# Printed-Source Analysis for PanLex

Jonathan Pool
The Long Now Foundation

PanLex



http://panlex.org

First PanLex Summer Internship

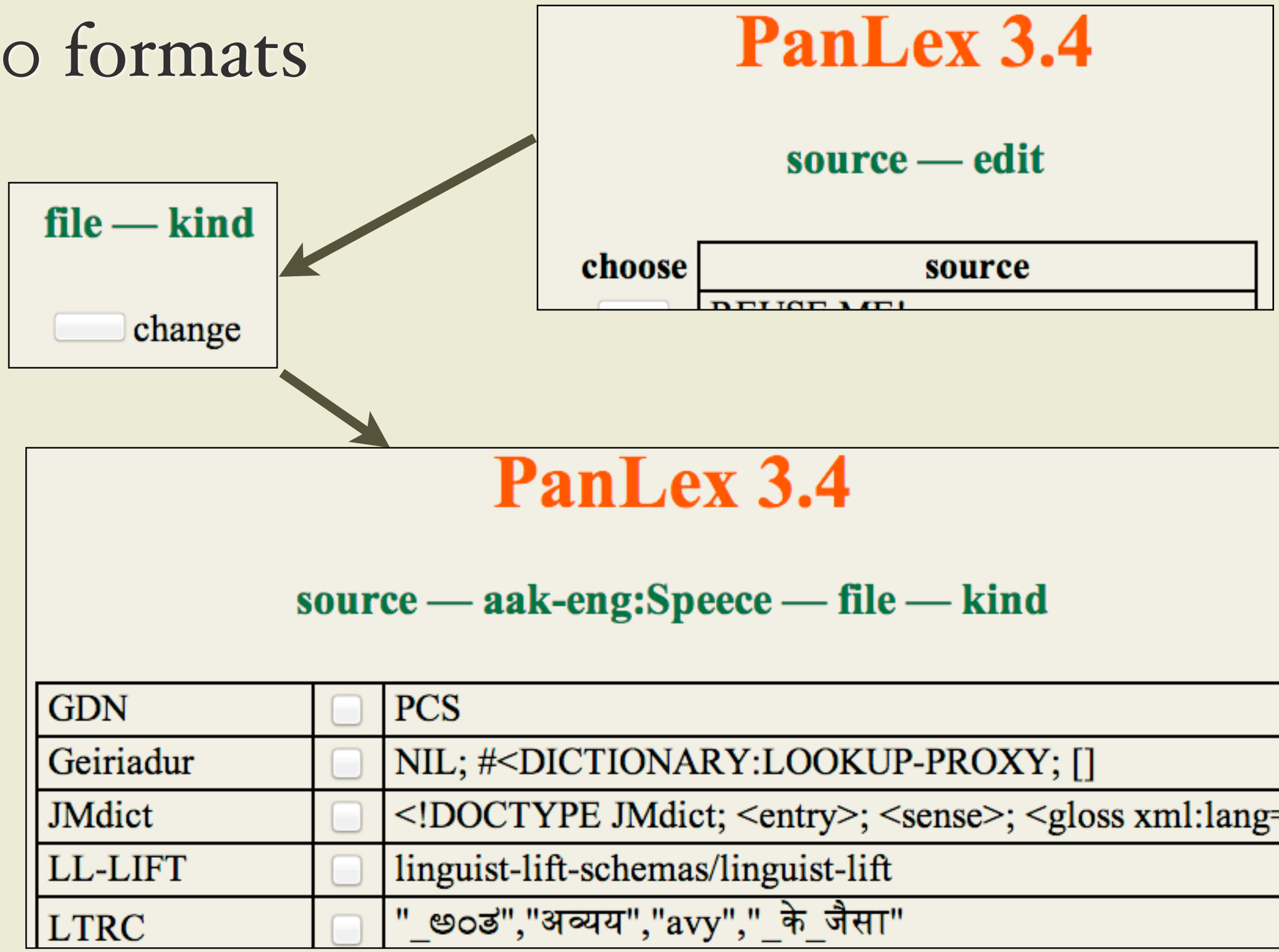
29 July 02013

# Summary

- Printed sources
- PanLex collection at Internet Archive
- Analytical tasks
- Tools
- Opportunities

# Printed sources

70 formats



# Printed sources

## Printed formats

| | | |
|---|---|---|
| pdf | ☐ | |
| pdf-img | ☐ | |
| pdf-lock/encrypt | ☐ | |
| png | ☐ | |
| po | ☐ | |
| popdict | ☐ | |
| prc | ☐ | |
| prt | ☐ | ✏ @ ⊕ |
| prt@Plx | ☐ | ✏ |
| ps | ☐ | |

# Printed sources

## Printed sources in the source archive

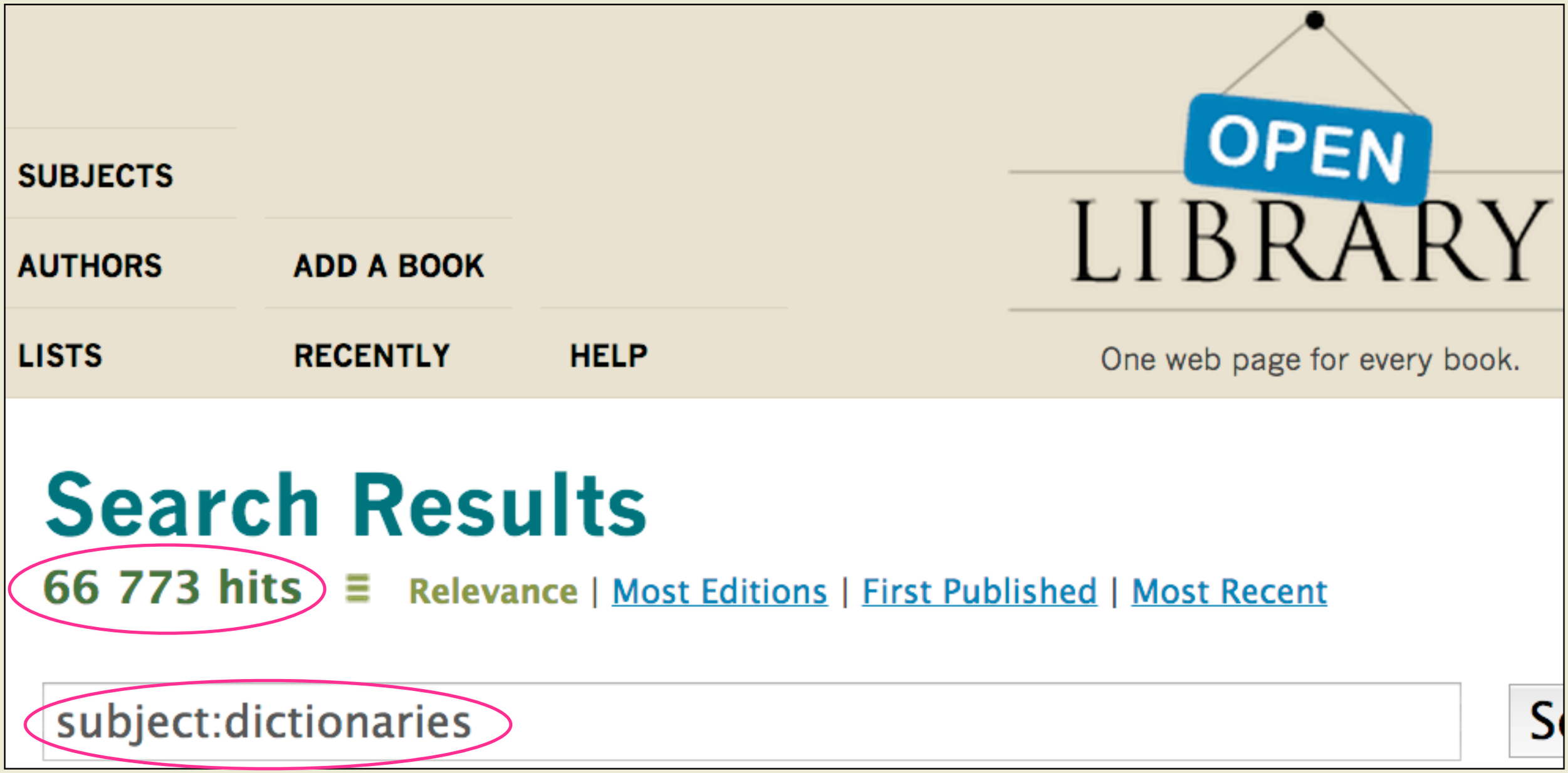
```
plx=# select count (ap) from ap;
 count
-------
  3993
```

274 printed-only

```
plx=# select count (ap) from
(select ap from ap except select
af.ap from af, fm where fm.fm =
af.fm and (tt like 'pdf-%' or tt
like 'prt%')) as tbl;
 count
-------
  3719
```

# Printed sources

## Printed sources in the world



# Printed sources

## Printed sources in the world


# Internet Archive PanLex collection

# Internet Archive PanLex collection

- A comparative vocabulary of Formosan languages and dialects - Ogawa, Naoyoshi
  **Keywords:** Austronesische talen
  **Downloads:** 2
- The Chimwiini lexicon exemplified - Kisseberth, Charles W
  Includes bibliographical references (p. xxxiv)
  **Keywords:** Mwini dialect; Swahili language
  **Downloads:** 3
- A classified vocabulary of the Punu language - 880-01 Yukawa, Yasutoshi, 1941-
  Includes index
  **Keywords:** Punu language; Punu language; Punu language; English language; French language; Punu (African people); Punu
  **Downloads:** 1
- Degema-English dictionary with English index - Kari, Ethelbert E
  Includes bibliographical references (p. xliii-xliv) and index
  **Keywords:** Degema language; English language; Lengua degema; Lengua inglesa; Lengua degema; Lengua inglesa
  **Downloads:** 2
- EG-Woerterbuch mathematischer Begriffe = EK-vortaro de mathematikaj terminoj - Hilgers-Yashovardhan, R. edt
  **Keywords:** Mathematik; Europäische Gemeinschaft; Mathematik; Wissenschaftliche Bibliothek
  **Downloads:** 1
- Gendai Uigurugo shojiten = Concise modern Uyghur-Japanese dictionary - 880-01 Sugawara, Jun, 1966-
  Parallel title also in Uighur
  **Keywords:** Uighur language; Uigurugo
  **Downloads:** 4
- 『ヒンディー語動詞基礎語彙集』 = Hindi - Japanese dictionary of selected verbs with illustrative sentences - 町田和彦/Kazuhiko Machida
  **Downloads:** 4
- A Jita vocabulary - Kagaya, Ryohei
  **Keywords:** Jita language; Jita language; Jita language; English language; Japanese language; Jita; Engels; Japans
  **Downloads:** 2
- Guyish-i Gilaki-i Lahijan : vazhahnamah va parahi vizhagiha-yi avayi va sakhtavazhahi - Jahangiri, Nader
  jild-i 1. Vizhagiha-yi avayi va sakhtavazhahi -- jild-i 2. Vazhagan-i Alif ta Sin -- jild-i 3. Vazhagan-i Shin ta Ya
  **Keywords:** Gilaki language; Gilaki language; Gilaki; Gilaki language
  **Downloads:** 1
- Gendai Chibettogo doshi jiten : Rasa hogen = A verb dictionary of the modern spoken Tibetan of Lhasa, Tibetan-Japanese - 880-01 Hoshi, Izumi
  Dictionary entries in Tibetan script with romanization
  **Keywords:** Tibetan language
  **Downloads:** 1
- Kainanto hogen kiso goishu - 880-01 Nakajima, Motoki, 1942-
  Includes bibliographical references (p. 8) and index
  **Keywords:** Chinese language
  **Downloads:** 1
- AA諸言語教育基本語彙表（入門期の学習に必要な基礎語彙600項目試案） - 梅田博之
  **Downloads:** 4
- The Pyen (or Phen) language : its classified lexicon - 880-01 Shintani, Tadahiko, 1946-
  **Downloads:** 3

# Internet Archive PanLex collection

## Gendai Uigurugo shojiten = Concise modern Uyghur-Japanese dictionary (2009)

**Title (alternate script):** = Concise modern Uyghur-Japanese dictionary
**Author:** 880-01 Sugawara, Jun, 1966-
**Subject:** Uighur language; Uigurugo
**Publisher:** Fuchu-shi : Tokyo Gaikokugo Daigaku Ajia Afurika Gengo Bunka Kenkyujo
**Language:** Japanese; Uighur
**Digitizing sponsor:** Jonathan Pool
**Book contributor:** The Long Now Foundation, PanLex Project
**Collection:** lendinglibrary; browserlending; panlex; longnow; americana
**Notes:** some content may be lost due to the binding of the book.

**Full catalog record:** MARCXML

This book has an editable web page on Open Library.

### Description

Parallel title also in Uighur

Includes bibliographical references (p. xxii-xxiv)

# Internet Archive PanLex collection

**Selected metadata**

| | |
|---|---|
| **Isbn:** | 9784863370265 |
| **Lccn:** | 2010428072 |
| **Page-progression:** | lr |
| **Scanningcenter:** | sanfrancisco |
| **Mediatype:** | texts |
| **Shiptracking:** | y0599 |
| **Identifier:** | gendaiuigurugosh008800 |
| **Scanner:** | scribe5.sanfrancisco.archive.org |
| **Ppi:** | 500 |
| **Camera:** | Canon EOS 5D Mark II |
| **Operator:** | associate-lan-zhu@... |
| **Scandate:** | 20130408202946 |
| **Republisher:** | associate-karina-martinez@... |
| **Imagecount:** | 788 |
| **Identifier-access:** | http://archive.org/details/gendaiuigurugosh008800 |
| **Identifier-ark:** | ark:/13960/t8ff5b645 |
| **Bookplateleaf:** | 0004 |
| **Ocr:** | ABBYY FineReader 8.0 |
| **Sponsordate:** | 20130430 |

# Internet Archive PanLex collection

## PanLex 3.4

**source — 3317 — jpn-uig:菅原**

| | |
|---|---|
| number | 3317 |
| PanLex — beginning | 2011-04-25 |
| name | jpn-uig:菅原 |
| World Wide Web | http://www.aa.tufs.ac.jp/en/publications/lexicon |
| ISBN | 9784863370265 |
| author | 菅原純(編) |
| title | 『現代ウイグル語小辞典』 = Concise modern Uyghur-Japanese dictionary |
| publisher | アジア・アフリカ言語文化研究所 |
| year | 2009 |
| person — good | 6 |
| person — number | |
| fact — other | アジア・アフリカ基礎語彙集/Asian and African Lexicon, 53 (B026) |
| permission — kind | na |
| right — text | |
| right — person — name | |
| permission — email — address | |

| | |
|---|---|
| good — number — edit — necessary — 1/0 | 0 |
| file — difficult | 8 |
| file — submit — necessary — 1/0 | 1 |
| expression — edit — necessary — 1/0 | 0 |
| expression — edit — done — language — (aaa,bbb,ccc) | |
| file — address | jpn-uig-菅原 |
| fact — other | book donated by ILCAA |

translation — count

**language**

| | |
|---|---|
| jpn-000 | 日本語 |
| uig-000 | ئۇيغۇرچە |
| uig-001 | Uyghurche |

**file — kind**

prt@Plx

**edit — person**

# Internet Archive PanLex collection

| | |
|---|---|
| file — difficult | 8 |
| file — submit — necessary — 1/0 | 1 |
| expression — edit — necessary — 1/0 | 0 |
| expression — edit — done — language — (aaa,bbb,ccc) | |
| file — address | jpn-uig-菅原 |

# Internet Archive PanLex collection

stop

beginning

back

## PanLex 3.4

[PanLem (usred6w)]
[9.1.9]

you — testintern

source — edit — permission

| change | source | language — necessary | file — kind | difficult | title |
|---|---|---|---|---|---|
| ± | nan-eng:Nakajima | nan-000 | pdf-img | 8 | 『海南島方言基礎語彙集』 |
| ± | eng-jpn-rus-niv:丹菊逸治 | niv-000; jpn-002 | pdf-img | 8 | ニヴフ語サハリン方言基礎語彙集（ノグリキ周辺地域）＝Basic vocabulary of the Sakhalin dialect of Nivkh language : Nogliki dialect |
| ± | nsz-eng:Kroeber | nsz-000 | pdf-img | 4 | The Valley Nisenan |
| ± | ood-eng:Allison | ood-000 | pdf-img | 4 | O'otham Ñiokĭ Haichu A:ga |
| ± | eng-mya-shn-pce:Shin | pce-002; pce-001; mya-002; shn-000; pce-003; pce-004; pce-005 | pdf-img | 7 | The Palaung Language: Comparative Lexicon of its Southern Dialects (I) |

# Analytical tasks

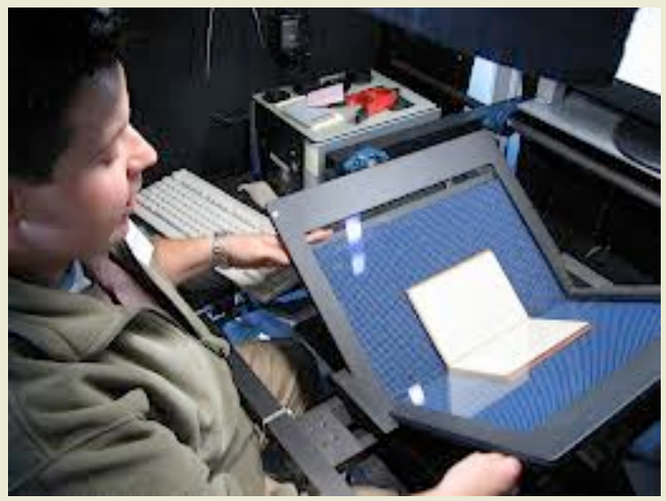
## Same as for digital text, plus:

- Page imaging
- Spatial structural analysis
- Text identification and sequencing
- Script identification
- Font identification
- Face identification
- Character segmentation and sequencing
- Character identification

# Analytical tasks

## Page imaging

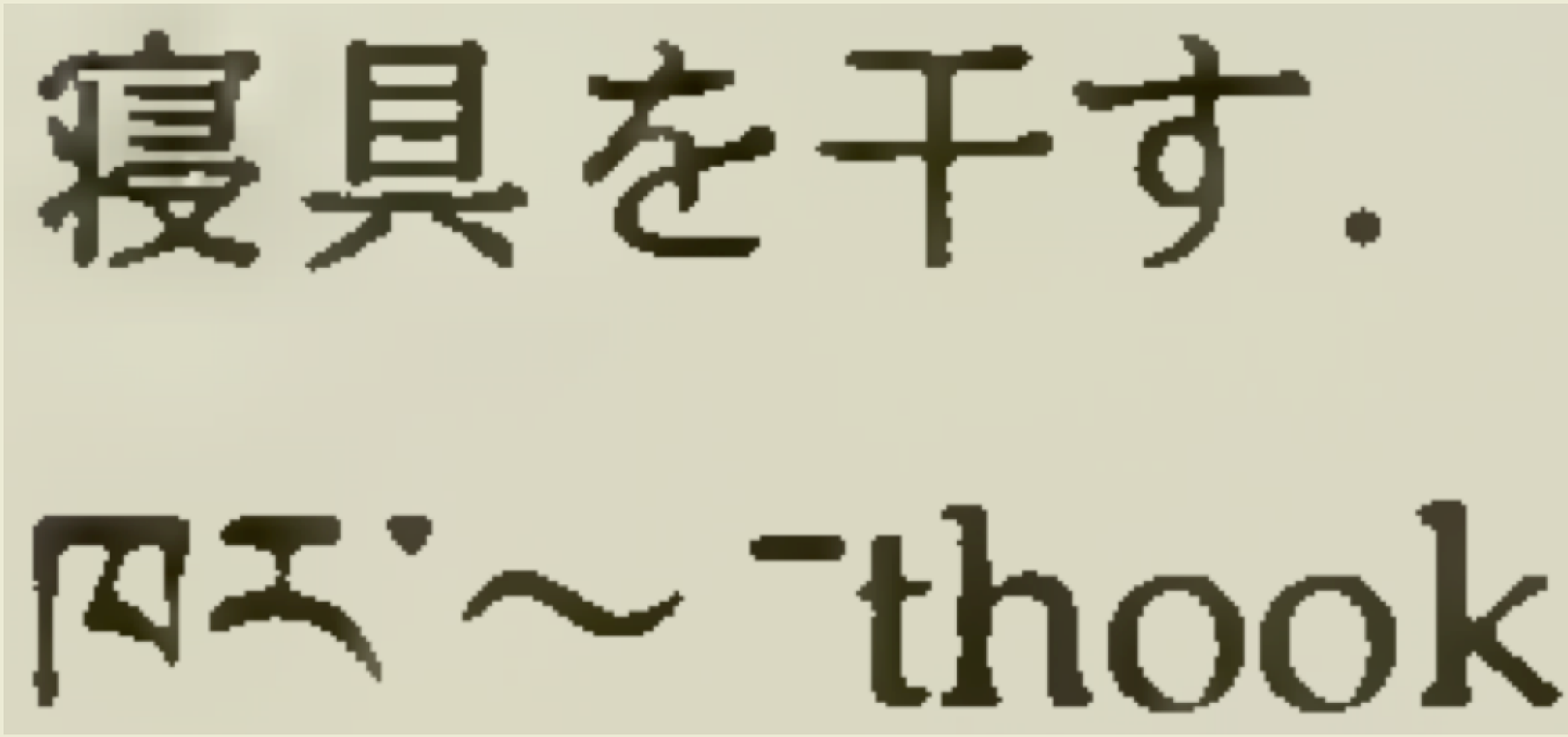
- Produces image file(s)
- Internet Archive produces JPG2000, PDF/A, Djvu, and Ebook files (in color)
- PanLex collection: format changes from “prt@Plx” to “PDF-img”



# Analytical tasks

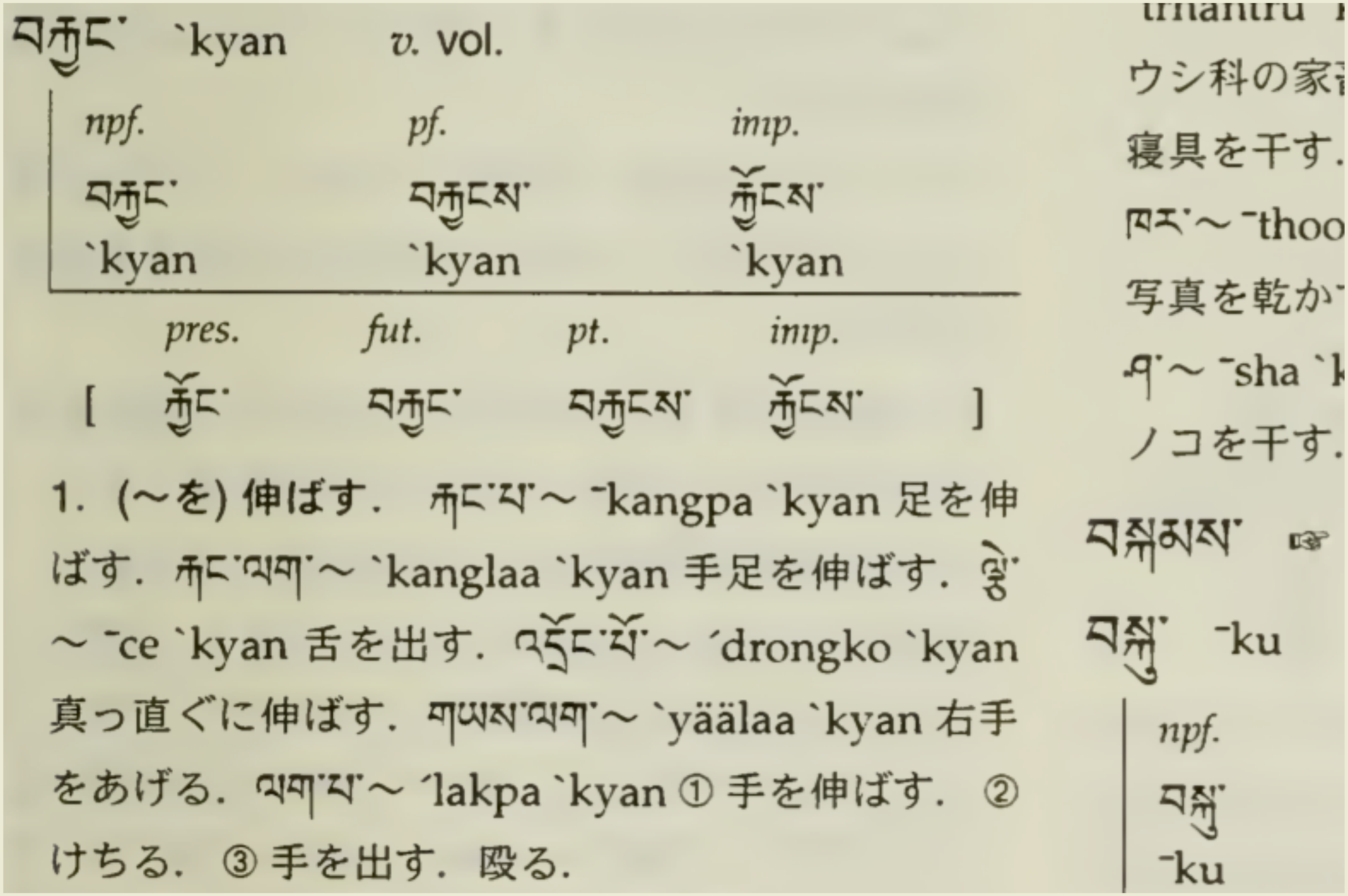
## Page imaging

- Internet Archive resolution



# Analytical tasks

## Spatial structural analysis



བཀྱང་ `kyan *v.* vol.

| *npf.* | *pf.* | *imp.* |
|---|---|---|
| བཀྱང་ | བཀྱངས་ | ཀྱངས་ |
| `kyan | `kyan | `kyan |

| *pres.* | *fut.* | *pt.* | *imp.* |
|---|---|---|---|
| ཀྱང་ | བཀྱང་ | བཀྱངས་ | ཀྱངས་ |

1. (～を) 伸ばす. ཞང་པ་～ ¯kangpa `kyan 足を伸ばす. ཞང་ལག་～ `kanglaa `kyan 手足を伸ばす. ལྕེ་～ ¯ce `kyan 舌を出す. འདྲོང་པོ་～ ´drongko `kyan 真っ直ぐに伸ばす. གཡས་ལག་～ `yäälaa `kyan 右手をあげる. ལག་པ་～ ´lakpa `kyan ① 手を伸ばす. ② けちる. ③ 手を出す. 殴る.

thantru

ウシ科の家

寝具を干す.

ཐང་～ ¯thoo

写真を乾か

ཤ་～ ¯sha `k

ノコを干す.

བསྐམས་ ☞

བསྐུ་ ¯ku

| *npf.* |
|---|
| བསྐུ་ |
| ¯ku |

# Analytical tasks

## Text identification and sequencing

**F**

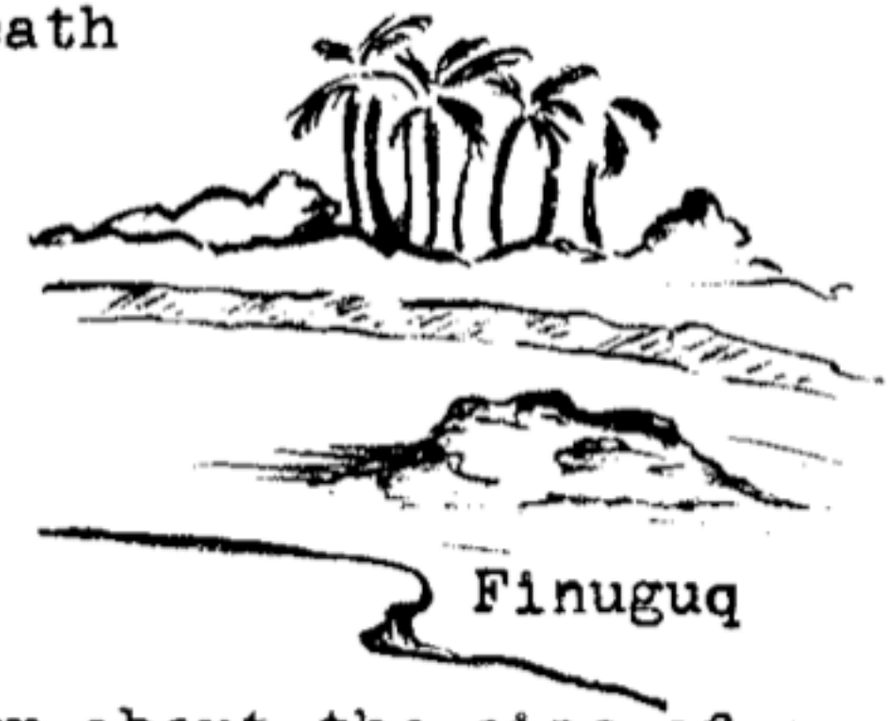
FETFET  Whistle by drawing breath over lower teeth.

FINUGUQ (pulô)  Island (in river).

FOFOY  (tutubí)  Dragonfly.

FOOG  (sipol, sutsot, paswít) Whistle.

FOOLOK  (diwatà)  Type of fairy about the size of a

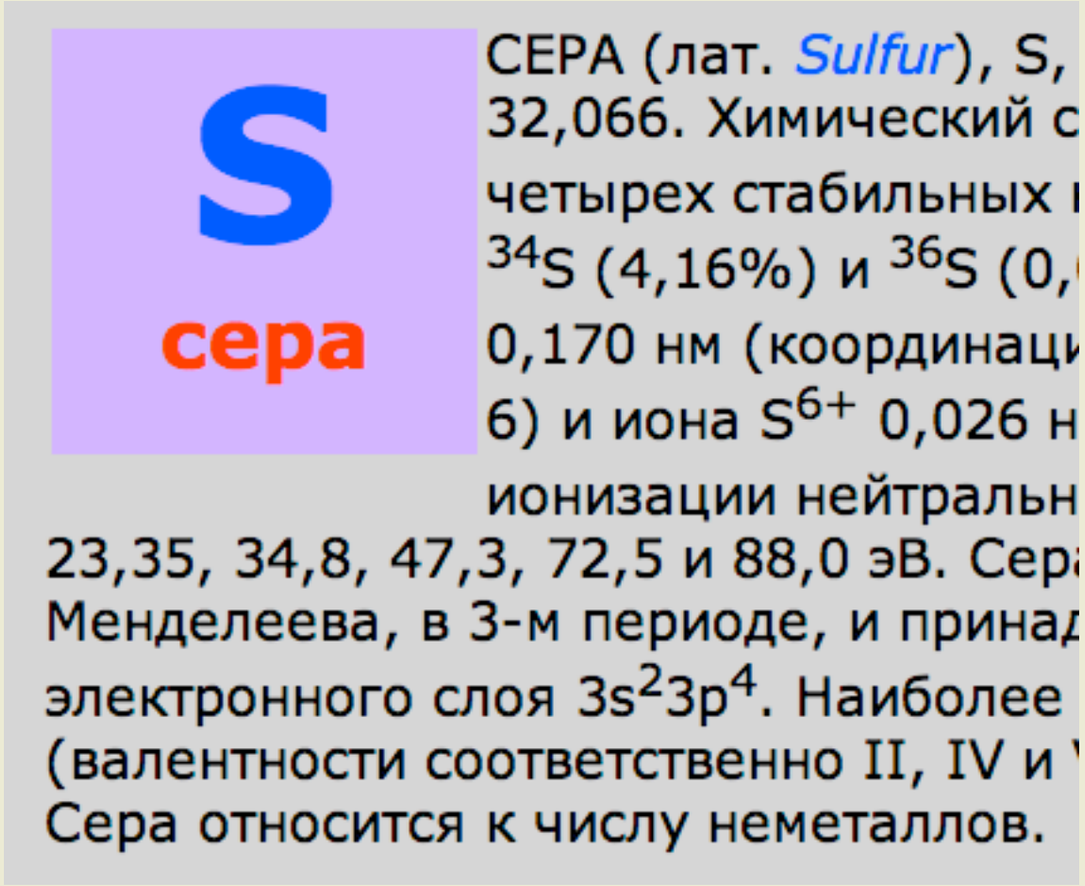


Finuguq

# Analytical tasks

## Script identification

CEPA

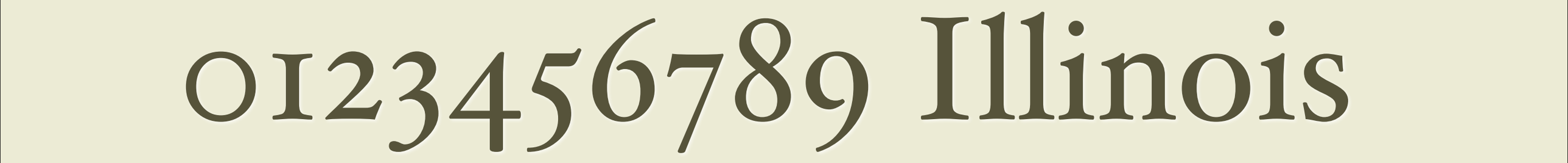


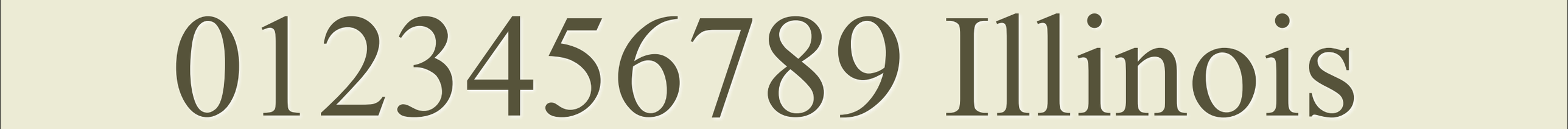

CEPA

CEPA Center for European Policy Analysis

# Analytical tasks

## Font identification




# Analytical tasks

## Face identification

0123456789 Illinois

*0123456789 Illinois*

# Analytical tasks

## Character segmentation and sequencing


m m

# Analytical tasks

## Character identification

one ell

# Tools

## Image-to-text conversion (“optical character recognition”) tools: competitors

- Nicomsoft
- Tesseract
- OCRopus
- Abby FineReader
- Adobe Acrobat
- OmniPage
- Readiris
- GOCR
- LEADTOOLS
- Many others

# Tools

## Image-to-text conversion ("optical character recognition") tools: properties

- Commercial versus open-source
- Script-specific versus multiscriptal
- Language-specific versus multilingual
- Diaglottal versus synglottal (language-mixing)
- Immutable versus trainable
- Isolated versus integratable
- Fossilized versus actively developed

# Tools

## Prior research

- Christa Mabee, "Optical Character Recognition in Multilingual Text: A Brief Survey", 02012
- Christa Mabee, "Report on Internship Project: Optical Character Recognition in Multilingual Text", 02012
- Marcin Heliński, Miłosz Kmieciak, Tomasz Parkoła, "Report on the comparison of Tesseract and ABBYY FineReader OCR engines", 02012 (PanLex copy)
- Ray Smith, Daria Antonova, and Dar-Shyang Lee, "Adapting the Tesseract Open Source OCR Engine for Multilingual OCR", 02009
- Burcu Karagol-Ayan, "Resource Generation from Structured Documents for Low-density Languages", 02007
- Tapas Kanungo and Song Mao, "Stochastic Language Models for Style-Directed Layout Analysis of Document Images", 02003

# Tools

## Prior research: conclusions

- Page imaging is hard
- Tool training is necessary but expensive
- Non-dictionary corpus integration has not been tried
- PanLex requires multiscriptal, multilingual, synglottal, trainable, integratable, actively developed tools
- Few (≅ 2) tools even aspire to all of these properties
- Most likely candidates now: ABBYY FineReader 11 Professional Edition, ABBYY FineReader Engine, Tesseract

# Tools

## ABBYY FineReader 11 Professional Edition

- Application
- Commercial
- SKU FRPFWA11E
- Windows 8 / 7 / Vista / XP; Windows Server 2012 / 2008 / 2008 R2 / 2003
- Express edition (Windows and Macintosh) and web service are more limited
- Price per license: list $300, street $270+, direct (through 31 July) $180 (or $120 for “upgrade”)

# Tools

## ABBYY FineReader Engine

- SDK
- Commercial
- Version 10: Windows 7 / Vista / XP / 2000; Windows Server 2003
- Version 9: Linux (GCC 2.95–4)
- Version 8: Macintosh (OS X 10.4–10.8)
- Prices not published

# Tools

## Tesseract

- SDK and command-line application
- Open-source (Apache License)
- Partly sponsored by Google
- Version 3.02: Linux, OS X
- Additional OSs: Windows with VC++ or CygWin, iOS, Android
- Used by about 40 other tools and projects

# Tools

## Script identification

| | *AFRP 11* | *AFRE 10* | *AFRE 9* | *AFRE 8* | *Tesseract* |
|---|---|---|---|---|---|
| *Latin* | ◉ | ◉ | ◉ | ◉ | ◉ |
| *Cyrillic* | ◉ | ◉ | ◉ | ◉ | ◉ |
| *Greek* | ◉ | ◉ | ◉ | ◉ | ◉ |
| *Armenian* | ◉ | ◉ | ◉ | ◉ | |
| *Hebrew* | ◉ | ◉ | | ◉ | ◉ |
| *Han* | ◉ | ◉ | ◉ | | ◉ |
| *Hiragana/Katakana* | ◉ | ◉ | ◉ | | ◉ |
| *Hangul* | ◉ | ◉ | ◉ | | ◉ |
| *Thai* | ◉ | ◉ | ◉ | | ◉ |
| *Arabic* | ◉ | (◉) | | | ◉ |
| *Devanagari* | | | | | ◉ |
| *Telugu* | | | | | ◉ |
| *Tamil* | | | | | ◉ |
| *Malayalam* | | | | | ◉ |
| *Kannada* | | | | | ◉ |
| *Bengali* | | | | | ◉ |
| *Cherokee* | | | | | ◉ |

# Tools

## Quality comparison

- Heliński et al.:
  - FineReader was better on average than Tesseract, but worse on some tasks
  - Test was entirely on Polish (thus Latin-script) text
- Mabee:
  - FineReader was better on average than Tesseract
  - Test was on Latin, Thai, Arabic, and Tibetan scripts

# Tools

## Training comparison

- Both FineReader and Tesseract can (in principle) be trained, but:
  - Bugs prevented Mabee from training FineReader
  - Mabee found Tesseract training tedious and useless
- Approximate training costs for 1 font per Heliński et al.:
  - FineReader by ABBYY: $40,000
  - FineReader in-house: 25 person-days
  - Tesseract in-house: 8 person-days

# Tools

## Extensibility comparison

- Tesseract operates on more scripts than FineReader
- But both omit Georgian, Sinhala, Lao, Tibetan, Burmese, Khmer, Ethiopic, etc.
- So both need to be extended for PanLex
- Both permit defining new languages and providing new dictionaries

# Tools

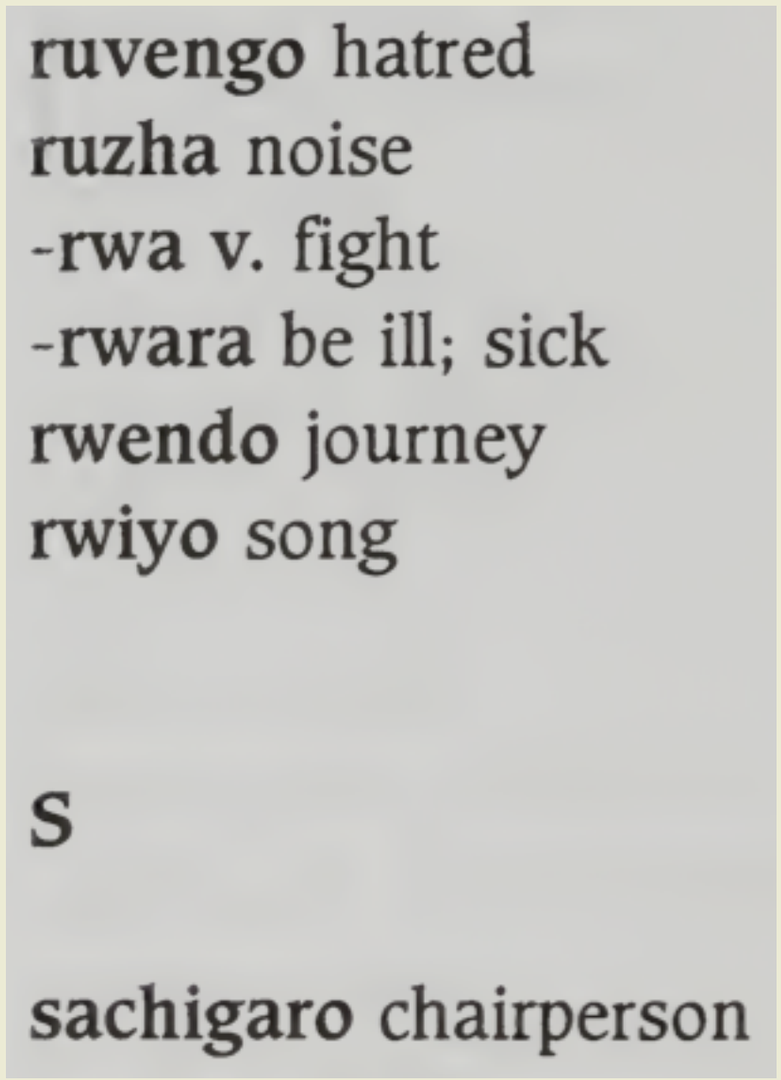
## Integratability comparison

- Desiderata for PanLex:
  - Dynamic integration of expression corpus (ex table)
  - Dynamic integration of approved characters (cu and cp tables)
- ABBYY FineReader 11 Professional Edition: no
- ABBYY FineReader Engine: maybe
- Tesseract: probably

# Tools

## Performance on an easy source

- FineReader



**ruvengo** hatred
**ruzha** noise
-**rwa** v. fight
-**rwara** be ill; sick
**rwendo** journey
**rwiyo** song

**S**

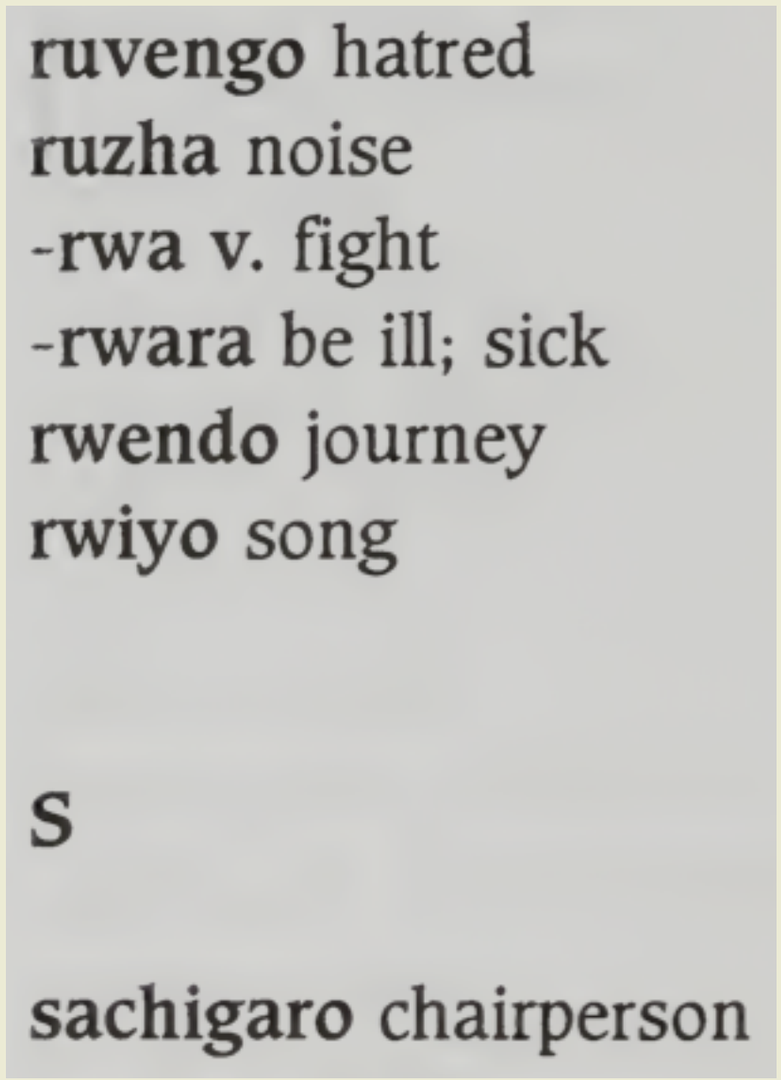
**sachigaro** chairperson

```
ruvengo hatred
ruzha noise
-rwa v. fight
-rwara be ill; sick
rwendo journey
rwiyo song



sachigaro chairperson
```

# Tools

## Performance on an easy source

- Tesseract

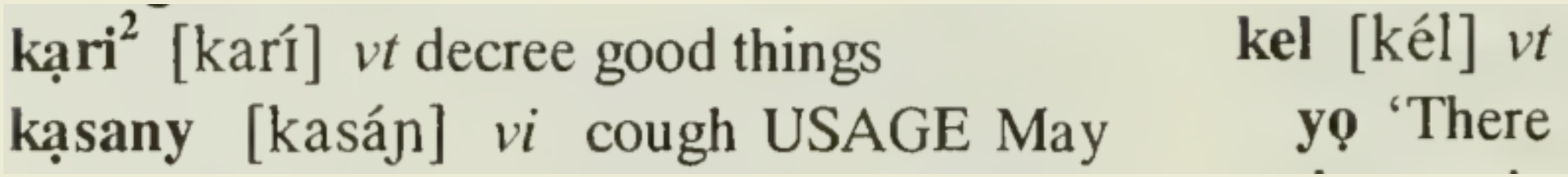



```
ruvengo hatred
ruzha noise
-rwa v. fight
-rwara be ill; sick
rwendo journey
rwiyo song
S
sachigaro chairperson
```

# Tools

## Performance on a more complex source

- FineReader

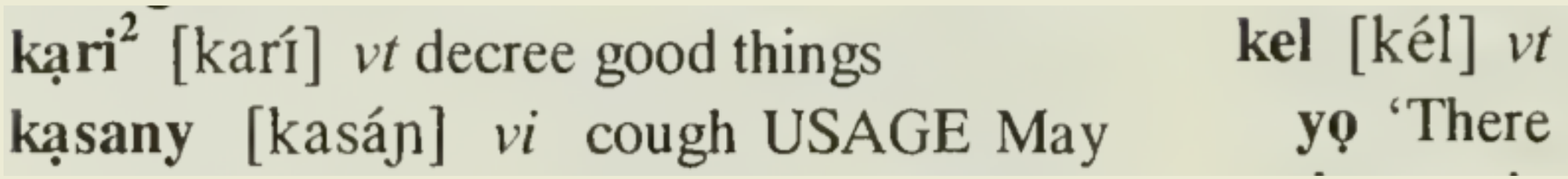



```
kari 2 [kari] vt decree good things
kasany [kasan] vi cough USAGE May
```

# Tools

## Performance on a more complex source

- Tesseract




```
kariz [kari] vt decree good things
kasany [kasap] vi cough USAGE May
```

# Tools

## Performance on a very complex source

- FineReader

| *npf.* | *pf.* | *imp.* |
|---|---|---|
| བསྒྲུལ་ | བསྒྲུལ་ | སྒྲུལ་, སྒྲོལ་ |
| ⁻kyää | ⁻kyää | ⁻kyää, ⁻kyöö |

| *pres.* | *fut.* | *pt.* | *imp.* |
|---|---|---|---|
| [ སྒྲོལ་ | བསྒྲུལ་ | བསྒྲུལ་ | སྒྲོལ་ ] |

```
  npf.  P/
  "kyaa "kyaa
imp.
"kyaa, "kyoo
[    gar          qgar       qsp-         gar          ]
```

# Tools

## Performance on a very complex source

- FineReader (“Finnish”)

| *npf.* | *pf.* | *imp.* |
|---|---|---|
| བསྒྲལ་ | བསྒྲལ་ | སྒྲལ་, སྒྲོལ་ |
| ⁻kyää | ⁻kyää | ⁻kyää, ⁻kyöö |

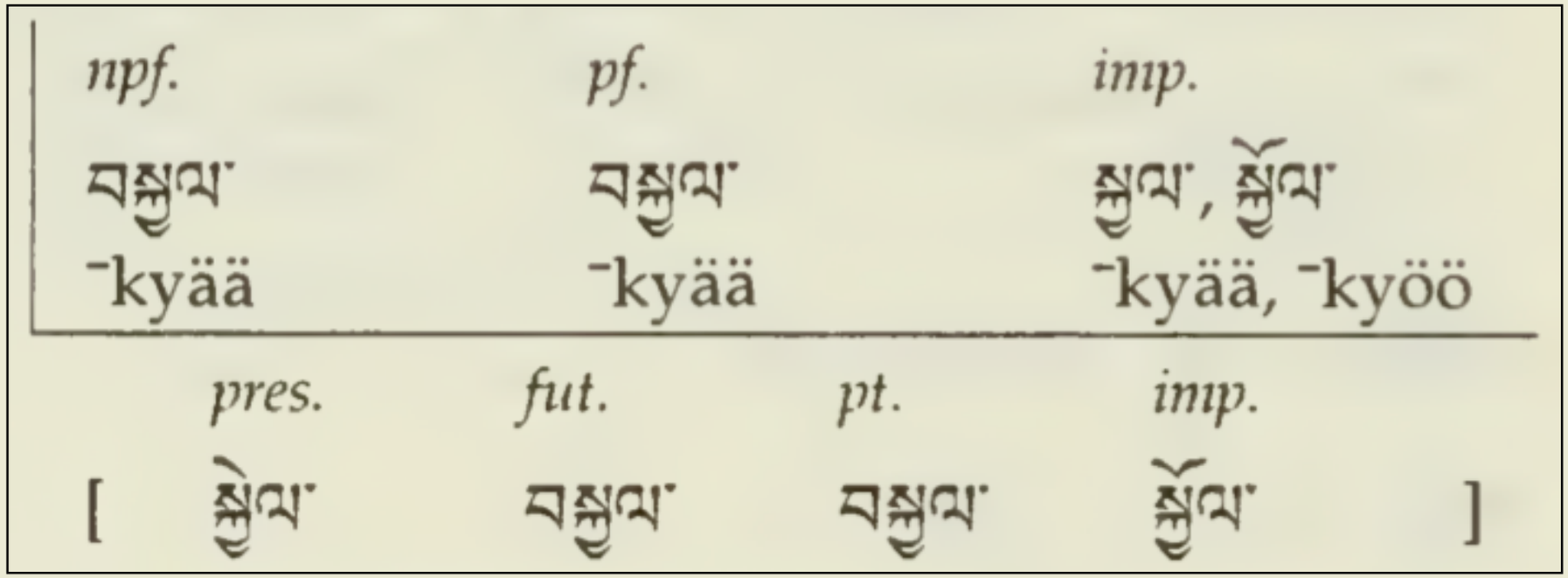
| | *pres.* | *fut.* | *pt.* | *imp.* | |
|---|---|---|---|---|---|
| [ | སྒྲོལ་ | བསྒྲལ་ | བསྒྲལ་ | སྒྲོལ་ | ] |

```
   npf. Pf-
   qg^' "kyää   qspr "kyää
imp.
go,-,lq-
"kyää, "kyöö
[      lar          qgof         qgq-         laj-          ]
```

# Tools

## Performance on a very complex source

- FineReader (“Finnish”, HTML output)



| *npf.* | *pf.* | *imp.* |
|---|---|---|
| བསྒྲུལ་ | བསྒྲུལ་ | སྒྲུལ་, སྒྲོལ་ |
| ⁻kyää | ⁻kyää | ⁻kyää, ⁻kyöö |

| | *pres.* | *fut.* | *pt.* | *imp.* | |
|---|---|---|---|---|---|
| [ | སྒྲོལ་ | བསྒྲུལ་ | བསྒྲུལ་ | སྒྲོལ་ | ] |

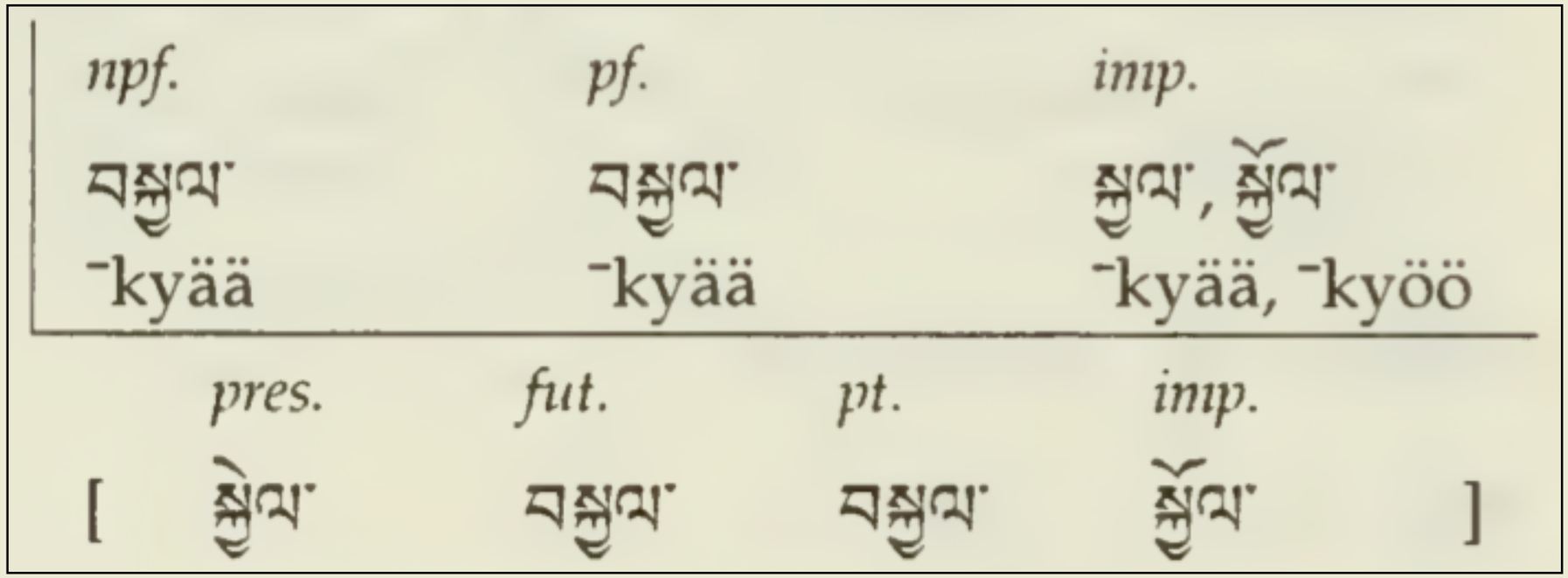
| *npf.* | *Pf-* | *imp.* |
|---|---|---|
| qg^' | qspr | go,-,\|q- |
| "kyää | "kyää | "kyää, "kyöö |

| | | | | |
|---|---|---|---|---|
| [ \|ar | qgof | qgq- | \|aj- | ] |

# Tools

## Performance on a very complex source

- Tesseract



| *npf.* | *pf.* | *imp.* |
|---|---|---|
| བསྒྲུལ་ | བསྒྲུལ་ | སྒྲུལ་, སྒྱོལ་ |
| ⁻kyää | ⁻kyää | ⁻kyää, ⁻kyöö |

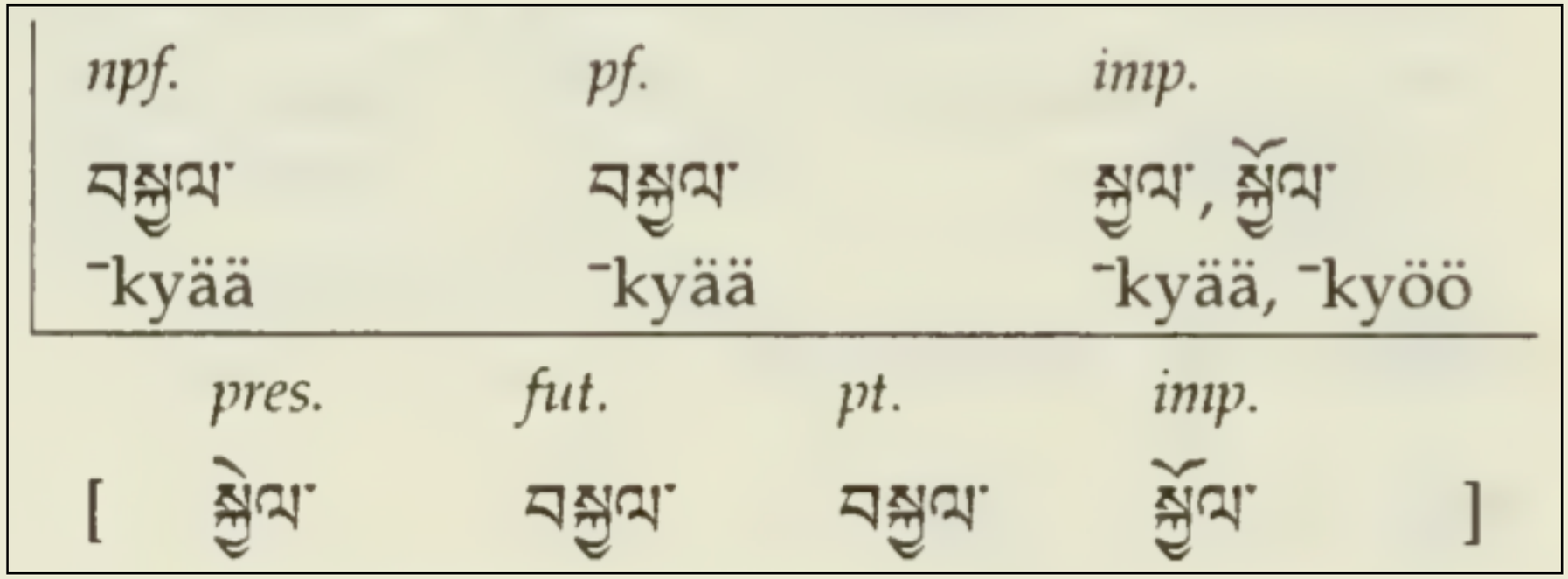
| | *pres.* | *fut.* | *pt.* | *imp.* | |
|---|---|---|---|---|---|
| [ | སྐྱེལ་ | བསྒྲུལ་ | བསྒྲུལ་ | སྒྱོལ་ | ] |

```
npf. pf. Imp.
magma" r1504" gm", §m'
'kyééi 'kyéi5 ‘kyéiéi, ‘ky66
pres. fut. pt. mzp.
[ gar agar qgw"
\F
```

# Tools

## Performance on a very complex source

- Tesseract (“Finnish”)



| *npf.* | *pf.* | *imp.* |
|---|---|---|
| བསྒྲུལ་ | བསྒྲུལ་ | སྒྲུལ་, སྒྲོལ་ |
| ⁻kyää | ⁻kyää | ⁻kyää, ⁻kyöö |

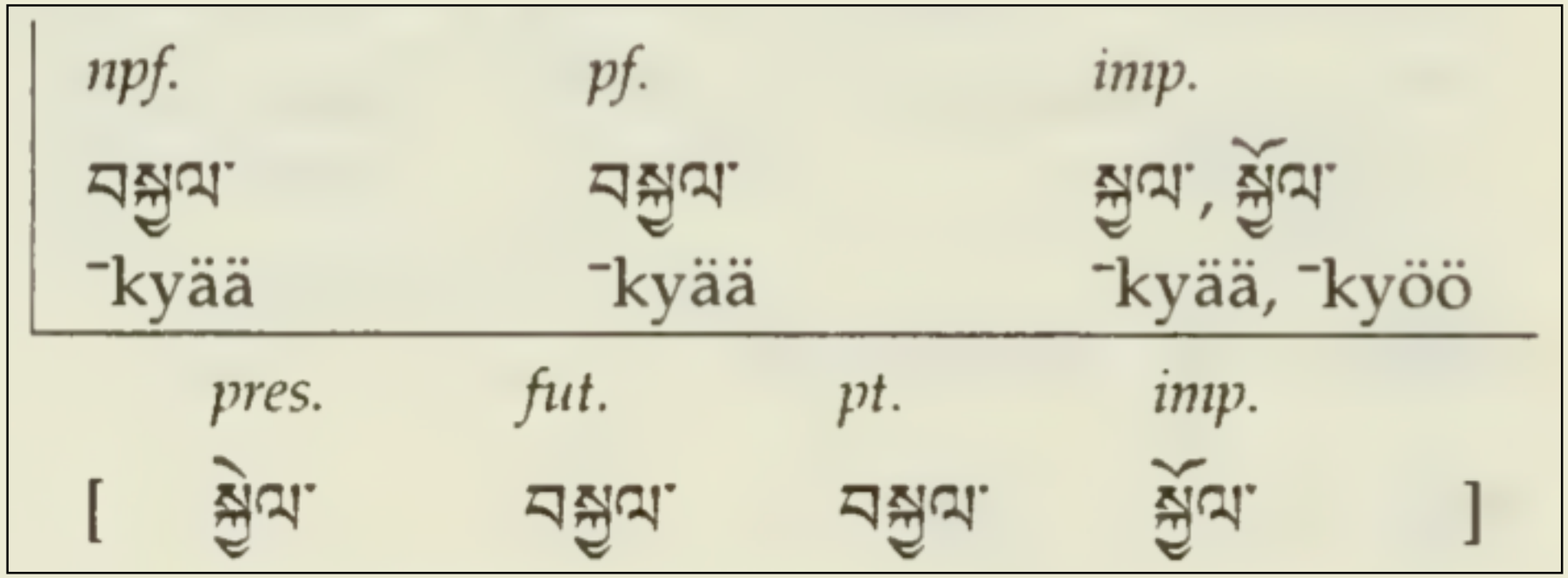
| | *pres.* | *fut.* | *pt.* | *imp.* | |
|---|---|---|---|---|---|
| [ | སྒྲེལ་ | བསྒྲུལ་ | བསྒྲུལ་ | སྒྲོལ་ | ] |

```
npf. pf. ımp.
mgflı' rıgfıı' gfır, šfıı'
-kyää -kyää -kyää, -kyöö
pres. fııl. pt. mıp.
[ àfıı' ngfıı' fıgw'
`f
```

# Tools

## Performance on a very complex source

- Tesseract (“Finnish”, HTML output)



| *npf.* | *pf.* | *imp.* |
|---|---|---|
| བསྒྲུལ་ | བསྒྲུལ་ | སྒྲུལ་, སྒྲོལ་ |
| ⁻kyää | ⁻kyää | ⁻kyää, ⁻kyöö |

| | *pres.* | *fut.* | *pt.* | *imp.* | |
|---|---|---|---|---|---|
| [ | སྒྲེལ་ | བསྒྲུལ་ | བསྒྲུལ་ | སྒྲོལ་ | ] |

npf. pf. ımp. mgflı' rıgfıı' gfir, šfıı'

-kyää -kyää -kyää, -kyöö

pres. fııl. pt. mıp.

[ àfıı' ngfıı' fiıgw'

`f

# Opportunities

## PanLex and image-to-text conversion tools

- Benefits to PanLex from tool extensions
- Benefits from PanLex for tool extensions
- Which tool(s) to use and support?
- Multiple tools?
  - Advantages: federation, complementarity, independence
  - Disadvantages: expense, forfeiture of expertise

# Opportunities

## Partnerships for tool development

- Tool developers
- Possessors of lexical data
- Tool consumers
  - Internet Archive
  - HathiTrust
  - Google
  - Project Gutenberg