Translation Inference in PanLex

Jonathan Pool
The Long Now Foundation

PanLex

http://panlex.org

First PanLex Summer Internship

9 August 02013

Summary

- Translation inference in PanDictionary
- Responsibility for translation inference
- Translation inference in PanLem
- Future work

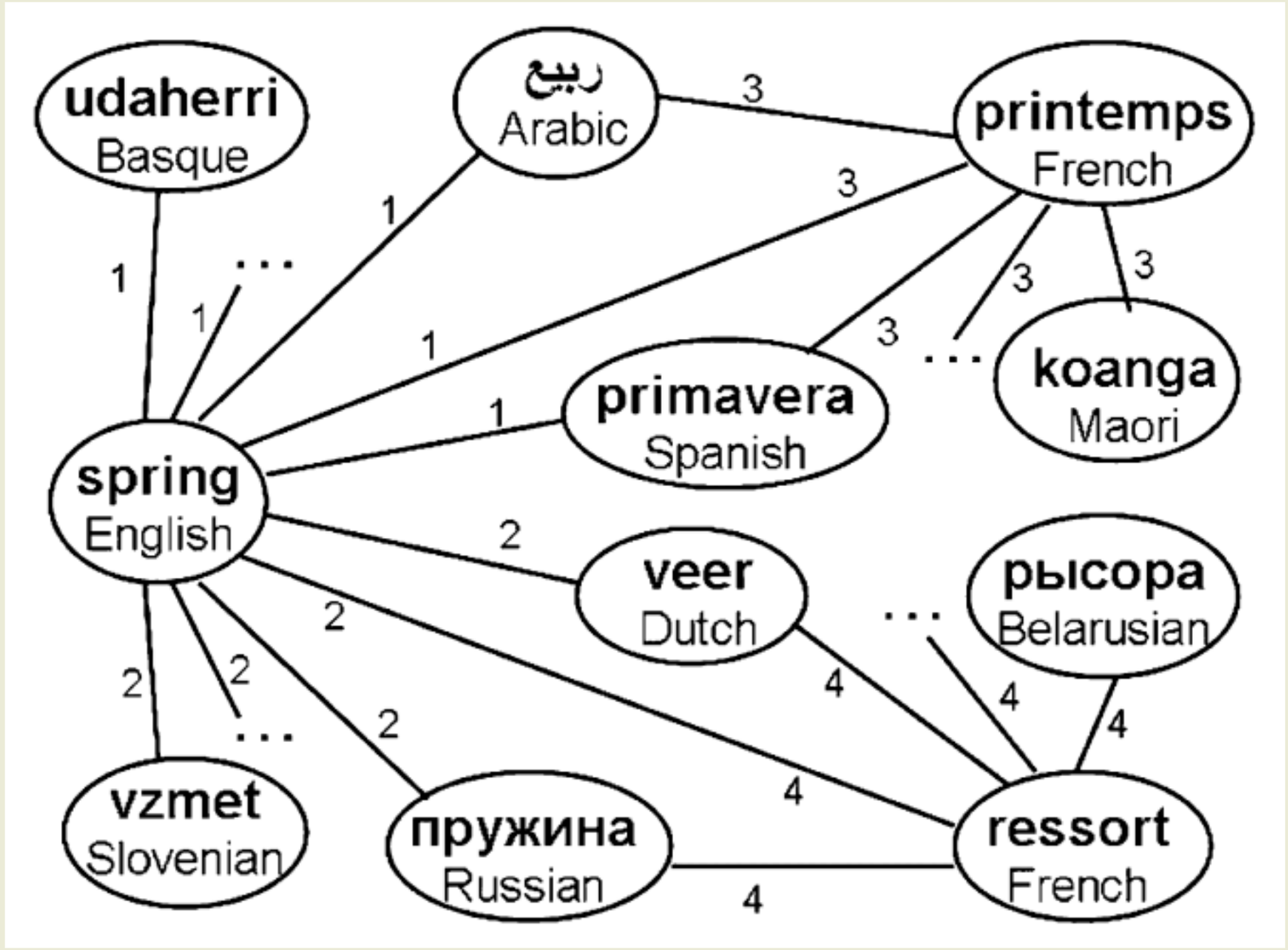
Translation inference in PanDictionary

TransGraph: Database from which PanLex was derived

PanDictionary: Translation-inference system based on TransGraph:

- Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Michael Skinner, and Jeff Bilmes, "Compiling a Massive, Multilingual Dictionary via Probabilistic Inference", 02009.
- Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Kobi Reiter, Michael Skinner, and Jeff Bilmes, "Panlingual Lexical Translation via Probabilistic Inference", 02010.

Translation inference in PanDictionary



PanDictionary infers that meanings 1 and 3 are identical, so “udaherri” is a translation of “koanga”.

Responsibility for translation inference

- PanLex project:
  - Produces data on *direct* (attested) translations.
  - Provides basic inference for demonstration and exploration.
- Researchers and developers:
  - Investigate methods of inference from direct translations.
  - Produce state-of-the-art inference tools, applications, and services.

Translation inference in PanLem

PanLem:

- An expert UI for PanLex
- Provides basic translation inference

What kind of translation inference?

- Distance-2 indirect translation (1 intermediate expression).
- Heterogeneous (2 translations by distinct sources).

Why distance = 2?

- Complexity: 3+ expensive in time and storage.
- Validity: “Semantic drift” grows with distance.

Why heterogeneous?

- Validity: Semantic drift more likely if homogeneous.

Translation inference in PanLem

Indirect translation

Validity problem with homogeneous indirect translation:

- Example:

English	Russian	French	German
world	мир	monde	Welt
peace	мир	paix	Frieden

“world” = “мир” = “paix”, but “world” ≠ “paix”
Distinct entries in a source are usually semantically different.
Entries in distinct sources are less often so.

Translation inference in PanLem

Indirect translation

PanLem also estimates indirect-translation qualities. Method:

- Find all heterogeneous distance-2 paths between 2 expressions.
- For each, multiply its 2 sources' estimated qualities (0–9).
- Add the products.

Translation inference in PanLem

Best 10 translations with scores

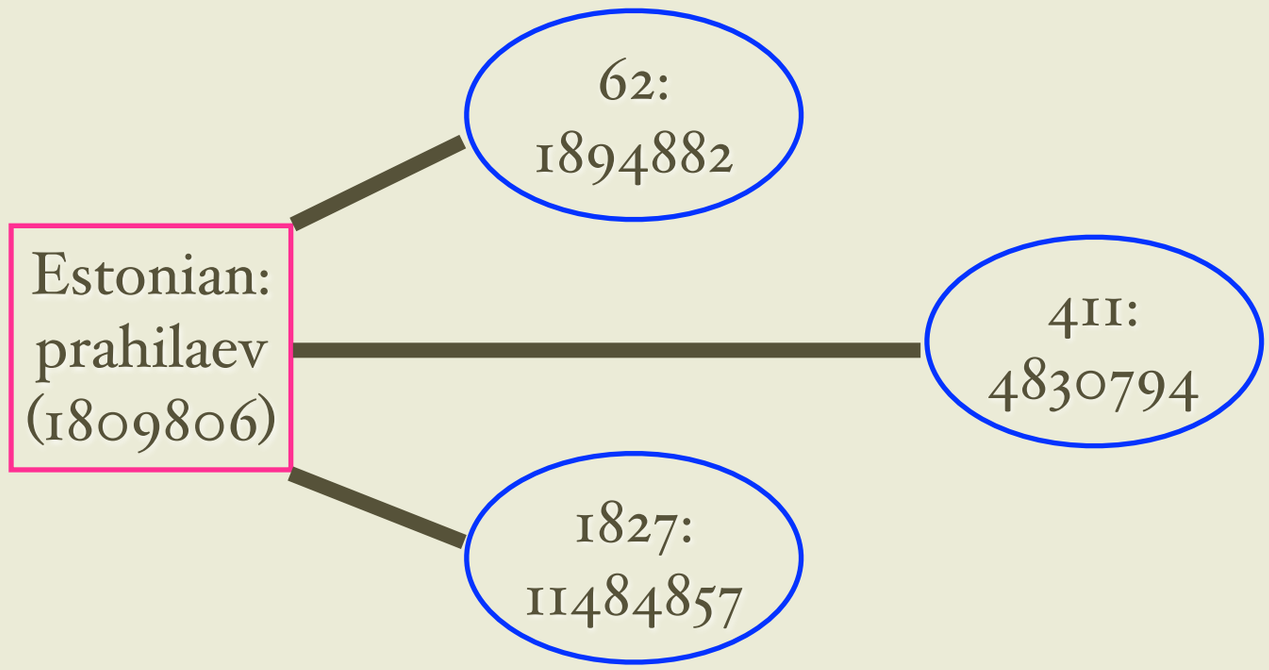
- Parameters
  - Expression to be translated: *E*
  - Language variety into which to translate it: *L*
  - Limit on the count of translations: *N*
- Example
  - *E* = Estonian: “prahilaev”
  - *L* = Russian
  - *N* = 10

Translation inference in PanLem

Best 10 translations with scores

- Step 1. Create a table of the meanings of *E* and those meanings' sources.

ap	mn
62	1894882
411	4830794
1827	11484857




```
$dbh->do(
    'create temporary table temp0mn on commit drop as '
    . "select ap, mn.mn from dn, mn where dn.ex = $_[0] and mn.mn = dn.mn"
);
# Create a temporary table of all meanings of the specified expression and their sources.
```

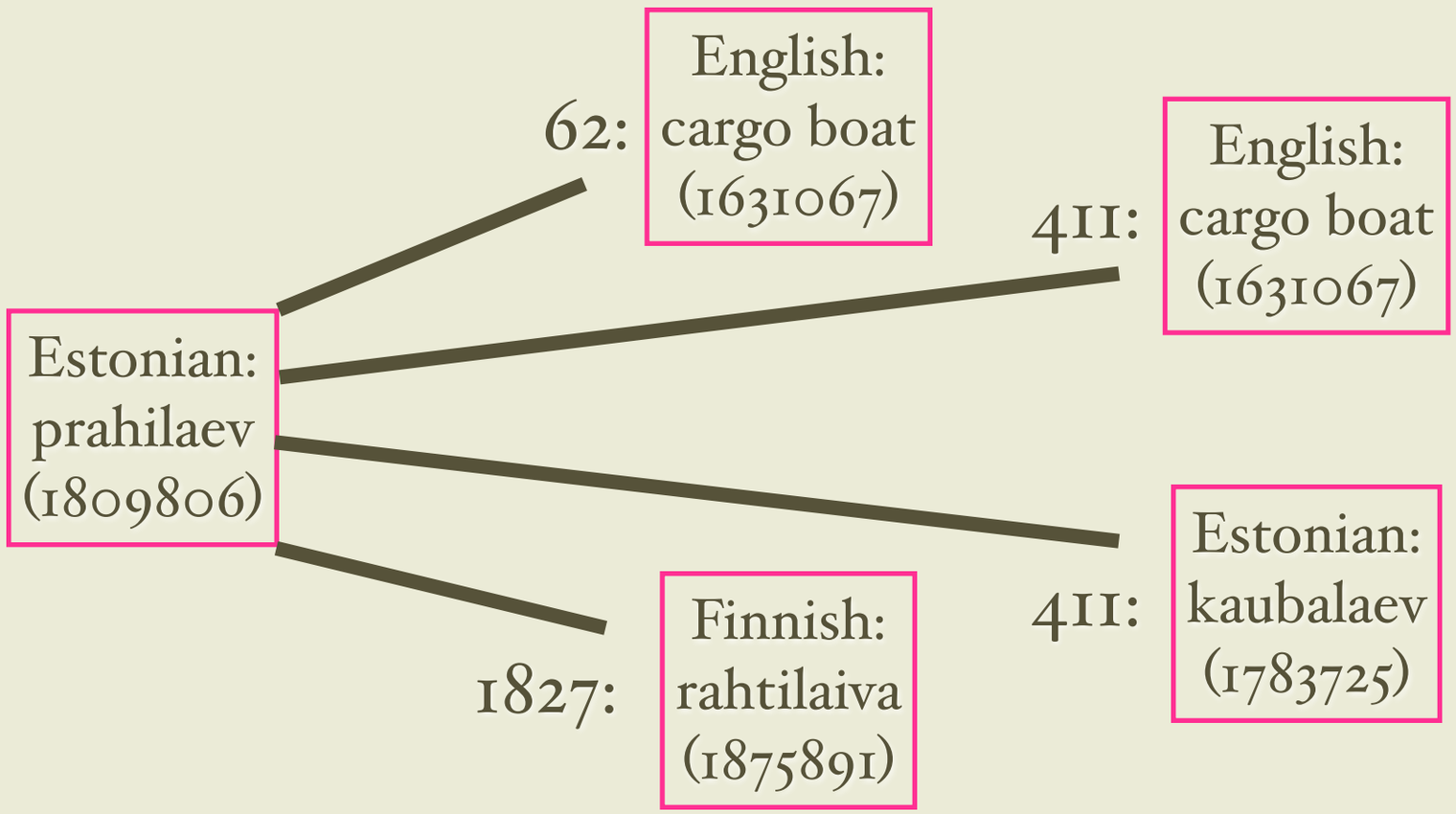
This and subsequent code is from subroutine TrtrQ in state file tvviz5w.pl

Translation inference in PanLem

Best 10 translations with scores

- Step 2. Create a table of the direct translations of *E* and their sources.

ap	ex
62	1631067
411	1631067
411	1783725
1827	1875891



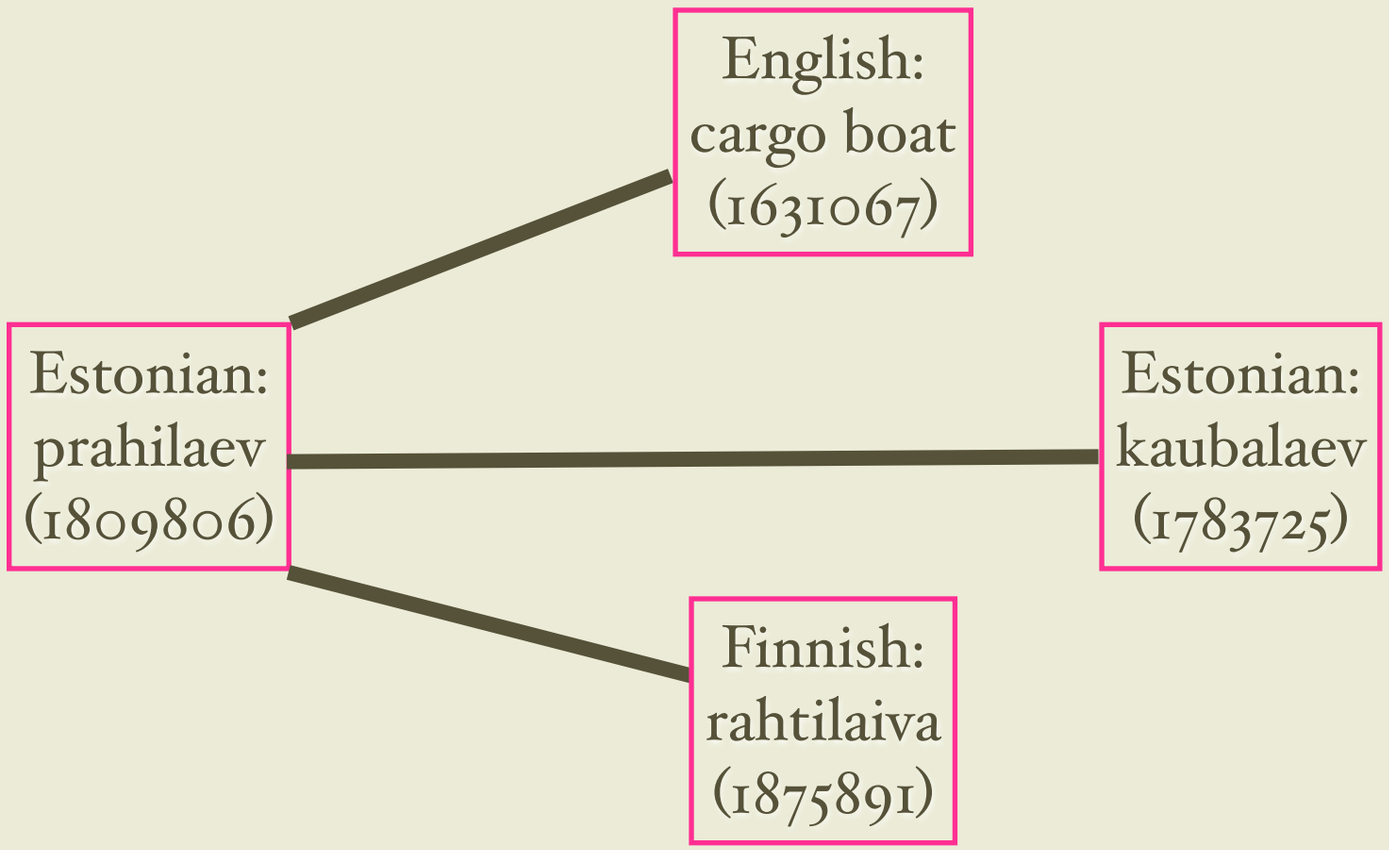

```
$dbh->do(
    'create temporary table temp0tr on commit drop as '
    . "select distinct ap, ex from temp0mn, dn where dn.mn = temp0mn.mn and ex != $_[0]"
);
# Create a temporary table of all distinct combinations of those meanings' sources and those
# sources' direct translations of the specified expression, i.e. expressions with those
# meanings except the specified expression.
```

Translation inference in PanLem

Best 10 translations with scores

- Step 3. Create a table of the direct translations of *E*.

ex
1631067
1783725
1875891

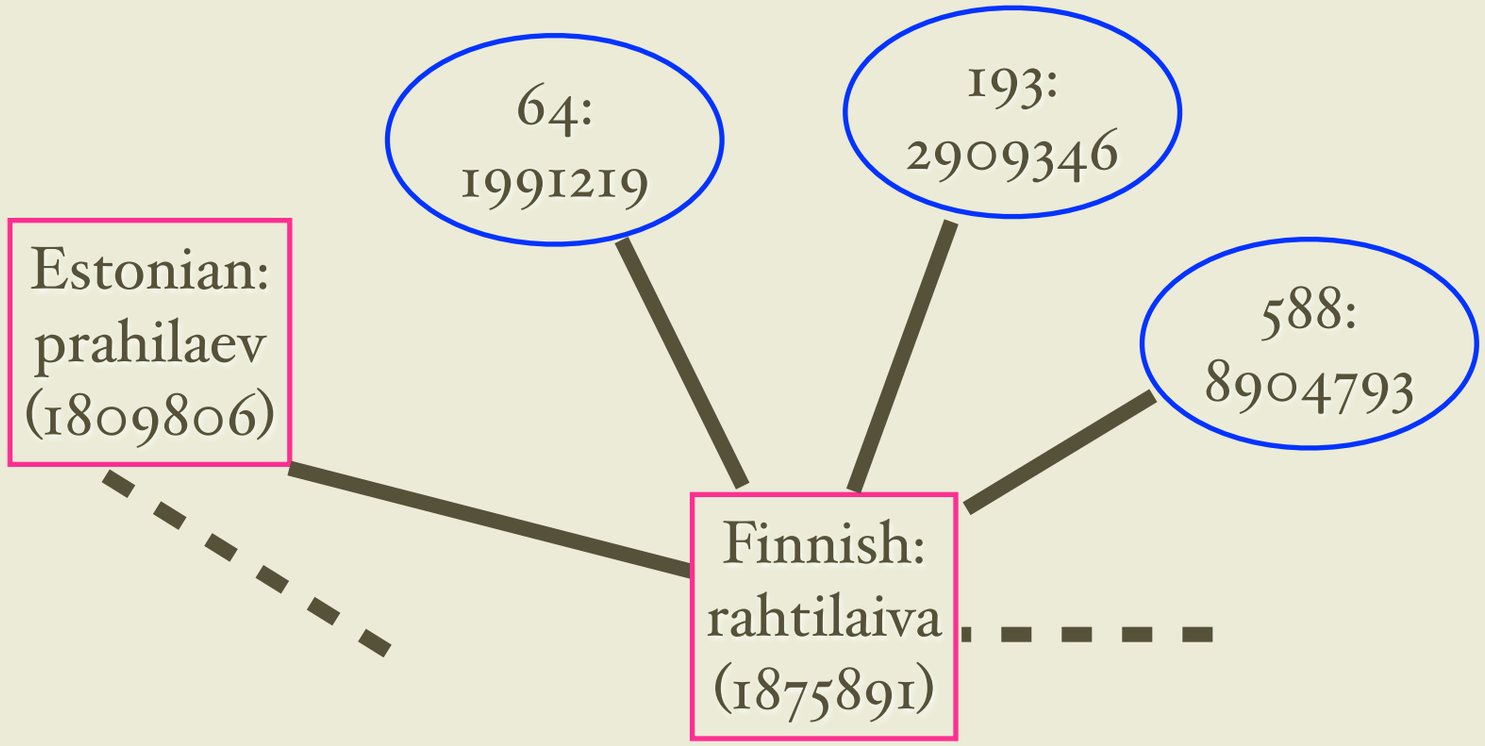



```
$dbh->do('create temporary table temp0ex on commit drop as select distinct ex from temp0tr');
# Create a temporary table of all those distinct direct translations of the specified expression.
```

Translation inference in PanLem

Best 10 translations with scores

- Step 4. Create a table of the direct translations of *E*, their meanings, and their sources.



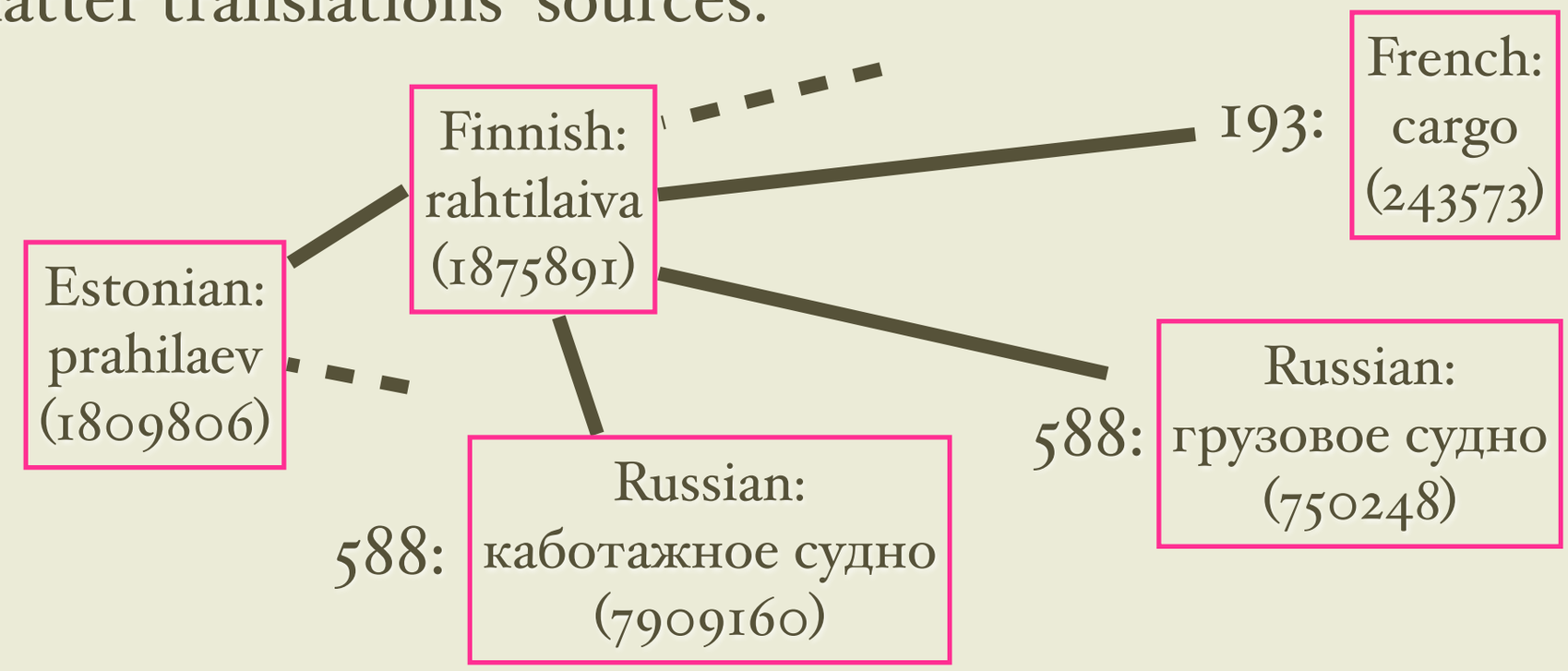
ex	ap	mn
1631067	62	1894882
1631067	68	2046115
1631067	116	2436134
1631067	116	2449647
1631067	123	2485030
1631067	411	4830794
1631067	422	5044265
1631067	563	7340025
1631067	571	7900041
1631067	1523	18351618
1631067	1890	14324463
1631067	2562	17179450
1631067	2566	17275450
1783725	62	1883065
1783725	62	1894885
1783725	62	1894895
1783725	62	1922661
1783725	411	4821827
1783725	411	4827719
1783725	411	4830794
1783725	411	4830797
1783725	411	4830807
1783725	411	4852058
1783725	411	4868066
1783725	411	4891307
1783725	1827	11477382
1875891	64	1991219
1875891	193	2909346
1875891	588	8904793
1875891	1441	11716326
1875891	1710	15489575
1875891	1710	15493201
1875891	1827	11484857

```
$dbh->do(
    'create temporary table temp1mn on commit drop as '
    . 'select dn.ex, ap, dn.mn from temp0ex, dn, mn '
    . 'where dn.ex = temp0ex.ex and mn.mn = dn.mn'
);
# Create a temporary table of those direct translations, all their meanings,
# and those meanings' sources.
```

Translation inference in PanLem

Best 10 translations with scores

- Step 5. Create a table of the direct translations of *E*, their direct translations other than *E*, and the latter translations' sources.




```
$dbh->do(
    'create temporary table temp1tr on commit drop as '
    . 'select distinct temp1mn.ex as ex1, ap, dn.ex as ex2 from temp1mn, dn '
    . "where dn.mn = temp1mn.mn and dn.ex != temp1mn.ex and dn.ex != $_[0]"
);
# Create a temporary table of all distinct combinations of those direct
# translations, their meanings' sources, and all their direct translations
# differing from both the specified expression and its direct translations.
```

ex1	ap	ex2
1631067	68	243573
1631067	116	2344398
1631067	116	2353233
1631067	123	2056015
1631067	411	1783725
1631067	422	7654272
1631067	563	750248
1631067	571	750248
1631067	1523	2056015
1631067	1890	13947545
1631067	2562	368212
1631067	2562	368215
1631067	2562	419960
1631067	2562	1131413
1631067	2562	12556265
1631067	2562	15730838
1631067	2566	12556265
1783725	62	342897
1783725	62	368212
1783725	62	419960
1783725	62	1631070
1783725	411	342897
1783725	411	360936
1783725	411	368212
1783725	411	419960
1783725	411	476971
1783725	411	599318
1783725	411	1046968
1783725	411	1631067
1783725	411	1631070
1783725	411	1779343
1783725	411	1783727
1783725	411	1783914
1783725	411	1783918
1783725	411	1809810
1783725	411	1820710
1783725	411	1824654
1783725	411	7600471
1783725	1827	232308
1875891	64	368212
1875891	193	243573
1875891	588	750248
1875891	588	7909160
1875891	1441	419960
1875891	1710	1433254
1875891	1710	14830495

Translation inference in PanLem

Best 10 translations with scores

- Step 6. Create a table of the sources and expressions producing the heterogeneous indirect translations of *E* in the specified language variety.

ap0	ex1	ap	ex2
62	1631067	563	750248
62	1631067	571	750248
411	1631067	563	750248
411	1631067	571	750248
1827	1875891	588	750248
1827	1875891	588	7909160

English: cargo boat (1631067)

Finnish: rahtilaiva (1875891)

Russian: грузовое судно (750248)

Russian: каботажное судно (7909160)

```
$dbh->do(
    'create temporary table temp2tr on commit drop as '
    . 'select temp0tr.ap as ap0, temp1tr.* from temp0tr, temp1tr, ex where ex1 = temp0tr.ex '
    . "and temp1tr.ap != temp0tr.ap and ex.ex = ex2 and lv = $_[1]"
);
# Create a temporary table of the combinations of sources and expressions constituting
# 2-source, 2-hop translations of the specified expression into the specified variety.
```

Translation inference in PanLem

Best 10 translations with scores

- Step 7. Introduction.

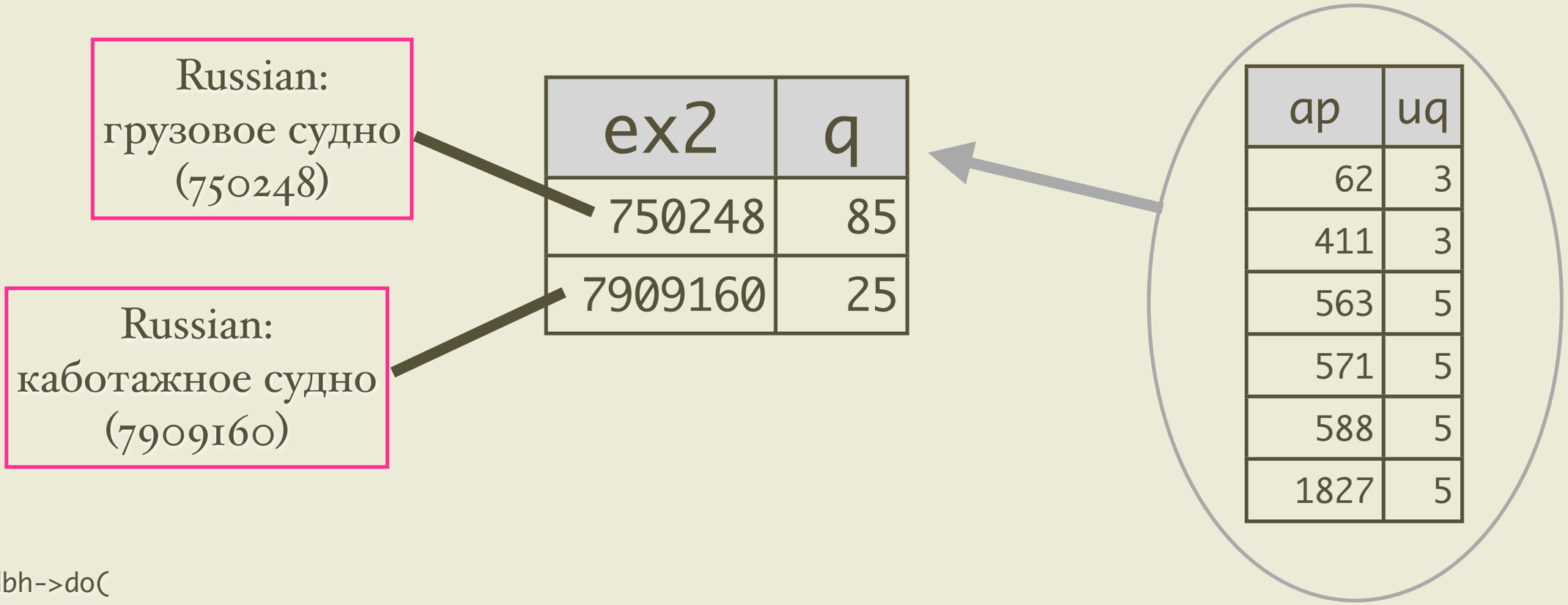
Heterogeneous indirect translation quality

Let E and G be expressions, let G be a heterogloss of E, let $\{A, B\}$, $\{C, D\}$, ... $\{Y, Z\}$ be the source pairs of the heterogeneous indirect translations between E and G, and let $q(I)$ be the editor-estimated quality of source I. Then the *quality* of G as a heterogloss of E is $q(A)q(B) + q(C)q(D) + ... + q(Y)q(Z)$.

Translation inference in PanLem

Best 10 translations with scores

- Step 7. Create a table of the heterogeneous indirect translations of *E* in the specified language variety and their qualities.




```
$dbh->do(
    'create temporary table tempq on commit drop as '
    . 'select ex2, sum (ap0.uq * ap1.uq) as q from temp2tr, ap as ap0, ap as ap1 '
    . "where ap0.ap = ap0 and ap1.ap = temp2tr.ap group by ex2 order by q desc limit $_[2]"
);
# Create a temporary table of those combinations' distinct target expressions and each one's
# quality estimate, defined as the sum of the products of the estimated qualities of the 2
# sources of the combinations with it as their target expression.
```

Translation inference in PanLem

Best 10 translations with scores

- Future improvements
  - Combine direct and indirect quality-rated translation
  - Divide *uq* by 10 or more to weight direct translations more
  - 3 user alternatives: all direct, all direct + indirect separately, best