2.0

Release date: 2012, September 26
`http://www.unicode.org/versions/Unicode6.2.0/`
[…] the total number of characters assigned in the standard […]
110,117

Co to jest znak (character)?

Terminologia

(abstract) character = znak (piśmienny)
(encoded) character = (piśmienny) znak kodowy
code point = współrzędna kodowa

Unicode 6.2.0

Release date: 2012, September 26
`http://www.unicode.org/versions/Unicode6.2.0/`
[…] the total number of characters assigned in the standard […]
110,117
(That is the traditional count

Co to jest znak (character)?

Terminologia

(abstract) character = znak (piśmienny)
(encoded) character = (piśmienny) znak kodowy
code point = współrzędna kodowa

Unicode 6.2.0

Release date: 2012, September 26
`http://www.unicode.org/versions/Unicode6.2.0/`
[…] the total number of characters assigned in the standard […]
110,117
(That is the traditional count, which totals up graphic and format characters

Co to jest znak (character)?

Terminologia

(abstract) character = znak (piśmienny)
(encoded) character = (piśmienny) znak kodowy
code point = współrzędna kodowa

Unicode 6.2.0

Release date: 2012, September 26
`http://www.unicode.org/versions/Unicode6.2.0/`
[…] the total number of characters assigned in the standard […]
110,117
(That is the traditional count, which totals up graphic and format characters, but omits surrogate code points, ISO control codes, noncharacters, and private-use allocations.)

Co to jest znak (character)?

Janusz S. Bień, 2004

http://bc.klf.uw.edu.pl/114/

[…] znaki piśmienne to pojęcie pierwotne, zdefiniowane przez wyliczenie.

Co to jest znak (character)?

Janusz S. Bień, 2004

`http://bc.klf.uw.edu.pl/114/`

[…] znaki piśmienne to pojęcie pierwotne,
zdefiniowane przez wyliczenie.

(piśmienne) znaki kodowe to pojęcie pierwotne,
zdefiniowane przez wyliczenie

Co to jest znak (character)?

Janusz S. Bień, 2004

`http://bc.klf.uw.edu.pl/114/`
[…] znaki piśmienne to pojęcie pierwotne,
zdefiniowane przez wyliczenie.

(piśmienne) znaki kodowe to pojęcie pierwotne,
zdefiniowane przez wyliczenie

Przykłady znaków kodowych

Co to jest znak (character)?

Janusz S. Bień, 2004

`http://bc.klf.uw.edu.pl/114/`

[…] znaki piśmienne to pojęcie pierwotne, zdefiniowane przez wyliczenie.

(piśmienne) znaki kodowe to pojęcie pierwotne, zdefiniowane przez wyliczenie

Przykłady znaków kodowych

Co to jest znak (character)?

Janusz S. Bień, 2004

`http://bc.klf.uw.edu.pl/114/`

[…] znaki piśmienne to pojęcie pierwotne,
zdefiniowane przez wyliczenie.

(piśmienne) znaki kodowe to pojęcie pierwotne,
zdefiniowane przez wyliczenie

Przykłady znaków kodowych



Co to jest znak (character)?

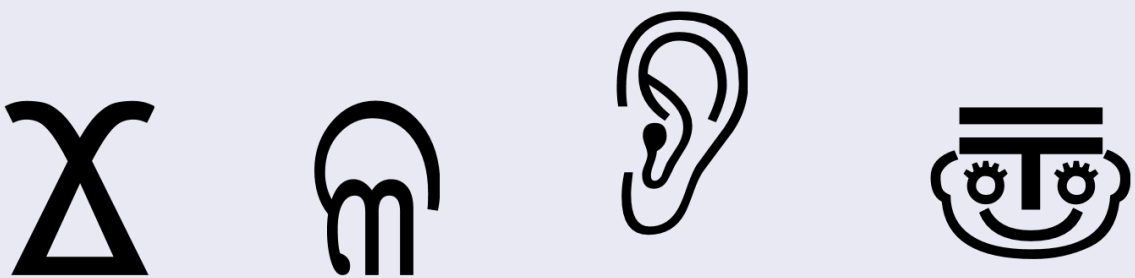
Janusz S. Bień, 2004

`http://bc.klf.uw.edu.pl/114/`

[…] znaki piśmienne to pojęcie pierwotne,
zdefiniowane przez wyliczenie.

(piśmienne) znaki kodowe to pojęcie pierwotne,
zdefiniowane przez wyliczenie

Przykłady znaków kodowych



Co to jest znak (character)?

Janusz S. Bień, 2004

`http://bc.klf.uw.edu.pl/114/`

[…] znaki piśmienne to pojęcie pierwotne,
zdefiniowane przez wyliczenie.

(piśmienne) znaki kodowe to pojęcie pierwotne,
zdefiniowane przez wyliczenie

Przykłady znaków kodowych

Co to jest znak (character)?

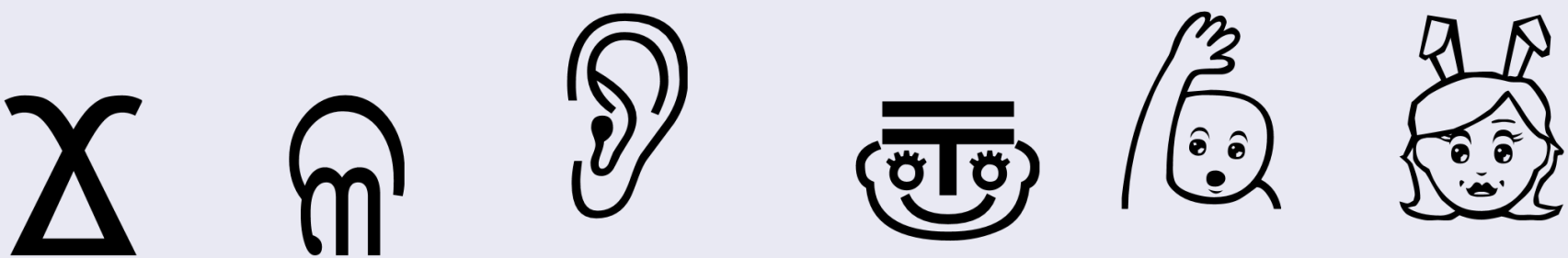
Janusz S. Bień, 2004

`http://bc.klf.uw.edu.pl/114/`

[…] znaki piśmienne to pojęcie pierwotne,
zdefiniowane przez wyliczenie.

(piśmienne) znaki kodowe to pojęcie pierwotne,
zdefiniowane przez wyliczenie

Przykłady znaków kodowych



Co to jest znak (character)?

Tekstony [JSB]: współrzędne kodowe

- ’LATIN SMALL LETTER N’ (U+006E)
- ’COMBINING ACUTE ACCENT’ (U+0301)
- ’LATIN SMALL LETTER N WITH ACUTE’ (U+0144)

Co to jest znak (character)?

Tekstony [JSB]: współrzędne kodowe

- ’LATIN SMALL LETTER N’ (U+006E)
- ’COMBINING ACUTE ACCENT’ (U+0301)
- ’LATIN SMALL LETTER N WITH ACUTE’ (U+0144)

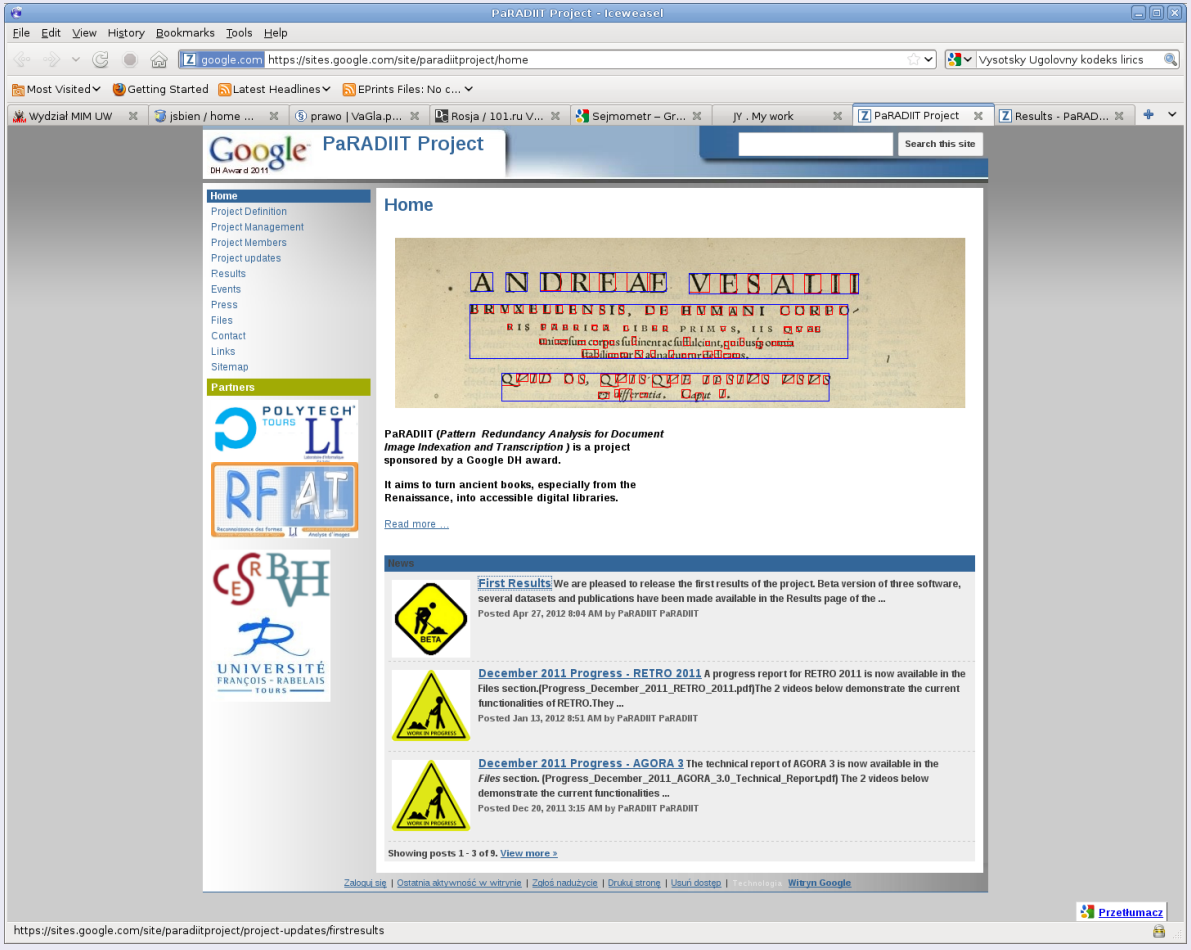
Tekstele [JSB]: klasy równoważności

«LATIN SMALL LETTER N WITH ACUTE»

- U+006E, U+0301
- U+0144

Prace planowane

Analiza podobnych projektów



Prace planowane

Analiza podobnych projektów

http://www.nauka.gov.pl/fileadmin/user_upload/Nauka/NPRH/20111027_NPRH_modul_1-1.pdf

Narodowy Program Rozwoju Humanistyki
Moduł 1.1 pozycja 138

Grant 11H 11 023180

Korpus polszczyzny XVI wieku. Etap I:
Digitalizacja źródeł oraz stworzenie narzędzi informatycznych
i udostępnienie materiałów testowych korpusu

Instytut Badań Literackich PAN,

dr Patrycja Potoniec
[mgr Krzysztof Opaliński]

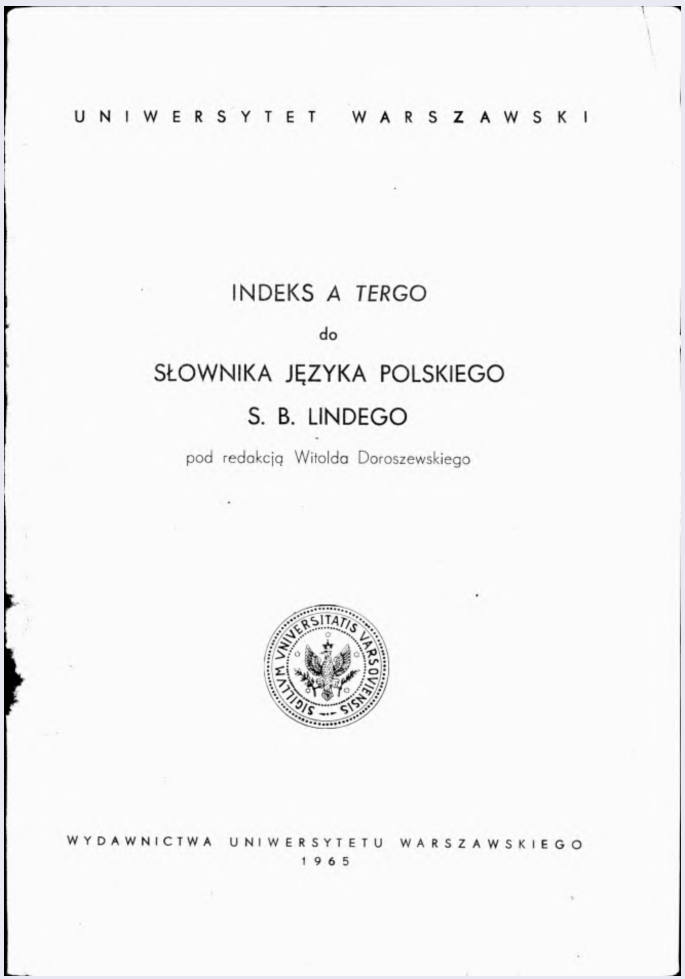
1 239 614 zł

Słownik Lindego (1807-1814 [2. wyd. 1854-1861])

AGREST, u. m. a) jakiekolwiek jagody kwaśne niedojźrzałe ſauere unreiſe Trauben und Beeren. *Carn.* agres; *Graec.* ἀχράς, *pyrus silvestris*, *Lat.* agresta, vinum acre; *Gall.* aigras; Bosn. egresc, ogresta, gresc; *Ital.* agresta = uva acris, acerba, (*Bosn.* zagresciti, zakisseliti = acrefacere); *Croat.* jegrist; *Dalm.* egrist, gres, ogresta, vinika; *Hung.* egres; (*Ross.* пародокъ). Czekał aby ziemia zrodziła jagody winne alić zrodziła agrest. 1. *Leop. Jes.* 5, 2, (3. *Leop.* płonki). Agrest, to jest wino dzikie albo leśne. *Urs. Gr.* 132. b) sok z takich jagód wytłoczony, der aus ſolchen unreifen Beeren gepreßte herbe Wein. Z niedojrzałych winnych jagód wytłacza sie sok agrest *(Verjus)* cierpki i kwaśny, do przypraw w kuchni. *Kl. Dyk.* 3, 168. *Zaw. Gos. Sleszk. Ped.* 407. *Sien.* 187. *Cresc.* 298. *Spicz.* 96. — 2) Agrest, krzak i owoc tego krzaku, ribes grossularia *Linn. Carn.* agres, berberion agresove; *Vind.* agress, oistniza, kosmatizhi, kosmazhizhki, kuseji; *Boh.* angresst, srstka; *Ross.* крыжовникъ; die Stachelbeere, der Stachelbeerſtrauch, (Öſter. Agras). *Kluk. Dyk. Jundz.* AGRESTOWY, a, e, z agrestu, od agrestu, 1) Agreſt = (ſauer). Wino agrestowe cierpkie. *Spicz.* 96. 2) jagody agrestowe w potrawach bywają używane, *ribes. Kl. Ros.* 1, 151. Stachelbeeren. =

Prace planowane

Indeks *a tergo* do słownika Lindego



UNIWERSYTET WARSZAWSKI

INDEKS *A TERGO*

do

SŁOWNIKA JĘZYKA POLSKIEGO

S. B. LINDEGO

pod redakcją Witolda Doroszewskiego

WYDAWNICTWA UNIWERSYTETU WARSZAWSKIEGO
1965

-ewać [100] -iwać

podwiewać
przewiewać
chwiewać
*zachwiewać
ochwiewać
*wychwiewać
*rozchwiewać
owiewać
dowiewać
powiewać
*wwiewać
zob. wwiać
*wywiewać
przywiewać
zwiewać
rozwiewać
wzwiewać
*ziewać
*zaziewać
dziewać
nadziewać
*zadziewać
zob. zadziać
odziewać
od-ziewać
naodziewać
dodziewać
*przeodziewać
podziewać
*zapodziewać
zob. zapodziać
*spodziewać
zob. spodziać
*rozpodziewać
*przyodziewać
wdziewać
*zawdziewać
zob. zawdziać
*przewdziewać
powdziewać
*przywdziewać
*wydziewać
zdziewać
nazdziewać
*rozdziewać
wzdziewać
*przeziewać
*wyziewać
nalewać
ponalewać
zalewać
*oblewać

nadlewać
*namdlewać
*zemdlewać
*omdlewać
omglewać
poomdlewać
odlewać
podlewać
popodlewać
przelewać
nadtruchlewać
bolewać
*zabolewać
pobolewać
ubolewać
spółubolewać
dolewać
pooplewać
polewać
*zatlewać
ulewać
wlewać
wylewać
przylewać
zlewać
rozlewać
*wzlewać
zob. wzlać
*zagorewać
zob. zagorzeć
*przygorewać
*rozgorewać
zob. rozgorzeć
zawrewać
nagrzewać
zagrzewać
*odgrzewać
*podgrzewać
*przegrzewać
ogrzewać
*dogrzewać
*pogrzewać
*wygrzewać
*przygrzewać
*zgrzewać
*rozgrzewać
*wzgrzewać
zob. wzgrzać
*pomrzewać
*zaźrzewać
*przeźrzewać

*dojźrzewać
dojrzewać
*dozrzewać
dozdrzewać
uźrzewać
*wyźrzewać
*wygaiwać
łaiwać
*odłaiwać
*wyłaiwać
*utaiwać
*zaszpeciwać
*zachciwać
*zażółciwać
(przekrzciwać)
miłościwać
mościwać
*pościwać
*dokleiwać
*wykleiwać
*wysmagiwać
naciągiwać
*zaciągiwać
*zawściągiwać
naurągiwać
biegiwać
*zabiegiwać
zob. zabiedz
*dobiegiwać
*ubiegiwać
*wybiegiwać
*przybiegiwać
*zbiegiwać
*rozbiegiwać
*wysięgiwać
*wyłęgiwać
*wyłągiwać
zob. wyłądz
*wypręgiwać
zob. wyprężyć
*zaprzęgiwać
zob. zaprządz
*wyprzęgiwać
*wyprzągiwać
zob. wyprządz
poodździwigiwać
*wydźwigiwać
*zalgiwać
zob. zalgnąć
*wylgiwać
*zelgiwać
*wyczołgiwać

*odtargiwać
*podtargiwać
poprzetargiwać
*wytargiwać
powytargiwać
*roztargiwać
*sługiwać
zob. sługować
*nasługiwać
*zasługiwać
zob. zasłużyć
nadsługiwać
*odsługiwać
*dosługiwać
*posługiwać
*przysługiwać
*zamrugiwać
*zastrugiwać
*przestrugiwać
*wystrugiwać
*wydrygiwać
*wystrzygiwać
*podwierzgiwać
*wywierzgiwać
*wybryzgiwać
*wykichiwać
*zasłuchiwać
zob. zasłychnąć
*dosłuchiwać
*wysłuchiwać
poruchiwać
kiwać
naskakiwać
zaskakiwać
obskakiwać
nadskakiwać
*odskakiwać
zob. odskoczyć
podskakiwać
*zeskakiwać
*przeskakiwać
*doskakiwać
*uskakiwać
*wskakiwać
*wyskakiwać
powyskakiwać
*przyskakiwać
*zaplakiwać
*odplakiwać
*przeplakiwać
*oplakiwać

Prace planowane

Joanna Bilińska

Describing Linde's Dictionary of Polish for Digitalisation Purposes. In: Electronic lexicography in the 21st century: new applications for new users (eLEX2011), 10-12.11.2011, Bled, Slovenia
http://bc.klf.uw.edu.pl/216/

Joanna Bilińska

Составление перечня сокращенных названий языко
в в рамках проекта дигитализации
«Словаря польского языка» С.Б.Линде
Информационные технологии и письменное наследие,
El'Manuscript-12, Petrozawodsk (Rosja), 3-8 września 2012 r.
http://bc.klf.uw.edu.pl/301/

Dygitalizacje

http://eprints.wbl.klf.uw.edu.pl/view/creators/Linde=3ASamuel_Bogumi==0142=3A=3A.html
http://eprints.wbl.klf.uw.edu.pl/61/

Kontakt

jsbien@uw.edu.pl
jsbien@mimuw.edu.pl

nmpt-ann@mimuw.edu.pl
http://lists.mimuw.edu.pl/listinfo/nmpt-ann
nmpt-l@mimuw.edu.pl
http://lists.mimuw.edu.pl/listinfo/nmpt-l