Making English-Norwegian parallel corpora from websites

Report on work carried out by Uni Research on behalf of Nasjonalbiblioteket.

Paul Meurer and Andrew Salway, 26 February 2018. Last revision: 2 April 2018.

1. Summary

The aim of the work reported here is to assess the potential for creating English-Norwegian parallel text corpora using material from (mainly) public websites. Section 2 gives an overview of the parallel text that appears to be available from 21 public websites for at least one of the language pairs English-Bokmål, English-Nynorsk or Bokmål-Nynorsk, and it mentions some other possible sources. Section 3 analyses the most promising websites in terms of how pairs of web pages can be identified and aligned, the degree of actual parallelism between the content of aligned pages, and the quantity of textual material available for a parallel corpus. Section 4 is dedicated to the task of sentence alignment which seems to be the most challenging step in making parallel corpora. Then in Section 5 we step through each of the stages for making parallel corpora from two public Norwegian websites (nav.no and nyinorge.no). This results in two gold standard parallel corpora (one English-Bokmål, one English-Nynorsk) and provides some insights and suggestions about the automated processing of website material into parallel corpora. The main findings and deliverables are described in Section 6.

2. Overview of websites

Public sites

The starting point for gathering websites was the site www.norge.no, a portal to other pages of the public sector. We examined the following 21 sites that are linked to from www.norge.no:

www.skatteetaten.no, www.bufdir.no, www.regjeringen.no, www.nav.no, www.politi.no, www.forbrukerrådet.no, www.posten.no, www.husbanken.no, www.domstol.no, www.kartverket.no, helsenorge.no, www.ssb.no, www.innovasjonsnorge.no, www.vitnemalsportalen.no, www.nyinorge.no, www.studyinnorway.no, www.visitnorway.no, www.arbeidstilsynet.no, www.samordnaopptak.no, forsvaret.no, www.lanekassen.no

Several public websites not mentioned on www.norge.no were also included:

www.vegvesen.no, www.toll.no, www.imdi.no, www.udi.no, www.miljodirektoratet.no, www.workinnorway.no, www.statped.no

For each of the listed websites we browsed through a sample of pages in order to determine the extent to which parallel texts are available. Table 1 below lists the web sites that seem most promising as sources of parallel corpora and notes the extent to which we have processed material from the site, e.g. “harvested”, “harvested and page aligned” or “converted to parallel corpus”. In the table the code ‘nor’ stands for a mixture of Bokmål and Nynorsk, in the case there is no clear separation between Bokmål and Nynorsk.

Table 1. Potential sources for parallel material

Site	Languages	Status
www.norge.no	nor, eng	Harvested
www.skatteetaten.no	nor, eng	Converted to parallel corpus
www.regjeringen.no	nor, eng, smi	Harvested
www.nav.no	nor, eng, smi	Converted to parallel corpus
www.nyinorge.no	nob, eng	Converted to parallel corpus
www.forbrukerradet.no	nor, eng	Harvested
www.domstol.no	nob, nno, eng, smi	Harvested and page aligned
www.toll.no	nor, eng	Harvested and page aligned
www.ssb.no	nor, eng	Harvested
www.vitnemalsportalen.no	nor, eng	Harvested
www.arbeidstilsynet.no	nor, eng, pol	Not harvested
www.udi.no	nob, nno, eng	Harvested
www.workinnorway.no	nor, eng	Not harvested

In the following sites we found pages in different languages but at most these pages were only marginally aligned: www.bufdir.no (nob, nno, eng), www.husbanken.no (nor, eng), www.kartverket.no (nor, eng), www.innovasjonsnorge.no (nor, eng), www.visitnorway.no (nor, eng, and 11 other languages), www.samordnetopptak.no (nor, eng), www.lanekassen.no (nob, nno, eng), www.vegvesen.no (nob, nno, eng), www.imdi.no (nob, eng), www.miljodirektoratet.no (nor, eng, smi).

Finally, the following websites appear to be almost entirely monolingual: www.politi.no (nor), www.posten.no (nor), helsenorge.no (nor), forsvaret.no (nor), www.studyinnorway.no (eng).

Further analysis of the 21 websites is provided in an accompanying spreadsheet (see Section 6).

Other potential sources of parallel text

Corporate web sites. We identified several corporate websites which contain sizable amounts of parallel pages (Bokmål and English): www.statoil.no, www.hydro.no, www.orkla.no and www.exxonmobil.no. Among these, Statoil has informally given permission for their content to be used as a source of parallel material; this should be formalized in form of a contract. The website www.orkla.no has a liberal copyright (see https://www.orkla.no/ansvarserklaering/) which should mean that their content can be used as a source of parallel material. We did not receive a response from Hydro to our request. We were not sure who to contact at Exxon Mobile. Three other corporate websites were examined but found not to contain suitable material for a parallel corpus: www.dnb.no, www.telenor.no, www.akerasa.no.

Opus/Opensubtitles. http://opus.nlpl.eu/OpenSubtitles2018.php contains 8,625,000 aligned Norwegian-English sentence pairs compiled form www.opensubtitles.org. The material is open source and free.

Web sites harvested by Tilde. Tilde in Latvia has harvested many Norwegian potentially parallel web sites and translation memories. Clearing of copyright issues is ongoing, but the material has not been processed further.

3. Further analysis of potential sources

Having identified the set of potential sources (Table 1) by browsing the websites, we then looked at these in more detail in order to assess: (i) how parallel pages are linked together, and hence whether we can automatically make pairs of texts (page alignment); (ii) the extent to which parallel pages are actually parallel, e.g. are they parallel at the sentence level or not, with regards to sentence alignment; and, (iii) how much material there is on each site.

Page alignment

It turned out that each of the sites had its own idiosyncrasies and peculiarities, so no general method can be given to accomplish the task of page alignment, i.e. making pairs/triples of parallel texts. (One could conceive a machine learning approach to aligning English and Norwegian pages that could work for any set of bilingual texts, but that would be for a research project and hence not feasible in this pilot project).

That said, we observed two types of page linking schemes that are found on the examined sites: (i) linking by an explicit link to the corresponding page (via an html link, or a Javascript function call using an interpretable address, or similar); and, (ii) by similar directory structure.

The first type is clearly the easiest one to exploit, as it is mechanical and explicit.

The second type is easy to use if the paths leading to parallel pages are identical except for a language-specific prefix. However, in most instances, the path leading to the English translation of a Norwegian page is the English translation of the path to the Norwegian page (which uses Norwegian path names, often composed of the page titles). This makes an automatic matching

of corresponding pages quite difficult. We have not attempted so far to devise an algorithm that could accomplish this, and it might even be most appropriate to do the matching manually.

Degree of parallelism

Even when English and Norwegian pages are obviously and overtly aligned (say by a language-switch link), it does not follow that the pages are translations of each other. Often they are (e.g. nyinorge.no) but often they are also only partial translations, for a variety of reasons. The English version might be adapted to the special needs of a non-Norwegian audience, or recent changes to the Norwegian text might not have found their way into the English version. For example, both of these cases can be found in nav.no. Such incomplete partial parallelism poses challenges to the task of sentence alignment, as discussed below. In other cases, the English texts paraphrases or summarizes the content of the Norwegian text; this is common in regjeringen.no and domstolen.no. There are also websites where there are linked Norwegian and English pages but their content is independent (not parallel), e.g. bufdir.no. Furthermore, in other cases (regjeringen.no), English HTML text might be linked to PDF text on the Norwegian side. We have not looked into converting PDF files into text files suitable for making corpora, but prior experience tells us that this can be challenging depending on the nature of the PDF content.

Text extraction and size estimation

In order to estimate the amount of available parallel text, we harvested the html pages of all of the promising sites and extracted the relevant portions of the texts. This was done using the open-source generic text extraction tool jusText[1], which strips off boilerplate and returns the relevant text with reasonable accuracy for a rough estimate of text size. In addition, pairs of pages were identified where it was possible without too much manual work, i.e. mainly using page links as detailed above. In this way, the amount of parallel text on each website could be estimated.

The largest site by far is www.regjeringen.no with 10 million English words, although it is not easy to get an overview on how parallel the texts are. This is followed by domstol.no (eng: 800,000 words, nob: 700,000 words, nno: 120,000 words), skatteetaten.no (eng: 310,000 words, nob: 250,000 words, nno: 220,000 words) and nav.no (eng: 100,000 words, nob: 84,000 words, nno: 65,000 words). (In these numbers, only supposedly parallel texts are included.) These four websites comprise the vast majority of the available material. The remaining sites together comprise about 800 parallel pages, amounting to around 150,000 words per language.

For the sites deemed feasible for parallel text extraction, we have harvested and aligned most of the parallel pages (see Table 1). Our expectation is that extracting the relevant text content is not too time consuming using a similar method as utilized for nav.no (see Section 4). An exception are those pages whose content (on the Norwegian side) is found in PDF documents they are linked to: this primarily concerns links from regjeringen.no and domstolen.no to

[1] https://github.com/miso-belica/jusText

lovdata.no. Here, a suitable technique will have to be developed unless off-the-shelf tools are good enough.

4. Sentence alignment

The most challenging task for making parallel corpora will normally be sentence alignment. In the first step, the paragraphs of the texts were split into sentences, using a finite-state-based sentence splitter. To assess the potential for an automated solution we evaluated three widely used open-source tools – hunalign[2], yasa[3] and champollion[4] – using the extracted and aligned nav.no and nyinorge.no pages. Among the tools, hunalign had satisfactory performance for creating parallel corpora (see Evaluation figures below). Where the tools failed, the reason seems to be that the texts contain noisy data. They are only partially parallel: often, sections are missing on one or the other side, or corresponding sections are not translations of each other. This gets the tools out of sync, and wrong alignments are produced. The confidence scores assigned to the aligned pairs by the tools are not always reliable. This is especially true for yasa (where we weren't able to make sense of the scores), and to a lesser extent for hunalign. (Champollion does not give confidence scores.) In particular, sentence pairs that clearly are unrelated sometimes get high scores if they happen to fill a gap between aligned sentences and have comparable length.

Although hunalign and yasa are language-independent, the performance of hunalign can in principle be improved by augmenting it with a dictionary of bilingual word or phrase pairs.(Champollion has to be endowed with a large dictionary to be useful at all.) We tried to do this for hunalign by extracting a word list from the LEXIN Norwegian-English dictionary, but this had no positive effects on performance; to the contrary, precision even dropped. One probable reason is that the LEXIN-derived word list is too small to be useful. Better results might be obtained with access to more comprehensive Norwegian-English word lists.

We also tried to base alignment on lemmatized text, to make the word-list information more efficient, but this had no positive effect either.

We ran champollion using the same dictionary data, and a stemmer generated from the fullform word list for Bokmål in Norsk ordbank, but performance was not satisfactory. Here, some more efforts should be made to find out if improvements are possible.

Evaluation figures

Evaluation of hunalign, yasa and champollion was done against manually aligned data from nav.no (excluding champollion) and nyinorge.no (see Section 5). The applied methodology, which is based on sentence-level alignment (every pair of aligned sentences does count), is that proposed in Langlais et al., "Methods and Practical Issues in Evaluating Alignment Techniques", 1989.

---

[2] https://github.com/danielvarga/hunalign
[3] https://github.com/rali-udem/yasa
[4] http://champollion.sourceforge.net

For nav.no the results were: hunalign precision = 0.94, recall = 0.96; yasa precision = 0.85, recall = 0.86.

For nyinorge.no the results were: hunalign precision = 0.98, recall = 0.94; yasa precision = 0.91, recall = 0.87; champollion precision = 0.91, recall = 0.87.

It is evident that hunalign performs significantly better than the other two programs. The texts pose various challenges to alignment programs, most notably missing sections, non-translational equivalents, and also transposition of sentences and paragraphs. As our experience with nav.no and nyinorge.no shows, if perfectly aligned sentences are required for a corpus then most of the work in corpus creation will have to be done in the sentence alignment step. Even if sentence alignment is to be done wholly automatically then at least a sample of sentences should be manually aligned for each website so that the performance of the automatic technique can be checked. (NB. the different performance for the two websites above).

5. Making corpora from nav.no, nyinorge.no and skatteetaten.no

We chose the sites nav.no and nyinorge.no for testing the feasibility of running the corpus building task end-to-end and to produce two gold standard parallel corpora (one English-Bokmål, one English-Nynorsk). Based on this experience, the large but nicely parallel site skatteetaten.no was also processed.

Harvesting

The sites nav.no and nyinorge.no were harvested using the program wget. This resulted in a hierarchical directory structure that replicates the path structure of the page URLs. The actual page content is found in the index.html files located in the respective leaf directories. (This implies that files other than index.html can be disregarded.)

Page linking

The sites nav.no and nyinorge.no exemplify both kinds of page linking, i.e. linking by an explicit link in the html/javascript, and by using similar path structures for pages in different languages.

www.nav.no

Since an English page can correspond both to a Bokmål and to a Nynorsk page, it makes sense to take the English pages as starting point, and to try to find their Bokmål and/or Nynorsk equivalents.

The page URLs themselves cannot easily and automatically be brought into correspondence, as the English URL is a translation of the Norwegian URL, e.g.:
`https://www.nav.no/en/Home/Benefits+and+services/Relatert+informasjon/child-support-child-maintenance`

But here, the Bokmål or New Norwegian page corresponding to an English page, if available, can be found in the language switcher drop down menu, represented as a element containing <a> elements for the languages. Searching for the string 'Bokmål (Content language option)' readily yields the <a> elements containing the page URL for the Bokmål page in the @href attribute value, and similarly for Nynorsk (in the example below we have simplified the code for exposition):

```
<div class="content-languages" role="button">
     <ul>
      <li><a class="active-lang" title="English (Content language option)">English</a></li>
      <li><a href="https://www.nav.no/no/Person/Familie/Barne+og+ektefellebidrag/Barnebidrag/Barnebidrag" title="Bokmål (Content language option)">Bokmål</a></li>
      <li><a href="https://www.nav.no/no/Person/Familie/Barne+og+ektefellebidrag/Nynorsk/barnebidrag" title="Nynorsk (Content language option)">Nynorsk</a></li>
     </ul></div>
```

www.nyinorge.no

On this site, no automatic correspondence between page URLs is available; one has to rely solely on the translational correspondence between the URLs. The only feasible way of matching the URLs was to do this manually. Thus, we extracted all English and Norwegian (Bokmål) page URLs and tried to match them by hand. This was quite time-consuming and not entirely straight-forward because of often similar URL strings that could not be matched correctly without looking at the actual pages.

The problem was compounded by the fact that some pages had no translation at all, and most of the pages had (near) duplicates. For detecting the duplicates the program neardup[5] was used. Unfortunately, the program reliably finds near duplicates only if the files are sufficiently large so it had to be applied to the html pages, and not to the extracted text files. Although all duplicates found by the program were true (near) duplicates, it did miss many duplicates. Those were identified manually after the alignment was done.

Text extraction

It turned out that jusText was not reliable enough to extract just the relevant parts of the pages (title, table of contents and body text). It often omitted relevant parts of the text, and worse it did it differently for Norwegian and English pages. (See evaluation below.) So, we wrote special purpose scripts for the two websites that took specific features of the HTML coding of the pages

[5] http://www.softcorporation.com/products/neardup/

into account in order to extract the content and some structural information with much more accuracy. Obviously, this script will have to be adapted to each further website.

The scripts take advantage of HTML attributes (class and id) attached to DIV, P, SPAN and other elements to identify relevant text. In nyinorge.no, all article text is included in a `<div class="mainContent">` element; material outside this element can safely be ignored. Inside this element, the title and body text are contained in easily distinguishable elements, viz. `<div id="intro">` and `<div class="article>`. A few elements contained therein have to be excluded (e.g. `<a class="printPage">`), but all in all, extraction can be accomplished by using a few simple deterministic rules, avoiding all the guesswork that justext has to do to cope with HTML pages of arbitrary provenance. The result is very clean and reliable. For nav.no, a very similar approach led to equally reliable extraction results.

Evaluation. We calculated performance figures for justext, measured against the custom scripts, whose extraction results were hand-checked and taken as gold standard.

nav/nob: precision 0.95, recall 0.68

nav/nno: precision 0.96, recall 0.67

nav/eng: precision 0.99, recall 0.73

nyinorge/nob: precision 0.92, recall 0.82

nyinorge/eng: precision 0.88, recall 0.84

Although the precision of justext is satisfactory in most cases, recall is not (even after tweaking some parameters), which justifies the extra work to write custom extraction scripts.

Sentence alignment

In order to get a perfectly aligned parallel corpus which could be used for evaluating automatic sentence alignment (see Section 4), we decided to do the alignment semi-automatically, i.e. by reviewing all the alignment pairs suggested by automatic processing with hunalign and then correcting where necessary. This was done in a web-based graphical tool that was written for that purpose. Both one-to-one, one-to-many, many-to-one and many-to-many alignments were created.

Results

As stated, at the end of the project work, the site skatteetaten.no was processed (and manually proof-read) in the same way.

The result are two parallel gold corpora (of approx. 350,000 words (nno) resp. 450,000 (eng) for nob-eng and 290,000 words (nno) resp. 350,000 (eng) for nno-eng) that can be used to test and evaluate other automatic or semi-automatic alignment techniques.

Sentences that could not be aligned are excluded from the corpora.

6. Main findings and deliverables

Main findings

Not all Norwegian public web sites contain English-Norwegian parallel texts. Those that do always have Bokmål, but Nynorsk is often missing.

The easiest way to process the pages is to harvest the whole site's HTML content and to process the files locally.

Pages can in most cases be automatically aligned using links in the HTML code or similar directory structure. In the latter case, problems can arise when target URLs are translations of source URLs. Here, manual alignment might be the only feasible method.

Using a generic tool (jusText) to extract the HTML page content results in unsatisfactory recall; writing custom extraction scripts seems to be best.

In a few sites, HTML text is linked to PDF files; here, a suitable extraction process will have to be devised.

Among the tested tools for sentence alignment, hunalign performs best and has satisfactory precision and recall. Since there is considerable variation among the sites, it is however not clear how general these findings are. Therefore, results should be manually checked for every site.

Deliverables

This report.

Two parallel corpora (see the Appendix for details):

- PubBEPC: A Bokmål-English parallel corpus (material from nav.no and nyinorge.no) (organized as two linked physical corpora: PubBERC/Nob and PubBERC/Eng). This has a token count of 359,000 nob / 449,000 eng and a sentence count of 26,700.
- PubNEPC: A New Norwegian-English parallel corpus (material from nav.no, as nyinorge.no only contains Bokmål) (organized as two linked physical corpora: PubNERC/Nno and PubNERC/Eng). This has a token count of 290,000 nno / 354,000 eng and a sentence count of 21,000.

CMDI-metadata for the two corpora can be found at hdl:11495/DE47-A8F8-D1C3-1 and hdl:11495/DE47-B352-6F3F-2.

A more detailed analysis of the 21 public websites that were examined can be found at http://clarino.uib.no/iness/resources/misc/parallel.xlsx.

Harvested material. The page aligned part of the harvested material (nav, nyinorge, norge, ssb, toll, skatteetaten; see Table 1) is available on request.

Appendix: Corpus documentation

The two parallel corpora that were created are both accessible for analysis in the online corpus analysis system Corpuscle - clarino.uib.no/corpuscle. The corpus files can be downloaded from http://clarino.uib.no/korpuskel/download?corpus=parallel-nob/nob&location=align-nob-eng.tar.gz and http://clarino.uib.no/korpuskel/download?corpus=parallel-nno/nno&location=align-nno-eng.tar.gz .

The corpora have the following positional attributes:

word	the corpus word token
alignment	alignment information (a pointer to the corresponding alignment unit in the target corpus
site	the text source (e.g., nav or nyinorge)

The following structural units are defined:

document	the text unit, text coming from a single web page
sa	a sentence alignment unit, grouping sentences that are aligned to the same group of target sentences
s	a single sentence