

# PanLex Source Analysis

Jonathan Pool  
The Long Now Foundation

**PanLex**



<http://panlex.org>

First PanLex Summer Internship

8 July 02013

# Summary

- Final source file format
- Linguality
- Classification
- Definitions
- Word classes
- Lemmatization
- Complexity triage
- Intellectual property

# Final source file format

Simple text:  
only expressions

```
.  
Ø  
  
eng-000  
chronological  
epo-000  
tempa  
epo-000  
kiama  
  
eng-000  
spatial  
...
```

# Final source file format

Full text:  
anything

```
:  
Ø  
  
df  
eng-000  
someone with a duty of care  
dm  
eng-000  
law  
ex  
eng-000  
fiduciary  
...
```

# Linguality

Varilingual

•  
Ø

eng-000

chronological

epo-000

tempa

epo-000

kiama

eng-000

spatial

...

# Linguality

Centrilingual

·  
1

eng-000

chronological

epo-000

tempa

spatial

eng-000

locational

...

# Linguality

Bilingual

•  
2

eng-000

epo-000

chronological

tempa

kiama

spatial

loka

# Classification

## Meaning identifier vs. expression

```
<entry id="d1e65446">
<trait name="original-id" value="13546"/>
<!--formtype=word--><lexical-unit>
  <form lang="pot">
    <text>mkom</text>
  </form>
</lexical-unit>
<sense>
  <grammatical-info value="Noun:Animate:Singular"/>
  <definition>
    <form lang="eng">
      <text>ice</text>
    </form>
  </definition>
</sense>
</entry>
<entry id="d1e65465">
```



# Classification

## Meaning identifier vs. expression

### 9. Wordlists

The following wordlists are based on the SWADESH "First 100" Diagnostic WORD LIST. Above each list is given the name of the village and clan of the speaker. Aspiration [<sup>h</sup>] has not been marked.

	Nyasoso/ Mwetug	Eboko Bajog/ Mbwogmut	Ngonmin/ Mwetan
1. I	mě	mě	mě
2. thou	wě	wě	wě
3. we	sé	sé	sé
4. this	ăně (cl.1)		ănén
5. that	ănén (cl.1)		ăníní
6. who	ḥzě	ḥzě	ḥzě
7. what	cě		c̣ě
8. not	sāké		sāké

# Classification

## Meaning identifier vs. expression

All Ingush forms are given in the [Ingush Practical Orthography](#).

1	I	so (nominative case)
1a	me	aaz (ergative case)
2	you	hwo (nominative case)
2a		wa (ergative case)
3	we (exclusive)	txo
3a	we (inclusive)	vaj
4	this	jer
5	that	yz
6	who	mala
6a		hwan- (oblique stem, e.g. erg. hwanuu)
7	what	fy
7a		sie(n)- (oblique stem, e.g. gen. sien)
8		-(a)c (suffix, in most indicative categories)
8a	not	ca (preposed particle, in some moods)
8b		ma (preposed particle, in imperative)

art-eng-inh:S1LI

# Classification

Meaning identifier vs. expression

art- 244	Ars Signorum
art- 245	Swadesh 100
art- 246	Fitusa
art- 247	Euransi

# Classification

## Meaning identifier vs. expression

mi  
2  
ex  
eng-000  
thou  
ex  
bss-000  
wè

ex  
art-245  
2  
ex  
eng-000  
thou  
ex  
bss-000  
wè

# Classification

## Definition vs. expression

349.	Widow	寡婦	nimakalkalima so nakem
350.	Wife	妻子	mavakes
351.	Wind	風	sazowsaw
352.	Wine(rice)	酒	saki
353.	Wing	翅膀	panid
354.	Winnow	去米糠	manigi so vias no mogis

eng-cmn-tao:Rau

# Classification

## Definition vs. domain

base (chemistry): негіздер  
base (military): база

# Classification

## Definition vs. domain

df  
eng-000  
base (chemistry)  
ex  
eng-000  
base  
ex  
kaz-000  
негіздер

dm  
eng-000  
chemistry  
ex  
eng-000  
base  
ex  
kaz-000  
негіздер

# Definitions

base (chemistry): негіздер.  
ерітінділерінде бір немесе бірнеше  
гидроксид иондарын түзіп,  
диссоцияланатын күрделі заттар.



# Definitions

base (chemistry): негіздер. ерітінділерінде бір немесе бірнеше гидроксид иондарын түзіп, диссоциацияланатын күрделі заттар.

dm

eng-000

chemistry

df

kaz-000

ерітінділерінде бір немесе бірнеше гидроксид иондарын ...

ex

eng-000

base

ex

kaz-000

негіздер

# Definitions

enusogr	small grey monkey with elongated ears
onecee	kind of tent
tonocel	fear of being kidnapped
groson	kind of riddle

# Definitions

## Personal editing

enusogr	(small grey) monkey (with elongated ears)
onecee	(kind of) tent
tonocel	fear (of being kidnapped), apagophobia
groson	(kind of) riddle

# Definitions

Automated analysis

```
s/\tkind of /\t(kind of) /;
```

enusogr    small grey monkey with elongated ears

onecee    (kind of) tent

tonocel    fear of being kidnapped

groson    (kind of) riddle

# Definitions

## Serialization: exdftag

```
'exdftag'      => { cols => [0, 1], re => '(:?:\([^()]+\)| ([^ () ]+ )', subre => '[][/,;?!~]' },
```

```
# maxchar: maximum character count  
permitted in an expression, or  
' ' if none; default ' '.
```

```
example: 25.
```

```
# maxword: maximum word count  
permitted in an expression, or  
' ' if none; default ' '.
```

```
example: 3.
```

# Definitions

Serialization: exdftag

'exdftag'  $\Rightarrow$  { ... },

maxchar  $\Rightarrow$  30, maxword  $\Rightarrow$  3

*result*

«ex»nimakalkalima so nakem  
*22 characters, 3 words*

«ex»

«ex»manigi so vias no mogis  
*23 characters, 5 words*

«df»

# Definitions

## Serialization: exdftag

'exdftag' => { cols => [0, 1], re => '(?:\([^()]+\)|\([^ () ]+\) )', subre => '[][/,;?!~]' },

# re: regex matching a  
definitional part of an  
expression, or '' if none.

'exdftag' => { ... },

re => '(?:\([<sup>blue</sup>^()<sup>blue</sup>]+\<sup>blue</sup>\)|(<sup>purple</sup>[^ () ]+\<sup>purple</sup>))'

# Definitions

Serialization: exdftag

'exdftag'  $\Rightarrow$  { ... },

re  $\Rightarrow$  '(?:\([^\(\)]+\)| ([^ () ]+))'

«ex» (kind of) tent

(kind of) tent	<i>matches</i>
<i>result</i>	«df» (kind of) tent «ex» tent



# Definitions

## Serialization: normalize

```
'normalize' => { col => 0, uid => 'eng-000', min => 50, mindeg => 10 },
```

# min: minimum score (0 or more) a proposed expression must have in order to be accepted outright as an expression. Every proposed expression with a lower (or no) score is to be replaced with the highest-scoring expression sharing its language variety and degradation, if any such expression has a higher score than it does.

# Definitions

## Serialization: normalize

```
'normalize'    => { col => 0, uid => 'eng-000', min => 50, mindeg => 10 },
```

```
# mindeg: minimum score a proposed  
           expression that is not accepted  
           outright as an expression, or its  
           replacement, must have in order to be  
           accepted as an expression.
```

# Definitions

Serialization: normalize

'normalize' => { ... },

uid => 'eng-000',

min => 50,

mindeg => 10

# Definitions

## Serialization: normalize

```
'normalize'    => { ... },
```

```
uid => 'eng-000',
```

```
min => 50,
```

```
mindeg => 10
```

⟨⟨ex⟩⟩man of war

man of war	<i>score = 49 (&lt; 50)</i>
man-of-war	<i>score = 171 (≥ 10)</i>
<i>result</i>	⟨⟨ex⟩⟩man-of-war

# Definitions

## Serialization: normalize

```
'normalize'    => { ... },
```

```
uid => 'eng-000',
```

```
min => 50,
```

```
mindeg => 10
```

⟨⟨ex⟩⟩Polish

Polish	$score = 358 (\geq 50)$
<i>result</i>	⟨⟨ex⟩⟩Polish

(“polish” score = 822, but it isn’t checked.)

# Definitions

## Serialization: normalize

```
'normalize'    => { col => 0, uid => 'eng-000', min => 50, mindeg => 10 },
```

```
# dftag: definition tag, if proposed
expressions not accepted as
expressions and not having
replacements accepted as expressions
are to be converted to definitions;
' ' (blank) if they are to be converted
to pre-normalized expressions.
default '⌞df⌟'.
```

# Definitions

## Serialization: normalize

```
'normalize'      => { ... },
```

```
uid => 'eng-000',
```

```
min => 50,
```

```
mindeg => 10
```

⌞ex⌟ limousine liberal

limousine liberal	<i>score</i> = 5 (< 50)
<i>result</i>	⌞df⌟ limousine liberal

# Definitions

## Serialization: normalize

```
'normalize'      => { ... },
```

```
uid => 'eng-000',
```

```
min => 50,
```

```
mindeg => 10,
```

```
dftag => ''
```

⌞ex⌟limousine liberal

limousine liberal	$score = 5 (< 50)$
<i>result</i>	⌞exp⌟limousine liberal



# Word classes

pomotu [n.f.]

elentah [v.i.]

windshield

fall asleep

# Word classes

pomotu [n.f.]

elentah [v.i.]

windshield

fall asleep

ex

xyz-999

pomotu

wc

noun

md

gram

f.

ex

eng-000

windshield

# Word classes

## Tabularization

pomotu [n.f.]

windshield

elentah [v.i.]

fall asleep

pomotu	n.f.	windshield
elentah	v.i.	fall asleep

# Word classes

Serialization: wctag

pomotu	n.f.	windshield
elentah	v.i.	fall asleep

'wctag'  $\Rightarrow$  { col  $\Rightarrow$  1 },

wc.txt

n.f.	noun:f.
n.idiom.	noun:n.idiom.
...	
v.i.	verb:v.i.

# Word classes

Serialization: wctag

'wctag'  $\Rightarrow$  { col  $\Rightarrow$  1 },

wc.txt

n.f.	noun:f.
n.idiom.	noun:n.idiom.
...	
v.i.	verb:v.i.

<i>original</i>	n.f.
<i>wc.txt output</i>	noun:f.
<i>wctag output</i>	⌞wc⌟noun⌞md:gram⌟f.

# Lemmatization

Elevate

a windshield

to fall asleep

have faith in

to the best of s.o.'s knowledge

be wise

Thank you

# Lemmatization

Straightforward cases: tabularization

Elevate	elevate
a windshield	«wc:noun» windshield
to fall asleep	«wc:verb» fall asleep

# Lemmatization

Straightforward cases: serialization

'wcshift'  $\Rightarrow$  { col  $\Rightarrow$  1 },

<i>original</i>	⌞wc:noun⌟windshield
<i>extag output</i>	⌞ex⌟⌞wc:noun⌟windshield
<i>wcshift output</i>	⌞ex⌟windshield⌞wc⌟noun



# Complexity triage

**hatak**, see *atak*.

**hatak**, **hatak**, n., a man; Matt. 9: 2; 18: 12; a person; a husband (*hutuk at ikimiksho*, she has no husband); a being in some senses, but not in all; a human being; mankind; folk; folks; a mortal; mortals; a subject; an inhabitant; *Kenan hatak*, Matt. 10: 4; a red man; a native; an Indian, being used to distinguish the red men from the whites; *hatak at ikbi*, a red man made it; *hatuk an*, a man, Matt. 15: 11. This word implies nationality; a man of the nation to which the speaker or the hearer belongs. *hatuk inkoi*, an Indian mile; *na hollo inkoi*, an English

cho-eng:Byington

# Complexity triage

malady; *abeka, illilli*  
male; *hatak nakni, nakni*  
malefactor, *hatak yoshoba*  
malevolence, *imanukfila okpulo*  
malevolent, *imanukfila okpulo*  
malice; *nukkilli, nukoa*  
malice, to bear; *anukkilli*  
malicious; *nukkilli, nukoa*  
malign; *nukkilli, nukoa*  
malignant; *nukkilli, nukoa*  
maligner, *isht yopula*  
malignity, *imanukfila okpulo*  
mall; *anowa, iti isht boa*

eng-cho:Byington

# Intellectual property

Claim A: language as property

Signed Agreement:

I understand that the Wôpanâak **Language is protected by copyright**, is not for sale and may not be used on any items whatsoever that are intended for sale, trade, or licensing, nor may it be replicated on items that will be sold, traded or licensed. I further understand that any violation of this agreement **will be prosecuted** to the full extent of the law.

[http://www.wlrp.org/Request\\_for\\_Linguistic\\_Analysis.2010.pdf](http://www.wlrp.org/Request_for_Linguistic_Analysis.2010.pdf)

# Intellectual property

## Claim B: limits on use of documentation

This dictionary, or part of it, is not to be used for **commercial purposes**. “Mirroring” on other web sites is not permitted. The dictionary may be copied freely for personal use. Shorter excerpts of the content may be quoted, as long as the source is referenced, including the URL to the web site. Permission may be given to use entire dictionaries for special non-profit scientific, artistic or similar purposes.

<http://www.alternative-dictionaries.net/dictionary/Basque/Basque.pdf>

# Intellectual property

Claim C: required attribution

I make this document freely available. You are free to copy it and give it to others - provided your copies **refer** to the original Dinka-English document, and **acknowledge** the SIL linguists and Blench, who put that document together.

<http://www.rogerblench.info/Language/Nilo-Saharan/Nilotic/English-Dinka%20glossary%207%20May.pdf>

# Intellectual property

## Controversy: language as property

... each side shall take a firm yes or no position on **whether computer programming languages are copyrightable.**

6	IN THE UNITED STATES DISTRICT COURT	
7	FOR THE NORTHERN DISTRICT OF CALIFORNIA	
8		
9		
10	ORACLE AMERICA, INC.,	No. C 10-03561 WHA
11	Plaintiff,	
12	v.	<b>SUPPLEMENTAL REQUEST</b>
13	GOOGLE INC.,	<b>FOR FURTHER BRIEFING</b>
14	Defendant.	
15	<hr/>	
16	In the briefs due next Thursday, each side shall take a firm yes or no position on whether	
17	computer programming languages are copyrightable. In addition, each side shall include	
18	whether it has ever taken an inconsistent position before any other court or agency, including the	
19	PTO, and if so, append those inconsistent statements.	
20		
21	<b>IT IS SO ORDERED.</b>	
22		
23	Dated: April 6, 2012.	
24		WILLIAM ALSUP UNITED STATES DISTRICT JUDGE

<http://docs.justia.com/cases/federal/district-courts/california/candce/3:2010cv03561/231846/874/>



# Intellectual property

Controversy: language as property

The common consensus among developers is that **Java, as a language, cannot be copyrighted**. ReadWriteWeb's former Web master, Jared Smith, said it best:

**"It's akin to copyrighting English."**

<http://readwrite.com/2012/04/16/oracles-uphill-battle-to-claim>

# Intellectual property

## PanLex: questions

- Is PanLex a derivative work?

<http://www.copyright.gov/circs/circ14.pdf>

- Is PanLex transformative?

[http://www.law.uci.edu/pdf/treese/reese\\_fair\\_use\\_transformative.pdf](http://www.law.uci.edu/pdf/treese/reese_fair_use_transformative.pdf)

- What laws and jurisdictions govern PanLex?

[http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2181671](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2181671)

- Will open linguistic data become the norm?

<http://linguistics.okfn.org/2011/05/20/the-open-linguistics-working-group/>



# Intellectual property

PanLex: working assumptions

- PanLex is a novel compilation of facts
- PanLex is transformative
- U.S. law governs PanLex
- PanLex uses of sources are fair uses
- Claims that PanLex has infringed will be rare

<http://panlex.org/tech/doc/deriv/sourcing.shtml>

# Intellectual property

## Transformativeness: one recent case

The Authors Guild v. HathiTrust, U.S. District Court for Southern District of N.Y., Opinion and Order, 10 Oct. 02012

- “A transformative use may be one that actually changes the original work. However, a transformative use can also be one that serves an entirely different purpose.” (p. 16)
- “Several courts have upheld wholesale copying of works where the use and purpose for the copies was clearly distinguishable from those of the original.” (p. 16)
- “Copying factual works is more likely fair use than copying creative works.” (p. 18)
- “The third fair-use factor considers whether the amount of copying was reasonable in relation to the purpose. ... Sometimes it is necessary to copy entire works.” (p. 18)
- “A copyright holder cannot preempt a transformative market.” (p. 20)

<http://thepublicindex.org/filings/hathitrust>

# Intellectual property

## PanLex: statement to users

“PanLex is a growing collection of information about lexical translations. You may access it at no cost. We do not restrict your use of it, and **to the best of our knowledge the publication of information from PanLex will not violate any rights** in the works from which PanLex has drawn information. When you reproduce information from PanLex, we would appreciate attribution to the PanLex project, its sponsor The Long Now Foundation, and the project website, <http://panlex.org>. We also welcome your reports on how you have used PanLex and suggestions for improvement. Please let us know if you wish to discuss agreements for additional services, support, or collaboration, such as updates, high-volume query processing, translation-inference research, user-interface customization, more detailed provenance documentation, **releases of claims by third parties**, or additional quality control.”