Tvärslå – defining an XML exchange format and then building an on-line Nordic dictionary

Viggo Kann
KTH CSC
viggo@nada.kth.se

Joachim Hollman
Algoritmica HB
joachim@algoritmica.se

June 25, 2007

Abstract

Tvärslå is a dynamically expandable multilingual on-line dictionary, composed of all dictionaries used and developed in the Nordisk netordbog (Nordic Web Dictionary) project. Currently the languages included are Swedish, Danish, Norwegian, Icelandic, Finnish and English. Tvärslå can be used both interactively and called by the Tvärsök system [1]. This article describes the functionality of Tvärslå and how the system was constructed, beginning in choosing an XML format suitable for exchanging dictionaries within the project.

1 Introduction

The purpose of the Nordisk netordbog (Nordic Web Dictionary) project is to collect and create dictionaries between the Nordic languages, make them searchable on the web, and use these resources to automatically translate web queries in one of the languages to the other languages, in order to find matching web pages in the other languages. The motivation for this is the fact that people in the Nordic countries most often can read texts written in other Nordic languages but are not able to construct search queries in any other language than their own. For the Nordic council of ministers this is a real problem, since the web pages on their web site are written in some but not all of the Nordic languages, and a user searching on the web site should find information also if it is written in another Nordic language. The council therefore has funded the Nordisk netordbog project with partners in all the Nordic countries.

This paper describes the part of the project called *Tvärslå* (Swedish for cross-lookup). Tvärslå is a dictionary lookup system capable of dynamically handling dictionaries in many languages. The dictionaries may be bilingual, multilingual, unilingual (containing synonyms) or any combination of these. Since existing dictionaries are encoded in very different formats, and since we wanted to be able to use data from existing dictionaries when constructing new dictionaries, for example as in [4], we had to define a format suitable for encoding electronic dictionaries.

Unfortunately there is no standard format for dictionary exchange. TEI, the Text Encoding Initiative, had a work group for *computational lexica* in 1991–1993[1], but it did not present any result. There have been a few attempts to define formats, namely TBX, TermBase eXchange[2], which is a terminology exchange format, and OLIF, Open Lexicon Interchange Format[3]. The closest thing to a common standard is probably the TEI standard for

[1] http://www.tei-c.org/Vault/AI/ai6w04.txt

[2] http://www.lisa.org/standards/tbx/

[3] http://www.olif.net/

print dictionaries[4]. We agreed on using the TEI standard as a starting point, and simplify it into a format useful to the project and hopefully also other similar projects.

2 XML format for dictionary exchange

There are two major problems with the TEI XML standard for print dictionaries. First, it is too big: the definition (DTD) is 7.000 lines. Second, the standard allows several ways to express the same thing, which is a problem at least when writing a parser for the DTD. It was clear that we had to scale down the TEI standard to be able to use it in the project. We wanted to make it as simple as possible, yet complex enough to be able to express what was needed in the project. And the need in the project was not only for Tvärslå, but for using the data to create multilingual dictionaries. An example of a dictionary entry (for the Swedish word *jätte*) in our resulting scaled down XML is seen in Figure 1.

```
<entry>
  <form>
     <orth>jätte</orth>
  </form>
  <gramGrp>
     <pos>noun</pos>
  </gramGrp>
  <def type="explanation">sagofigur som är
     mycket större än en människa</def>
  <trans lang="en">giant</trans>
  <index>jätten</index>
  <index>jättar</index>
</entry>
```

Figure 1: Example of the coding of an entry from a dictionary using our XML format.

Before the dictionary entries there is a header with information about the dictionary, see the example in Figure 2.

```
<teiHeader type="dictionary" id="lexin"
           date.created="2006-05-31">
<fileDesc>
  <titleStmt>
    <title>Lexin svensk-engelsk ordbok
    </title>
    <author>Språkrådet</author>
    <principal>Viggo Kann</principal>
  </titleStmt>
  <publicationStmt>
   <availability><p lang="sv">
      fritt inom projektet
   </p></availability>
  </publicationStmt>
  <sourceDesc><bibl>
      http://lexin.nada.kth.se
  </bibl></sourceDesc>
</fileDesc>
<profileDesc> <langUsage>
  <language id="sv" usage="source">
   svenska</language>
  <language id="en" usage="target">
   engelska</language>
</langUsage> </profileDesc>
</teiHeader>
```

Figure 2: Example of the coding of a dictionary header using our XML format.

3 Differences between TEI and our XML

We simplified the TEI P4 Print dictionaries standard a lot (from 178 kbyte to 6 kbyte) by removing superfluous elements and attributes. The only changes that are not completely compatible with the TEI standard are the following.

- Inside the *def* element we allow the element *trans* (for a translation of the definition).
- The *def* element has the new attribute type with value *definition* or *explanation*.
- The *index* element is not empty.
- A few requiredness restrictions have been changed in either direction.

[4] http://www.tei-c.org/P4X/DI.html

The DTD defining our format is available from the project web page[5]. There is also a small Python program that automatically encodes tab separated term lists in our XML format. This program makes it extremely easy to transform simple term lists to the XML format.

4 The dictionaries in Tvärslå

The main source of dictionaries used in the project is Lexin, which was primarily produced to meet the need of immigrant education, in Sweden funded by the Swedish national agency for school improvement. Lexin has later propagated to the other Scandinavian countries.

The following dictionaries are currently part of the Nordisk netordbog project:

- Swedish-Finnish Lexin dictionary (30.000 entries)[6]
- Swedish-English Lexin dictionary (32.000 entries)[7]
- English-Swedish Lexin dictionary (48.000 entries)[8]
- Danish-Swedish Lexin dictionary (4.000 entries)[9]
- Icelandic-English-Swedish Lexin dictionary (15.000 entries)[10]
- Norwegian (Bokmål and Nynorsk)-English-Swedish Lexin dictionary (20.000 entries)[11]
- People's synonym dictionary (45.000 pairs of Swedish synonyms)[12] [3]
- Danish-English term list, constructed in the project (3.000 entries)
- Scandinavian dictionary (Nordic council of ministers, 3.500 Swedish entries, 2.900 Danish entries, 3.500 Norwegian entries)[13]
- Scandinavian public administration terms, constructed in the project (Icelandic-Danish-Swedish-Norwegian-Finnish-English, 2.000 entries)
- Term lists constructed in the ScanLex project (Icelandic-Danish-Swedish-Norwegian-Finnish-English, several thousands of entries)[14]

More dictionaries will be added to Tvärslå as soon as they are available.

5 Functionality of Tvärslå

In the beginning of the project we agreed on the functionality of the Tvärslå system:

- Look up words in any of the Nordic languages and English.
- Possible to specify which language the search word is given in, or say that it can be in any language.
- Translate to specified language or any language.
- Look up forward or backward in the direction of the dictionaries (setting). Term lists of course have no direction.
- Correct misspellings (setting).
- Add dictionaries dynamically.
- See which dictionaries are loaded.

[5] http://www.csc.kth.se/tcs/projects/netordbog/

[6] http://lexin.nada.kth.se/sve-fin.html

[7] http://lexin.nada.kth.se/sve-eng.html

[8] http://lexin.nada.kth.se/sve-eng.html

[9] http://lexin.emu.dk/

[10] http://www.lexis.hi.is/lexin_ny.html

[11] http://decentius.hit.uib.no/lexin.html

[12] http://lexin.nada.kth.se/synlex.html

[13] http://www.nordskol.org/ordbog/

[14] http://uit.no/scandiasyn/scanlex/

- See which dictionary a translation originates from.

- Work both interactively and as a web service.

The Tvärslå user interface, available on `http://ordbok.nada.kth.se`, is shown in Figure 3. Figure 4 shows an example of the result of a lookup. Clicking on the name of a dictionary shows the information page about that dictionary (see Figure 5) composed of the information in the header part of the XML (see Figure 2).

6 Implementation of Tvärslå

The on-line dictionary is implemented using a *servlet* in the programming language Java. Simplified, a servlet can be said to be a program that is run on a web server and that dynamically creates web pages as responses to external parameterized requests[15]. It is important to notice that the servlet is always run on the server side and not in the web browser on the client. It is possible to create web pages dynamically in several other ways. Traditionally, so called CGI programs have been used, but there are severe problems with the response time and scalability of such solutions. Several popular web sites, for example Swedish Lexin[16], have recently substituted a CGI solution for a servlet solution.

Internally the servlet is divided into four parts:

1. Loading of the dictionaries in binary form as a separate hash table for every language pair.

2. Parsing of the parameters of a call.

3. Lookup in the dictionaries using the hash tables.

4. Presentation of the translations.

The first step is only performed once, when the servlet is initialized. Step 2, 3 and 4 are performed for every call to the dictionary. In step 2 the search word, source language(s), target language(s) and settings are collected. Step 3 can further be divided into three phases:

3.1 For each dictionary that corresponds to a valid combination of a source and a target language, the search word is looked up. If there is a translation the result of the lookup is stored.

3.2 If the user wanted the lookup to be performed also in the reverse direction step 3.1 is done again, with source and target languages swapped.

3.3 If no translation was found and the user allowed spelling correction, the search word is considered misspelled and all possible spelling corrections are checked as in steps 3.1 and 3.2. The spelling corrections are generated using the Stava method [2] using an edit distance metric. First corrections on distance 1 (differing in one letter) from the misspelled word are looked up, and only if there were no hits on these words, spelling corrections on distance 2 (differing in two letters) are looked up.

Finally, in step 4 all translations found in step 3 are transformed to an HTML text that is returned to the web browser of the user.

Every time a new dictionary is added to the system, either a new version of an existing dictionary or a completely new dictionary, the binary dictionaries that are affected have to be recreated. This is done by a Java program that parses the dictionaries in XML and produces one binary index for each language pair.

7 Tvärslå SOAP Web Service

The Tvärslå dictionary can also be accessed as a Web service[17]. This means that anyone can write a

[15] `http://java.sun.com/products/servlet/overview.html`

[16] `http://lexin.nada.kth.se`

[17] `http://en.wikipedia.org/wiki/Web_services`

program (in any language) that looks up words in the dictionary using a simple application program interface.

The interface consists of only two methods:

1. Look up a word

2. Look up an array of words

Both functions return an array of result objects (source and target language, dictionary, question, answer). The web service is implemented using the Java platform Axis[18]. There is a detailed description of the interface on the web[19].

8 Conclusions

We have shown that a very simplified variant (6 kbyte definition compared to 178 kbyte) of the TEI standard for print dictionaries is suitable for encoding dictionaries that are to be exchanged within a project and to be made searchable on-line. Tvärslå, an efficient and dynamically extendable multilingual on-line dictionary has been constructed in order to present the dictionaries that have been collected and constructed within the Nordisk netordbog project. Currently (June 2007) about 500 Tvärslå lookups are made daily. Table 1 shows the distribution of the searches with respect to the number of hits in the Tvärslå dictionaries. Table 2 shows how common the different languages are as source and target languages in Tvärslå searches.

It is very easy to extend Tvärslå with new dictionaries, as long as they are encoded in the XML format.

# of hits	distribution of searches
0	46%
1	16%
2	10%
3	6%
4	4%
5	2.5%
6	2.0%
7	1.3%
8	1.1%
9	0.9%
> 9	10%

Table 1: Distribution of the number of hits.

Language	Source	Target
all	13%	27%
Swedish	42%	32%
Danish	20%	16%
Norwegian	18%	17%
English	5%	4%
Icelandic	1.5%	2%
Finnish	1.0%	1.5%

Table 2: Distribution of questions to Tvärslå with respect to different source and target languages.

References

[1] H. Dalianis, M. Rimka, and V. Kann. Using Uplug and SiteSeeker to construct a cross language search engine for Scandinavian. In these Proceedings, 2007.

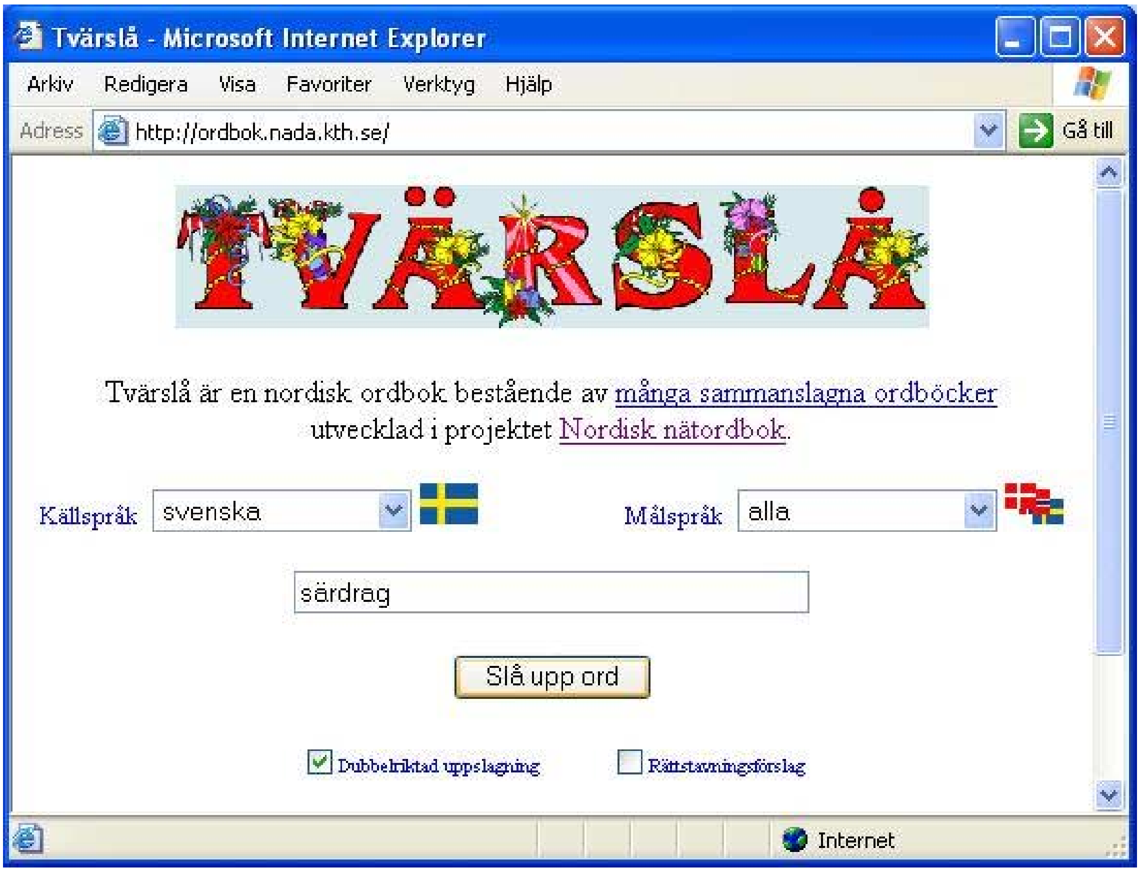
[2] R. Domeij, J. Hollman, and V. Kann. Detection of spelling errors in Swedish not using a word list en clair. *J. Quantitative Linguistics,* 1:195–201, 1994.

[3] V. Kann and M. Rosell. Free construction of a free Swedish dictionary of synonyms. Nodalida 2005, Joensuu, 2005. See also `http://www.csc.kth.se/tcs/projects/infomat/rapporter/kannrosell05.pdf`

[4] J. Sjöbergh. Creating a free digital Japanese-Swedish lexicon. In Proceedings of PACLING 2005, pages 296–300, Tokyo, 2005.

[18] `http://ws.apache.org/axis/`

[19] `http://ordbok.nada.kth.se:8070/axis/services/NordicDictionaries?wsdl`



Adress http://ordbok.nada.kth.se/

Tvärslå är en nordisk ordbok bestående av många sammanslagna ordböcker utvecklad i projektet Nordisk nätordbok.

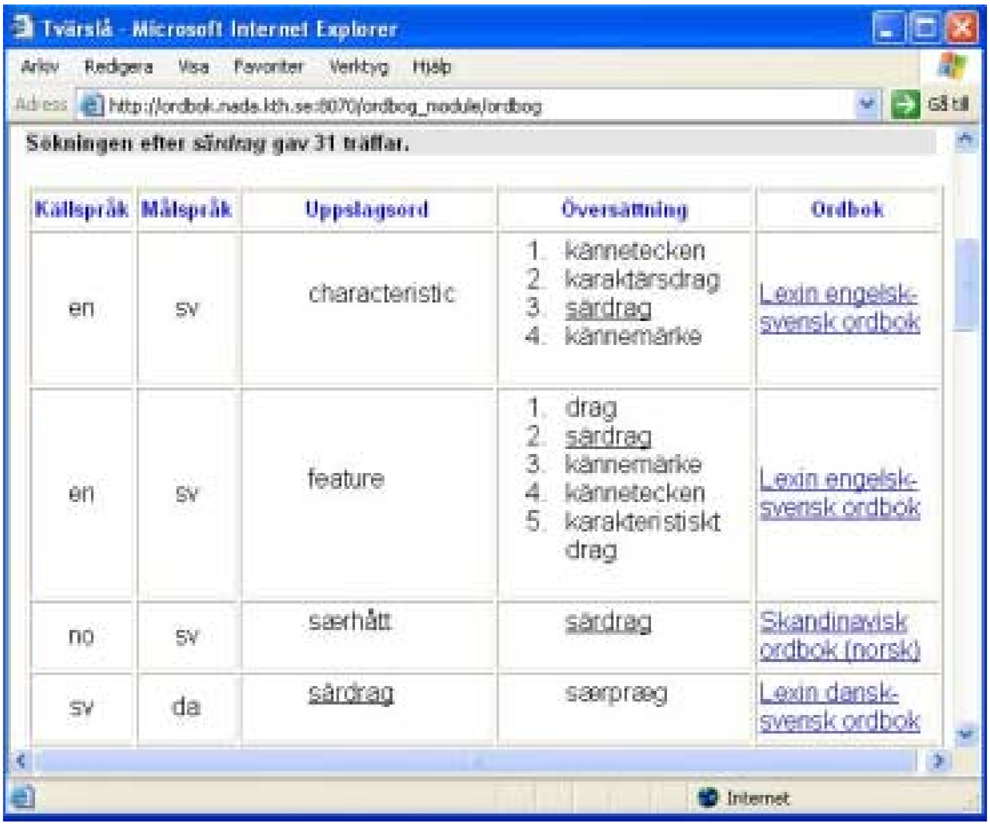
Källspråk svenska

Målspråk alla

särdrag

☑ Dubbelriktad uppslagning □ Rättstavningsförslag

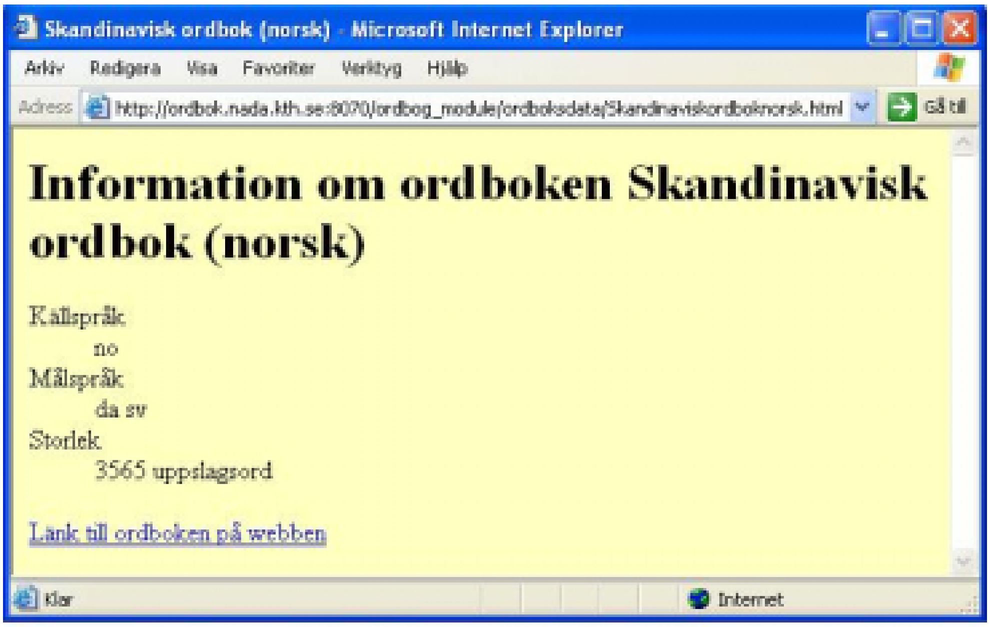
Figure 3: The Tvärslå interface. The Swedish word *särdrag* is looked up.


Tvärslå - Microsoft Internet Explorer
Arkiv Redigera Visa Favoriter Verktyg Hjälp
Adress http://ordbok.nada.kth.se:8070/ordbog_module/ordbog
Gå till
Internet

Sökningen efter särdrag gav 31 träffar.

Källspråk	Målspråk	Uppslagsord	Översättning	Ordbok
en	sv	characteristic	1. kännetecken 2. karaktärsdrag 3. särdrag 4. kännemärke	Lexin engelsk-svensk ordbok
en	sv	feature	1. drag 2. särdrag 3. kännemärke 4. kännetecken 5. karakteristiskt drag	Lexin engelsk-svensk ordbok
no	sv	særhått	särdrag	Skandinavisk ordbok (norsk)
sv	da	särdrag	særpræg	Lexin dansk-svensk ordbok

Figure 4: The result of the lookup of *särdrag*.


Skandinavisk ordbok (norsk) - Microsoft Internet Explorer
Arkiv Redigera Visa Favoriter Verktyg Hjälp
Adress http://ordbok.nada.kth.se:8070/ordbog_module/ordboksdata/Skandinaviskordboknorsk.html
Gå till
Klar
Internet

Information om ordboken Skandinavisk ordbok (norsk)

Källspråk
    no
Målspråk
    da sv
Storlek
    3565 uppslagsord

Länk till ordboken på webben

Figure 5: A dictionary information page of Tvärslå.