



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Geographic visualisation of place names in Swedish literary texts

Dana Dannélls, Lars Borin, Leif-Jöran Olsson

Språkbanken
Department of Swedish
University of Gothenburg

Named Entity Recognition in Digital Humanities
Workshop
June 9-10 2015



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Geographical Information System (GIS)

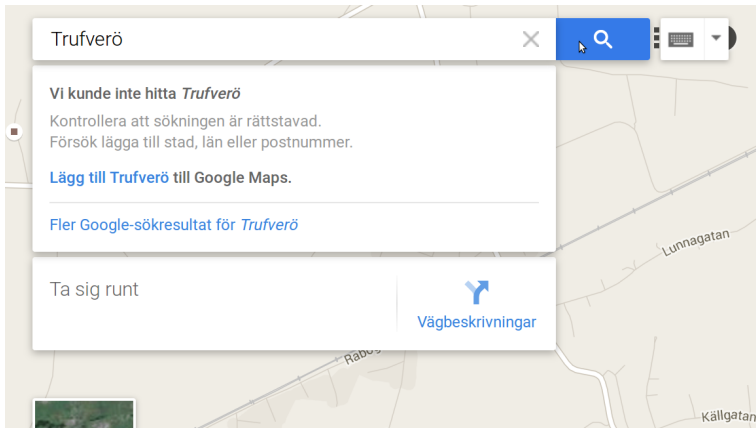
- ▶ System for capturing, storing, checking and displaying data.
- ▶ Data is usually presented in a form of point, line pixel, or polygon can be combined with data that are in table form, or already in map form.
- ▶ It is well suited to mapping data, but also allows to explicitly research the geographic aspects of the data and change over time (favored in DH).
- ▶ Multiple layers of information can be displayed on a single map (rivers, roads, pollution, population, vegetation, etc.)
 - ▶ Google Maps



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Google Maps





GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Motivation

Geographical locations which are found in older literary texts – e.g. no longer existing places or older name variants – are usually not available.

The maps available on the internet are often non-distributable.

We want to have meaningful data so we can answer questions like:

- “where does the plot of the story take place?”
- “what are the spelling variants of a place name for a certain period?”
- “how has the location of places changed over time?”



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Challenges

- ▶ How to recognize place names in historical texts
 - ▶ lack of a standard orthography
 - ▶ morphological variation
- ▶ How to render digital maps to present these historical locations
 - ▶ missing place names in databases
 - ▶ missing place name coordinates



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Språkbanken

Språkbanken, 'the Swedish Language Bank', is a research unit which focuses on developing open linguistic resources and tools for use by researchers and online visitors from different research fields.

The corpus resources offer access to a vast amount of written historical and literary texts.

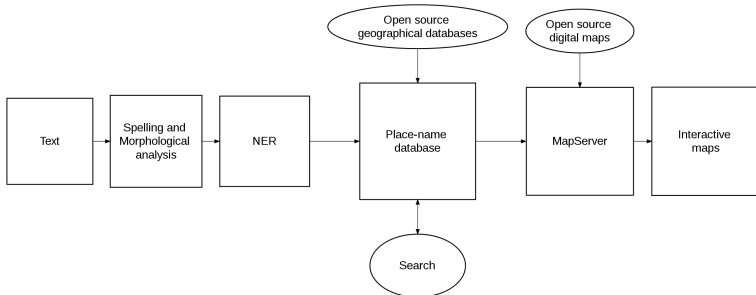
The lexicon resources offer access to modern and historical lexicons.



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Method overview





Spelling variation of place names

In text collections from the 18th and 19th centuries, we find the place names 'Lapland' and 'Laplandiya' which are spelling variants of the province **Lapland**.

1734 ÅRS FÖRARBETEN

eller ock ett häradsting om åhret hålles och hållas kan, såsom uti **Lapland** och Kåpparbärgs lähn eller flerestädes.

NYSVENSKA, ÖVRIGT – FORNSVENSKA TEXTBANKENS MATERIAL

iska länder i gemen, det lämpa utlänningarna på oss, men wi på **Lapland**.

AFTONBLADET 1850-TALET

olyper og Echidnodermer .» M: r Wolley 1 » om Lemsensvserme i **Lapland** 2 » om de forskjellige Siags Rovdyr og Rovfugle og om de sidstes Forplantning .»

KALMAR 1870-TALET

it i tillfälle att fe en i England utförd litografi, föreställande » The **Lapland** gigante, the tallest woman in the world » famt Hennes närmaste » nförwandter.

KALMAR 1910-TALET

n i Titanic-katastrofen utspelades i måndags morse, då ångaren **Lapland** ankom till Plymouth med de 107 man af Titanics besättning, hvilka räddades vid

TIDNING FÖR WENERSBORGS STAD OCH LÄN 1890-TALET

" taga upp ånga för att ej drifva på land, medan - ångfartyget " **Lapland** ", tillhörande Donald-Currie-linien, ej kunde motstå stormen utan måste söka sk

LITTERATURBANKEN (stödjer ej utökad kontext)

Denmark, Norway, Sweden, Ice- land and **Lapland**.

) [3869 Lachesis Lapponica or a tour in **Lapland**, now first published from the original manuscript journal of the celebrated Linnx
erbi, 1773-1846, författade Travels through Sweden, Finland and **Lapland**, 1802 (övers.

12 Majj 1741 prest- vigd til Missionarius i **Lapland** tillika med en annan sin Landsman och äfven Medelpadus Mag: Petrus Holmbon
än thet samma täckes tekna / Om Lemlar menar iag them man i **Lapland** finner Vid: Terser: an- Och man för några Åår haar seet (som een notat: in Exod: §
Inde dictum putant Sewaberg uti Wermeland / ock Saivovari uti **Lapland** / fördy Hedningarna hafwa haft sin Helgedom ther.



Spelling variation

Levenshtein distance calculations combined with a more specific linguistically informed method for distinguishing not only between different spelling variants but also between different variants given a certain period.

e → ä :	0.2	<i>Stre<u>ng</u>näs Str<u>äng</u>näs</i>
W → V :	0.27	<i><u>W</u>retstorp <u>V</u>retstorp</i>
fv → v :	0.31	<i>Skäl<u>f</u>vum Skäl<u>v</u>um</i>
mp → m :	0.45	<i>h<u>am</u>p<u>n</u> h<u>a</u>m<u>n</u></i>

(Ahlberg & Bouma, 2012; Adesam et al., 2012)



[Logga in](#)

Utveckling

lag nn del 45.8 🐦 🐦

bo 5.6

Spara.



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

Named entity recognizer (NER)

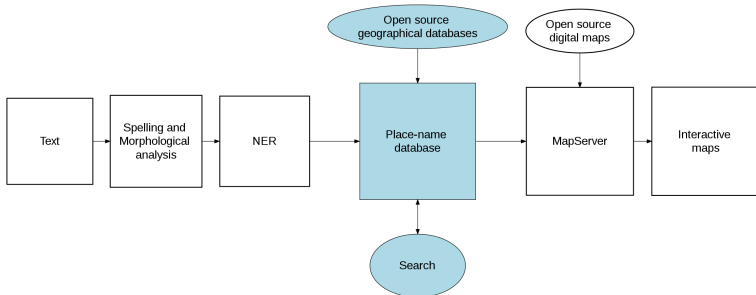
- ▶ Automatically extracts names across large collections of texts.
- ▶ Based on modern domain independent gazetteers.
- ▶ Some of the place names appearing in old literary texts are not always recognized.
- ▶ NER is combined with a place name lexicon for specific time periods.



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Placename database





GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

GeoNames geographical database

geonameid :	integer id
name :	name of geographical point (utf8)
asciiname :	name of geographical point (ascii)
alternatenames :	alternatenames
latitude :	latitude in decimal degrees
longitude :	longitude in decimal degrees
feature class :	see codes
feature code :	see codes
country code :	ISO-3166 2-letter country code
cc2 :	alternate country codes
admin1 code :	fipscode
admin2 code :	code for 2nd administrative division
admin3 code :	code for 3rd administrative division
admin4 code :	code for 4th administrative division
population :	bigint (8 byte int)
elevation :	in meters, integer
gtopo30 :	average elevation of 30'x30'
timezone :	the timezone id
modification date :	date of last modification



GeoNames data

36	Europe/Stockholm	2012-01-17				
604726	Lappeasuvanto	Lappeasuvanto	Lappeasuvanto,Lappiosuvanto	67.48333		
	Stockholm	2012-01-17				
604727	Lappeasuvanto	Lappeasuvanto		67.48333	21.15	S F
604728	Lappea Lappea	Lappea,Lappekoski		67.16667	23.56667	F
604729	Lappberget	Lappberget		66.3	22.78333	T F
604730	Lappberg	Lappberg		67.58333	20.2	S F
604731	Lappbäcken	Lappbacken		65.81667	23.41667	F
604732	Lapmovare	Lapmovare		65.95	20.06667	T F
604733	Lapland Lapland	Lapland,Laplandiya,Lappi,Lapland		68	25	L
604734	Lapisnenä	Lapisnena		68.68333	21.31667	1
604735	Lapisäive	Lapisäive	Laapisoivi,Lapisäive,Lapisäive	68.7		2
	Stockholm	2012-01-17				
604736	Lapiojoki	Lapiojoki		66.03333	23.73333	F
604737	Lantjärv	Lantjarv	Landtjarv,Landtjärv,Lantjarv,Lantjärv			F
	Europe/Stockholm	2012-01-17				
604738	Läntinen Matkajärvi	Lantinen Matkajarvi	La Matkajarvi,Lantinen Ma			
14		0	224	Europe/Stockholm		2
604739	Lansjärv	Lansjaerv	Over Lansjarv,Övre Lansjarv,Över Lansjärv			
68	96	Europe/Stockholm	2011-10-30			

Problem: spelling variation for specific time periods
and no longer existing place names.



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

No longer existing place names

Extracted from our corpora resources and soon also from Lantmäteriet (the Swedish mapping, cadastral, and land registration authority).

Example 1: The capital of Norway is being referred to as 'Christiania' when mentioned in novels between 1624 and 1877 and as 'Kristiania' from 1877 to 1925, and after that as 'Oslo'.

Example 2: When the name 'Danzig' appears with its German name in a Swedish novel that is written before 1980, it is likely to refer to the Polish city 'Gdansk'.



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Språkbanken's place name database

Språkbanken's database differs from the GeoNames database in at least three ways:

- (1) fewer redundant place locations;
- (2) spelling variants found for particular place names and time periods;
- (3) explicit information about place names from different time periods.



Coordinate search I

getcoordinates.php < Växjö, Gävle, Karlstad

```
--<geo>
- <geoname xml:id="K3dfabcc74bc5bb9f03b3c7e1b493e783">
  <name> Växjö</name>
  <lat>56.8777</lat>
  <long>14.8091</long>
  <feat-class>P</feat-class>
  <feat-code>PPLA</feat-code>
  <country-code>SE</country-code>
  <modification-date>2010-03-28</modification-date>
</geoname>
- <geoname xml:id="Kfb867951a140821508179e5807404e38">
  <name> Gävle</name>
  <lat>58.7333</lat>
  <long>16.6833</long>
  <feat-class>S</feat-class>
  <feat-code>FRM</feat-code>
  <country-code>SE</country-code>
  <modification-date>1994-01-14</modification-date>
</geoname>
- <geoname xml:id="K46c09cbe0829cd8dea35fb1e0c3cdca9">
  <name> Karlstad </name>
  <lat>59.4447</lat>
  <long>13.3375</long>
  <feat-class>S</feat-class>
  <feat-code>AIRP</feat-code>
  <country-code>SE</country-code>
  <modification-date>2007-07-17</modification-date>
</geoname>
- ...
```



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Coordinate search II

getcoordinates.php < Berget

```
-<geo>
  -<geoname xml:id="K461c39d660ecf629a269867cf0eda6c8">
    <name> Berget</name>
    <lat>59.9333</lat>
    <long>13.2667</long>
    <feat-class>T</feat-class>
    <feat-code>HLL</feat-code>
    <country-code>SE</country-code>
    <modification-date>1994-01-14</modification-date>
  </geoname>
  -<geoname xml:id="Kc262b0ee7f9c759d7576ac131f5af9ed">
    <name> Berget</name>
    <lat>57.8906</lat>
    <long>16.4272</long>
    <feat-class>P</feat-class>
    <feat-code>PPL</feat-code>
    <country-code>SE</country-code>
    <modification-date>2001-05-11</modification-date>
  </geoname>
  -<geoname xml:id="K5ea48bc17a75dd3d70d8f59553fe0245">
    <name> Berget</name>
    <lat>57.9</lat>
    <long>12.9167</long>
    <feat-class>S</feat-class>
    <feat-code>FRMS</feat-code>
    <country-code>SE</country-code>
    <modification-date>1994-01-14</modification-date>
  </geoname>
  -<geoname xml:id="Kc8e86e48d290e36a3e40bcc03363e1de">
    <name> Berget</name>
    <lat>57.5667</lat>
    <long>12.0833</long>
    <feat-class>P</feat-class>
```



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

GIS at Språkbanken

- ▶ The open source MapServer platform (Kropla 2005).
- ▶ The geographical data is derived from Open Street Map dataset.
- ▶ The development environment has a user interface.
- ▶ Generate interactive maps, static and dynamic.



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Place-name visualization from Swedish literary texts

Det går an from 1838 by Carl Jonas Love Almqvist mentions more than 10 place names:

*Stockholm, Riddarholmsstranden, Mälaren,
Södertelje, Strengnäs, Granfjärden,
Glanshammar, Trufverö, Västerås, Kungsör,
Westgötaland, Wenern, ...*

Nils Holgerssons underbara resa from 1962 by Selma Lagerlöf mentions more than 50 place names:

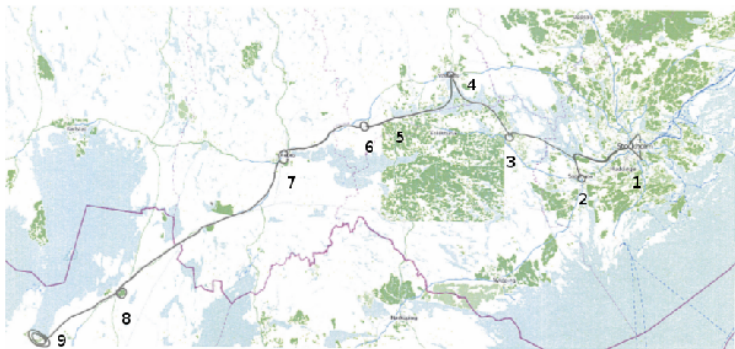
*Fjällbacka, Frösön, Garpenberg, Glimminge,
Grövelsjön, Gullöfallet, Görälven, Göta
kanal, Göteborg, Haga, Lappland, Lidingön,
Skara, ...*



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Static map generated for *Det går an*

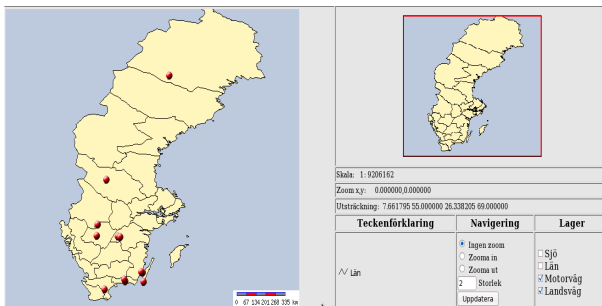




GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Dynamic map generated for *Nils Holgerssons underbara resa*





GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Conclusions

- ▶ We address some of the challenges with orthographic and morphological variation, missing place names, and missing place name coordinates.
- ▶ These challenges form a central part in the development of methods and tools for the automatic analysis of historical Swedish literary texts at our research unit.
- ▶ MapServer offers new opportunities for visualizing geographical information of place names found in our corpora.



GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Thank you!