Deriving a Bi-lingual Dictionary from Raw Transcription Data

Peter Juel Henrichsen

Center for Computational Modelling of Language
Copenhagen Business School, Denmark
pjuel@id.cbs.dk

Abstract

We present a bigram-based method for deriving bi-lingual dictionary entries from two corpora of spontaneous speech (as represented in transcriptions). In contrast to e.g. [1], our method does not require translated or otherwise aligned texts; the corpora representing the source and target languages may be unrelated wrt. size, vocabulary richness, frequency distribution, and activity type. Examples are given using Danish and Swedish transcription data (and hints of English). We conclude with a discussion of the use of corpus-driven methods in language preservation and litigation projects.

1. Introduction

Is it possible to automatically derive a bi-lingual dictionary from two independent transcription corpora with no annotations of any kind? If yes, to what extent do the translations comply with a standard dictionary? What can be learned about the relations between spoken and written language in the process?

In this paper we address these questions by presenting and discussing two corpus-driven algorithms. The first one, called 'Siblings', is used for word type clustering within one language, preparing the grounds for the second algorithm, 'Cousins', used for bi-lingual type-to-type mapping. Both algorithms are bigram-based.

We believe that corpus-driven methods like the ones proposed here could be of good use in language preservation projects on a tight budget.

The paper is structured as follows. We first introduce the concept of word pair proximity. We then present the transfer function together with some applications and results. Finally, we discuss how these and similar methods could be adapted for use in field linguistics enabling the linguist to exploit his transcriptions without the need for semantic or grammatical annotation.

2. Reference corpora

Table 1: Reference corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Style</th>
<th>Size (tok.)</th>
<th>Utterance length</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAN</td>
<td>Labovian interv.</td>
<td>1,335,252</td>
<td>5.6 tok./utterance</td>
</tr>
<tr>
<td>SWE</td>
<td>various activities</td>
<td>1,308,098</td>
<td>6.9 tok./utterance</td>
</tr>
<tr>
<td>swe</td>
<td>8% of SWE</td>
<td>100,023</td>
<td>6.9 tok./utterance</td>
</tr>
<tr>
<td>ENG</td>
<td>Labovian interv.</td>
<td>104,617</td>
<td>6.4 tok./utterance</td>
</tr>
<tr>
<td>NEWS</td>
<td>Dan. newspaper</td>
<td>1,335,266</td>
<td>-</td>
</tr>
</tbody>
</table>

DAN is derived from the Danish spoken language corpus BySoc (The Copenhagen Study in Urban Sociolinguistics) consisting of 80 long, informal conversations ([2], [3]). SWE is derived from the Swedish corpus GSLC (Göteborg Spoken Language Corpus) containing about 350 transcribed dialogues covering a wide range of social activities ([4]). ENG is derived from the SLX corpus of Classic Sociolinguistic Interviews ([5])

We removed all non-orthographic elements in DAN, SWE and ENG in order to level out the notational discrepancies in the source corpora wrt. prosodic markup, non-verbal communication, external event, etc. Thus the corpus suite contains orthographically controlled lexical words only, segmented with one utterance per line.

NEWS is a fragment of corpus Berlingske-99 of Danish newspaper text ([3]), one sentence/line, no interpunction.

3. Word type clustering

We first introduce the concept of word pair proximity. Consider an example. In Table 2 are shown the five most frequent bigrams of type [W hun] in corpus DAN.

Table 2: Five bigrams

<table>
<thead>
<tr>
<th>Bigram</th>
<th>C</th>
<th>Freq.</th>
<th>Eng. translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>INIT hun</td>
<td>1,160</td>
<td>20.5%</td>
<td>INIT she</td>
</tr>
<tr>
<td>og hun</td>
<td>271</td>
<td>4.8%</td>
<td>and she</td>
</tr>
<tr>
<td>er hun</td>
<td>202</td>
<td>3.6%</td>
<td>is she</td>
</tr>
<tr>
<td>men hun</td>
<td>201</td>
<td>3.6%</td>
<td>but she</td>
</tr>
<tr>
<td>har hun</td>
<td>194</td>
<td>3.4%</td>
<td>has she</td>
</tr>
</tbody>
</table>

Notice in particular bigram [INIT hun]. We adopt the convention that each utterance begins with INIT (utterance-begin) and ends with FIN (end-of-utterance). The most frequent [W hun] bigram thus has ‘hun’ as utterance initial token, the second most frequent being [og hun], etc. Now compare the corresponding bigrams [W han] (he) and [W og] (and).

Table 3: Ten bigrams

<table>
<thead>
<tr>
<th>Bigram</th>
<th>C</th>
<th>Freq.</th>
<th>Bigram</th>
<th>C</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>INIT han</td>
<td>2,325</td>
<td>20.1%</td>
<td>INIT og</td>
<td>16,865</td>
<td>48.0%</td>
</tr>
<tr>
<td>og han</td>
<td>536</td>
<td>4.8%</td>
<td>og og</td>
<td>229</td>
<td>0.7%</td>
</tr>
<tr>
<td>er han</td>
<td>438</td>
<td>3.9%</td>
<td>er og</td>
<td>23</td>
<td>0.07%</td>
</tr>
<tr>
<td>men han</td>
<td>391</td>
<td>3.5%</td>
<td>men og</td>
<td>19</td>
<td>0.05%</td>
</tr>
<tr>
<td>har han</td>
<td>336</td>
<td>3.0%</td>
<td>har og</td>
<td>14</td>
<td>0.04%</td>
</tr>
</tbody>
</table>

About 100 lines of the SLX transcription covering a recurrent reading test have been omitted in ENG.

The utterance definition is not strictly constant. In DAN, an utterance is a string of tokens delimited by any of these events: turn shift, pause (silence), non-verbal communication, passage marked as unintelligible, and any external interruption. In SWE and ENG, the boundary coincides with the turn shift. This together with SWE’s deviating style may explain the differences in utterance length (cf. Table 1)
As seen, [men han] accounts for 3.5% of all bigrams [W han] closely matching the 3.6% figure of [men hun]. In contrast, [men og] covers just 0.05% of [W og] having a mere 19 occurrences (even though type 'og' outnumbers 'han' and 'hun' by far). In short, 'hun' and 'han' seem to prefer similar (left) contexts, while 'og' is completely different.

Generalizing this observation we compute the proximity of two types X and Y as

\[ \text{Prox}(X,Y) = \left( \sum_{z \in \text{Voc}} C_z \cdot \left( 1 - \frac{|L_z - L_i|}{L_z + L_i} \right) \right)^{-1} \]

where Voc is the set of all types in corpus K, L_i is the number of occurrences in K of bigram [z X], L_z is the number of occurrences in K of [z Y], and R_i = \{ [Y z] \}^3, C_z. and C_i is the number of occurrences in K of types X, z, and z', respectively.

Prox values range between 0 and 1 (valid input). Kindred words score high, while unrelated words score low.

\[ \text{Prox}(\text{hun}', \text{han}', \text{DAN}) = 0.680 \]
\[ \text{Prox}(\text{hun}', \text{og}', \text{DAN}) = 0.073 \]

To verify Prox as an indicator of grammatical kinship, we study a large number of word pairs. For each type X in DAN we let Y run over all types in DAN and compute a sorted list of (X,Y) proximity values. Below is a selection of Xs picked from various grammatical categories, each shown with its five closest related Ys (sorted after decreasing Prox value).

### Table 5: Open-class siblings

<table>
<thead>
<tr>
<th>Type X</th>
<th>sjovt</th>
<th>fnadSV</th>
<th>tfids</th>
<th>homembIC</th>
<th>rejser</th>
<th>travelbRESS</th>
<th>storebrobr</th>
<th>elder</th>
<th>brother</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closest Y</td>
<td>skaegt</td>
<td>funny</td>
<td>christe-</td>
<td>nedbIC</td>
<td>kommer</td>
<td>comerv</td>
<td>kommer</td>
<td>brobr</td>
<td>brother</td>
</tr>
<tr>
<td>2nd</td>
<td>satt</td>
<td>nice</td>
<td>startet</td>
<td>startedbIC</td>
<td>går</td>
<td>gr9PTC-walkbIC</td>
<td>lillebrobr</td>
<td>younger</td>
<td>brother</td>
</tr>
<tr>
<td>3rd</td>
<td>hyggeligt</td>
<td>cozy</td>
<td>uddannet</td>
<td>trainedbIC</td>
<td>ryger</td>
<td>r9PTC,-smokebIC</td>
<td>saster</td>
<td>sister</td>
<td></td>
</tr>
<tr>
<td>4th</td>
<td>spændende</td>
<td>exiting</td>
<td>ansat</td>
<td>employedbIC</td>
<td>tager</td>
<td>g9PTC,-takey</td>
<td>storessister</td>
<td>elder</td>
<td>sister</td>
</tr>
<tr>
<td>5th</td>
<td>fint</td>
<td>fine</td>
<td>gift</td>
<td>marriedbIC</td>
<td>korer</td>
<td>drivebPRESS</td>
<td>far</td>
<td>father</td>
<td></td>
</tr>
</tbody>
</table>

Notice the cultural finger prints: The best substitute for 'født' (born) is 'døbt' (christened), not e.g. conceived. In the same vein, the best substitute for 'sondag' (Sunday) is 'lørdag' (Saturday), 'fredag' (Friday) being no. 2, 'torsdag' (Thursday) no. 51 only! The closest match to 'penergo' (money) is 'børn' (children), followed by 'mennesker', 'ting', 'biler', 'bajere' and 'problemer' in that order (people, things, girls, cars, beers, and – problems).

Xs scoring low for all Ys are likely to be grammatical particles, types scoring high for some Ys typically belong to highly structured paradigms. The particle 'at' (infinitive marker/subordinating conjunction) thus selects 'om' (subordinating conjunction) as its closest Y, but the measured proximity is low.

\[ \text{Prox}(\text{at}, \text{om}) = 0.137 \] (tolthat, if/whether/about)

\[ \text{Prox}(\text{en}, \text{han}) = 0.680 \] (she, he)

\[ \text{Prox}(\text{en}, \text{var}) = 0.651 \] (is, was)

\[ \text{Prox}(\text{mm}, \text{ja}) = 0.684 \] (uhuh, yes)

See [6] for a detailed analysis of the DAN Siblings log (i.e. all pairs of word types in DAN sorted by Prox) showing word pair proximity to be a highly reliable kinship indicator. [7] has a discussion on how to implement Siblings-based word clustering in a computationally efficient way.

### 4. The transfer function

The second algorithm produces a word-to-word dictionary for two cognate languages (for non-cognate languages word mapping, especially of function words, hardly makes sense).

The translation process needs to be 'seeded' by a small number of known translations serving to mediate between the contexts of types in the source and target languages. So we first produce a small collection of controlled mappings of highly frequent types in DAN onto types in SWE, such as:

### Table 4: Closed-class siblings

<table>
<thead>
<tr>
<th>Type X</th>
<th>hun</th>
<th>she</th>
<th>mange</th>
<th>many</th>
<th>derude</th>
<th>out</th>
<th>there</th>
<th>nei</th>
<th>no</th>
<th>otte</th>
<th>eight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closest Y</td>
<td>han</td>
<td>noge</td>
<td>down</td>
<td>nay</td>
<td>ni</td>
<td>nine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd</td>
<td>they</td>
<td>nogen</td>
<td>down</td>
<td>nay</td>
<td>seks</td>
<td>six</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd</td>
<td>jeg</td>
<td>there</td>
<td>over</td>
<td>ah</td>
<td>syv</td>
<td>seven</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th</td>
<td>vi</td>
<td>two</td>
<td>up</td>
<td>well</td>
<td>tolv</td>
<td>twelve</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5th</td>
<td>du</td>
<td>meget</td>
<td>her</td>
<td>ja</td>
<td>fire</td>
<td>four</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Observe that kinship declines gracefully with decreasing Prox values. Type 'hun' (she) picks 'han' (he) first, agreeing with 'hun' on person, number, and case. Second pick is 'de' (they) sharing person and case, but not number, then 'jeg' sharing number and case, but not person, etc.

\footnote{For perspicuity, we ignore any illegal 0s. A stricter Prox definition would have \( z (z') \) run over types occurring in the left (right) context of X only. Note that each token in K has a left and a right context by definition since any utterance initial (final) token T participates in a bigram [INIT T] (T FIN). INIT and FIN are thus treated as types included in Voc (i.e. eligible for \( z, z' \), but not for X, Y).}
are needed to get good translations of types up to about rank 200. In this range, the closed class items prevail. The Cousins Prox formula is a straightforward generalization of the Siblings formula.

\[
\text{Prox}(A,B,\text{K1},\text{K2}) = \sum_{z \in \text{Voc}_{\text{K1}}} C_z \cdot \left(1 - \frac{L_z - L_z}{L_z + L_z} \right) \sum_{z \in \text{Voc}_{\text{K1}}} C_z' \cdot \left(1 - \frac{R_z - R_z}{R_z + R_z} \right)
\]

where \(\text{Voc}_{\text{K1}}\) is the set of all types in corpus \(\text{K1}\) and

\[
L_z = \text{occurrences in } \text{K1} \text{ of bigram } [z:A] \\
L_z = \text{occurrences in } \text{K2} \text{ of bigram } [\text{SEED}(z):B] \\
R_z = \text{occurrences in } \text{K1} \text{ of bigram } [A \ z'] \\
R_z = \text{occurrences in } \text{K2} \text{ of bigram } [B \ \text{SEED}(z')]
\]

The new Prox is a function of four arguments: two types \(A\) and \(B\), and two corpora. If \(A\) occurs in \(\text{K1}\) and \(B\) in \(\text{K2}\), Prox measures their mutual proximity. With a SEED function of ten entries, the translation capacity is found to be good up to about rank 100 with about 80% correct or almost-correct translations. But also for shorter SEED lists – even including the empty list – the translations are good enough to be useful. Shown below are the first 20 types of DAN together with their translations derived from SWE using very few seeds only (0, 2, and 4 SEED entries respectively).

**Table 6: Cousins ranked 1-20**

| Rank | Danish | Swedish | SEED
|------|--------|---------|------|
| 1    | det    | sÅt!    | det
| 2    | ja     | nO!     | ja
| 3    | og     | mEN!    | och
| 4    | jeg    | hAN!    | jag
| 5    | er     | var!    | var! är
| 6    | sÅ!    | det>oCh!| inte
| 7    | der    | nu>!    | det
| 8    | ikke   | hÄr>ocksÅ!| inte
| 9    | var    | inte>Sn> (j3)\! | var
| 10   | i      | dom>den> oM> (7)\! | om>sÅ> (7)\! | att
| 11   | har    | skul>!   | skul>! har
| 12   | at     | (18)\! | (18)\! | att
| 13   | mm     | !       | m
| 14   | ik\!   | !       | va
| 15   | men    | !       | men
| 16   | jo     | eller>! | nu
| 17   | du     | dom>den> | !
| 18   | en     | (12)>! | (13)>! | en
| 19   | pÅ    | du>den> mEd> (5)>! | med>du> till>! | pÅ
| 20   | vi     | för>et>et>! | !

Example: The cell containing "hÄr>ocksÅ!" (line #8) reads: first bid is 'här', 2nd is 'också', 3rd (and correct) is 'inte'. Italic font is used for \(B_s\) in the same POS as the correct translation ('också, also, is a sentential adverb on a par with 'inte, not). 'it' means correct form. Dictionary approved translations are shown in column B\(\text{Dict}\). We used [8] as reference dictionary while translations of idiomatic speech types not in the dictionary, such as 'må' and 'va', were confirmed by four Scandinavian linguists, two Danish and two Swedish.

Let us pick out some details for closer study. Notice the \(B\text{SEED}=0\) session (3rd column). Even with access to the raw utterance segmentation only, judgments are found to be quite good. About half of the top-20 types are translated correctly or almost correctly (e.g. 'var' (was) for 'är' (is) in line #5).

With a SEED list of just 4 controlled translations ('det', 'ja', 'og', 'jeg'), 17 out of 20 types get a correct or almost correct translation.

Certain categories are harder to translate than others, notably grammatical particles (#12), prepositions (#10, #19), and determiners (#18); but notice that in many cases fair substitutes are offered (italicized in Table 6), e.g. 'om' for 'at' (both subordinating conj.) and 'med' for 'pa' (both prep.). As was the case with Siblings clustering, Cousins translations are often sensitive even when not strictly correct.

Observe that correct translations are not necessarily preserved when adding more items to the SEED list (cf. #5, #11) – even if the overall correctness figure is of course improving with increasing \(\text{SEED}\).

As seen in Table 6, most highly frequent Danish types have etymological equivalents in Swedish (a notable exception being #7 'der'). This is however not always the case. Table 7 shows an assorted collection of translations from a \(\text{SEED}\) = 20 session using \(K_1=\text{DAN}\) and \(K_2=\text{SWE}\). None of these translations were present in the SEED list, i.e. they are genuine products of the translation session.

**Table 7: Assorted cousins**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Danish</th>
<th>Swedish</th>
<th>Eng. translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>#32</td>
<td>nÅ</td>
<td>javisst</td>
<td>yeah, sure, uhuh</td>
</tr>
<tr>
<td>#45</td>
<td>skål</td>
<td>fär</td>
<td>must</td>
</tr>
<tr>
<td>#65</td>
<td>nok</td>
<td>faktiskt</td>
<td>most likely, sort-of</td>
</tr>
<tr>
<td>#71</td>
<td>I</td>
<td>ni</td>
<td>youDOS, DUSC</td>
</tr>
<tr>
<td>#72</td>
<td>bare</td>
<td>liksom</td>
<td>just, kind-of, like</td>
</tr>
<tr>
<td>#77</td>
<td>synes</td>
<td>tycker</td>
<td>recogres, thinkres</td>
</tr>
<tr>
<td>#96</td>
<td>helt</td>
<td>alldeles</td>
<td>completely, really</td>
</tr>
<tr>
<td>#103</td>
<td>huske</td>
<td>ihåg</td>
<td>remember, mind</td>
</tr>
<tr>
<td>#110</td>
<td>vel</td>
<td>faktiskt</td>
<td>sort-of, yknow</td>
</tr>
<tr>
<td>#126</td>
<td>kun</td>
<td>bara</td>
<td>only, just</td>
</tr>
<tr>
<td>#146</td>
<td>hvoridan</td>
<td>hur</td>
<td>how</td>
</tr>
<tr>
<td>#147</td>
<td>ellers</td>
<td>däremot</td>
<td>else, though</td>
</tr>
</tbody>
</table>

These translations are all correct (approved using the same criteria as before). They are furthermore interesting by being etymologically unrelated. In many cases, the etymologically corresponding type, if any, is a so-called false friend being only superficially similar. The ability of the translation engine to see through false friendships may be helpful to the field linguist aligning the vocabularies of two cognate languages.

4.1. Extending the corpus suite

We have performed a series of translation sessions with various corpora in place of \(\text{SWE}\) – such as swe, ENG and NEWS (see Table 1). In this section we present some of our findings in a condensed form. The ENG and swe data are hitherto unpublished; the SWE/NEWS data also appear in [6].

\(4\) At present \(\text{SEED}\) entries are always picked from the top of the frequency list: 1st entry is 'det'-'det', 2nd (if any) is 'ja-'ja', 3rd 'og-''och', 4th 'jeg-''jeg' etc. Entries could also be selected individually; an interesting question, then, is how the distribution of \(\text{SEED}\) items over parts-of-speech effects the translations produced. What kind of seed is most fertile?
As mentioned before, Siblings and Cousins sessions based on DAN and SWE often arrive at results that are sensible even when not strictly correct. Therefore we also compare the production of almost correct translations, by which we mean translations that meet one of these requirements:

- 1st bid is correct (approved using [8] and [9])
- 2nd bid is correct (same criteria)
- 1st bid deviates minimally⁵ from the correct form

Several interesting conclusions can be read off these graphs.

Supplying a small number of SEED entries has a very significant impact on the translation quality. Beyond 10 the improvement is much slower (see however note 4).

For cognate languages like Danish and Swedish, the translation method is seen to down-scale nicely from corpora of 1M to about 100k (compare SWE and swe). Below 100k, results deteriorate fairly quickly (cf. [6]).

Observe what happens when we replace SWE of written Swedish by NEWS of written Danish (newspaper articles). One might think that the Danish-Danish translation task is easier than the Danish-Swedish task since in this case SEED is simply the identity function. All seeded translations are then perfect by definition. Nevertheless the generated translations turn out to be extremely poor – far worse than the Danish-English ones (compare NEWS and ENG), even though English is only distantly related to Danish.

We conclude that, concerning context selection, there is a profound difference between spoken and written style.

5. Discussion

For the traditional field linguist, the easiest categories to establish are the concrete nouns, closely followed by the content verbs and adjectives – in short: the open classes – as these can be determined to a large extent by deixis (“What is the name of the thing I am holding?”, “What am I doing now?”, “What do these two objects have in common?”). Much more recalcitrant are the function words, since most linguistically naive speakers have difficulties explaining their meaning and use – especially to a foreigner. See [10], [6] for discussions of the challenges and hardships in the field.

Prox based methods, on the other hand, produce a dictionary which is often unreliable for content words (and low-frequency function words as well). In our DAN-SWE experiments, the translations of low frequency nouns and verbs are in general not much better than chance. However, our method shows good performance in translating highly frequent function words: personal pronouns, connectives, discourse tags, feedback particles, auxiliary verbs, etc.

Thus the weaknesses of the two methods are to some extent complementary, and perhaps they could be made to cancel out each other. Combining traditional field methods with easy-to-handle techniques based on spoken language elicitation, low-quality transcription⁶, and simple statistics may therefore comprise a workable strategy for low budget language description.⁷

We have demonstrated how simple statistical tools can be used for exploiting raw transcriptions unaccompanied by semantic or grammatical knowledge – turning the only ubiquitous data source into valuable linguistic information.

6. References


⁵”Minimal deviation” implies inclusion in the same POS. Furthermore: For pronouns, all morphological feature values shared but one. For verbs, same stem but wrong tense, or vice versa. For conj., interjections, prep., adv., feedback particles, numerals: same POS. For nouns and adjectives (very few in #1-100): GND, NUM, DEF shared.

⁶Prox based methods are highly error-resistant (see [6])

⁷Siblings based methods are currently being tested on transcription corpora for South African languages ([11])