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 literation 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 our reference corpora of spoken Danish, Swedish, and English. Then we 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*

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

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])[1].

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.[2]

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*

Bigram	C	Freq.	Eng. translation
INIT hun	1,160	**20.5%**	INIT *she*
og hun	271	**4.8%**	*and she*
er hun	202	**3.6%**	*is she*
men hun	201	**3.6%**	*but she*
har hun	194	**3.4%**	*has she*

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*

Bigram	C	Freq.	Bigram	C	Freq.
INIT han	2,325	**20.1%**	INIT og	16,865	**48.0%**
og han	536	**4.8%**	og og	229	**0.7%**
er han	438	**3.9%**	er og	23	**0.07%**
men han	391	**3.5%**	men og	19	**0.05%**
har han	336	**3.0%**	har og	14	**0.04%**

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

[2]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

$$Prox(X,Y,K) = \frac{\sum_{z \in Voc} C_z \cdot (1 - \frac{|L_1 - L_2|}{L_1 + L_2})}{C_x} \cdot \frac{\sum_{z' \in Voc} C_{z'} \cdot (1 - \frac{|R_1 - R_2|}{R_1 + R_2})}{C_x} \quad (1)$$

where *Voc* is the set of all types in corpus *K*, L_1 is the number of occurrences in *K* of bigram [*z X*], L_2 of [*z Y*], R_1 of [*X z'*], and R_2 of [*Y z'*].[3] C_x, C_z and $C_{z'}$ is the number of occurrences in *K* of types *X*, *z,* and *z'*, respectively.

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

$$\begin{aligned} Prox(\text{'hun', 'han', DAN}) &= 0.680 \\ Prox(\text{'hun', 'og', DAN}) &= 0.073 \end{aligned} \quad (2)$$

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 *X*s picked from various grammatical categories, each shown with its five closest related *Y*s (sorted after decreasing *Prox* value).

Table 4: *Closed-class siblings*

Type *X*	hun *she*	mange *many*	derude *out there*	nej *no*	otte *eight*
Closest *Y*	han *he*	nogle *some*	dernede *down there*	næ *nay*	ni *nine*
2nd	de *they*	nogen *any/ some*	derinde *in there*	næh *nay*	seks *six*
3rd	jeg *I*	flere *more*$_{\text{COUNT}}$	derovre *over there*	ah *oh/aha*	syv *seven*
4th	vi *we*	to *two*	deroppe *up there*	nå *well*	tolv *twelwe*
5th	du *you*$_{\text{NOM,SG}}$	meget *much*	her *here*	ja *yes*	fire *four*

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.

[3]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*).

Table 5: *Open-class siblings*

Type *X*	sjovt *fun*$_{\text{ADV}}$	født *born*$_{\text{PTC}}$	rejser *travel*$_{\text{PRES}}$	storebror *elder brother*
Closest *Y*	skægt *funny*	døbt *christe-ned*$_{\text{PTC}}$	kommer *come*$_{\text{PRES}}$	bror *brother*
2nd	rart *nice*	startet *started*$_{\text{PTC}}$	går *go*$_{\text{PRES}}$, *walk*$_{\text{PRES}}$	lillebror *younger brother*
3rd	hyggeligt *cozy*	uddannet *trained*$_{\text{PTC}}$	ryger *rush*$_{\text{PRES}}$, *smoke*$_{\text{PRES}}$	søster *sister*
4th	spændende *exiting*	ansat *employed*$_{\text{PTC}}$	tager *go*$_{\text{PRES}}$, *take*$_{\text{PRES}}$	storesøster *elder sister*
5th	fint *fine*	gift *married*$_{\text{PTC}}$	kører *drive*$_{\text{PRES}}$	far *father*

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 'søndag' (*Sunday*) is 'lørdag' (*Saturday),* 'fredag' (*Friday*) being no. 2, 'torsdag' (*Thursday*) no. 4, and 'mandag' (*Monday*) no. 51 only! The closest match to 'penge' (*money*) is 'børn' (*children*), followed by 'mennesker', 'ting', 'piger', 'biler', 'bajere' og 'problemer' in that order (*people*, *things*, *girls*, *cars*, *beers*, and – *problems*).

*X*s scoring low for all *Y*s are likely to be grammatical particles, types scoring high for some *Y*s 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.

Closest *Y*s (3)

Prox(at,om)	=	0.137	(*to*/*that*, *if*/*whether*/*about*)
Prox(hun,han)	=	0.680	(*she*, *he*)
Prox(er,var)	=	0.651	(*is*, *was*)
Prox(mm,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 language (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:

Danish		**Swedish**	(4)
det	→	det	*it*/*that*/*this*/*the*$_{\text{UTT,SG}}$
ja	→	ja	*yes*
og	→	och	*and*

Only a handful, 5-20 say, of such translations (called *SEED*)

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.

$$Prox(A,B,K1,K2) = \frac{\sum_{z \in Voc_{K1}} C_z \cdot (1 - \frac{|L_1 - L_2|}{L_1 + L_2})}{C_x} \cdot \frac{\sum_{z' \in Voc_{K1}} C_{z'} \cdot (1 - \frac{|R_1 - R_2|}{R_1 + R_2})}{C_x} \quad (5)$$

where Voc_{K1} is the set of all types in corpus *K1* and

L_1 = occurrences in *K1* of bigram [*z A*]
L_2 = occurrences in *K2* of bigram [*SEED(z) B*]
R_1 = occurrences in *K1* of bigram [*A z'*]
R_2 = occurrences in *K2* of bigram [*B SEED(z')*]

The new *Prox* is a function of four arguments: two types *A* and *B*, and *two* corpora. If *A* occurs in *K1* and *B* in *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	***A***	$\boldsymbol{B}_{\|SEED\|=0}$	$\boldsymbol{B}_{\|SEED\|=2}$	$\boldsymbol{B}_{\|SEED\|=4}$	$\boldsymbol{B}_{\mathbf{Dict}}$
#1	det	*så*>**!**	**!**	**!**	det
#2	ja	*m*>**!**	**!**	**!**	ja
#3	og	*men*>**!**	*men*>**!**	**!**	och
#4	jeg	*han*>**!**	*han*>**!**	**!**	jag
#5	er	**!**	*var*>**!**	*var*>**!**	är
#6	så	*det*>*och*>**!**	**!**	**!**	så
#7	der	nu>**!**	**!**	**!**	det
#8	ikke	här>*också*>**!**	nu>*också*>**!**	**!**	inte
#9	var	inte>*är*> *(3)*>**!**	**!**	**!**	var
#10	i	dom>den> *om>(7)*>**!**	du>dom> som>*till*>**!**	du>*till*> dom> som>**!**	i
#11	har	**!**	*skulle*>**!**	*skulle*>**!**	har
#12	at	*(18)*>**!**	*(18)*>**!**	*om*>*som*> *så*>*(7)*>**!**	att
#13	mm	**!**	**!**	**!**	m
#14	ik'	**!**	**!**	**!**	va
#15	men	**!**	**!**	**!**	men
#16	jo	eller>**!**	**!**	**!**	nu
#17	du	*dom*>*den*>**!**	**!**	**!**	du
#18	en	*(12)*>**!**	*(13)*>**!**	*(13)*>**!**	en
#19	på	du>den> *med>(5)*>**!**	du>*med*> *till*>**!**	*med*>du> **!**	på
#20	vi	för>*en*>*ett*>**!**	**!**	**!**	vi

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*). '**!**' means correct form. Dictionary approved translations are shown in column $\boldsymbol{B}_{\mathbf{Dict.}}$ We used [8] as reference dictionary while translations of idiomatic speech types not in the dictionary, such as 'mm' and 'va', were confirmed by four Scandinavian linguists, two Danish and two Swedish.

Let us pick out some details for closer study. Notice the |*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 'på' (both prep.). As was the case with Siblings clustering, Cousins translations are often sensible 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 |*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 |*SEED*| = 20 session[4] using *K1*=DAN and *K2*=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*

Rank	**Danish**	**Swedish**	**≈ Eng. translation**
#32	nå	javisst	*yeah, sure, uhuh*
#45	skal	får	*must*
#65	nok	faktiskt	*most likely, sort-of*
#71	I	ni	*you*$_{\text{NOM,PLUR}}$
#72	bare	liksom	*just, kind-of, like*
#77	synes	tycker	*recon*$_{\text{PRES}}$, *think*$_{\text{PRES}}$
#96	helt	alldeles	*completely, really*
#103	huske	ihåg	*remember, mind*
#110	vel	faktiskt	*just, sort-of, y'know*
#126	kun	bara	*only, just*
#146	hvordan	hur	*how*
#147	ellers	däremot	*else, though*

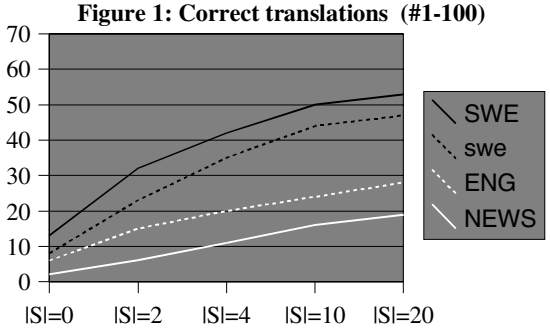
These translations are all correct (approved using the same criteria as before). They are furthermore interesting by being etymologically unrelated. In many cases, the etymologically corresponding type, if any, is a so called *false friend* being only superficially similar. The ability of the translation 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 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 *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'→'jag' etc. Entries could also be selected individually; an interesting question, then, is how the distribution of *SEED* items over parts-of-speech effects the translations produced. What kind of seed is most fertile?

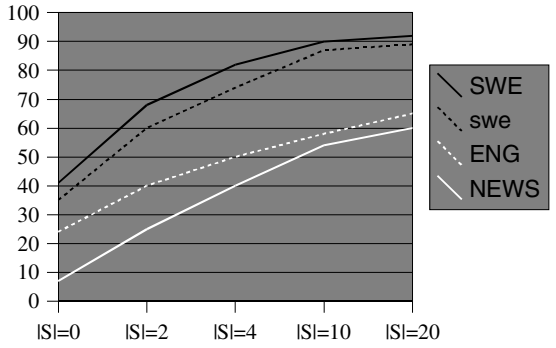
Figure 1: Correct translations (#1-100)



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 minimally[5] from the correct form

Figure 2: Almost correct translations (#1-100)



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 *spoken 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] "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.

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 transcription[6], 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

[1] Chen, B.-X. & Du, L.-M. (2002) *Automatic Construction of English-Chinese Translation Lexicon from Parallel Spoken Language Corpora*; ISCSLP 2002

[2] Gregersen, F. et al (1991) *The Copenhagen Study in Urban Sociolinguistics*; vol 1+2, Copenhagen: Reitzel

[3] Henrichsen, P.J. (2002) *Some Frequency based Differences between Spoken and Written Danish*; Gothenburg Pap. in Theoretical Linguistics 88

[4] Allwood, J. et al (2001) *Annotations and Tools for an Activity Based Spoken Language Corpus*; SIGdial-2001

[5] Labov, W. (1984) *Field Methods of the Project on Linguistic Change and Variation*; Prentice Hall

[6] Henrichsen, P.J. (2004) *Siblings and Cousins – Statistical Methods for Spoken Language Analysis*; Acta Linguistica Hafniensia 36

[7] Grönquist, L. et al (2003) *A Method for Finding Word Clusters in Spoken Language*; CL2003 (Lancaster)

[8] Molde, B. (1980, 2000) *Dansk-Svenska Ordbog*; Norstedts Förlag; 726pp

[9] Vinterberg, H. & Bodelsen, C.A. (1998) *Dansk-Engelsk Ordbog*, Gyldendal; 2610pp

[10] Newman, P.; M. Ratliff (eds) (2001) *Linguistic Fieldwork*; Cambridge Univ.

[11] Allwood J. et al (2003) Developing a tag set and tagger for the African languages of South Africa, J. of Southern African Linguistics and Applied Language Studies 21(4)

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

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