Title:

# Applying Language Technology to Detect Shift Effects

Authors:

**John Nerbonne**

University of Groningen

j.nerbonne@rug.nl

**Timo Lauttamus**

University of Oulu

timo.lauttamus@oulu.fi

**Wybo Wiersma**

University of Groningen

wybo@logilogi.org

**Lisa Lena Opas-Hänninen**

University of Oulu

lisa.lena.opas-hanninen@oulu.fi

**Abstract**

We discuss an application of a technique from language technology to tag a corpus automatically and to detect syntactic differences between two varieties of Finnish Australian English, one spoken by the first generation and the other by the second generation. The technique utilizes frequency profiles of trigrams of part-of-speech categories as indicators of syntactic distance between the varieties. We then examine potential shift effects in language contact. The results show that we can attribute some interlanguage features in the first generation to Finnish substratum transfer. However, there are other features ascribable to more universal properties of the language faculty or to "vernacular" primitives. We conclude that language technology also provides other techniques for measuring or detecting linguistic phenomena more generally.

## 1 Introduction

The present paper[1] applies techniques from language technology, i.e. application-oriented computational linguistics, to detect syntactic differences between two different varieties of English, those spoken by first and second generation Finnish Australians. It also examines the degree to which the syntax of the first generation differs from that of the second, presumably due to the language shift that the first generation group made later in life and the traces it has left in their English. This line of research naturally attempts not only to detect differences of various kinds, but also to interpret their likely sources, including both first language interference but also more general tendencies, called "vernacular primitives" by Chambers (2003: 265-266). To explain differential usage by the two groups, we also draw on the strategies, processes and developmental patterns that second language learners usually evince in their interlanguage regardless of their mother tongue (Færch & Kasper 1983; Larsen-Freeman & Long 1991; Ellis 1994). To forestall a potential misunderstanding, we note that we propose how to automate the **detection** of the concrete syntactic differences, but not their **interpretation** (possible causes). The paper will summarize the findings concerning the Finnish emigrants that Lauttamus, Nerbonnne & Wiersma (2007) report on at length in order to give the reader a sense of the potential of the technique.

A second purpose of the paper is to reflect and generalize on the success of this technique borrowed from language technology in order to suggest that language technology might be a promising source in which to seek techniques for measuring or detecting linguistic phenomena more generally. Language technology has developed a number of techniques which expose the latent structure in language use. We harness

[1] This project is partly funded by the Academy of Finland (project # 113501).

one of those in the study of language contact, namely tagging words with their syntactic categories (parts-of-speech, hence POS), in an effort to detect the syntactic differences in the speech of juvenile vs. adult emigrants from Finland to Australia. We shall note other promising opportunities, but the purpose of the reflection is naturally not to claim that the language technology is a panacea for problems of linguistic analysis, but rather to stimulate readers to look toward language technology to explore issues in contact linguistics.

The first part of the paper summarizes the work on detecting syntactic interference among the Finnish emigrants to Australia, and the second makes the more programmatic argument that language technology should not be regarded as a set of tools for applications, but rather as a set of generic tools for exposing linguistic structure. Our paper does not focus on language contact exclusively as this has been influenced by globalization, but the contact effects we focus on do result from a substantial migration from one side of the earth to another. Our intention is to contribute to general techniques for the detection of syntactic differences.

## 2 Detecting Syntactic Differences: Techniques

Syntactic theory uses analysis trees showing constituent structure and dependency structure to represent syntactic structure, so a natural tool to consider for the task of detecting syntactic differences would be a parser – a program which assigns the syntactic structure appropriate for an input sentence (given a specific grammar). We decided, however, against the use of a parser, and for the more primitive technique of part-of-speech tagging (explained below) because, even though automatic parsing is

already producing fair results for the edited prose of newspapers, we suspected that it would be likely to function very poorly on the conversational transcripts of second language learners. Both the conversation style of the transcripts and the frequent errors of learners would be obstacles. We return below to the selection of corpora and its motivation.

### 2.1 Tagging

We detect syntactic differences in two corpora in a fairly simple way (Lauttamus et al. 2007). We first TAG the two corpora automatically, i.e. we automatically detect for each word its syntactic category, or, as it is commonly referred to, its part-of-speech (POS). Below we provide an example:

| (1) | the | cat | is |
|---|---|---|---|
| | ART (def) | N (com, sing) | V (cop, pres) |
| | on | the | mat |
| | PREP (ge) | ART (def) | N (com, sing) |

We tagged the corpora using the set of POS tags developed for the TOSCA-ICE, which consists of 270 POS tags (Garside et al. 1997), of which 75 were never instantiated in our material. Since we aim to contribute to the study of language contact and second language learning, we chose a linguistically sensitive set, that is, a large set designed by linguists, not computer scientists. In a sample of 1,000 words we found that the tagger was correct for 87% of words, 74% of the bigrams (a sequence of two words), and 65% of the trigrams (a sequence of three words). The accuracy is poor compared to

newspaper texts, but we are dealing with conversation, including the conversation of non-natives. Since parsing is substantially less accurate than POS tagging, we feel that this accuracy level confirms the wisdom of not trying to use the more informative technique of full parsing.

The POS tags are then collected into ordered triples, the TRIGRAMS ART(def)-N(com, sing)-V(cop, pres), ..., PREP(ge)-ART(def)-N(com, sing). We use POS trigrams, rather than single tags, as indications of syntactic structure in order to obtain fuller reflection of the complete syntactic structure, much of which is determined once the syntactic categories of words are known. In making this last assumption, we follow most syntactic theory, which postulates that hierarchical structure is (mostly) predictable given the knowledge of lexical categories, in particular given the lexical 'head'. Sells (1982, sec. 2.2, 5.3, 4.1) shows how this assumption was common to theories in the 1980s, and it is still recognized as useful (if imperfect given the autonomy of "constructions", which Fillmore & Kay (1999) demonstrate). So if syntactic heads have a privileged status in determining a "projection" of syntactic structure, then we will detect syntactic differences in two varieties by quantifying the distribution of parts-of-speech in context.

## 2.2 Comparison

We then collect all the POS trigrams found in the corpora (13,784 different POS trigrams in the case of the Finnish Australian data), and count how frequently each occurs in both of the corpora. We then compare this 2 X 13,784 element table, asking two questions. First, we wish to know whether the distribution in the two rows might

be expected by chance, in other words, whether there is a statistically significant difference in the distributions. Second, in case the overall distributions differ significantly (*p*-values at or below 0.05), we calculate which frequent POS trigrams are responsible for the skewed distribution. We suppress the technical details in this presentation, referring the interested reader to Nerbonne & Wiersma (2006).

In connection with the second goal, we examine the 200 POS trigrams that contribute the most to the skewing of the distribution between the two corpora. Both the relative differences in corpora (i.e. which percentage of a given POS trigram occurs in one corpus as opposed to another) and also the overall frequencies of the trigram are taken into account. We weight more frequent POS trigrams more heavily because more frequent patterns are likely to be perceptively salient, and also because we are most certain of them. We turn to an examination of the Finnish Australian data below.

### 2.3 Discussion

By analyzing differences in the frequencies of POS trigrams, we importantly identify not only deviant syntactic uses ("errors"), but also the overuse and underuse of linguistic structures, whose importance is emphasized by researchers on second language acquisition (Coseriu 1970; Ellis 1994: 304-306 uses for underuse 'underrepresentation' and overuse 'over-indulgence'; de Bot et al. 2005: A3, B3). According to these studies, it is misleading to consider only errors, as second language learners likewise tend to overuse certain possibilities and tend to avoid (and therefore underuse) others. For example, de Bot et al. (2005) suggest that non-transparent constructions are systematically avoided even by very good second language learners.

We like to emphasize that our work assumes, **not** that syntax consists solely of part-of-speech sequences, but only that differences in part-of-speech sequences are indicative of syntactic differences in general. It is important to emphasize that we do not claim to have developed a technique that probes all conceivable syntactic differences **directly**, but rather a technique that detects traces of differences in superficial syntax. Those differences might naturally have causes in deeper levels of syntactic structure. In a contribution with more room for reflection, we would expand on how we derive inspiration from other indirect measurement techniques such as the measurement of latitude via differences in the local solar noon with respect to Greenwich mean time, or the measurement of temperature via the expansion of a fluid.

Uriel Weinreich (1953: 63) noted the difficulty of aggregating over language contact effects:

> No easy way of measuring or characterizing the total impact of one language on another in the speech of bilinguals has been, or probably can be devised. The only possible procedure is to describe the various forms of interference and to tabulate their frequency.

Our proposed technique for detecting syntactic differences does indeed aggregate over many indicators of syntactic difference, in a way that makes progress toward assessing the "total impact" in Weinreich's sense, albeit with respect to a single linguistic level, namely syntax. We do not develop a true **measure** of syntactic difference here as that would require further calibration and validation, preferably cross-linguistically, but we

do claim to detect differences in the frequency with which different constructions are used.

If such a measure could be validated and calibrated, it would be important not only in the study of language contact but also in the study of second language acquisition. We might then look afresh at issues such as the time course of second language acquisition, the relative importance of factors influencing the degree of difference such as the mother tongue of the speakers, other languages they know, the length and time of their experience in the second language, the role of formal instruction, etc. It would make the data of such studies amenable to the more powerful statistical analysis reserved for numerical data.

### 2.4 Previous work

Aarts & Granger (1998) suggest focusing on tag sequences in learner corpora, just as we do. We add to their suggestion statistical analysis using permutation statistics, which allows us to test whether two varieties vary significantly. We discuss similar technical work below, none of which has focused on analyzing language contact, however.

## 3 The Australian English of Finnish Emigrants

We shall describe the differences between the English of those who emigrated as adults and those who emigrated as children (juveniles). After studying the transcripts,

we assume that the latter's English is near native, and so we focus below on the English of those who emigrated as adults.

### 3.1 Linguistic Situation of the Adult Emigrants

We note that the linguistic development of the two Finnish groups in Australia is best described as language shift. We are therefore concerned with bigenerational bilingualism as a series of stages in the assimilation of the Finnish ethnic minorities into a linguistically, socially and culturally English-dominant speech community, which inevitably entails the loss of the variety of Finnish used in the speech communities and Anglicization among these ethnic groups. We note that language shift seems to take place no later than during the 2nd generation of various ethnic groups in the US, with the exceptions of Spanish and Navajo (Karttunen 1977; Veltman 1983; Smits 1996; Klintborg 1999). The evidence from Hirvonen (2001) also supports this; American Finnish does not seem to survive as a viable means of communication beyond the second generation.

The situation is similar in Australia. Clyne & Kipp (2006: 18) note that "high-shift" groups in Australia tend to be ones who are culturally closer to Anglo-Australians in contrast with some "low-shift" groups with different "core values such as religion, historical consciousness, and family cohesion". The evidence in Lauttamus et al. (2007) suggests that also Finnish Australians represent those language groups that shift to English very rapidly in the second generation. It appears that even members of the 1st generation of immigrants may demonstrate a variety of achievements, including native-like ability (cf. Piller 2002), that members of the 2nd generation speak natively and that

language attrition does not wait till the 3rd generation but begins with the 1st generation (cf. Waas 1996; Schmid 2002, 2004; Cook 2003; Jarvis 2003). Consequently, we expect to find most of the evidence for syntactic interference (substratum transfer) in the English of first generation Finnish Australians, as the second generation has already shifted to English without any interference from Finnish. The findings in Lauttamus et al. (2007) all point in the direction that second generation Finnish Australians speak (almost) natively, with very little Finnish interference in their English. This is corroborated by findings in some other studies, such as Lahti (1999) and Kemppainen (2000) on lexical features, Mannila (1999) on segmental features, Laakkonen (2000) on rhythm, and Markos (2004) on hesitation phenomena.

Like similar groups in the United States (cf. Lauttamus & Hirvonen 1995), the adult immigrants typically go on speaking Finnish at home as long as they live, and carry on most of their social lives in that language, leaving Finnish their dominant language. They struggle to learn English, with varying success, e.g. usually retaining a noticeable foreign accent. But they are marginally bilingual, as most of them can communicate successfully in English in some situations.

We contrast their situation with that of their children. The immigrant parents speak their native language to their children, so this generation usually learns the ethnic tongue as their first language. The oldest child may not learn any English until school, but the younger children often learn English earlier, from older siblings and friends. During their teens the children become more or less fluent bilinguals. Their bilingualism is usually English-dominant: they tend to speak English to each other, and it is sometimes difficult to detect any foreign features at all in their English. As they grow older and move out of the Finnish communities, their immigrant language starts to

deteriorate from lack of regular reinforcement. Even if this generation marries within its ethnic group, as is frequently the case, English nonetheless becomes the language of the household, and only English is spoken to the following generation.

The language contact scholarship distinguishes situations of SHIFT from MAINTENANCE (Thomason and Kaufmann, 1988; Van Coetsem, 1988). The adult emigrant group, our focus here, maintains Finnish, but, more to the point, shifts to English, the subject of our research. Their Finnish is linguistically dominant, while English is socially dominant throughout Australia. In a situation of adult language shift, we expect interference from the native (Finnish) in the acquired (English) language, beginning with pronunciation (phonology) and morphosyntax. Lexical interference is comparatively weak.

### 3.2 Finnish Australian English Corpus (FAEC)

Greg Watson of the University of Joensuu compiled a corpus of English conversations with Finns who had emigrated to Australia nearly thirty years earlier (Watson 1996). This corpus was kindly put at our disposal. All the respondents were Finnish native speakers. We divided them into two groups, "adults", or adult emigrants, who were over 18 upon arrival in Australia, and "juveniles", the children of the adults, who were all under 17 at the time of emigration. We distinguish between adult immigrants and immigrant children based on Lenneberg's (1967) well-known critical age hypothesis, which suggests a possible biological explanation for successful L2 acquisition between age two and puberty. Note that 'adult' vs. 'juvenile' refers only to the age at emigration: all the respondents were over 30 at the time of the interviews.

The adults were 30 years old on arrival (on average), and 58.5 at the time of the one-hour interview, and the juveniles were 6 and 36, respectively. There were 62 adult and 28 juvenile interviews, and there were roughly equal numbers of males and females. The interviews were transcribed in regular orthography by trained language students and later checked by Watson. Speakers were not tested for English proficiency, but it is clear from a quick view of the data that the juveniles' English is considerably better than that of the adults'. The juveniles had gone to school in Australia, and the adults in Finland. Our corpora contained 305,000 words in total.

## 4 Differences observed

The following section summarizes some of the material in Lauttamus et al. (2007). The evidence from our syntactic analysis using the POS-tag trigrams and a permutation test like the one described in detail in Nerbonne & Wiersma (2006) shows that there are differences between the adults and the juveniles at a statistical significance level of 0.01. Our report focuses first on the aggregate effects of syntactic distance between the two groups of speakers and then we move on to discuss more specific "syntactic contaminants" in the English of the adults. The role of the language technology, specifically the POS-tags and the permutation test used to identify differing elements in the distribution, is that of detection. We also attempt to interpret the differences, but we have not enlisted language technology for this purpose.

### 4.1 General Effects

Some of the significant syntactic differences found in the data might be attributed to the lower level of fluency in the adults. Their language exhibits the following:

(a) Overuse of **hesitation phenomena** (pauses, filled pauses, repeats, false starts etc.), arising from difficulties in speech processing and lexical access.

(b) Overuse of **parataxis** (particularly with ***and*** and ***but***) as opposed to hypotaxis.

(c) Underuse of **contracted forms** that the juveniles use easily and naturally, e.g. *I've been running, I'd like to go, I'll finish my degree.* Adults mostly use full forms such as *I have been, she will be.*

(d) Reduced repertoire of **discourse markers** such as *you know, you see, I mean.* Adults do use *you know* (with other hesitation phenomena), but as a time-gaining device rather than as a genuine discourse marker. In contrast, the juvenile emigrants use a more varied repertoire of markers, which often function as appeals to the interviewer.

(e) Avoidance of **complex verb clusters.** Juvenile, but not adult, emigrants use structures such as *I would have had it, I still probably would have ended up getting married.*

(f) Avoidance of **prepositional and phrasal verbs.** In contrast, juvenile emigrants have no difficulty with verbs such as *I* ***ran out of*** *money, I just* ***opted out for*** *an operation.*

(g) Underuse of the existential ***there***. The adults either avoid using the existential or attempt to express it without the word *there* (cf. section 4.2.3). We include this in the list of general differences as an example of a general difficulty that speakers have with peculiar English constructions.

We extracted the properties above by investigating frequent POS-trigrams that differed significantly in one group as opposed to another (individual *p*-values at or below 0.05), also using 90% relative frequency as a threshold, i.e. where 90% occurred in the one group or the other (once we applied a correction for overall difference in corpus size). This means that **avoidance** does not imply total absence of a feature in a group. Nor do we wish to suggest that adults are **consciously** avoiding certain constructions such as hypotaxis. The differences in usage patterns could arise through other strategies.

The ability to identify these sources of deviation in the use of English by the adult Finnish emigrants confirms our contention that the comparison of POS-trigram frequencies indeed reflects the syntactic distance between the two varieties of English and, consequently, aggregate effects of the difference in the two groups' English proficiency. The shift to English has indisputably proceeded along different paths in the two groups, the adults (still) showing features of "learner" language, or shift with interference, and the juveniles those of shift without interference.

### 4.2 Specific syntactic effects

We turn to differences in specific constructions. In examining these, we shall interpret them on the basis of our knowledge of standard (acrolectal) Finnish and English, which is a risky undertaking. We shall likewise entertain interpretations based on what we know about non-standard (basilectal) varieties of English and Finnish, but our knowledge is less than perfect here.

In examining the following differences in POS-trigram frequency, we will be asking whether the observed syntactic deviations from the norms of standard (acrolectal) English may be ascribed to contact effects from a Finnish substratum, to more universal, 'natural' tendencies in non-standard varieties in general, or to other factors. Modesty compels us to note that we are aware that there are many further sources of influence which might explain why the language of second language learners differs from that of native speakers. Lauttamus et al. (2007) discusses this in more detail.

To illustrate how explanations compete, consider the fact that different adult speakers fail to enforce subject-verb concord, thus *They all doned here, they, - they wasn't raw [kangaroo] skin.* The subject-verb nonconcord in *they wasn't* ('they weren't') is all putatively a vernacular universal. But in non-acrolectal Finnish, subject-verb nonconcord is also frequent, e.g. *ne meni Groningeniin* ('they went to Groningen', *ne*, plural of *se* 'it', + *meni* 'went' 3rd person sg), which shows subject-verb nonconcord in person and number, as opposed to standard Finnish: *he menivät Groningeniin* (*he* 'they' + *meni+vät* 'went' + 3rd person pl). Just as some vernacular Englishes, non-acrolectal Finnish also violates the standard subject-verb concord rule.

To support a potential role of the 'vernacular' approach in our analyses, we refer to Fenyvesi & Zsigri (2006: 143). They suggest that less educated speakers of English (such as the adults), who have usually learnt their L2 via listening, rely on *auditory* input, whereas more educated immigrant language speakers (such as our juveniles), who have acquired their L2 also through reading and writing, and therefore been exposed to a more or less codified standard (acrolectal) variety, rely on *visual* input as well. The

fact that the adults in our study have mainly been exposed to spoken, *basilectal* (Australian) English is likely to give rise to some general vernacular features.

We discuss two patterns in detail, one we attribute to the Finnish substrate, and the other to general simplifying tendencies. We note several others more briefly, hoping to provide the flavor of the previous work.

### 4.2.1 Article usage

The adults demonstrate overuse (and underuse) of the indefinite and definite articles, *a(n)* and *the*, characteristic of a learner whose L1 has no article system (such as Finnish), as exemplified in the following:

(2a) in that time /in a Finland/ because wasn't very

(2b) first we go /to the Finland/

(3) we been /in a Brisbane/ Brisbane because ah

(4) in /the Brisbane and/

(5) I had /a different birds/ in Finland

In example (5) the indefinite article occurs with a countable, plural noun head, a very unlikely overuse for a native speaker, however informal. The juveniles do not show similar linguistic behavior being more proficient in English.

Finnish Americans also overuse the articles, particularly the indefinite article, using it, for example, with proper nouns. Pietilä (1989: 167-168) shows that Finnish Americans, particularly elderly, first generation speakers often supply redundant

definite articles, such as those in (2b) and (4) rather than indefinites, such as (2a) or (3). Pavlenko & Jarvis (2002: 207) show that most of the L1-influenced article errors committed by Russian L2 users of English were omissions and that only a few involved oversuppliance of the definite article. (Similarly to Finnish, Russian has no article system.)

Finnish does allow for the use of the demonstrative pronouns *this* and *that* to mark definiteness instead of the definite article, which may explain overuses we found in English demonstratives used by the adult emigrants:

(6) it's /this taxation is/ really something in Finland

(7) I watch /that ah news/ and 'Current Affair'

In these contexts there is no apparent need to use the demonstratives, e.g. a need to contrast one news (broadcast) with another. We note, however, that in a potential Finnish variant of (6), *juuri tämä verotus* […] *Suomessa…*('[it's] the very taxation […] in Finland') it would be quite acceptable to use the demonstrative *tämä* 'this' to make the reference not only definite but also specific. We conclude tentatively that the adult overuse of the demonstratives also originates from Finnish substratum transfer. This is consonant with the fact that adults may also overuse *that one* in expressions such as *I don't /remember that one/ either, I can't /explain that one/, I can't really /compare that one/*, where the NP *that one* has more or less the same function as the pronoun *it.*

To summarize, we ascribe the deviant usage of the articles in the English of the adult Finnish Australians to substratum transfer from Finnish (which has no article system to express (in)definiteness and specificity). Because Finnish has no articles, we

might think that there is nothing to transfer (cf. Arabski 1979). However, we agree with Ellis (1994: 306-315), who argues that the absence of a feature in the first language may have as much influence on the second language as the presence of a different feature. In addition to contact-induced effects, it appears that general HYPERCORRECTION (or overgeneralization), common in 'learner' language, may be a contributing factor. In this light, an uncertainty of article usage in speakers whose L1 has no articles is "universal".

### 4.2.2 Acquired formulae

The distribution of the POS-trigrams also revealed that the adults have acquired some formulae such as *that's* and *what's* without mastering their grammar:

(8) ah /*that's is*/ not my occupation

(9) I think /*that's is*/ a no good

(10) um /*that's is*/ a same um

(11) and /*that's a*/ causing discomfort in

(12) oh /*what's is*/ on that

*That's* and *what's*, acquired as fixed phrases, have apparently been processed as single elements. The fact that they are then combined with full copulas or progressive auxiliaries indicates that the speaker has not mastered the grammar of the reduced-form clitic. We also found examples such as *what's a that sign, what's a that seven or something.* Ellis (1994: 20), for one, argues that learners often produce formulae or ready-made chunks as their initial utterances. Acquired formulae cannot be ascribed to

substratum transfer, as they tend to be recurrent in any interlanguage. In particular, however, we know of no plausible Finnish model for this difference.

### 4.2.3 Other deviant patterns detected via POS-tag distribution differences

In this section we note some of the other deviant patterns discussed at length in Lauttamus et al. (2007). We discuss them here to add a sense of the value of the technique.

#### Omission of the progressive auxiliary *be*

Adults frequently omit the progressive auxiliary verb *be* (present and past) while the juveniles do not. The adults produce examples such as *when we /drivin' in the/ road.* Absence of the copula *be* is one of the alleged vernacular universals (Chambers 2004), and we also found more numerous examples of that in the adults' speech. Even though Finnish has no formal contrast between the progressive and non-progressive aspect, so that it might be expected to be problematic, still that does not seem to explain this specific error, which we therefore would ascribe to more universal properties of language contact rather than to substratum transfer from Finnish, specifically the difficulty learners have in acquiring unstressed elements. Pietilä (1989: 180-181) also notes that the most frequent verb form error in the English of the first generation Finnish Americans is the omission of the primary *be* in the progressive.

#### Omission of existential *there* and anaphoric *it*

We noted an example of the omission of the existential (expletive) *there* in 4.1 above, which we class here with problems with the anaphoric *it* in subject position. We find examples such as *and summer /time when Ø is/ a people*, where it is apparent that the speaker is aiming at *when <u>there</u> is/are people* or *because is [tough]* where '*because <u>it</u> is [tough]'*. These examples can be explained in terms of substratum influence from Finnish, which would assign the subject argument of the copula verb *be* to the NP *(a) people*, or to the AP *[tough]*, and, consequently, would not mark the subject in the position before the copula.

**Absence of prepositions**

The adults tend to leave out prepositions with motion verbs such as *move*, *go*, *come*, as exemplified in *and they move me <u>/other room</u> where/*. Since Finnish has no prepositions, this looks like a straightforward case of substratum influence, but the case is complicated by the fact that many vernacular varieties of English tend to leave out prepositions in expressing spatial relations with motion verbs (cf. e.g. Linn 1988).

**Deviant word order**

The adults also demonstrate deviant word order, particularly with adverbials, which are often placed before the object, as exemplified in *I /don't like really/ any old age*. As pre-object placement of the equivalent adverbials in Finnish would be quite acceptable, we suspect that this is contact-induced. Similarly time adverbials were

found to be positioned differently, with the adults favoring a front position for frame adverbials such as *fifteen years ago we drivin' around*, but not so drastically as to result in out-and-out errors. This is a feature that can be ascribed to Finnish substratum transfer as well, since in Finnish a time adverbial often appears in this position, without being accompanied by focus as in English. We conjecture that the adults are overusing this construction ignoring its pragmatic conditioning.

***not* in pre-verbal position**

The adults produce negated utterances where the *not* is placed in pre-verbal position, as in *but uh /we not cook/ that way* (without the supporting verb *do*). This can be ascribed to Finnish substratum transfer, because Finnish always has the negative item (*ei*, inflected in person and number like any verb in standard Finnish) in pre-verbal position. Here again, there is a plausible alternative explanation from universal tendencies in second language syntax. Ellis (1994: 99-101, 421-422) notes that "there is strong evidence that in the early stages of L2 acquisition learners opt for preverbal negation, even where the L1 manifests postverbal negation" (p. 421).

### 4.2.4 Conclusions with respect to results

Lauttamus et al. (2007) note as well that differences became visible with respect to the use of *what* as a relative pronoun (*cars what they built*), the overuse of the simple present (*when we come in Australia thirty years ago*), and in the tendency to form all

measure terms with plurals; we found only one example in which the singular is used (*three **foot** wide standing up*). The conclusion there was that the computational techniques had been useful in detecting deviant patterns, but the paper was focused on the interpretation of differences found.

We note that our work has focused on detecting differences using rough, shallow syntax annotation. A calibrated metric of difference would provide a numeric score of syntactic difference in a way that would allow us to compare across languages, e.g. to compare the English of Finnish immigrants to the English of Chinese immigrants, and we have not attempted that to date. It is clear that this would be interesting from several points of view, including second language learning and language contact studies. One might think that the $R^2$ scores (or $\chi^2$ scores) used internally might serve this purpose, but they both depend on corpus size, which is a poor property for a candidate measure. Minimally we would need to correct for corpus size in order to use them. This will be future work.

## 5 Language technology offers tools to study language contact

There are several similar uses of language technology (LT) in detecting differences in language use, especially in information retrieval, authorship detection, and forensic linguistics, in general in all those fields where TEXT CLASSIFICATION plays a role. Information retrieval classifies texts into relevant vs. irrelevant (with respect to a user query), authorship detection classifies texts according to their authors, and forensic linguistics does the same (among other things). Nerbonne (2007) reviews especially the work on authorship, surely the most challenging classification task. The

suggestion here is that we might approach language contact studies from a similar perspective, attempting to design systems that classify texts into native and contact-affected, naturally always from the perspective of a single language (we are unlikely to detect contact effects cross-linguistically). It almost goes without saying that for language contact studies the real interest is less in the sheer ability to classify and more in the linguistic features that form the basis of the classification, but the other fields have likewise been interested in the linguistic basis of successful classifiers, so language-contact studies is by no means alone in the wish to identify concrete linguistic effects. The well-informed reader may object that a great deal of text classification focuses on lexical features, but we would note that syntactic features enjoy growing popularity (see below as well).

As a brief aside, let us add that there has been a significant infusion of LT techniques in the field of dialectology, which (often) attempts to separate varieties into classes of dialects spoken in dialect areas, and which has therefore made use of classifying techniques similar to those used in text classification. Nerbonne (to appear) provides an overview of LT techniques which have been brought to bear in measuring varietal differences, and Spruit (2008) applies LT techniques to the problem of classifying Dutch dialects on the basis of syntax. Spruit had the luxury of basing his work on an elaborate database on the syntactic properties of the Dutch dialects, the *Syntactic Atlas of Netherlandic Dutch* (SAND). We are not aware of similar resources available in language contact studies.

If the study reported on in the first part of this paper appears promising, we suggest that further investigations into the use of LT in contact studies would be fruitful.

To give some idea of what might be possible, let us discuss points at which the present article might have been different, perhaps better.

For example, we might have attempted to detect syntactic structure at a more abstract level. We chose to use POS-tags, the syntactic categories of the words in the text. But LT offers at least three more possibilities, that of CHUNKING, that of PARTIAL PARSING, and that of FULL PARSING. Chunking attempts to recognize non-recursive syntactic constituents, *[the man] with [the red hair] in [the camel-hair jacket] waved as he [passed by],* while partial parsing attempts to infer more abstract structure where possible: *[$_S$ [$_{NP}$[the man] with [the red hair]] in [the camel-hair jacket] waved] as [$_S$ he [passed by]].* Note that the latter example is only partially parsed in that the phrase *[in the camel-hair jacket],* which in fact modifies *man,* but which might mistakenly be parsed to modify *[hair],* is not attached in the tree. In general, chunking is easier and therefore more accurate than partial parsing, which, however, is a more complete account of the latent hierarchical structure in the sentence. There is a trade-off between the resolution or discrimination of the technique and its accuracy.

Baayen, van Halteren & Tweedie (1996) work with full parses on an authorship recognition task, while Hirst & Feiguina (2007) apply partial parsing in a similar study, obtaining results that allow them to distinguish a notoriously difficult author pair, the Brontë sisters. The point of citing them here is to emphasize that LT methods are being applied to practical problems even today: one should not regard them merely as promising possibilities for the future.

Our study might also have attempted to use these more discriminating techniques, but we were dissuaded by the fact that the more sensitive techniques have more difficulty in analyzing unedited, indeed, very spontaneous, text, which has the

added difficulty of being ridden with second language errors. But whether this is in fact the case is an empirical question waiting for a future research project.

It is clear that the technical view may add to the value of the work. Hirst & Feiguina (2007) are at pains to establish that their technique can work for even short texts (500 wd. and fewer), and this could be an enormous advantage in considering other applications, e.g. to the pedagogical question of identifying foreign influences in the writing of second language users.

## 6 Conclusion

In this paper we argue that by using frequency profiles of trigrams of POS categories as indicators of syntactic distance between two different groups of speakers we can detect the "total impact" of L1 on L2 in SLA. Our findings show syntactic contamination from Finnish in the English of the adult first generation speakers, and, moreover, we were able to identify several syntactic areas in which the adult emigrants differed significantly from their native-like children. Some of the features found in the data can be explained by means of contact-induced influence whereas others may be primarily ascribed to "learner" language or to more universally determined properties of the language faculty. We close the paper with an appeal to researchers in the study of language contact to look to language technology for tools to reveal the latent structures in language use, especially syntactic and phonological structure.

## References

Aarts, J. & S. Granger. 1998. Tag sequences in learner corpora. A key to interlanguage grammar and discourse. *Learner English on computer*, S. Granger (ed.), 132-141. London: Longman.

Arabski, J. 1979. *Errors as indicators of the development of interlanguage.* Katowice: Universytet Slaski.

Baayen, H., H. van Halteren & F. Tweedie. 1996. Outside the cave of shadows. Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3): 121-131.

Bot de, K., W. Lowie & M. Verspoor. 2005. *Second language acquisition: An advanced resource book.* London: Routledge.

Chambers, J.K. 2003. *Sociolinguistic theory. Linguistic variation and its social implications.* Oxford: Blackwell.

Chambers, J.K. 2004. Dynamic typology and vernacular universals. *Dialectology meets Typology*, B. Kortmann (ed.), 127-145. Berlin: Mouton.

Clyne, M. & S. Kipp. 2006. Australia's community languages. *International J. Soc. Lang.* 180: 7-21.

Cook, V. (ed.). 2003. *Effects of the second language on the first.* Clevedon, UK: Multilingual Matters.

Coseriu, E. 1970. *Probleme der kontrastiven Grammatik.* Düsseldorf : Schwann.

Ellis, R. 1994. *The study of second language acquisition.* Oxford: Oxford University Press.

Fenyvesi, A. & G. Zsigri. 2006. The role of perception in loanword adaptation. The fate of initial unstressed syllables in American Finnish and American Hungarian. *SKY Journal of Linguistics* 19: 131-146.

Fillmore, C. & P. Kay. 1999. Grammatical construction and linguistic generalizations. The *what's x doing y* construction. *Language* 75: 1-33.

Færch, C. & G. Kasper. 1983. *Strategies in interlanguage communication*. London: Longman.

Garside, R., G. Leech & T. McEmery. 1997. *Corpus annotation. Linguistic information from computer text corpora.* London: Longman.

Hirst, G. & O. Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary & Linguistic Computing* 22(4): 405-419.

Hirvonen, P. 2001. Doni finished – meni läpi – highskoulun. Borrowing, code-switching and language shift in American Finnish. *Global Eurolinguistics. European languages in North America – Migration, maintenance and death,* P. Sture Ureland (ed.), 297-324. Tübingen: Niemeyer.

Jarvis, S. 2003. Probing the effects of the L2 on the L1. A case study. *Effects of the second language on the first*, V. Cook (ed.), 81-102. Clevedon, UK: Multilingual Matters.

Karttunen, F. 1977. Finnish in America. A case study in monogenerational language change. *Sociocultural dimensions of language change*, B.G. Blount & M. Sanches (eds.), 173-184. New York: Academic Press.

Kemppainen, J. 2000. Lexical features in the spoken English of Finnish Australians. MA thesis, University of Oulu.

Klintborg, S. 1999. *The transience of American Swedish.* Lund: Lund University Press.

Laakkonen, K. 2000. A study of the realization of rhythm in the English of Finnish Australians. MA thesis, University of Oulu.

Lahti, H. 1999. Lexical errors in the spoken English of Finnish Australians. MA thesis, University of Oulu.

Larsen-Freeman, D. & M.H. Long. 1991. *An introduction to second language acquisition research.* London: Longman.

Lauttamus, T. & P. Hirvonen. 1995. English interference in the lexis of American Finnish. *The New Courant* 3: 55-65. Department of English, University of Helsinki: Helsinki University Press.

Lauttamus, T., J. Nerbonne & W. Wiersma. 2007. Detecting syntactic contamination in emigrants. The English of Finnish Australians. *SKY Journal of Linguistics* 21: 273-307.

Lenneberg, E. 1967. *Biological foundations of language.* New York: John Wiley.

Linn, M. 1988. The origin and development of the Iron Range Dialect in Northern Minnesota. *Studia Anglica Posnaniensia* XXI: 75-87.

Mannila, T. 1999. A study of phonic interference from Finnish in the English of Finnish Australians. MA thesis, University of Oulu.

Markos, M. 2004. '*No, no swearing, no swearing allowed'*. A comparative study of hesitation phenomena in the spoken English of two generations of Finnish Australians. MA thesis, University of Oulu.

Nerbonne, J. 2007. The exact analysis of text. Foreword to the 3rd ed. of F. Mosteller & D. Wallace *Inference and disputed authorship: The Federalist,* xi-xx. Stanford: CSLI.

Nerbonne, J. (to appear). Techniques for measuring dialect differences. *Theories and methods*, J. E. Schmidt & P. Auer (eds.). Mouton: Berlin [Language and Space].

Nerbonne, J. & W. Wiersma. 2006. A measure of aggregate syntactic distance. *Linguistic distances*, J. Nerbonne & E. Hinrichs (eds.), 82-90. Shroudsburg, PA: ACL.

Pavlenko, A. & S. Jarvis. 2002. Bidirectional transfer. *Applied Linguistics* 23: 190-214.

Pietilä, P. 1989. *The English of Finnish Americans with reference to social and psychological background factors and with special reference to age*. Turun yliopiston julkaisuja, Sarja B, Osa 188. Turku: Turun yliopisto.

Piller, I. 2002. Passing for a native speaker. Identity and success in second language learning. *Journal of Sociolinguistics* 6: 179-206.

Sells, P. 1982. *Lectures on contemporary syntactic theories.* Stanford: CSLI.

Smits, C. 1996. *Disintegration of inflection. The case of Iowa Dutch.* The Hague: Holland Academic Graphics.

Schmid, M. 2002. *First language attrition, use and maintenance. The case of German Jews in Anglophone countries*. Amsterdam: John Benjamins.

Schmid, M. 2004. Identity and first language attrition. A historical approach. *Estudios de Siciolingüística* 5: 41-58.

Spruit, M. 2008. *Quantitative perspectives on syntactic variation in Dutch dialects*. Ph.D. diss, University of Amsterdam.

Thomason, S.G. & T. Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Van Coetsem, F. 1988. *Loan phonology and the two transfer types in language contact*. Dordrecht: Foris.

Veltman, C. 1983. *Language shift in the United States.* Berlin: Mouton.

Waas, M. 1996. *Language attrition Downunder. German speakers in Australia*. Frankfurt: Peter Lang.

Watson, G. 1996. The Finnish-Australian English corpus. *ICAME Journal: Computers in English Linguistics* 20: 41-70.

Weinreich, U. [1953] 1974. *Languages in contact*. The Hague: Mouton.