

Working in Darkness

corpus based tools for low-budget field linguistics

Peter Juel Henriksen
CMOL¹
pjuel@id.cbs.dk

Abstract

We suggest a method using *speech corpus elicitation* for language preservation and literation projects. We present a simple algorithmic tool called Siblings-and-Cousins for deriving lexical information from a corpus of (possibly error-prone) transcriptions of informal conversations. The formal properties of the algorithm are discussed and exemplified in a 'staged experiment' using two well-described languages (Danish and Swedish) as stand-ins for two endangered languages. Finally, we reflect on the use of spoken language corpora as an abundant information source in preservation and literation projects on a tight budget.

1. Introduction

Imagine yourself in this situation: You are a field linguist about to make contact with a target group of language users – a village called Alpha. You don't know the first thing about their language, Alphish. You hardly share any language with any local. Alphish has no written code, most or all speakers are illiterate, and the language is undescribed in the literature. Your time and budget are limited, and so actually *learning* the language is not an option. Your technical equipment is restricted to a portable computer and a sound recorder. Also, time is running out: A written code for Alphish must be established within a short time and the Alphas literated for their language to survive. What can you do?

No matter whether your task is to found an official orthography, establish a language technology, or merely prepare a linguistic description, your first step is to determine the range of available information sources. Traditional field work relies on sources such as:²

- Previously published descriptions
- Knowledge of cognate languages and dialects
- Extensive interviews of a single (or a few) informants who are carefully selected among the most 'linguistically conscious' subjects and who are often given a

¹ Center for Computational Modelling of Language, Copenhagen Business School, Denmark

² cf. Newman et al (2001)

- rudimentary education
- Introspection
- Articulatory experiments (using e.g. palatograph and laryngoscope)
- Tape/video recordings of gatherings (conversation, feast, ceremony...)

However, in the working situation outlined above – let us call it *darkness* – most traditional methods are not applicable. There is no literature, no time for training and interrogation of informants, no money for advanced equipment, etc.

As we will argue, under these conditions a good starting point would be the recording and transcription of a small number of informal conversations – the only ubiquitous language activity type. The recording sessions can be initiated immediately since they don't depend on a comprehension of Alphish.

We demonstrate a computational method for deriving a fairly accurate preliminary dictionary covering the most frequent word forms, from a transcription corpus that may be rife with flaws and inconsistencies (unavoidable when working in darkness). The only precondition is a rudimentary knowledge of a cognate language, call it Betish, possibly un-written as well. This knowledge could well be the result of a similar literacy project in a nearby community Beta.

Finally, we discuss the pros, cons, and prospects of the corpus based approach.

2. The recording session

The preferred activity type is *informal conversation* with two or three participants. In order to secure a good recording quality, the recording sessions should take place indoor with everybody sitting down, not manipulating objects in a noisy way. The linguist's attitude should be "a fly on the wall". Microphones with good signal/noise ratio (fairly inexpensive) and digital recording equipment are recommended (the cheapest options are cd or mini disk recorders, or recording to a computer directly).

3. Transcription

How can you transcribe a conversation in a language unknown to you? The short answer is: you can't, since you cannot segment the sound stream into words, much less into morphemes.

A better answer goes as follows. For the present purposes, you do not need anything near a perfect transcription (which is next to impossible in any case for an un-written language). All you need is a moderate degree of consistency. Some factors come in handy:

1. Informal conversation is often rich in single word utterances, allowing a 'kick start' of lexical registrations
2. Very basic linguist-informant interviews concerning word and utterance boundaries can be possible, even when traditional interviews about word semantics are not (due to the darkness)
3. The formal method to be introduced is very robust towards transcription errors (such as false word boundaries and collapsed phonetic distinctions) – as long as they just provide a general noise background, i.e. are distributed fairly evenly over the corpus and over the word types.
4. In general, phonetic recognition is easier than determination of meaning, since

homonymy and polysemy are not a problem³

4. A formal method for dictionary elicitation

In this section we present a formal method for deriving an Alphish-Betish dictionary using only

1. a sketchy transcription of Alphish informal conversations
2. a sketchy transcription of Betish informal conversations
3. a rudimentary understanding of Betish (just a few words)

In order to verify our findings, we have two actually well-described languages play the role of Alphish (the target language) and Betish (the reference language). First we proceed as if nothing was known about Alphish and only rudimentary lexical knowledge existed of the other, and then we evaluate the results using a standard dictionary. As stand-ins for Alphish and Betish we pick Danish and Swedish, respectively. These two languages meet the requirements:

- distinct languages (politically, historically, and in the 'public understanding')
- cognate languages (linguistically)
- adequate transcription corpora available (including informal speech)⁴
- bilingual dictionary available (for verification)

Danish and Swedish are, lexically and grammatically, quite similar (Allwood 1997, Henrichsen 2002, Henrichsen and Allwood *forthcom.*). Yet Danes and Swedes often have difficulties communicating orally, and they typically switch to English when conversing. So the languages are cognate, but this fact is to some extent hidden to the language users due to the pronounced prosodic and phonological differences. We speculate that this situation could be typical in large parts of the third world too.

In the following sections we introduce the formal components, viz. the algorithm Siblings for type clustering, and the algorithm Cousins for lexical mapping.

The Alphish and Betish (alias Danish and Swedish) corpora are referred to as α and β , respectively.

4.1 Algorithm 1: Siblings

We first introduce the concept of word pair proximity.

Consider the four most frequent bigrams of type " $_ he$ " in a typical English text⁵, aligned with the corresponding bigrams " $_ she$ " and " $_ on$ ".

³For the sake of experiment, the author has transcribed a few minutes of vivid conversation in Turkish and Finnish – two typologically distinct, non-Indoeuropean languages completely unknown to me. Using a non-linguist for contact person (asking only simple yes-no questions concerning segmentation), I succeeded in producing transcriptions adequate for the present purposes.

⁴*Danish reference corpus*: Corpus BySoc, <http://www.id.cbs.dk/~pjuel/BySoc>, 1.3 mio wds, informal dialogues, cf. Henrichsen 1998. *Swedish reference corpus*: GSLC, Gothenburg Spoken Language Corpus (1.3 mio running wds, 20 different activity types including informal dialogues, Allwood 2001.

⁵ Agatha Christie: "The Mysterious Affair at Styles"

8.9%	"that <i>he</i> "	8.4%	"that <i>she</i> "	2.0%	"that <i>on</i> "
6.0%	"as <i>he</i> "	7.6%	"as <i>she</i> "	0.1%	"as <i>on</i> "
2.9%	"and <i>he</i> "	3.6%	"and <i>she</i> "	0.6%	"and <i>on</i> "
2.4%	"when <i>he</i> "	3.2%	"when <i>she</i> "	0.0%	"when <i>on</i> "

As seen, the bigram "that *he*" accounts for 8.9% of all bigrams of the form $[X \textit{he}]$, and so forth. Observe that *he* and *she* prefer rather similar contexts, while *on* is totally different. So, a very simple way of clustering kindred word types would be to compare their selection of immediate contexts. Generalizing this observation, we compute the proximity of two types X and Y as

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

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_2 of $[z Y]$, R_1 of $[X z]$, and R_2 of $[Y z]$.⁶ $Prox$ values range between 0 and 1 (for valid input). Kindred words – such as Alphish 'hun' (*she*) and Alphish 'han' (*he*) – score high, as opposed to unrelated words such as 'hun' and 'og' (*she, and*).

$$\begin{aligned} Prox(\text{'hun'}, \text{'han'}, \alpha) &= 0.6893 \\ Prox(\text{'hun'}, \text{'og'}, \alpha) &= 0.0745 \end{aligned}$$

In order to verify $Prox$ as an indicator of grammatical kinship, we study a large number of word pairs. For each type X in α we let Y run over all types in α and compute a sorted list of (X, Y) proximity values. Shown in the table are a selection of X s picked from various grammatical categories; each X is shown with its five closest related Y s (sorted after decreasing $Prox$ value).

Table 1. Closed class "siblings"

Type X	'hun' (she)	'mange' (many)	'derude' (out there)	'ja' (yes)	'otte' (eight)
Closest Y	'han' (he)	'nogle' (some)	'dernede' (down there)	'jaja' (yes yes)	'ni' (nine)
2nd	'de' (they)	'nogen' (some/any)	'derovre' (over there)	'nå' (well)	'syv' (seven)
3rd	'jeg' (I)	'flere' (several)	'derinde' (in there)	'nej' (no)	'seks' (six)
4th	'vi' (we)	'to' (two)	'deroppe' (up there)	'jojo' (oh yes)	'fire' (four)
5th	'du' (you _{NOM,SG})	'tre' (three)	'herude' (out here)	'næ' (not really)	'elleve' (eleven)

⁶For simplicity, 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). Furthermore, we assume that each token in K has a left (right) context; an initial (final) token participates trivially in a bigram $[\text{INIT } X]$ ($[X \text{ FIN}]$), where INIT and FIN are treated as types included in Voc (i.e. eligible for z, z' , but not for X, Y).

Table 2. Open class "siblings"

Type X	'sjovt' (fun _{ADV})	'født' (born _{PTC})	'synes' (think _{PRES} , recon _{PRES})	'rejser' (travel _{PRES})	'storebror' (elder- brother)
Closest Y	'skægt' (funny)	'døbt' (christened _{PTC})	'tror' (believe _{PRES} , think _{PRES})	'kommer' (come _{PRES})	'bror' (brother)
2nd	'rart' (nice)	'opvokset' (grown-up _{PTC})	'mener' (think _{PRES} , mean _{PRES})	'går' (go _{PRES} , walk _{PRES})	'lillebror' (younger- brother)
3rd	'spændende' (exiting)	'uddannet' (trained _{PTC})	'forstår' (understand _{PRES})	'ryger' (rush _{PRES} , smoke _{PRES})	'søster' (sister)
4th	'hyggeligt' (cozy)	'gift' (married _{PTC})	'ved' (know _{PRES})	'tager' (go _{PRES} , take _{PRES})	'far' (father)
5th	'morsomt' (amusing)	'ansat' (employed _{PTC})	'syntes' (thought _{PAST})	'flytter' (move _{PRES})	'mor' (mother)

Observe that kinship declines gracefully with decreasing *Prox* values. For instance, type 'hun' (*she*) selects 'han' (*he*) as its preferred substitute. 'Han' is morphologically closely related, agreeing with 'hun' on person, number, and case. Second pick is 'de' (*they*), sharing person and case, but not number. Third is 'jeg', sharing number and case, but not person, etc.

Notice in passing the 'cultural finger prints': The best substitute for 'born' is 'christened', not e.g. 'conceived'. The best substitute for 'Sunday' is 'Saturday', while 'Friday' is number 2, 'Thursday' number 5 only; the best substitute for 'money' is 'children', number two being 'things', number three 'people'. The best substitute for 'sentence' is 'book'.

Prox is found to be a very reliable kinship indicator (when used on highly frequent types). Not only are the clusterings in almost all cases in concord with intuition, the specific *Prox* values also indicate reliably the strength of the relation. Xs scoring low for all Ys are likely to be grammatical particles, types scoring high 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 Ys

Prox('at','om') = 0.159 (*to/that, if/whether/about*)

Prox('hun','han') = 0.689 (*she, he*)

Prox('er','var') = 0.660 (*is, was*)

Prox('mm','ja') = 0.722 (*uhuh, yes*)

In order to approach the darkness conditions, we then redid the calculations using a tiny subcorpus of α (corresponding to 5 hours of speech), in which 33% of all tokens were corrupted (introducing wrong ligatures, collapsed distinctions, etc. (details are in Henrichsen 2004). For the most frequent 100 words, the quality of the clustering was found to decrease only slightly. This suggests that the Siblings algorithm or similar methods could be applicable in darkness, providing the field linguist with an early, rough sketch of the grammatical and morphological groupings of the target language.

Notice however, that the method is suitable for frequent types only. Types below rank 200 or so (content words and rarer function words) are not judged reliably by *Prox*. We return to this issue in the final section.

4. 2 Algorithm 2: Cousins

The second algorithm produces a word-to-word dictionary translating the most frequent types of Alphish to types in Betish, of which we already possess a rudimentary dictionary. First, we produce (as best we can) a list containing a small number of word type pairs (*A,B*) such that type *A* is the Alphish translation of Betish type *B*.

<i>Alphish (Danish)</i>		<i>Betish (Swedish)</i>
ja	↔	ja
og	↔	och
jeg	↔	jag
at	↔	att
ham	↔	honom
meget	↔	mycket

Only a handful (5-20, say) of such translations of highly frequent types are needed; but a basic knowledge of Betish (required) together with the α clustering information (produced using the Siblings algorithm) should be of help. The lexical fragment is used in the translation function *SEED* mapping Alphish types onto Betish types (e.g. *SEED*('og') = 'och').

The Cousins proximity formula is a straightforward generalization of the Siblings formula.

$$Prox(X, Y, K1, K2) = \frac{\sum_{z \in Voc_{K1}} C_z \cdot \left(1 - \frac{|L_1 - L_2|}{L_1 + L_2}\right)}{C_x} \cdot \frac{\sum_{z' \in Voc_{K2}} C_{z'} \cdot \left(1 - \frac{|R_1 - R_2|}{R_1 + R_2}\right)}{C_x}$$

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

- L_1 = occurrences in *K1* of bigram [*z X*]
- L_2 = occurrences in *K2* of bigram [*SEED(z) Y*]
- R_1 = occurrences in *K1* of bigram [*X z'*]
- R_2 = occurrences in *K2* of bigram [*Y SEED(z')*]

The new *Prox* is a function of four arguments: two types *X* and *Y*, and *two* corpora. If *X* occurs in *K1* and *Y* in *K2*, *Prox* measures their mutual proximity, the *SEED* list mediating between the *X* contexts and *Y* contexts. With a *SEED* list of twenty entries, the translation capacity is found to be very 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. Similarly to the Siblings clusterings, Cousins translations are usually sensible even when they are not

strictly correct.

Table 3. Swe-Danish translations (closest As, measured as $Prox(B,A,\beta,\alpha)$)

Rank(A)	A	$B_{ SEED =0}$	$B_{ SEED =2}$	$B_{ SEED =4}$	Dictionary
#1	det	så> det	det	det	det
#2	ja	m> ja	ja	ja	ja
#3	og	men> och	men> och	och	och
#4	jeg	han> jag	han> jag	jag	jag
#5	er	är	var> är	var> är	är
#6	så	det>och> så	så	så	så
#7	der	nu> det	det	det	det
#8	ikke	här>också> inte	nu>också> inte	inte	inte
#9	var	inte>är>(3)> v ar	var	var	var
#10	i	dom>den> om>(7)> i	du>dom> som>till> i	du>till>dom> som> i	i
#11	har	har	skulle> har	skulle> har	har
#12	at	(18)> att	(18)> att	om>som>så> (7)> att	att
#13	mm	m	m	m	m
#14	ik'	va	va	va	va
#15	men	men	men	men	men
#16	jo	eller> nu	nu	nu	ju, nu
#17	du	dom>den> du	du	du	du
#18	en	(12)> en	(13)> en	(13)> en	en
#19	på	du>den> med>(5)> på	du>med> till> på	med>du> på	på
#20	vi	för>en>ett> vi	vi	vi	vi

Types in **bold** are considered as correct translations. Example: The cell containing "det>och>så" reads: "first bid is 'det', 2nd is 'och', 3rd (and identical to the correct translation) is 'så'."

As reference dictionary we used Norstedts Dansk-Svenska Ordbok, the largest Danish-Swedish contemporary dictionary published in Scandinavia. For As and Bs not present in the dictionary (idiomatic speech types like 'mm' and 'va'), translations are conferred with four Scandinavian linguists (two native speakers of Danish, and two of Swedish).

Notice in particular the $SEED=0$ column. Even with access to the raw utterance segmentation only, the Cousins algorithm's judgments are quite impressive – about half of the top-20 types are translated correctly or almost correctly.

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 translate harder than others, notably grammatical particles (#12), prepositions (#10, #19), and determiners (#18); but notice that in many cases fair substitutes are offered (e.g. 'om' for 'at', 'med' for 'på').

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 $|SEED|$.

Table 4 below shows an assorted collection of Cousins translations from a $|SEED| = 20$ session (all translations shown are absent from the SEED list, i.e. are genuine products of the translation session).

Table 4. Assorted 'Cousins'

Rank	Danish	Swedish
#32	nå	javisst
#45	skal	får
#65	nok	faktiskt
#71	I	ni
#72	bare	liksom
#77	synes	tycker
#96	helt	alldeles
#103	huske	ihåg
#110	vel	faktiskt
#126	kun	bara
#146	hvordan	hur
#147	ellers	däremot

Readers with a knowledge of Scandinavian languages will acknowledge that these translations are all satisfactory (most of them are typical for spoken language). They are furthermore interesting by being etymologically unrelated, that is, the word pairs in the table are cognate only in the dimensions of distribution and semantics. In many cases, the etymologically corresponding type (if any) is a so called 'false friend' being only superficially similar. The ability of the Cousins algorithm to see through false friendships could be of importance saving some frustrations for the field linguist.

Of course, the Cousins algorithm has a drawback in its need of a *SEED* list of known translations. However, nothing in principle prevents the formulation of an algorithm that does not need a reference dictionary (as documented in the $|SEED|=0$ session presented above). We are currently working on such algorithms.

4.3 Reflections on robustness

We have redone all calculations using a number of alternative proximity formulae (F1 is the formula used in 4.1).

$$\begin{array}{ll}
 \text{F1. } Prox = \frac{\sum A}{C} \cdot \frac{\sum B}{C} & \text{F2. } Prox = \frac{\sum A + \sum B}{2C} \\
 \text{F3. } Prox = \frac{(\sum A \sum B)^2}{C^4} & \text{F4. } Prox = \frac{\sqrt{\sum A \sum B}}{C}
 \end{array}$$

All results reported in this paper are found to be theory robust, in the sense that even a radical shift in the definition of proximity (e.g. replacing a multiplication by an addition, compare F1 and F2) leaves the overall results virtually unaffected (Henrichsen 2004).

Grönqvist (2003) has a discussion on the properties of F1 versus F2.

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.

The Siblings-Cousins method, on the other hand, produces a dictionary which is unreliable for content words (and low-frequency function words as well). In the Swedish-Danish experiments, the translations of low frequency nouns and verbs are in general not much better than chance⁷. However, assuming that the reference language and the target language are indeed cognate, the algorithm shows very good performance in translating highly frequent function words – such as personal pronouns, connectives, discourse tags, feed back particles, and auxiliary verbs.

As seen, the weaknesses of the two methods are more or less 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, and simple statistics may therefore comprise a workable strategy for low budget literacy. The formal methods presented here are just the first results of a possible future research programme. We have shown that simple statistical tools can be used for exploiting raw transcriptions unaccompanied by semantic knowledge – turning the only ubiquitous data source into valuable linguistic information.

The main point we want to make here, then, is a suggestion. If the field linguist must work in darkness, a reasonable first step is to establish a transcription corpus of informal conversations.

⁷ Certain content words are translated fairly well, especially highly frequent adjectives and adverbs, and types belonging to structured semantic paradigms, such as names of weekdays, family relations, and so forth (cf. table 2).

References

- Allwood, J. (1997) *Some Frequency based Differences between Spoken and Written Swedish*; XVIth Scandinavian Conference of Linguistics, Dept. of Linguistics, Univ. of Turku
- Allwood, J.; L. Grönqvist; E. Ahlsén; M. Gunnarsson (2001) *Annotations and Tools for an Activity Based Spoken Language Corpus*; proceed. of SIGdial-2001
- Grönqvist, L.; M. Gunnarsson (2003) *A Method for Finding Word Clusters in Spoken Language*; Proceedings CL2003 (Lancaster Univ.)
- Henrichsen, P.J. (1998) *Peeking Into the Danish Living Room - Internet Access to a Large Speech Corpus*; proceed. of NODALIDA-98; pp.109-119
- Henrichsen, P.J. (2002) *Some Frequency based Differences between Spoken and Written Danish*; Gothenburg Pap. in Theoretical Ling. 88
- Henrichsen, P.J. (2004) *Automatic Speech-to-speech Translation Based on Extremely Narrow Contexts*; Acta Linguistica Hafniensia, to appear
- Henrichsen, P.J.; J. Allwood (forthcom.) *Comparing Swedish and Danish Spoken and Written Languages*; submitted
- Molde, B. (1980, 2000) *Dansk-Svenska Ordbog*; Norstedts Förlag; 726pp
- Newman, P.; M. Ratliff (2001) *Linguistic Fieldwork*; Cambridge Univ. Press