Timo Lauttamus, John Nerbonne and Wybo Wiersma

Detecting Syntactic Contamination in Emigrants: The English of Finnish Australians[1]

Abstract

The paper discusses an application of a technique to tag a corpus containing the English of Finnish Australians automatically and to analyse the frequency vectors of part-of-speech (POS) trigrams using a permutation test. Our goal is to detect the linguistic sources of the syntactic variation between two groups, the 'Adults,' who had received their school education in Finland, and the 'Juveniles,' who were educated in Australia. The idea of the technique is to utilise frequency profiles of trigrams of POS categories as indicators of syntactic distance between the groups and then examine potential effects of language contact and language ('vernacular') universals in SLA. The results show that some features we describe as 'contaminating' the interlanguage of the Adults can be best attributed to Finnish substratum transfer. However, there are other features in our data that may also be ascribed to more "universal" primitives or universal properties of the language faculty. As we have no evidence of potential contamination at the early stages of the Juveniles' L2 acquisition, we cannot yet prove or refute our hypothesis about the strength of contact influence as opposed to that of the other factors.

1. Introduction

The present paper applies computational techniques to obtain an aggregate measure of syntactic distance between two different varieties of English spoken by first and second- generation Finnish Australians and examines the degree of what we call syntactic 'contamination' in the English of the older emigrants (Adults). Our goal is to detect the linguistic sources of the variation between the two groups of speakers and interpret the findings from (at least) two perspectives, universal vs. contact influence. In our reading, the notion of "universal" is concerned, *not* with hypotheses about Chomskyan universals and their applicability to second language acquisition (SLA) or with hypotheses about potential processing constraints in any detail, *but* rather with more general properties of the language

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

faculty and natural tendencies in the grammar, called 'vernacular primitives' by Chambers (2003: 265–266). To explain differential usage by the two groups, we also draw upon 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, Thomason 2001).

Our method of detecting the linguistic sources of the variation between the two groups of speakers relies on a comparison of two sets of data. Making inferences about deviant usage in the groups by utilising our knowledge of standard (acrolectal) Finnish and English has its obvious limitations. To avoid a bias towards the acrolect, we also make deductions about observed 'contamination' on the basis of what we know about non-standard (basilectal) varieties of English and Finnish. This is important because we also consider a potential impact of vernacular primitives on the data to be analysed. We note that some features in an acrolect (such as the *–s* inflection in the third-person present tense in English) are often in violation of natural tendencies in the grammar, whereas others, recurrent in all vernaculars (basilects), e.g. subject-verb nonconcord, seem to be in violation of standard varieties. And we need to consider both types of features, since first-generation Finnish Australians are primarily exposed to a variety of English that is best characterised as a (spoken) basilect, while second-generation speakers, being educated in Australia, are also exposed to a (spoken and written) acrolect.

Finnish emigrants to Australia seem to represent those language groups that shift to English very rapidly in the second generation. 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." Although the authors do not mention Finnish Australians, we argue that the high-shift groups also include them. Consequently, we expect to find most of the evidence for syntactic contamination in the English of first-generation Finnish Australians, as the second generation may have already shifted to English without any interference from Finnish.

The fundamental idea of the technique proposed is to tag the material to be investigated automatically and analyse the frequency vectors of POS (part-of-speech) trigrams using a permutation test. An analysis of a corpus

containing the English of Finnish emigrants to Australia is promising in that the procedure described in detail in section 4 works well in distinguishing two different groups of speakers and also in highlighting syntactic deviations between the two groups. Using frequency profiles of trigrams of POS categories as indicators of syntactic distance between the groups, we can also interpret potential effects of language contact and/or language ('vernacular') universals more economically and efficiently in SLA.[2]

Most of the cross-linguistic research into SLA (see, e.g. Odlin 1989, 1990, 2006a, 2006b) has so far focused on examining typical second-language learners' errors, such as absence of the copula, absence of prepositions, different (deviant) uses of articles, loss of inflectional endings, and non-English word order, and on making inferences about them to explain potential substrate influence. Our aim, however, is also to detect a wider range of syntactic differences, including, e.g. the overuse of particular patterns and the eschewing of non-transparent or "difficult" constructions in one group of speakers as opposed to the other. In accordance with Nerbonne & Wiersma (2006), we therefore argue that by applying the procedure proposed in the following sections we are now in a better position than before to measure the "total impact of one language on another in the speech of bilinguals" and determine the aggregate effects of contact in the way that Weinreich (1953: 63) considered difficult to assay.

2. Finnish Australian English Corpus (FAEC)

We apply the procedure described in section 4 to a corpus compiled in 1994 by Greg Watson of the University of Joensuu, Finland (Watson 1995, 1996). The informants were all Finnish emigrants to Australia, and they are collectively classified in this report according to two criteria: (1) the 'Adults' (group 'A,' adult immigrants), who were over the age of 18 upon arrival in Australia; (2) the 'Juveniles' (group 'J,' juvenile immigrant

[2] Väyrynen (2005: 34-35, referring to a discussion in Hunnicutt & Carlberger 2001: 259) argues that the (word) trigram language model has been successful, because trigrams can, for example, simultaneously reflect the syntactic, semantic and pragmatic levels of language use. Although n-gram language models cannot account for syntactic long-distance dependencies, they can capture local word occurrence constraints efficiently (cf. Brill & Mooney 1997: 19, Sanders 2007: 1).

children of these adults), who were born in Finland and were all under the age of 17 at the time of emigration.[3]

The corpus studied for this report consists of 62 adult interviews and 28 immigrant child interviews, each lasting approximately 65 to 70 minutes. The average age of the Adults was 30 at the time of arrival, and 58.5 at the time of the interview, as opposed to the Juveniles, who were, on average, 6 at the time of arrival and 36 at the time of the interview. We will refer to those who emigrated as children as Juveniles and the interviews with them as Juvenile interviews even though their average age was 36 at the time of the interview. The genders are almost equally represented in the two groups. The interviews were transcribed in regular orthography by trained language students and later checked by the compiler of the corpus. We used only those sections of the interviews which consist of relatively free conversation, i.e. a total of 305,000 word tokens. We distinguish between adult immigrants and immigrant children based on Lenneberg's (1967) critical age hypothesis, which suggests a possible biological explanation for successful L2 acquisition between age two and puberty.[4]

Neither group of speakers was formally tested for their proficiency of English. By observing the data, however, we can confidently say that the average level of the Adults' English is considerably lower than that of the Juveniles', who had received their school education in Australia, as opposed to the Adults, who were educated in Finland.

[3] We would still like to find an opportunity to examine a comparison between non-immigrant English and extremely fluent immigrant English. We have no comparable data yet which would allow us to compare the Juveniles' language to that of native Australians. An investigation of a third group of Finnish Australians born in Australia must be postponed until we have all their informal interviews transcribed, digitised, and analysed.

[4] We are aware of the debate between the proponents of the 'critical,' or 'sensitive' period hypothesis (cf. e.g. Long 2005; Singleton & Ryan 2004) and those who favour not only biological, but also other, more general factors (such as socio-psychological and experiential variables) which may hinder the acquisition of native-like proficiency in L2 (cf. Moyer 2004).

3. Defining the language contact situation of Finnish Australians

The following description of the language contact situation of Finnish Australians is, by and large, based on a similar account of a number of immigrant groups of European origin in the United States (cf. Lauttamus & Hirvonen 1995), and supported by the biographical interview data elicited from our informants in Australia.

The immigrant generation (and particularly the Adults) will typically go on speaking Finnish at home as long as they live, and carry on most of their social life in that language. They struggle to learn English, with varying success. Even the best learners usually retain a distinct foreign accent and some other foreign features in their English. But we can say that they are *marginally bilingual*, as most of them can communicate successfully in English in some situations at least, although their immigrant language is clearly dominant.

The immigrant parents will also speak their native language to their children (the Juveniles), so this generation usually learns the ethnic tongue as their first language. The oldest child may not learn any English until he or she goes to school. The younger children often start learning English earlier, from their older siblings and friends. At any rate, during their *teen years* the second-generation children become more or less *fluent* bilinguals. Their bilingualism is usually English-dominant: they prefer 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.

In the second generation it is still common to marry within the ethnic group. But even if the spouses share an ethnic heritage, they are usually not comfortable enough in the ethnic language to use it for everyday communication with each other. Therefore they will not speak it to their children, either. For an ethnic language to stay alive as a viable means of communication for a longer time than two generations would require a continuous influx of new immigrants but, as Watson (1995: 229) points out, "since the late 1970's onwards immigration to Australia has been quite restricted, to all nationalities." Clyne & Kipp (2006: 18) also note that the language groups "with a very low shift to English" are those that have recently arrived from Southeast and East Asia and Africa, whereas the

high-shift groups tend to be ones "for whom there is not a big cultural distance from Anglo-Australians." Although there is no mention of Finnish Australians in Clyne & Kipp, we argue that they belong to the high-shift group.

The Finnish language in Australia is shaped by two factors, (1) isolation from the development of the Finnish language (an agglutinative one) in Finland, and (2) the powerful influence of English (an analytic language), language of the dominant culture. The former factor causes retention in Australian Finnish of some forms that have fallen out of use in Finland, and also considerable uncertainty among Finnish Australians as to what is "correct." The latter factor causes innovation in the form of lexical borrowing and of loan translations (cf. Kovács 2001b), but also, especially in the second generation of speakers, phonological, morphological and syntactic changes. Some give up speaking their ethnic language altogether.

From the *typological* point of view, the following generalisations can be made on the basis of the description above. Clearly, the basic pattern is one of *maintenance* of the ethnic language by the Finnish immigrant generation (Adults) and the subsequent *shift* from Finnish to English by the second generation (Juveniles). The first generation was characterised as "marginally bilingual." In contrast with the fluent bilinguals of the second generation, they can also be regarded as *non-fluent* bilinguals, or as L2 learners with some degree of L1 (Finnish) interference. The characterisation of the language contact described above implies that Finnish is *linguistically dominant* over English for the first generation (Adults), whereas English is *socially dominant* over Finnish.

Table 1 summarises the predictions that can be made about the linguistic levels affected by the two transfer types, *maintenance* and *shift*, in the transfer situations of the Adults.

	Interference in Finnish	Interference in English
	English (L2) → Finnish (L1)	Finnish (L1) → English (L2)
	sl → *RL*	*SL* → *rl*
	MAINTENANCE	*SHIFT*
Lexicon	strong	weak
Phonology	weak	strong
Morphosyntax	weak	moderate/unclear

Table 1. The two transfer situations and the linguistic levels predicted to be affected by interference in the (Australian) Finnish-(Australian) English language contact among the Adults (cf. Lauttamus & Hirvonen 1995: 59; Thomason & Kaufman 1988, Van Coetsem 1988, 1995). 'sl'/'SL' is source language, and 'rl'/'RL' recipient language. The linguistically dominant language of the speech community is typed in upper case. The attributes 'strong,' 'moderate' and 'weak' refer to the degree of the predicted interference.

One of the two transfer situations (maintenance) can therefore be specified as *sl* → *RL*. In this situation, which is typical of the Adults, English is the source language (*sl*) and Finnish the recipient language (*rl*). Characteristic of this transfer situation is *lexical borrowing,* whereby loan words are phonologically and morphologically adapted to the patterns of the *rl*. The levels of phonology, morphology and syntax ('morphosyntax') of Australian Finnish spoken by the immigrant generation seem to be in general resistant to interference from Australian English. As Kovács (2001a: 98) points out, borrowed words in Australian Finnish show "complete phonological (as well as morphological and syntactic) integration into the Finnish language system." The crucial feature is that first-generation Finnish Australians still maintain their own native language, Finnish. Since the focus of the present paper is on language shift, we will not examine the maintenance situation any further.

In the second of the two transfer situations of the first-generation Finnish Australians (shift), particularly that of the Adults, the interference from their native Finnish in their acquired English does not begin with vocabulary but with sounds (phonology) and morphosyntax. This pattern of interference, *SL* → *rl*, is characteristic of language shift. Evidence from the English spoken by first-generation Finnish Americans demonstrates that the phonological and morphosyntactic patterns often deviate from standard (American) English in the manner typical of 'learner language' or

interlanguage (cf. Pietilä 1989: 152–189; Hirvonen 1988). A similar linguistic behaviour is expected particularly of the Adult speakers of Finnish Australian English.

The column 'shift' represents the levels affected by interference from Finnish in English. The English spoken by first-generation Finnish Australians is primarily affected in its phonology, to a lesser extent in its morphosyntax, while lexical interference is only weak. It is also expected that the Adults show more contact-induced effects in their speech than the Juveniles, since the former only temporarily shift to English, as in the case of an interview, whereas the latter have already shifted to English as their language of everyday communication.

The fact that lexical interference from Finnish in English is weak could be explained as follows. The restricted variety of English spoken by the Adults is almost invariably used for out-group communication only. Given that (Australian) English is socially (but not linguistically) dominant over Finnish, massive lexical interference from Finnish would be detrimental to successful communication with monolingual English speakers. The expected direction of lexical interference is thus from *the socially dominant language into the socially subordinate one.* As shift-induced interference is mainly phonological and morphosyntactic, we will not examine lexical interference any further.

As Lauttamus & Hirvonen (1995) argue, from a synchronic point of view the transfer situation *SL* $\rightarrow$ *rl* described above, along with other comparable interlanguage situations, contains features of *shift with interference*. A distinction must, however, be made between the synchronic description of the transfer situation and the actual *outcome* of the shift. As evidenced by Lauttamus & Hirvonen's (1995) description, it can be expected that not only the second-generation Finnish Australians born in Australia but also the Juveniles, having received their school education in Australia, generally shift from the ethnic language to Australian English during their teen years. This enables them to become fluent bilinguals and achieve a virtually native-like competence in English. Accordingly, the column 'shift' in Table 1 only predicts which level of language is affected by substratum transfer from Finnish in the *process* of language shift among the Adults. With the two groups of speakers available, we have no evidence in our data of the exact time when the shift was completed among the Juveniles.

Given the fact that there is no evidence yet of any extensive Finnish substratum transfer in the English of the Juvenile speakers, we are led to the conclusion that second-generation Finnish Australians represent a typical case of *shift without interference*[5] similar to that of "urban immigrant groups of European origin in the United States" (Thomason & Kaufman 1988: 120), who maintain their own ethnic languages for the first generation, while their children (and grandchildren) shift to the English of the community as a whole with hardly any interference from the original languages. The issue of shift without interference for those second-generation speakers who were born in Australia remains unresolved until the interview material is transcribed, digitised and analysed. It has been argued, however, that immigrant languages in the US last for a maximum of two generations (cf. Karttunen 1977, Veltman 1983, Smits 1996, Klintborg 1999). One aim of the present project is to investigate whether this pattern is evident in our Australian data (cf. Clyne and Kipp 2006). More generally, it appears that even members of the first generation of immigrants demonstrate a variety of achievements, including native-like ability, and that members of the second generation speak natively (Piller 2002), and that language attrition does not wait till the third generation but begins with the first generation (cf. Waas 1996, Schmid 2002, 2004, Cook 2003, Jarvis 2003). Clyne (2003: 48), for one, also suggests that a person's higher education level may cause language shift if it results in more contact with the cultural life of the dominant group, and this is the expected pattern of behaviour of the Juveniles.

4. Detecting aggregate syntactic distance

We measure the difference in the syntaxes of the two groups, the Adult and the Juvenile immigrants to Australia, on the basis of the corpus described in section 2. We use those parts of the interviews spoken by the immigrants, and we divide this into the parts spoken by those who immigrated as juveniles and those who immigrated as adults.

[5] We use the terms *interference* and *transfer* as if they were freely interchangeable. Clyne (2003: 76) suggests a distinction between *transference* ('process') and *transfer* ('product').

We apply the computational and statistical techniques described by Nerbonne & Wiersma (2006), which we summarise here for ease of reference. Readers interested in technical detail are urged to consult the study directly.

4.1 Assigning syntactic categories

Although automatic parsing is already producing fair results for the edited prose of newspapers, it is not a promising avenue for parsing the conversational transcripts of second language learners. Both the conversation style of the transcripts and the frequent errors of learners would be obstacles. We can, however, assign minimal syntactic categories (part-of-speech, or POS, information) to the words, using a so-called "tagger." For this we used the TnT tagger (Brants 2000), which is freely available.[6]

We tagged the corpora using the tagset of 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 81.2% of 1-grams, 67.5% of the 2-grams, and 56.1% of the 3-grams. The accuracy is poor compared to newspaper texts, but we are dealing with conversation, including the conversation of non-natives.

In order to allow sensitivity to context, we collected the POS tags into trigrams, i.e. sequences of POS tags as they occur in corpora. For a sentence such as (1) *the cat sat on the mat*, the tagger assigns the following POS labels:

(1)	*the*	*cat*	*sat*	*on*	*the*	*mat*	**.**
	ART(def)	N(com, sing)	V(intr, past)	PREP(ge)	ART(def)	N(com, sing)	PUNC

[6] Sanders (2007: 1), who compares the present approach to an approach in which the sentences of a hand-corrected corpus are represented as parse trees rather than a vector of POS tags, argues that "the trigram approximation works well, but it does not necessarily capture all the information of syntactic structure such as long-distance movement."

And for a sentence such as (2) *we'll have a roast leg of lamb for breakfast…*(extracted from our data), the tagger assigns the following POS labels:

(2)

we	*'ll*	*have*	*a*	*roast*
PRON(pers, plu)	AUX(modal, pres, encl)	V(montr, infin)	ART(indef)	N(com, sing)

leg	*of*	*lamb*	*tomorrow (…)*
N(com, sing)	PREP(ge)	N(com, sing)	ADV(ge)

These are then collected into the trigrams as follows: (a) ART(def)-N(com, sing)-V(intr, past), ..., ART(def)-N(com, sing)-PUNC(per), and (b) PRON(pers, plu)-AUX(modal, pres, encl)-V(montr, infin),…, ART(indef)-N(com, sing)- N(com, sing),…, PREP(ge)- N(com, sing)-ADV(ge) …

We use POS trigrams as indications of syntactic structure, proceeding from the consensus in syntactic theory that a great deal of hierarchical structure is predictable given the knowledge of lexical categories, in particular given the lexical 'head.' Sells (1987, sec. 2.2, 5.3, 4.1) shows how this assumption was common to theories in the 1980s (Government and Binding theory, Generalized Phrase Structure Grammar, and Lexical Function Grammar), and the situation has changed little in the successor theories (Minimalism and Head-Driven Phrase Structure Grammar). Even though the consensus of twenty years ago has been relaxed in recognition of the autonomy of "constructions" (Fillmore & Kay 1999), it is still the case that syntactic heads have a privileged status in determining a "projection" of syntactic structure.

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, investigating whether the distribution in the two rows is of a sort one might expect by chance, and, in case it is not, calculating which frequent POS trigrams are responsible for the skewed distribution. Nerbonne & Wiersma (2006) describe the use of permutation tests for this purpose, but we suppress the detailed explanation of permutation tests here. We do like to emphasise that sheer corpus size should not lead to inflated estimates of statistical significance (Agresti 1996), which we see as a virtue in our approach.

Once we have ascertained whether two corpora differ in their syntax, we naturally also wish to understand what is responsible for the differences. For this purpose we have developed software to identify which POS-trigrams contribute most heavily to the difference measured by the permutation test. We then examine the 200 statistically significant POS-trigrams that have the most unequal division across the groups (i.e. not based on their absolute frequency, but on the relative size of the difference in their frequency between the two groups). These are the trigrams with the biggest R^2 value, and they therefore represent the most typical ones for each group. We turn to an examination of the Finnish-Australian data below.

4.2 Discussion

By analysing 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; de Bot et al. 2005: A3, B3). According to these experts 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. In a similar vein, Thomason (2001: 148) argues that learners often ignore or fail to learn certain target language distinctions that are opaque to them at early to middle stages of their learning process. It is this kind of SLA behaviour that we expect particularly of the Adults.

Some previous work, such as Poplack & Sankoff (1984), introduced techniques for studying lexical borrowing and its phonological effects, and Poplack et al. (1988) went on to exploit these advances in order to investigate the social conditions in which contact effects flourish best. We follow Aarts & Granger (1998) most closely, who suggest focusing on tag sequences in learner corpora, just as we do. We add to their suggestion a means of measuring the aggregate difference between two varieties, and show how we can test whether that difference is statistically significant.

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 measure sensitive to all conceivable syntactic differences, only a measure that correlates with syntactic differences as a whole.

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 measuring the aggregate degree of syntactic difference between two varieties attempts to measure the "total impact" in Weinreich's sense, albeit with respect to a single linguistic level, syntax.

If this 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. A numerical measure of syntactic difference might enable these fields to 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.

5. Syntactic analysis of the two varieties of Finnish Australian English

The evidence from our syntactic analysis using the POS tagger (tag trigrams) and a permutation test described in detail above shows that there are statistically significant differences between the Adults and the Juveniles. Our report focuses first on the aggregate effects of syntactic distance between the two groups of speakers and then we move on to discussing the data on what we call 'syntactic contamination' in the English of the Adults.

5.1 Aggregate syntactic differences between Adults (A) and Juveniles (J)

Some of the statistically significant syntactic differences found in the data that demonstrate aggregate effects of syntactic distance can be described in general terms as follows:

1. *Overuse* of **hesitation phenomena** by A as opposed to J (pauses, filled pauses, repeats, false starts etc.), arising from difficulties in speech processing in general and lexical access in particular.

2. *Overuse* of **parataxis** (particularly with ***and*** and ***but***) by A as opposed to hypotaxis, not only at phrasal level but also at clausal level.

3. *Underuse* of **contracted forms** by A as opposed to J, who use, quite fluently and naturally, forms such as *I've been running, I'd like to go, I'll finish my degree, we're very market-oriented*, whereas the Adults, while using some highly frequent contractions such as *don't* and *can't,* (occasionally) *it's* and *I'm*, mostly use full forms such as *I have been, it has been.*

4. *Differential usage* of **discourse markers** such as *you know, you see, I mean* by A as opposed to J in the sense that the Adults mostly use *you know* (followed by hesitation phenomena) as a time-gaining device in order to access the next lexical item rather than as a genuine discourse marker. In contrast, the Juvenile speakers use a more varied repertoire of markers, which often function as appeals to the interviewer.

5. Avoidance of **complex verb clusters** by A as opposed J, who frequently use structures such as I would have had it, I still probably would have ended up getting married.

6. Avoidance of **prepositional and phrasal verbs** by A as opposed to J, who use verbs such as *I* ***ran out of*** *money, I just* ***opted out for*** *an operation.*

7. Underuse of the existential (expletive) ***there*** by A as opposed to J, who almost invariably use the default singular with the existential in examples such as *I /mean there's/ something there*, *you /know there's/ no ads*, *I /mean there's/ no major changes.* In contrast, the Adults either underuse the existential or omit the existential in subject position (cf. section 5.2.4). The use of the default singular with *there* can be considered an English-specific *angloversal* (cf. Szmrecsanyi & Kortmann 2006) rather than a vernacular universal that is common to spoken vernaculars in general.

It should be remembered that we only used the POS-trigrams that were found to be statistically significant in one group as opposed to the other group, and that the top 200 trigrams that we analysed were sorted by their weight with 20 random samples from each group. We also used those POS-trigrams (which we call 'extreme') in which 90% of the occurrences were within the group that was investigated. The notion of *avoidance* does not therefore imply total absence of a feature in either group. Ellis (1994: 304–306) argues that avoidance ('underrepresentation') and overuse ('over-indulgence') can also result from transfer. Learners avoid using non-transparent L2 constructions because of the differences between their mother tongue and L2. The overuse of certain grammatical forms may also occur as "a consequence of the avoidance or underproduction of some 'difficult' structure" (Ellis 1994: 305). Accordingly, the overuse of parataxis in (2) by the Adults, for example, may partly result from their inability or unwillingness to produce subordinate constructions (hypotaxis).

Because our present analyses do not show whether or not the speakers involved are necessarily aware of what kind of target forms they are avoiding, we conclude that all the features above that demonstrate differential usage by the Adults are projections, not from direct contact between English and Finnish, but rather from the strategies and processes typically evinced by second-language acquirers regardless of their mother tongue. These include not only overuse, as in (1) and (2), underuse ('underrepresentation'), as in (3) and (4), avoidance, as in (5) and (6), but also over-generalization, simplification, and false hypotheses, which all may result in deviant usage. Our findings of the avoidance of complex verb clusters (5) and prepositional or phrasal verbs (6) in the Adults are in agreement with de Bot et al. (2005), who suggest that constructions which are not immediately transparent are systematically avoided by L2 learners.[7] Even though Finnish has no prepositional or phrasal verbs, and L1 influence may therefore exert itself, we argue that all these strategies and processes are best considered as second-language learners' "universals," and that Finnish Australians, the Adults in particular, also show them in their interlanguage.

[7] Thomason (2001: 148) argues that "errors of omission—failure to learn certain TL features—are among the most common of shift-induced interference features."

On the basis of the analysis above, we contend that the statistical evidence obtained from our data indeed reflects syntactic distance between the two varieties of English and, consequently, aggregate effects of the difference in the two groups' English proficiency. We also argue that the juvenile shifters, having no demonstrable syntactic contamination in their speech, must have already shifted to English as their language of everyday communication. The evidence above indisputably shows a differential shift to English in the two groups of speakers, the Adults (still) showing features of 'learner' language, and therefore those of 'temporary' or 'partial' shift, whereas the Juveniles those of shift without interference, having English as their linguistically and socially dominant language.

5.2 Syntactic contamination in the English of the Adults

In the following sections, we will be looking for a potential answer to the question of whether the observed syntactic deviations from the norms of standard (acrolectal) English should be ascribed to contact effects from a Finnish substratum, to more universal, 'natural' tendencies in non-standard varieties in general, or to other factors. Potential roles of these factors will be assessed when we analyse the syntactic contamination in the English of the Adults in more detail in the sections below. We admit, however, that our discussion does not exhaust even the kinds of explanations that are customarily examined for apparent "contact" or "second-language learner" effects. For example, we do *not* examine hypotheses about Chomskyan universals and their applicability to second language learning or hypotheses about potential processing constraints in any detail.

Filppula et al. (2006), for example, suggest that so-called 'vernacular universals' (Chambers 2004) and contact-induced patterns form a *continuum*, having at one end universal features where the case for contact is weak; at the other end are features where the role of contact is obvious and indisputable. In other words, we are concerned with a continuum varying in the degree to which the hypothesis can be explained without risking an alternative. Some features that are found in the data seem to represent the dominance of a (vernacular) universal over contact influence, such as the default singular in (7) above and absence of the copula *be* (in section 5.2.3), while others may be explained in terms of both factors, and there may still be other features that are contact-induced.

Features that support the theory of vernacular universals (Chambers 2004: 128–129; cf. Filppula et al. 2006) are, in addition to default singulars (subject-verb nonconcord), *-n'* in the present participle, morpheme-final consonant cluster simplification (*pos'office, han'ful*), final obstruent devoicing (*hundret* for *hundred*), and conjugation regularisation or leveling of irregular verb forms (*Mary heared the good news*), multiple negation or negative concord (*He didn't see nothing*) and absence of the copula (*We going as soon as possible*). We realise, of course, that most of these features recur ubiquitously all over the world, even outside of vernacular, e.g. in "child language, pidgins, creoles, and *interlanguage* varieties" and that they "cannot be merely English" (Chambers 2004: 128f.: italics ours).

We note that the theory of vernacular universals may, in fact, explain some of the constructions that we find in the data elicited from some of the Adult informants. For example, two different Adult speakers use the past form *doed* ('did') of the verb *do*, as in *when they get that, they doed whatever they like* (*doed* 'levelling' of the irregular past tense form of the verb *do*). Three different Adult speakers also use the verb form *doned* ('done' + *ed*), as in *They all doned here, they, - they wasn't raw [kangaroo] skin.* The same example also shows absence of the copula *be* ('they *were* all done here') and subject-verb nonconcord in *they wasn't* ('they weren't'), which are all candidate vernacular universals. 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). Similarly to some vernacular Englishes, also non-acrolectal Finnish 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 exposed to a more or less codified standard (acrolectal) variety, rely on *visual* input as well. The fact that the Adults have mainly been exposed to spoken, *basilectal* (Australian) English is likely to give rise to some general vernacular features.

The data relevant to our present syntactic analysis will be discussed from the point of view of both types of explanation, contact-induced and "universal." The following sections deal in more detail with some of the statistically significant syntactic differences that we find recurrent in our data, i.e. features of disfluent speech, article usage, omission of the primary verb *be* (together with the alleged vernacular universal *–n'* in the present participle), omission of *there* and *it*, absence of prepositions, deviant word order, position of the negator *not*, and a few others that we find interesting. The data to be analysed in the sections below were elicited from the clusters of POS-trigrams which had a similar constituent structure and which were shown to have statistically significant differences between the two groups of informants.

5.2.1 Disfluent speech

As pointed out in section 5.1, the Adults demonstrate typical features of disfluent speech (hesitation phenomena), such as (filled) pauses, repeats, false starts, repairs, incomplete or false syntactic structures, arising from difficulties in speech processing, and particularly in lexical access. Some of the trigrams, marked by '/' and embedded in their respective contexts, are exemplified in (3) to (7):

(3) *skin cancer and /um and uh/ and gene general*

(4) *but /ah I I/never been on*

(5) *clubs spades /hearts and uh/ uh cl oh [probably trying to access 'diamonds']*

(6) *(he) was a leading-hand um /leading-hand and ah/ last last last*

(7) *as in /a in a/ Finland because especially*

Features of disfluent speech are overwhelmingly the most frequent characteristics that distinguish the Adults from the Juveniles. They can occur at any syntactic boundary but mostly before nouns. They are, of course, typical of any kind of speech, native and non-native alike, but certainly more frequent in interlanguage or, more generally, in second

language acquisition (shift) where speakers demonstrate imperfect learning as they study an L2. In this view, then, features of disfluent speech may be seen as "second-language learners' universals." Evidence from the English spoken by first-generation Finnish Americans also shows similar patterns, but different from standard (American) English, in the manner typical of 'learner language' (cf. Pietilä 1989: 152–189; Hirvonen 1988, 1995). Pietilä (1989: 221) found a very significant difference between her elderly and younger adults in the use of hesitation phenomena, and we argue that the excess (overuse) of these reflects the Adults' lesser proficiency in English (cf. Lauttamus 1999: 105). Difficulties in controlling pause duration and placement seem to be common among all second-language learners irrespective of the target language. Paananen-Porkka (2007) argues that pausing, including filled pauses such as those in (3), (4), (5), and (6), seems to be the main source for the anomalies that she found in her investigation of English speech rhythm by Finnish comprehensive school students.

5.2.2 Article usage

The Adults also demonstrate overuse (and underuse) of the two articles, *a(n)* and *the*, characteristic of a learner whose L1 has no article system (such as Finnish), as exemplified in (8) to (12):

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

(9) *first we go /to the Finland/*

(10) *we been /in a Brisbane/ Brisbane because ah*

(11) *in /the Brisbane and/*

(12) *I had /a different birds/ in Finland*

Example (12) shows that the indefinite article is sometimes used with a plural, countable noun head. The fact that the Juveniles do not show similar linguistic behaviour implies that they are not only more proficient in their

English but also in their later stages of acquisition (language shift) than the Adults.

Evidence from the English of Finnish Americans also suggests that the overuse of the articles, particularly that of the indefinite article, is rather common in conjunction with proper nouns. Pietilä's (1989: 167–168) data show that it is the redundant definite article, as in (9) and (11) rather than the indefinite one, as in (8) and (10) that is used by Finnish Americans, especially by the elderly first-generation speakers. 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.) Pietilä's (1989: 165) data also suggest that most of the article errors committed by the first-generation Finnish Americans are omissions, which Thomason (2001: 148) considers to be among the most common of shift-induced interference features.

In Finnish, some of the functions of the articles are expressed, for example, in terms of case assignment, e.g. *luin kirjaa* ('I was reading a book'; *kirja**a*** partitive sg), as opposed to *luin kirjan* ('I read the book'; *kirja**n*** genitive sg). A frequently used variant of the example *luin kirjan* reads as *luin **sen** kirjan* ('I read it+GEN book+GEN'). The use of *se(n)* makes the reference explicitly definite and specific. (Some linguists therefore argue that *se* represents an article that is developing in Finnish.) In this vein, the Adults also show overuse of the demonstrative pronouns *this* and *that* to mark definiteness instead of the definite article:

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

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

In the contexts where we found trigrams such as these there is no apparent need to use the demonstratives. We note, however, that in a potential Finnish variant of (13), *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. There may be an uncertainty among the Adults as to expressing the distinction between definiteness and specificity in English, so that they rely on the use of the demonstrative in specific reference. It seems, then,

that the overuse of the demonstratives by the Adults may originate from Finnish substratum transfer. This is supported by the fact that some Adults 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 summarise, it seems to us that the deviant usage of the articles in the English of the Adult Finnish Australians can be ascribed to substratum transfer from Finnish (which has no article system to express (in)definiteness and specificity), and that it may best be described as indirect functional (shift-induced) interference. Because Finnish does not have articles, the absence of this structural feature in the mother tongue can be interpreted as evidence against L1 transfer in L2 (cf. Arabski 1979). However, we agree with Ellis (1994: 306–315), who strongly argues that the absence of a feature in the first language may have as much L1 influence on the second language as the presence of a different feature. In addition to contact-induced effects, it appears that general *hypercorrection* (or overgeneralisation), 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." Overuse of the definite article is also common in some Celtic-influenced varieties of English and extra-territorial Englishes (cf. Filppula et al. 2006).

5.2.3 Omission of primary (copula) *be*

Omission also of the primary verb *be* in the progressive (present and past) is frequent in the Adults as opposed to the Juveniles, as exemplified in (15) to (17):

(15) *when we /drivin' in the/ road*

(16) *no I just /workin' for seven/ oh eleven*

(17) *fifteen years ago /we drivin' round/*

Absence of the copula *be* is one of the alleged vernacular universals (Chambers 2004). We agree that it can be better ascribed to more universal properties of the language faculty rather than to substratum transfer from

Finnish, even though Finnish has no similar formal contrast between the progressive and non-progressive aspect. Learners often leave out unstressed elements (such as the primary verb *be* above) because they do not simply perceive them and, consequently, cannot process them. This explanation is supported by the evidence from Fenyvesi & Zsigri (2006), who strongly argue for the crucial role of perception in unstressed syllable deletion in loanwords (cf. Kenstowicz 2001). Working in the framework of Optimality Theory, they suggest that in languages such as Finnish and Hungarian where primary stress always falls on the first syllable "phonetic content preceding the stressed syllable may not be interpreted as part of the prosodic word" (p. 137). We note that in Finnish the word boundary always precedes primary stress, and if the onset of a word (such as *be* above) is not marked by stress, it is difficult, if not impossible, for a native speaker of Finnish to interpret the prosodic word. The suggestion that less educated speakers of English such as the Adults rely on *auditory* input (Fenyvesi & Zsigri 2006: 143) might also explain why the omission of other generally unstressed closed-class items is so frequent among first-generation speakers.

The English progressive form, as in (15) to (17), is rather difficult for native speakers of Finnish. This is supported by the evidence in Pietilä (1989: 180–181), who 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. To express the Finnish progressive aspect, native speakers of Finnish mainly use the simple present. The alternatives include a construction with a finite form of the verb *olla* 'be' followed by the third infinitive of the main verb in the inessive case, e.g. *Olin jo nukkumassa kun soitit*. 'I was already sleeping when you called' (literally 'I was already in sleeping'). Even though the formation and the use of this construction somewhat resemble those of the English *be* + *-ing* construction, it is unlikely that there should be any positive substratum transfer from Finnish to aid the use of the English progressive. From a perceptual point of view, we are tempted to believe that the use of the present participle form *–ing* alone is salient ('marked') enough to enable Finnish speakers of English to interpret the desired function of the progressive and mark the formal contrast between the progressive and non-progressive (cf. Opas-Hänninen et al. 2005).

In (15) to (17), the Adults also demonstrate the vernacular (basilectal) substitution of the alveolar nasal [n] represented by *–n'* for the standard (acrolectal) velar nasal represented by *–ng*. This substitution, frequently attested in many vernacular varieties of English, can therefore be accounted for in terms of the theory of vernacular universals rather than in terms of transfer from Finnish. However, since the standard (acrolectal) English velar nasal in final position of the word violates the (standard and vernacular) phonological rule of Finnish that does not allow velar nasals in that position, the substitution in question can also be explained in terms of Finnish substratum transfer.

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

Omission of the existential (expletive) *there* and the anaphoric *it* in subject position is also frequently attested in the English of the Adults. Some cases are exemplified in (18) to (20):

(18) *and summer /time when Ø is/ a people*

(19) *but not often /wine if Ø is/ some visitors come*

(20) *I don't like that /meat because Ø is/ I think*

Based on the contexts that these examples occur in, the speaker in (18) is apparently aiming at 'when *there* is/are people'; in (19) 'if *there* is/are some visitors coming,' and in (20) 'because *it* is (meat).' The examples in (18) and (19) 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* in (18), and to the NP *some visitors* in (19), and, consequently, would not mark the subject in the position before the copula. In (20), the absence of *it* is harder to explain since a pronoun would also be expected in Finnish. It may be attributed to a more universal look-back mechanism of the language faculty (Chambers 2006), so that the 'meat' that the missing 'it' would refer to is simply not repeated because it is already explicit in the linear order of the utterance and registered in short-term memory.

5.2.5 Absence of prepositions

The Adults tend to leave out prepositions with motion verbs such as *move*, *go*, *come*, as exemplified in (21) to (23):

(21) *and they move me /other room where/*

(22) *must go /work and uh/*

(23) *when we /came Australia that/*

Similar examples can be found in the speech of first-generation Finnish Americans (Pietilä 1989: 172–173). In expressing spatial relations with motion verbs, many vernacular varieties of English tend to leave out prepositions (cf. e.g. Linn 1988). This suggests that omission of prepositions may be a more general tendency, particularly with motion verbs such as those above, which intrinsically describe movement towards a location. It is also possible to propose Finnish substratum transfer to explain the omission in (21) to (23), i.e. lack of prepositions in Finnish. This argument is not, however, as convincing as it was in the case of Jarvis & Odlin (2000), who show that a prepositional choice that is not found in Swedish speakers of English is used by Finnish speakers of English in an example such as *Chaplin and girl sat **to** grass*. This kind of usage of the preposition *to* corresponds to the Finnish allative case inflection on *nurmikolle* ('to grass').

5.2.6 Deviant word order

The Adults also demonstrate deviant word order, particularly with adverbials, which are often placed before the object, as exemplified in (24) to (27):

(24) *I /don't like really/ any old age*

(25) *I don't /watch any more/ that one*

(26) *they /don't have maybe/ enough money*

(27) *fifteen years ago /we drivin' round/*

The pre-object placement of the equivalent adverbials in Finnish would be quite acceptable, and therefore we argue that it is contact-induced (shift-induced). Learners simply project their L1 structure (native word order potential) onto L2 patterns in constructing their version of L2 grammar. Pietilä (1989: 187–188) reports that elderly first-generation Finnish American L2 users of English have rather few word order errors, and that the most frequent type of error is the incorrect placement of the adverbial. Similarly to Finnish Australians and Americans, Russian L2 users of English also commit L1-based word order errors most of which involve adverbial placement (Pavlenko & Jarvis 2002: 208).

In example (27), discussed earlier in section 5.2.3, the time adverbial *fifteen years ago* is placed in pre-subject position. 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 placing the focus on an adjunct. In English the example is well-formed, but our statistical analysis demonstrates that the Adults used this much more than the Juveniles, whose English is native or near-native. We conjecture that the Adults are overusing this construction because they neglect its pragmatic conditioning. This is in agreement with some other studies which show evidence for substratum transfer involving focus structures that are communicatively motivated (cf. Odlin 2006a: 28). Since trigrams may not be ideal data to account for some contextual effects such as focus placement, more evidence from our corpora is certainly needed to corroborate this.

More generally, Odlin (1990: 107) argues that "there is no universal constraint on the transfer of basic word order" and that speakers from different backgrounds tend to display a preference for different word order patterns that directly reflect their L1s. He also argues that there is relatively little evidence for basic word order transfer in the literature because of the relative lack of study on beginning learners. Our data on the Adults, who are less exposed to English and rely more on their L1, and in whom word order transfer from Finnish is therefore more likely, indeed demonstrate more transfer effects on their English than the Juveniles.

5.2.7 Negator *not* in pre-verbal position

The Adults produce utterances where the negator *not* is placed in pre-verbal position, as exemplified in (28) to (30):

(28) *but uh /we not cook/ that way (without the primary verb do)*

(29) *I'm diabetic I /not can eat/ them*

(30) *not can say /not can say/ I not want*

This can be ascribed to Finnish substratum transfer (shift-induced interference), because Finnish always has the negative item (*ei*, inflected in person and number like any verb in standard Finnish) in pre-verbal position (as in 29 and 30, before the modal *can*). Although a substrate explanation may seem self-evident in (29) and (30), examples similar to (28) can be found in other non-standard varieties of English, e.g. in Spanish interlanguage English: *I no understand*, probably modeled on standard Spanish *(Yo) no entiendo* (Odlin 1989: 104–110). Examples (28) to (30) represent a developmental sequence described by Larsen-Freeman & Long (1991: 94) as 'internal pre-verbally negated strings,' which are common not only in Finnish but in learners of English from typologically different L1 backgrounds and therefore provide powerful evidence for language universals guiding, at least in part, interlanguage development. As Ellis (1994: 99-101, 421–422) points out, "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). In agreement with Larsen-Freeman & Long (1991: 106–107), we argue that L1 transfer occurs in parallel with general developmental processes, and it may strengthen the use of "a developmental form similar to an L1 structure" (such as a Finnish pre-verbally negated string).

5.2.8 Misuse of *what* as a relative

The misuse of the pronoun *what* as a relative pronoun or complementiser (31) by the Adults in (31) to (34) can be ascribed to substratum transfer from Finnish, where *mikä, mitä* sg 'what' or *mitkä* pl 'what' are used as

interrogative pronouns or relative pronouns, but only in some restricted contexts as a relative, e.g. *kaikki, mitä tiedän* 'all (that) I know.'

(31) *(it) was about twenty-five /minutes what they/ kept that balloon (that)*

(32) *(name) for other /games what we/ played (that or which)*

(33) *cars and all /machines what they/ built*

(34) *those Aussie /dishes what they/ eating*

We have not found a similar use of *what* as a relative in the 'Juveniles.' We argue that this usage of *what* as a relative is overgeneralised on the model of Finnish, even though we know that it is found in some substandard English.

5.2.9 Overuse of simple present

The Adults also extend the simple present (as opposed to the past tense and the progressive) to describe not only present but also past or future events, as exemplified in (35) to (38):

(35) *okay /we stay here/ we not go ('we'll stay here, we're not going')*

(36) *but /we wait till/ the ambulance come ('we waited/were waiting till the ambulance came')*

(37) *where /we live before/ this place ('where we lived before this place')*

(38) *when /we come in/ Australia ('when we came to Australia')*

This can be partly ascribed to Finnish substratum transfer, as Finnish has no equivalent progressive forms or ways to express future events in its repertoire and mainly uses the simple present in these functions (cf. section 5.2.3). Pietilä (1989: 176) also reports on the frequent use of the simple present to express past actions in the English of first-generation Finnish Americans. The fact that the present tense is used by the Adults in

reference to past events in (36, 37, 38; the use of the forms confirmed in their respective contexts) may also be explained in terms of a more universal tendency found in vernaculars to regularise morphology ('one form for all functions'). Some of the verb forms in the trigrams in (35), (36) and (38) are good examples of how our technique, which is sensitive to frequency differences, detects deviations that are not, strictly speaking, errors.

5.2.10 Terms of measurement

Both groups (A and J) almost invariably use the plural in the marking of nouns of measurement: *five miles; (a) hundred dollars; two hours* (noun as head of NP). We only found one example of the use of the singular: *three foot wide standing up*, which is, of course, also acceptable in standard (acrolectal) English. The use of the plural is also the default with plural quantifiers: *a couple of weeks, a few plays.*

The use of the plural as opposed to the singular is rather surprising because in many vernacular (non-standard) varieties of English the singular head is widely used because the singular apparently carries less cognitive cost; a notionally plural numeral already marks a phrase plural. The fact that we found little evidence of the singular is even more surprising because Finnish uses the singular in parallel cases: *viisi tuntia* 'five hours' (*tunti**a*** partitive sg), *kaksi viikkoa* 'two weeks' (*viikko**a*** partitive sg). Our findings may partly be explained by the fact that native speakers of Finnish are used to assigning inflection to the head of an NP (the partitive sg case [*viikko+a*], instead of the nominative sg [*viikko*]) for measurement and that the English *-(e)s* plural is apparently more salient phonetically (perceptually) than many other morphological endings.

5.2.11 Acquired formulae

There is also statistically significant evidence that the Adults have acquired some formulae such as *that's* and *what's*, as exemplified in (39) to (43):

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

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

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

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

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

That's and *what's*, acquired as fixed phrases, have apparently been processed as single elements. 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.

6. Discussion: contact-induced or universal?

In our syntactic analysis of the English of the Adult speakers of Finnish emigrants to Australia we have shown that a number of features that we describe as 'contaminating' the interlanguage can be attributed to Finnish substratum transfer. These features include (1) overuse (and underuse) of articles, (2) omission of the expletive *there*, (3) absence of a preposition such as *to* with motion verbs, (3) deviant word order with adverbials, (4) use of the negator *not* in pre-verbal position, (5) misuse of *what* as a relative or complementiser, and (6) overuse of the simple present to describe past and future events. What makes our argument for the role of contact somewhat less convincing is that almost all of these 'deviant' features might also be ascribed to more "universal" properties of the language faculty. However, our findings support other empirical evidence (reviewed, for example, by Larsen-Freeman & Long 1991: 96–113, and Ellis 1994: 299–345) that shows how the learner's L1 influences the course of L2 development at all levels of language, although transfer seems to be more conspicuous in phonology, lexis and discourse than in morphosyntax.

Nonetheless, there are other features in our data that may be ascribed to more "universal" primitives, such as the absence of the copula *be*, substitution of the *–n'* for *–ng* in the present participle, regularisation of morphology (use of the present tense form to describe past and future

events), and use of the default singular with *there* (by the 'Juveniles'), or to "language learners' universals," such as the overuse of hesitation phenomena, overuse of parataxis, underuse of contracted forms, and avoidance of complex verb clusters, prepositional and phrasal verbs. Since we have no evidence of potential contamination in the English of the Juvenile speakers at the early stages of their L2 acquisition, we are simply not in a position yet to prove or refute our hypothesis about the strength of contact influence as opposed to that of the other factors. The "high shift" to English (Clyne & Kipp 2006: 18) by the Juvenile speakers, without interference from Finnish, seems to support our argument that most of the features that we found in the data elicited from the Adult speakers appear to be temporary, even ephemeral, and have no permanent impact on the English of second-generation Finnish Australians. This is in agreement with the idea proposed by Larsen-Freeman & Long (1991: 107) that "beginners" (such as the Adults), who rely more on their L1 because of the limitations imposed on them by their L2, are initially more willing to transfer items from their native language.

7. Conclusion

In this paper we argue that using frequency profiles of trigrams of POS categories as indicators of syntactic distance between two different groups of speakers we can now give an estimate of the "total impact" of L1 on L2 syntax in SLA. Our findings show syntactic 'contamination' from Finnish in the English of the Adult first-generation speakers of Finnish ethnic origin. 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. In contrast with the Adult speakers of the first generation, the Juvenile speakers of the second generation, who acquired English much earlier, demonstrate a native or near-native command of English. This is in accordance with Riehl (2006), and many others, who point out that age (of onset) is a crucial determinant of successful L2 acquisition.

Our syntactic analyses of the two varieties of Finnish Australian English therefore strongly support the idea that language *shift* to English has already taken place among the majority of the Juvenile speakers as opposed to the Adult speakers, who still demonstrate typical morpho-

syntactic features of *temporary shift* and *imperfect learning* of English. We conjecture that the higher education level of the Juveniles in the language and culture of the dominant group may have accelerated language shift even further (cf. Clyne 2003: 48). Although we have a third corpus of the English of second generation speakers of Finnish ethnic origin born in Australia, our findings of the variety of English spoken by the Juvenile informants suggest that the corpus in question would not provide us with any deeper understanding of the general language development of the Finnish immigrant groups in Australia. We infer from the data of the Juveniles that potential residues of syntactic 'contamination' in the English of second-generation speakers can only be found by observing them systematically from the very inception of their L2 acquisition process up until they go to school, and that without any longitudinal study of their English we are unable to capture or intercept potential shift-induced transfer effects in progress.

References

Aarts, Jan & Granger, Sylviane (1998) Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In S. Granger (ed.), *Learner English on Computer*, pp. 132–141. London: Longman.

Agresti, Alan (1996) *An Introduction to Categorical Data Analysis.* New York: John Wiley.

Arabski, Janusz (1979) *Errors as Indicators of the Development of Interlanguage.* Katowice: Universytet Slaski.

Bot de, Kees, Lowie, Wander & Verspoor, Marjolijn (2005) *Second Language Acquisition: An Advanced Resource Book.* London: Routledge.

Brants, Thorsten (2000) TnT – a statistical part of speech tagger. In *6th Applied Natural Language Processing Conference*, pp. 224–231. Seattle: ACL.

Brill, Eric & Mooney, Raymond (1997) An overview of empirical Natural Language Processing. *AI Magazine* 18: 13–24.

Chambers, Jack (2003) *Sociolinguistic Theory: Linguistic Variation and its Social Implications.* Oxford: Blackwell.

—— (2004) Dynamic typology and vernacular universals. In B. Kortmann (ed.), *Dialectology meets Typology*, pp. 127–145. Berlin: Mouton.

—— (2006) Linguistic continuum from vernacular to standard. In *World Englishes.*

Clyne, Michael (2003) *Dynamics of Language Contact: English and Immigrant Languages.* Cambridge: Cambridge University Press.

Clyne, Michael & Kipp, Sandra (2006) Australia's community languages. *International J. Soc. Lang.* 180: 7–21.

Cook, Vivian (ed.) (2003) *Effects of the Second Language on the First.* Clevedon, UK: Multilingual Matters.

Coseriu, Eugeniu (1970) *Probleme der kontrastiven Grammatik.* Düsseldorf: Schwann.

Ellis, Rod (1994) *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

Fenyvesi, Anna & Zsigri, Gyula (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, Charles & Kay, Paul (1999) Grammatical construction and linguistic generalizations: the *what's x doing y* construction. *Language* 75: 1–33.

Filppula, Markku, Klemola, Juhani & Paulasto, Heli (2006) Digging for roots: universals and contacts in regional varieties of English. In *World Englishes.*

Færch, Claus & Kasper, Gabriele (1983) *Strategies in interlanguage communication*. London: Longman.

Garside, Roger, Leech, Geoffrey & McEmery, Tony (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman.

Hirvonen, Pekka (1988) American Finns as language learners. In M.G. Karni, O. Koivukangas and E.W. Laine (eds.), *Finns in North America.* Proceedings of Finn Forum III 5–8 September 1984, Turku, Finland. Migration Studies C9, pp. 335–354. Turku: Institute of Migration.

—— (1995) Phonological and morphological aspects of Finnish language attrition in the United States. In W. Viereck (ed.), *ZDL–Beiheft 77: Verhandlungen des Internationalen Dialektologenkongresses Bamberg 1990*, pp. 181–193. Band 4. Stuttgart: Franz Steiner.

Hunnicutt, Sheri & Carlberger, Johan (2001) Improving word prediction using Markov models and heuristic methods. *Augmentative & Alternative Communication* 17: 255–264.

Jarvis, Scott (2003) Probing the effects of the L2 on the L1: A case study. In V. Cook (ed.), *Effects of the Second Language on the First*, pp. 81–102. Clevedon, UK: Multilingual Matters.

Jarvis, Scott & Odlin, Terence (2000) Morphological type, spatial reference, and language transfer. *Studies in Second Language Acquisition* 22: 535–556.

Karttunen, Frances (1977) Finnish in America: A case study in monogenerational language change. In B. G. Blount & M. Sanches (eds.), *Sociocultural Dimensions of Language Change*, pp. 173–184. New York: Academic Press.

Kenstowicz, Michael (2001) The role of perception in loanword phonology. *Linguistique africaine* 20 [published in *Studies in African Linguistics* 32 (2003): 95–112].

Klintborg, Staffan (1999) *The transience of American Swedish.* Lund: Lund University Press.

Kovács, Magdolna (2001a) *Code–Switching and Language Shift in Australian Finnish in Comparison with Australian Hungarian.* Åbo: Åbo Akademi University Press.

—— (2001b) "Ja putsari toi lihan, pääkäri toi leivän." Australiansuomen leksikaaliset lainat ja kielen säilyttäminen (The effect of lexical borrowing on language maintenance in Australian Finnish). *Virittäjä* 3: 384–410.

Larsen-Freeman, Diane & Long, Michael H. (1991) *An Introduction to Second Language Acquisition Research.* London: Longman.

Lauttamus, Timo (1999) Fuzzy switch and loan types in the languages of Finnish Americans. *SKY Journal of Linguistics* 12: 87–109.

Lauttamus, Timo & Hirvonen, Pekka (1995) English interference in the lexis of American Finnish. *The New Courant* 3: 55–65. Department of English, University of Helsinki: Helsinki University Press.

Lenneberg, Eric (1967) *Biological Foundations of Language.* New York: John Wiley.

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

Long, Michael H. (2005) Problems with Supposed Counter–Evidence to the Critical Period Hypothesis. *International Review of Applied Linguistics in Language Teaching* 43, 287–317.

Moyer, Alene (2004) *Age, Accent, and Experience in Second Language Acquisition: An Integrated Approach to Critical Period Inquiry.* Clevedon: Multilingual Matters.

Nerbonne, John & Wiersma, Wybo (2006) A measure of aggregate syntactic distance. In J. Nerbonne & E. Hinrichs (eds.), *Linguistic Distances*, pp. 82–90. Shroudsburg, PA: ACL.

Odlin, Terence (1989) *Language Transfer: Cross–linguistic Influence in Language Learning.* Cambridge: Cambridge University Press.

—— (1990). Word–order transfer, metalinguistic awareness and constraints on foreign language learning. In B. VanPatten & J. Lee (eds.), *Second Language Acquisition – Foreign Language Learning*, pp. 95–117. Clevedon, Avon: Multilingual Matters.

—— (1992) Transferability and linguistic substrates. *Second Language Research* 8: 171–202.

—— (2006a) Could a contrastive analysis ever be complete? In J. Arabski (ed.), *Cross–linguistic Influence in the Second Language Lexicon*, pp. 22–35. Clevedon, U.K.: Multilingual Matters.

—— (2006b) Methods and inferences in the study of substrate influence. In *World Englishes.*

Opas–Hänninen, Lisa Lena, Hirvonen, Pekka, Juuso, Ilkka, & Lauttamus, Timo (2005) "Happen I not talking good English": The progressive aspect in the English of Finnish Australians. *Methods XII: Twelfth International Conference on Methods in Dialectology.* August 1 – 5, 2005. Workshop *Finno–Ugric Language Contacts (II)*. Université de Moncton, New Brunswick, Canada.

Paananen–Porkka, Minna (2007) *Speech Rhythm in an Interlanguage Perspective: Finnish Adolescents Speaking English.* Pragmatics, Ideology and Contacts Monographs. Doctoral diss. University of Helsinki.

Pavlenko, Aneta & Jarvis, Scott (2002) Bidirectional Transfer. *Applied Linguistics* 23: 190–214.

Pietilä, Päivi (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, Ingrid (2002) Passing for a native speaker: Identity and success in second language learning. *Journal of Sociolinguistics* 6: 179–206.

Poplack, Shana & Sankoff, David (1984) Borrowing: the synchrony of integration. *Linguistics* 22: 99–135.

Poplack, Shana, Sankoff, David & Miller, Christopher (1988) The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics* 26: 47–104.

Riehl, Claudia (2006) The bilingual brain. Online. [www.languages.unimelb.edu.au].

Sanders, Nathan C. (2007) Measuring Syntactic Difference in British English. *Proceedings of the ACL 2007 Student Research Workshop*, pp. 1–6. Prague, June 2007. Shroudsburg, PA.: ACL.

Schmid, Monika (2002) *First Language Attrition, Use and Maintenance: The case of German Jews in Anglophone Countries.* Amsterdam: John Benjamins.

—— (2004) Identity and first language attrition: A historical approach. *Estudios de Siciolingüística* 5: 41–58.

Sells, Peter (1982) *Lectures on Contemporary Syntactic Theories.* Stanford: CSLI.

Singleton, David & Ryan, Lisa (2004) *Language Acquisition: The Age Factor* (2nd ed.). Clevedon: Multilingual Matters.

Smits, Caroline (1996) *Disintegration of Inflection: The Case of Iowa Dutch.* The Hague: Holland Academic Graphics.

Szmrecsanyi, Benedikt & Kortmann, Bernd (2006) Vernacular universals and angloversals in a typological perspective. In *World Englishes.*

Thomason, Sarah G. (2001) *Language Contact: An Introduction.* Edinburgh: Edinburgh University Press.

Thomason, Sarah G. & Kaufman, Thomas (1988) *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.

Van Coetsem, Frans (1988) *Loan Phonology and the Two Transfer Types in Language Contact.* Dordrecht: Foris.

—— (1995) Outlining a model of the transmission phenomenon in language contact. *Leuvense Bijdragen* 84: 63–85.

Veltman, Calvin (1983) *Language Shift in the United States.* Berlin: Mouton.

Väyrynen, Pertti (2005) *Perspectives on the Utility of Linguistic Knowledge in English Word Prediction.* Oulu: Oulu University Press [also available at: http://herkules.oulu.fi/isbn951427850X/isbn951427850X.pdf].

Waas, Margit (1996) *Language Attrition Downunder: German Speakers in Australia.* Frankfurt: Peter Lang.

Watson, Greg (1995) A corpus of Finnish–Australian English: A preliminary report. In P. Muikku–Werner & K. Julkunen (eds.), *Kielten väliset kontaktit.* AFinLan

vuosikirja 1995. Suomen soveltavan kielitieteen yhdistyksen (AFinLa) julkaisuja 53, pp. 227–246.

—— (1996) The Finnish–Australian English corpus. *ICAME Journal: Computers in English Linguistics* 20: 41–70.

Weinreich, Uriel [1953] (1974) *Languages in Contact*. The Hague: Mouton.

World Englishes: Vernacular Universals vs. Contact–Induced Change. Abstracts and handouts. An International Symposium 1–3 September, 2006. University of Joensuu, Finland.

Contact information:

Timo Lauttamus
Faculty of Humanities
P.O.Box 1000
90014 University of Oulu
Finland
e-mail: timo(dot)lauttamus(at)oulu(dot)fi

John Nerbonne
Information Science
Rijksuniversiteit Groningen
P.O.Box 716
NL 9700 AS Groningen
The Netherlands
e-mail: j(dot)nerbonne(at)rug(dot)nl

Wybo Wiersma
Information Science
Rijksuniversiteit Groningen
P.O.Box 716
NL 9700 AS Groningen
The Netherlands
e-mail: wybo(at)logilogi(dot)org