Running-head: Lexical first language attrition

Title: Lexical access and lexical diversity in first language attrition*

Monika S Schmid[1] and Scott Jarvis[2]
[1]University of Essex/University of Groningen, [2]Ohio University

* Acknowledgement: The authors are particularly grateful for the kind help of Franck Bodmer Mory of the *Institut für deutsche Sprache* at Mannheim for assessing the frequency of the lemmas from our study against the COSMAS II Corpus.

Address for correspondence:
Monika S. Schmid
Department of Language and Linguistics
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
United Kingdom
mschmid@essex.ac.uk

Keywords: Language attrition, Lexical attrition, Mental lexicon, Methodology

Abstract

This paper presents an investigation of lexical first language (L1) attrition, assessing in what way a decrease in lexical accessibility manifests itself for long-term residents in a second language (L2) environment. We question the measures typically used in attrition studies (formal tasks and type-token ratios) and argue for an in-depth analysis of free spoken data, including factors such as lexical frequency and distributional measures. The study is based on controlled, elicited and free data from two populations of attriters of L1 German (L2 Dutch and English) and a control population (n=53 in each group). Group comparisons and a Discriminant Analysis show that lexical diversity, sophistication and the distribution of items across the text in free speech are better predictors of group membership than formal tasks or elicited narratives. Extralinguistic factors, such as the frequency of exposure and use or the length or residence, have no predictive power for our results.

(150 words)

When (monolingual) native speakers leave their country of origin, take up life in a different linguistic environment, become bilingual and consequently have less input in and make less use of their first language (L1), they often experience this language changing. This change can manifest itself in lexical access difficulties, disfluency phenomena, cross-linguistic interference, increased optionality in grammatical features and a foreign accent (among other things). The linguistic development experienced in such situations is commonly referred to as L1 attrition. It has often been described as a selective process (e.g. Sorace, 2005; Tsimpli, 2007) which affects different components of linguistic knowledge in a different order, at a different rate and to a different extent (Paradis, 2007), and which is sensitive to external factors such as age of emigration, length of residence and amount of L1/L2 exposure.

It has become almost axiomatic in language attrition research to assume that lexical-semantic knowledge is the most vulnerable part of the linguistic repertoire, deteriorating first, fastest and most dramatically as compared to, for example, grammar or phonetics (e.g. Hulsen, 2000; Köpke & Nespoulous, 2001; Köpke & Schmid, 2004; Montrul, 2008; Opitz, 2011, to name but a few). It is also commonly believed that lexical deterioration will be more or less linearly related to frequency of L1 use (e.g. Paradis, 2007). However, the empirical record regarding L1 lexical attrition, its relationship with other areas of attrition, and the factors that drive it, is quite weak. We argue that this is due to a paucity of studies that have approached L1 lexical attrition with a sufficiently stringent methodological design and sufficiently solid theoretical underpinnings. In particular, there are few studies comparing data elicited by means of multiple tasks or contrasting different theoretical predictions and assessing the impact of environmental variables, such as language use.

The present study focuses on lexical accessibility and explores the ways in which the twofold phenomenon of less exposure to the L1 and more exposure to an L2, may result in reduced L1 lexical accessibility. We also enquire whether the detectability of L1 lexical

reduction varies across different tasks. In order to obtain better insights into this phenomenon, the present study combines and compares different tasks and different measures and assesses their explanatory potential.

Language attrition in the lexicon

The term 'language attrition'[1] refers to changes in a native language that has either fallen into disuse or is used alongside an environmental one. In accordance with this definition, attrition is a process that is driven by two factors: (a) the presence, development and regular use of a second linguistic system, leading to crosslinguistic interference (CLI), competition and other effects associated with bilingualism, and (b) a decreased use of the attriting language, potentially leading to access problems (Schmid & Köpke, 2007).

A number of quantitative investigations have attempted to probe to what extent *lexical access* is compromised in L1 attrition due to long-term underuse, that is, whether attriters take longer to retrieve items from the mental lexicon or have a more restricted set of items at their disposal than unattrited monolinguals. The earliest such studies mainly relied on verbal fluency (VFT, e.g. Waas, 1996; Yağmur, 1997) or picture naming tasks (PNT, e.g. Ammerlaan, 1996; Hulsen, 2000). More recently, it has been argued that such experimental approaches should be combined with analyses of lexical access and lexical diversity in (relatively) unguided free speech (Schmid, 2004) as well as with investigations of disfluency phenomena (Schmid & Beers Fägersten, 2010). Such comparisons not only allow an assessment of lexical diversity across the full range of a speaker's repertoire (while controlled tasks are necessarily limited by the stimuli used), they may also show whether attrition effects are less (or more) pronounced in controlled tasks that allow the participant to fully focus on lexical retrieval as opposed to naturalistic language production situations where information

[1] In keeping with common practice, we reserve the term 'language attrition' for the attrition of a native language, while the attrition of later learned ones is referred to as second or foreign language attrition.

from all linguistic levels needs to be rapidly integrated in real time. If it could be shown that attrition effects are more pronounced in the latter type of data, this might indicate that attrition itself is largely the outcome of the increased cognitive load involved in managing two linguistic systems at the same time, as opposed to lexical access problems due to an increase in activation thresholds.

The role of L1 use

The assumption that (lexical) attrition is closely related to frequency of L1 use bears close resemblance to Michel Paradis' Activation Threshold Hypothesis (Paradis 1993; 2004; 2007; henceforth: ATH), This hypothesis takes as its point of departure the fact that accessing items stored in memory requires a certain amount of neuronal excitation. The level of energy necessary to retrieve any given item is determined by the frequency and recency with which it has previously been called upon, so that less frequent items and items that have not been used for a long time become harder to access. Attrition is therefore hypothesized to predominantly affect less frequent linguistic/lexical items, and to be more pronounced for speakers who do not use their L1 on a regular basis (Andersen, 1982, Paradis, 2007).

Furthermore, the activation threshold is not determined solely by activation but also by inhibition. Every time a speaker selects a target item from the lexicon, its competitors (semantic/phonological neighbours, cognates, translation equivalents etc.) have to be inhibited. This inhibition process raises the activation threshold, making the inhibited items harder to access subsequently (Green, 1986; 1998). The phenomena we can witness in the attritional process are thus not only dependent on the underuse of the attriting system (raising of the AT due to non-activation), but also on the presence, use and development of the environmental language (raising of the AT due to inhibition). Bilinguals routinely have to inhibit the language that is not chosen for activation (e.g. Bialystok, 2005), and this may then

lead to phenomena such as word finding difficulties etc. It is therefore important to take into consideration not only the amount of use of both languages, but also the contexts of use (formal/informal, monolingual/bilingual interlocutor) which affect the degree to which the other language is inhibited in these situations (Schmid, 2007).

In this context it is interesting to note that there is only one language use factor which has emerged as a significant predictor for attrition across a range of skills and linguistic levels in a number of studies (e.g. de Leeuw, Schmid & Mennen, 2010; Schmid, 2007; Schmid & Dusseldorp, 2010), namely the use of the L1 for professional purposes: The more the migrants use the L1 at work, the less L1 attrition they exhibit. The differential impact of informal vs. formal (i.e., professional) L1 use can be explained in terms of inhibition: in the daily life of a migrant, most interlocutors with whom the L1 is spoken informally (friends, family members) are also bilingual, so that code-switches and code-mixing do not need to be inhibited. In a professional context, however, it is usually not considered appropriate to mix the two languages, so that speakers who use their L1 at work will probably have more practice at inhibiting the L2 and resisting any ‘intrusions’ (lexical or otherwise) from this language system. This increased practice at inhibiting the environmental linguistic system may explain why such speakers are able to perform better in their L1 than speakers who do not have occasion to use their L1 in these types of formal situations (Schmid, 2007).

A further intriguing question in the context of the ATH is to what extent lexical attrition phenomena differ between contact languages of varying typological distance. If attrition is predominantly determined by frequency/recency of activation, speakers of two languages that share a large part of their lexicon should have an advantage over those migrants who are bilingual in typologically more distant languages, since the activation of cognates can be assumed to spread across languages (Berthele, 2011; Dijkstra, 2005; Jarvis, 2009). If, on the other hand, inhibition is the driving force of lexical access difficulties, speakers of more

distant languages can be expected to encounter fewer difficulties, as there will be less competition between translation equivalents which share no surface similarity. This implies that attrition studies would benefit from making comparisons of populations with either different L1s or different L2s instead of focusing on a single language combination (as has almost invariably been the case so far).

Measuring lexical diversity

It was pointed out above that it is desirable for investigations of lexical attrition to combine a range of tasks probing lexical accessibility. Controlled tasks, for example measuring a person's response times in naming a given picture stimulus, allow the participant to focus all attention on the lexical retrieval process. In free speech, on the other hand, language processing takes place across many linguistic levels simultaneously, and there may be trade-off effects, for example between morphosyntactic complexity/accuracy and lexical diversity, or between the use of less frequent linguistic items and fluency. This suggests that attriters, who may be experiencing difficulties with other parts of the linguistic repertoire, might show larger differences from controls in free speech than in controlled tasks.

Investigations of (elicited) free speech have to tackle the problem of how to measure lexical diversity in such data. This question has received considerable attention recently (for a comprehensive overview, see Jarvis, 2012, 2013a, 2013b). Traditionally, it has been addressed with type-token ratio (TTR) based measures, which relate the total number of words in a text to the total number of different lemmas (e.g. Jarvis, 2002; Schmid, 2011). Simple TTRs have been shown to be a problematic measure of lexical diversity in free speech in that they vary as a function of text length, since the rate of word repetition inevitably increases as the text grows longer. McCarthy and Jarvis (2007) demonstrate that this problem is persistent even for measures of lexical diversity which have been devised to overcome the

impact of length such as Guiraud's index, Yule's K, and VOCD (see McCarthy and Jarvis 2007; Jarvis and Daller 2013). While the effects of length on probability-based measures such as Yule's K and VOCD are relatively subtle, the principles of probability render them sensitive to a second factor: evenness, which is a matter of how evenly the tokens in a sample are distributed across types. Jarvis (2012, 2013a, 2013b) argues that length and evenness are inherent properties of diversity. This consideration is particularly relevant for situations such as language attrition: where lexical access is compromised, activation of specific items might be contextually determined and spread across a narrower range of the semantic field, leading to a more uneven distribution of the tokens across the entire text than would be the case for monolinguals.

The first problem (text length or volume) can be solved through the use of a measure of textual lexical diversity (MTLD), which has been applied in recent studies to avoid the effects of text length altogether (McCarthy & Jarvis 2010, 2013). It is calculated as the average number of running words in a text that remain above a certain type-token ratio (usually .72). Evenness can be assessed through measures used in the field of ecology in studies dealing with biodiversity. Although numerous indices of evenness exist (Smith & Wilson, 1996), the one we have adopted for the present study is the one based on Shannon's entropy index (Shannon, 1948), which has been described by Chao and Jost (in press) as the most appropriate general-purpose measure of diversity because it avoids giving too much weight to either rare or abundant species. Evenness is calculated as a ratio of the observed diversity of a sample (i.e., Shannon's index) divided by the maximum possible diversity of the sample that would occur if all types (or species) in the data were equally abundant (Pielou 1969).

So far, no measures that attempt to correct the problems of volume and evenness have been applied in language attrition studies, and lexical diversity in free speech has invariably

been described in terms of measures such as TTR, VOCD or Guiraud. These are also the measures included in the test battery proposed by Schmid (2004, 2011, see also www.let.rug.nl/languageattrition), which combines controlled and free speech tasks and has recently been applied in a number of investigations of the attrition of a variety of languages.

As can be seen in Table 1, the results from these studies appear to be somewhat inconsistent: with one exception, all investigations did find a statistically significant increase in (at least some) disfluency markers (e.g. empty pauses, filled pauses, repetitions). Lexical diversity (VOCD) in free speech and performance on the verbal fluency task, on the other hand, did not differ consistently across populations and studies.

/ insert Table 1 here /

A closer look at Table 1 reveals a relationship between findings and sample size for these measures: In investigations with relatively small population sizes (25 or lower), statistical significance is not reached (and, with the exception of Dostert, whose attriters outperform the controls on a number of measures, all authors note that the descriptive statistics do show better performance of the controls on most tasks). The lack of significant findings in the investigations by Cherciov, Opitz and Varga may therefore be a Type II error due to the limited sample size. All of the larger studies show consistent differences between attriters and controls in free speech and on the VFT. The PNT used by Yılmaz and Schmid (2013), on the other hand, does not reveal a significant difference between attriters and controls despite the comparatively large sample of 54 speakers in each population. This suggests that there may be population-specific constraints on lexical attrition, since the speakers investigated in this particular study (Turks in the Netherlands) are drawn from the numerically largest group of non-Western immigrants in the host country and may therefore have a larger L1 network

providing support for language maintenance than is the case in the other investigations. Similarly, the fact that Dostert did not find any attrition of L1 English may indicate that the global importance and presence of this language exerts a protective effect even in the absence of dense personal networks where the L1 is spoken.

Summary and research questions

Although lexical attrition has long been assumed to be one of the earliest and most noticeable symptoms of attrition at large, relatively few studies have attempted so far to gain a comprehensive view of what exactly this process entails. In particular with respect to the predictions concerning lexical reduction made by Andersen (1982) and echoed in the ATH, there is to date little solid evidence of whether and to what extent these hold true. To the extent that these predictions have been tested, this has generally been done exclusively through the use of controlled tasks. However, controlled tasks are often limited in their potential to probe the reduction of lexical accessibility, as they rely on certain characteristics of the stimuli (e.g. imageability in picture naming, limitations to a certain lexical field in semantic verbal fluency tasks). Previous studies that have examined lexical attrition in free speech have also suffered a second shortcoming in that they have so far relied on measures of type-token relationships, which do not allow insights into lexical frequency/sophistication or into characteristics of the distribution of items across the text (evenness).

A second issue that is of theoretical interest is the question of the role of activation vs. inhibition in lexical attrition: if attriters do have lexical access problems, are these linked to the fact that L1 items are more difficult to access (because they are used infrequently) or to problems involving the inhibition of closely related L2 items? This question may to some extent be resolved through a comparison of two groups of attriters who speak the same L1 but a different L2.

The present study address both of these issues and also explores the impact of personal background and language use factors on lexical attrition. It is guided by the following research questions:

RQ1: Lexical access: Is lexical access in the L1 compromised among long-term migrants (henceforth: attriters)? To what extent are such access problems differentially detectable by formal tasks vs. in spontaneous language use?

- RQ1a: Are attriters outperformed by predominantly monolingual speakers in their country of origin (henceforth: controls) on a controlled lexical access task?
- RQ1b: Do attriters use a less diverse lexical repertoire in free speech, as assessed by a number of diversity measures?
- RQ1c: Do attriters show an overall preference for more “common, high-frequent” lexical items and a tendency to underuse “less-common, low-frequency” items (as predicted by Andersen, 1982)?
- RQ1d: Can attrition effects across items of varying lexical frequency be predominantly ascribed to problems of activation due to non-use, or to problems of inhibition related with the increased use of L2?

RQ2: Measuring attrition: What measures of lexical diversity are best suited to detect lexical attrition in free speech?

- RQ2a: Are type-token based measurements good indicators of lexical attrition?
- RQ2b: What measurements are suitable to detect differential attrition effects across items of varying lexical frequency?
- RQ2c: What measures are suitable to detect attrition effects linked with the distribution, as opposed to the selection, of lexical items in the wider discourse?

RQ3: Extralinguistic variables: Which sociolinguistic and personal background variables have an impact on lexical attrition?

RQ3a: How does L1 use in a variety of settings (formal vs. informal, interactional vs. exposure etc.) affect lexical attrition?

RQ3b: How do personal background factors, such as age at emigration, length of residence, and level of education affect lexical attrition?

RQ3c: To what extent does the degree of similarity between L1 and L2 lexicons affect lexical attrition?

The study

Participants

The present investigation is based on two verbal fluency tasks as well as on free and elicited speech collected from 159 native speakers of German. 53 of these speakers emigrated to Anglophone Canada (Vancouver area) and 53 to the Netherlands at least 9 years before the data collection at age 14 or older. 53 had lived in Germany all their lives and never used a language other than German routinely. The three groups were matched for gender, age and education. Participant characteristics are summarized in Table 2.

/ insert Table 2 here /

A sociolinguistic questionnaire comprising 78 items was used to assess the bilingual speakers' history, linguistic habits and attitudes towards L1 and L2 (for details see www.let.rug.nl/languageattrition/SQ and Schmid & Dusseldorp, 2010) and also formed the basis for free speech production (see below). All variables were coded on a scale between 0 and 1, where 0 means no use of the L1 or a strong preference for the L2, while 1 indicates very high use or strong preference for the L1. In order to reduce these variables to a realistic

number of predictors for statistical analysis, the following compound factors were calculated, based on the procedures suggested by Schmid and Dusseldorp (2010):

Total L1 use: This variable was an average of a total of 17 questions pertaining to the use of German in informal settings (Cronbach α = .919). It comprised the following subfactors:

- overall frequency of use of L1 (1 question)
- frequency of use of L1 within the family (1 question)
- language used most frequently on a daily basis (1 question)
- frequency of use of L1 with the partner (4 questions)
- frequency of use of L1 with children (where applicable, 4 questions)
- frequency of use of L1 with friends (4 questions)
- frequency of visits to Germany (1 question)
- frequency of contacts (telephone, letters, email etc.) with Germany (1 question)

Affiliation with L1 language and culture: This variable was based on a total of 6 questions pertaining to linguistic and cultural affiliations and preferences (Cronbach α = .706). It comprised the following factors:

- importance of maintaining the L1
- preferred culture
- preferred language
- language and culture with which there are the strongest emotional ties
- language and culture with which speaker identifies most
- if all external constraints were removed, would speaker like to move back to Germany?

L1 at work: previous investigations of L1 attrition have established that the use of the L1 for professional purposes often emerges as the most important external predictor for

attrition across a range of skills (see above). This factor, based on one question (how frequently do you use German for professional purposes?) is therefore also included here.

The two attriting populations behave very similarly with respect to the amount of use they make of their L1 in their daily lives (none of the predictors differ significantly for the two populations, see Supplementary Table 1). Of course the geographical proximity of the Netherlands to Germany may impact to some extent on the opportunity that individual speakers have to use their L1 (e.g. in terms of how often they can visit their home country). However, in their close social networks, daily contacts and professional lives, both groups appear to use the L1 with approximately equal frequency, and they are also similar in their cultural and linguistic preferences.

While the two two bilingual populations are similar to each other on these factors, there is considerable individual variability, with some individuals in both groups having frequent contact and high affiliation with Germans while other use it only sporadically and have low affiliation.

Verbal fluency tasks

The investigation used two semantic verbal fluency (VF) tasks. In each task, the participant was given 60 seconds to produce as many items from a given semantic category as possible. The two categories used were a) *Fruit and Vegetables* and b) *Animals*. Overall production on the two tasks was averaged for each participant (VFTot). Furthermore, each task was divided into 6 segments of equal length (10 seconds), and the average productivity of each speaker in each segment was calculated.

Speech samples

Two speech samples were collected from each speaker. The first was a conversation about the individual's history and biography which typically lasted between 30 and 90 minutes. This interview allowed the participants to make use of the full range of their language skills on a variety of topics that most migrants frequently talk about, and thus imposed few contextual constraints. The second speech sample was a monologue-narrative in which participants were watched and subsequently narrated a 10-minute film sequence from the silent Charlie Chaplin movie 'Modern Times' (Perdue 1993; for the full elicitation procedure see Schmid, 2011). This task was assumed to be somewhat more challenging, as it required the participant to describe a specific set of events and items, while also imposing the extra cognitive task of remembering the narrative line.

Due to equipment failure and time constraints, there was a small number of participants from whom one of the two speech samples could not be elicited. The sociolinguistic interviews (n=153) comprised a total of 378,740 tokens (not counting filled pauses, false starts, repetitions and self-corrections). The distribution of these data across the three groups is given in Supplementary Table 2. The Charlie Chaplin film retellings (n=155) comprised a total of 110,270 tokens, for full details see Supplementary Table 3.

Lexical diversity analyses

Our approach to lexical diversity in the present study involves the use of type-token based measures, such as VOCD, as well as measures designed to solve the problematic areas discussed above. The full list of the diversity-related measures and indices used in this study comprises:

- Types (the number of different words in each sample)
- Tokens (the total number of words in each sample)

- VOCD (a probability-based measure of diversity derived through random sampling)
- MTLD (the mean number of running words that remains above a TTR threshold of .72)
- Shannon's index (a probability-based measure of diversity that takes both types and evenness into account)
- Evenness (how evenly tokens are spread across types; derived from Shannon's index)
- Effective types (the number of types in the sample adjusted downward in accordance with the degree of unevenness in the sample; see Chao & Jost, in press)
- Rarity (the mean frequency rank of the types in the sample; based on the COSMAS II corpus, which we describe in more detail below)
- Dispersion (the mean number of words between tokens of the same type)

We furthermore investigated the development of lexical sophistication (that is, the use of words with varying frequencies) in the attritional process. For both sets of files (interview and film retelling), our analysis of lexical diversity was based on the lemmatized version of the files created by means of the procedure for morphological tagging and disambiguation described above. The VOCD measure was calculated for this lemmatized corpus with the help of the CLAN program.

Lemmatization was checked manually in the lists that had been extracted from each file. From this list, all lexical items (nouns, full verbs and adjectives) were retained. This yielded a total of 8,715 lemmatized lexical types (91,242 tokens) in the interview files and 3,082 lemmatized lexical types (31,442 tokens) in the film retelling files.

Since one of the predictions for lexical attrition is that attriters will come to prefer more frequent words and underuse less frequent ones, overall frequency was assessed in two ways. Firstly, it was established how frequently each lemma was used within its relevant corpus (the interview and film retelling files, respectively). Secondly, the total frequency of use for each lemma was also measured in the COSMAS II corpus (Corpus Search Management and

Analysis, maintained by the *Institut für deutsche Sprache* at Mannheim and based on ca. 5.4 billion word forms)[2].

Lexical sophistication was then assessed by dividing all types into five frequency bands both on the basis of our own corpora and of the COSMAS II corpus. Each of these bands contained (as closely as possible) the same number of tokens (as suggested by Schmid, Verspoor & MacWhinney, 2011). An overview of these frequency bands is given in Table 3.

/ insert Table 3 here /

It is notable that the most frequent band in both corpora contains a number of topic-specific items. For the interview, these were words related to the migration experience, such as *Jahr*, 'year', *sprechen* 'talk', *Deutsch* 'German', *Englisch* 'English'; in the film retelling, among the most frequent words were nouns relating to important protagonists or items from the film, such as *Polizist* 'policeman' (the 10 minute film sequence features 6 different police officers), *Frau* 'woman', *Mädchen* 'girl', *Brot* 'bread' and *Haus* 'house'.

For each speaker it was then assessed which proportion of the lemmatized lexical items that s/he had used fell into each of these frequency bands. Furthermore, it was determined what proportion of the items each speaker had used were among the 50 most frequent in the relevant corpus (as suggested by Paul Meara, p.c.).

Cognates

---

[2] The authors would like to express their gratitude to the *Institut für deutsche Sprache* for making these data available, and in particular to Franck Bodmer Mory, who very kindly conducted the analyses for us. The COSMAS II corpus contains texts from Germany, Austria and Switzerland. The frequency with which lexical items were used in our own corpora (interview and film retelling) correlated significantly ($p < .001$) with the COSMAS II frequencies, and these correlations were highest when only the data from Germany were considered (interview: $r = .239$, film retelling: $r = .543$; for the data from Austria $r = .169/.511$; for the data from Switzerland $r = .196/.508$; for all data combined $r = .207/.537$). Since all participants originated from Germany, we based our further analyses on these data only.

In order to determine to what degree crosslinguistic similarities might have impacted on the attritional process it was assessed how many of the lexical items used in the corpus shared a similar form across the contact languages (German-English for the Canadian group, German-Dutch for the Netherlands group). The two combined corpora (interview and film retelling) contained a total of 10,481 lemmatized lexical types, which were classified for similarities in lexical form[3]. Of these, 4,445 (42.41%) were German-Dutch (but not German-English) cognates, 3,080 (29.39%) were German-Dutch-English cognates, 98 (0.94%) were German-English (but not German-Dutch) cognates, and 2,858 (27.27%) shared no similarities between German and either of the contact languages. This means that, of the lexical items (types) used in the present corpus, a total of 7,525 (71.80%) were similar in form between Dutch and German, and 3,178 (30.32%) were similar between English and German. It was then assessed for all bilingual speakers what proportion of the speech they had produced consisted of cognate items in their own L2.

Results

Verbal fluency

The total average score achieved in the two VF tasks and the average productivity of all participants in each 10-second segment are shown in Table 4. The descriptive statistics show that the German controls outperform the two bilingual populations in each segment of the task as well as on their total score. These differences were assessed by means of a multiple analysis of variance (MANOVA) and shown to be highly significant (Wilks' Lambda: $F = 3.135$, $p < .001$, $\eta^2 = .111$). Tests of between-subject effects for the individual dependent variables show that effect size has a tendency to decrease as the task proceeds (see Table 4).

[3] The classification was performed by a student assistant who is an early Dutch-German bilingual and a 3rd-year student of English. She rated those items as cognate which were predictable on the basis of regular phonological correspondences.

/ insert Table 4 here /

A further interesting effect is illustrated in Fig. 1: All groups begin at a fairly high level of productivity, which drops sharply in the next segments and then levels off. The differences between attriters and controls seem particularly pronounced in segment 2 and 3. Towards the end of the task, the attriters in the Netherlands appear to 'catch up' with the controls, whereas the Germans in Canada (who had started at a higher level of productivity than the participants with Dutch L2) show a further drop.

/ insert Fig. 1 here /

Free speech

Lexical diversity

Table 5 summarizes the descriptive group results for lexical diversity and sophistication based on the present corpus.

/ insert Table 5 here /

Multivariate analyses of variance by group were conducted, respectively, for the interview and film retelling data; these analyses are summarized in Table 6. Since Jarvis (2013) has pointed out that most measures of lexical diversity are affected to some extent by volume (ie., text length), the number of tokens produced in each sample was included as a covariate. With all variables represented in Table 5 included, Box's M was highly significant, indicating a violation of the homogeneity of the covariance matrix. Removing Shannon's index from the analyses reduced Box's M to acceptable levels for both datasets (interview: Box’s M 184.427,

p = .026; film retelling: Box's M = 174.482, p = .071).[4] Group differences were significant for both datasets at p < .01 (Wilks' Lambda F = 11.439, partial η^2 = .475 for the interview data and F = 2.082, partial η^2 = .140 for the film retelling data). The covariate, number of tokens, was highly significant at p<.001 for both datasets, with a very high effect size (interview: F = 204.138, partial η^2 = .942; film retelling: F = 208.357, partial η^2 = .942).

The analysis of the individual dependent variables (see Table 6) shows that volume consistently affects the first set of variables designed to assess lexical diversity and distribution, with the exception of the MTLD measure. VOCD is only affected in the film retelling but not in the interview data, all other measures (Evenness, Dispersion and Rarity) appear highly sensitive to volume. Where the frequency analyses are concerned, there are hardly any volume effects (with the exception of the highest frequency band in the interview data and the second one in the film retelling).

Turning to the group comparison, there appear to be few differences across populations regarding the two overall measures of lexical diversity, VOCD and MTLD. The only significant finding is that MTLD is somewhat lower for the Canadians in the film retelling data than for the other populations. Evenness, dispersion and rarity are affected consistently in the interview data (with a marginal significance for evenness among the Canadian group and (highly) significant levels on all other measures) but not at all in the film retelling.

In the interview, frequency band 1, 3 and 5 as well as the 50 most frequent items differ between populations, while for the film retelling data, only Frequency Band 2 and 5 as well as the 50 most frequent items are significantly different. Effect sizes for these group differences are, however, quite weak, ranging from η^2 = .080 (Frequency band 5 in film retelling) to η^2 = .162 (Frequency band 5 in interview), indicating that the group differences, while consistent, are hardly dramatic. Supplementary Figures 1a/b illustrate for the lexical frequency bands that

[4] MANOVA is robust in relation to homogeneity assumptions except when sample sizes are unequal and Box's M shows a significance of p < .001 (see Tabachnick & Fidell, 1996, p. 382).

the group differences are indeed the effect of the attriters differing from the controls in the direction which was predicted (ie. an overuse of the more frequent items and an underuse of the least frequent ones).

/ insert Table 6 here /

COSMAS II frequencies

Table 7 summarizes the results for the frequencies of items used by each speaker, based on the COSMAS II corpus. In these data, there do not seem to be straightforward tendencies among the attriters to overuse the more frequent or underuse the less frequent items, as was the case for the frequencies based only on the corpora at hand in the analyses above. The distribution across frequency bands is graphically represented in Supplementary Figures 2a (interview data) and 2b (film retelling data).

/ insert Table 7 here /

Multivariate analyses of variance by group were conducted, respectively, for the interview and film retelling data based on the COSMAS II frequencies summarized in Table 7. As in the analyses presented above, number of tokens was included as a covariate. The homogeneity of the covariance matrix was assessed by means of Box's M, which was not significant for either dataset (interview: Box's M 27.549, $p = .038$; film retelling: Box's M 52.926, $p = .188$). Group differences were significant for the interview data (Wilks' Lambda F -= 15.035; $p < .001$; partial $\eta^2 = .341$), but not for the film retelling data (Wilks' Lambda = 1.335; $p = .086$). The covariate, number of tokens, was marginally significant for the interview data (Wilks'

Lambda F = 2.559, $p < .05$, partial $\eta^2 = .081$) but not for the film retelling data (Wilks' Lambda F = 1.335, p = .245).

The analysis of the individual dependent variables in the interview data shows significant group differences for all frequency bands. Planned comparisons show that the attriters in Canada differ from the controls on frequency bands 2-5, showing a higher use of the high-frequency items in band 2 than the controls, but also somewhat higher proportions of the low-frequency items in bands 4 and 5. The attriters in the Netherlands are only different from the controls on the extremely high-frequency band 1, which they use more frequently. The full tests are summarized in Table 7 above.

Cognates

For all speech samples it was assessed what proportion of the lemmatized lexical items used consisted of German-English (GE-EN) and of German-Dutch (GE-NL) cognates. The use of GE-EN cognates between the controls and the Germans in Canada looks very similar, the controls use 48.97% in the interview and 44.93% in the film retelling, while the Canadian attriters have 48.01 and 46.30%, respectively. For the GE-NL cognates, a slightly higher use of cognates by the Dutch-German bilinguals can be observed, they account for 83.25% in the interview and 82.19% in the film retelling for the controls, and for 86.31% in the interview and 84.66% in the film retelling for the attriters in the Netherlands.

A multivariate analysis of variance by group was conducted for the use of GE-EN and GE-NL cognates in interview and film retelling. The homogeneity of the covariance matrix was assessed by means of Box's M, which was not significant (Box's M = 25.460, p = .223). Group differences were significant (Wilks' Lambda F = 4.602; $p < .001$; partial $\eta^2 = .142$).

Planned contrasts established no overuse of GE-EN cognates by the Germans in Canada, but significant overuse ($p < .01$) of GE-NL cognates by the Germans in the Netherlands. (The fact that the use of GE-EN cognates in the film retelling data was marginally significant for this group as well might be due to the fact that there is a substantial overlap between GE-EN and GE-NL cognates, see above.) Effect sizes were small, with a maximum η^2 of .172. Table 8 gives the details of these analyses, while Supplementary Figures 3a/b illustrate the group differences in the use of cognates.

/ insert Table 8 here /

Summary of group comparisons

The previous analyses have attempted to probe lexical productivity, lexical access and lexical diversity among two attriting populations by means of a range of tasks and measures. Where the controlled VF task was concerned, the analyses found robust differences between populations, in particular for the early segments of the task, indicating that the attriters were less productive in naming items belonging to the two categories and took longer to 'get into' the task. For the two free speech samples, it is interesting to see that the measures involving ratios of types and tokens (VOCD and MTLD) show no substantial difference between populations. Evenness, dispersion and rarity do differ, but only for the interview and not for the film retelling data. These differences may thus to some extent be related to text length even though this was included as a covariate in the analyses, or to the fact that the film retelling refers to a predetermined sequence of events and items and is thus more constrained. In both corpora, the attriters overuse the most and underuse the least frequent vocabulary when frequency is based on the present corpora alone, but not when it is based on a large

corpus of German data of all types. Cognates are overused by L2 speakers of a closely related language (Dutch) that has a similar lexicon to the L1, but not by L2 speakers of English, whose lexicon differs more substantially from German despite the historical and typological relationship between these two languages. The differences between the L2 Dutch and L2 English groups are summarized in Table 9.

/ insert Table 9 here /

These findings suggest a complex pattern of how the lexicon is affected in the attritional process. The following section will investigate to what extent and in what manner the different measures investigated here can best be combined in order to predict whether a particular speaker is an attriter or not.

Discriminant analysis

Linear discriminant analysis (DA) is similar in many respects to MANOVA, in other respects to regression analysis, and in yet other respects to factor analysis. As described by Huberty and Olejnik (2006), DA and MANOVA rely on similar statistical techniques but are in effect the opposite of each other: The grouping variable in a MANOVA becomes the dependent variable in a DA. The purpose of DA is to use outcome variables as predictors of the group membership of individual cases by constructing a model that best predicts the dependent variable. In this sense, DA is similar to multiple logistic regression, which also involves the use of a categorical (or nominal) dependent variable. Finally, DA is also similar to factor analysis in that both types of analysis identify clusters of variables along multiple dimensions.

The DA in the present study was run with SPSS 19.0. The participants' group membership was used as the dependent variable. The independent variables included the following measures, totalling 54 items:

VF task: the total average number of elements named in both tasks per participant, and the total average number of elements named in each of the 6 10-second segments in both tasks (total 7 items)

interview corpus: VOCD, MTLD, Effective types, Shannon, Rarity, Evenness, Dispersion, Frequency band 1-5 based on present corpus, Proportion of 50 most frequent items based on present corpus, average frequency of lexical items based on present corpus, Frequency band 1-5 based on COSMAS II corpus, Proportion of 50 most frequent items based on COSMAS II corpus, Average frequency of lexical items based on Cosmas II corpus (21 items)

film retelling corpus: VOCD, MTLD, Effective types, Shannon, Rarity, Evenness, Dispersion, Frequency band 1-5 based on present corpus, Proportion of 50 most frequent items based on present corpus, average frequency of lexical items based on present corpus, Frequency band 1-5 based on COSMAS II corpus, Proportion of 50 most frequent items based on COSMAS II corpus, Average frequency of lexical items based on Cosmas II corpus (21 items)

Fluency in film retelling: The proportion of disfluency markers (filled pauses (FP), empty pauses (EP), repetitions (REP) and self-corrections or retractions (RETR)) were described and analysed for the film retelling data by Schmid & Beers Fägersten (2010). These measures were included in the present analysis, as well as the number of words spoken per minute (excluding disfluency phenomena) in these re-tellings (5 items).

The results from the cognate analysis were not included in this analysis, since they are presumed to affect the three populations differentially (it is of no consequence to an English L2 speaker whether a particular German word has a cognate equivalent in Dutch, and vice versa, nor does cognate status matter for the monolinguals).

The DA method was set to stepwise, which meant that it chose only one variable at a time in accordance with how much that variable contributed to the strength of the model. The criteria used for variable entry and removal were the default Wilks' Lambda F values of 3.84 for entry and 2.71 for removal. This ensured that a variable would be added to the model only if it contributed significantly to the strength of the model, and that it would subsequently be removed as new variables were added if it no longer made a significant contribution to the developing model. Results were cross-validated using leave-one-out cross-validation, a procedure that iterately builds a model using all but one of the cases, and then tests the model's predictive accuracy blindly (without access to its group membership) on the case that was left out during that iteration. The number of iterations during the leave-one-out cross-validation phase is equal to the number of cases, which means that each case is used once as the test case. The results of the cross-validation phase show the number of cases whose group membership was predicted correctly.

Due to missing variables for some of the participants, the DA omitted 13 participants from the analysis. The stepwise procedure selected 11 predictor variables (these are listed in Supplementary Table 4). The overall cross-validated classification accuracy of the 11-variable model is 85.0%, meaning that the group memberships of 85.0% of the participants (125 out of 147) were predicted correctly based on a model consisting of just the 11 variables selected (see Table 10 for the numbers and percentages of participants in each group whose membership was predicted accurately during the cross-validation phase). This is significantly above the chance level of 33% ($df = 4$, $n = 147$, $X^2 = 181.885$, $p < .001$). As shown in the

classification matrix in Table 10, the model is best at identifying the German speakers still living in Germany (94.2% classification accuracy), followed by the attriters living in Canada (84.0%). It is least effective at identifying the attriters living in The Netherlands (75.6%), who are often misclassified as attriters living in Canada (20.0%) and in two cases as German speakers still living in Germany (4.4%).

/ insert Table 10 here /

As mentioned, DA explores relationships among variables along different dimensions—referred to as functions—and the number of functions is always one less than the number of groups. The results of the DA showed that Function 1 accounts for 90.1% of the variance in the data, and that it is most characterized by qualities related to lexical diversity in the interview task. Of the 11 variables included in the classification model, four are primarily affiliated with Function 1. These include (a) interview evenness, (b) interview effective types, (c) interview percentage of items in frequency band 3, and (d) interview MTLD. The seven other variables are more closely aligned with Function 2, which accounts for 9.9% of the variance, and is characterized primarily by qualities related to lexical sophistication in the interview task.

DA assigns specific weights (or canonical discriminant function coefficients) to each variable, and uses these to calculate overall function scores for each participant. The group means for the three groups in relation to both functions are plotted in Supplementary Figure 4, which shows the overall separation among the groups in relation to the 11 predictor variables in the model. These results suggest that the participants' patterns of lexical diversity in the interview task (i.e., Function 1) separate the German speakers living in Germany from the

attriters, whereas the participants' patterns of lexical sophistication in the same task (Function 2) separate the attriters in The Netherlands from the other two groups.[5]

It is interesting that all but one of the predictive variables in the DA model relate to lexical qualities in the interview task. Only one predictor is associated with the film retelling task, and this variable is more a matter of fluency (i.e., words per minute) than of word choice. The variables pertaining to the VF task, which is one of the instruments most often used in attrition studies (and on which the populations differed significantly in the individual analyses presented above), do not appear in the final model at all.

The impact of extralinguistic variables

The analyses presented above have revealed a number of differences between the attriting and the control populations. The discriminant analysis provided a model which included those factors that collectively have the most predictive power in assigning group membership to each individual participant. For each speaker, a new variable was calculated, representing his or her cumulative score on the two functions that emerged from the DA.

In order to assess to what extent external factors such as the biological age of the speaker, the length of residence in the country of migration, as well frequency of use and attitudes towards the L1 and attitudes towards this language, might have had an impact on these scores (as was asked in RQ2), multiple linear regressions were conducted on the data from the bilingual speakers. The outcome variables for these analyses were those measures for each of the three tasks for which effect size had been strongest in the MANOVAs:

- the total number of items on the VFT
- Dispersion in the interview
- the proportion of items in Frequency Band 1 in the interview

[5] This interpretation is supported by the participants' aggregated function scores rather than by a comparison of the group means for individual variables.

- the proportion of items in Frequency Band 2 in the film retelling

In addition, the two functions that were calculated by the DA were used as outcome variables. The predictors entered into the models were the extralinguistic factors discussed above, namely:

- age at emigration
- length of residence
- overall L1 use
- affiliation with the L1 language and culture
- L1 use for professional purposes

One further variable was included in these models, based on the prediction made by Segalowitz (1991) that a higher L2 proficiency may lead to reduced automaticity in the L1. The original test battery contained a measure of overall L2 proficiency, namely a C-Test. The C-Test is a version of the cloze test which is used frequently in research on both second language acquisition and first language attrition. It consists of short texts in which parts of words are deleted according to a predetermined schema (see Schmid, 2011 for an in-depth discussion), and each correctly completed word is awarded one point. Our C-Test used five short texts with a maximum possible score of 100. Although the two attriting populations completed two different versions of this test (English for the Germans in Canada and Dutch for the Germans in the Netherlands), performance on this task was exactly the same: both groups achieved a mean score of 73.23 (English group: st.dev 16.31, range 18-96; Dutch group: st.dev. 15.57, range 19-99).

Of the six models arrived at by these linear regression analyses, only VFT emerged as having been significantly impacted by the predictors (for the full details of the analyses, see Supplementary Table 5), with a weak impact of L1 use for professional purposes and L2 proficiency (speakers who used German in their workplace and who achieved higher scores

on the L2 C-Test were somewhat more productive on the VFT). Age at testing was removed in all analyses as having no predictive power overall, and for all of the other analyses, none of the predictors reached significance. This implies that the variance present on those measures that can best discriminate the attriting populations from the monolingual controls is not affected by factors pertaining to age of emigration, length of residence, amount of L1 use, attitudes, or proficiency in L2. This finding may be surprising and appear counterintuitive, but it does underscore the results from a number of earlier studies, all of which failed to establish a link between such extralinguistic factors and language attrition (e.g. Schmid & Dusseldorp, 2010; Yilmaz and Schmid, 2013; Varga, 2012).

Discussion

The aim of the present study was to provide an analysis of first language attrition in the domain of the lexicon which would go beyond what previous studies have achieved in both breadth and depth. In order to achieve this, a set of data was analysed that comprised both controlled, formal tasks probing lexical access (Verbal Fluency tasks) and spontaneous speech elicited by means of a film retelling and a semi-structured interview. Our first set of research questions addressed the possible loss of lexical accessibility in a situation where the speaker has little opportunity to use her L1 and relies mainly on the L2 in everyday interactions. In particular, we were interested in the impact of activation vs. inhibition on the attritional process.

Secondly, we aimed to establish what types of measures might be most suited to detect L1 attrition in free speech. We augmented type-token based measures with an analysis of measures that take into account the rarity and distribution of items across the discourse, as well as assessing the proportion of items of varying lexical frequency. We also assumed that in free speech, the distribution of items might be more uneven for attriters than for controls,

due to the limited accessibility of the entire range and a more local spread of activation across semantic fields.

A last set of research questions related to the impact of individual external background factors such as age at emigration and L1 exposure as well as the degree of lexical differences between languages (English vs. Dutch in comparison to German) on the degree of attrition.

Verbal fluency task

The statistical comparisons showed that the attriters were indeed less productive than the controls on a verbal fluency task, indicating that they were able to access fewer items within a given lexical category and specified time limit. With an overall effect size of $\eta^2 = .151$, however, these findings indicate only a relatively weak difference in accessibility. Whether these are due to the fact that bilinguals have a much larger repertoire and furthermore have to suppress L2 items in order to be able to produce items in the L1, or whether they can actually be considered effects of an attritional process that has made lexical access more difficult and effortful, is difficult to establish based on the present data. The fact that bilinguals become slower at naming objects in their first language very shortly after the onset of bilingualism (and so probably not due to attrition) has been demonstrated e.g. by Mägiste (1979) and Linck, Kroll & Sundermann (2009). Future work might address this question by investigating fluent bilinguals in the country of origin.

It is interesting to observe, however, that the attriters in Canada start out closer to the native norm than the attriters in the Netherlands. During the last third of the task, the effect reverses and it is the Dutch L2 speakers who catch up with the native norm, while the Canadians drop back. It is possible that this crossover effect is due to the differential amounts of competition and consequently inhibition: for the L2 English speakers, competition from the L2 is less strong than for the attriters in the Netherlands whose L2 shares more cognates with

their L1 (as was shown above, less than one third of English items but more than two thirds of Dutch items are cognates in German). Presumably, speakers begin by naming those items that are most accessible in their minds, so that in the early stages of the task, what slows bilinguals down will be predominantly the added cognitive effort of inhibiting competitors. This will disadvantage speakers whose languages share a large part of their vocabulary. Later on, when the speaker has to 'dig deeper', it may become easier to activate less easily accessible items if they are supported by similar forms in both languages.

The verbal fluency task is a very popular task in language attrition studies, since it is easy to construct, administer and score, does not rely on specialised equipment (as do e.g. picture naming and other reaction time tasks), needs no stimuli etc. However, despite the robust group differences detected by the group comparisons, none of the variables pertaining to this task were selected in the Discriminant Analysis. In other words, the individual results from this task had very limited predictive power when it came to classifying an individual participant as an attriter or control speaker. This means that the use of the verbal fluency task in order to detect lexical attrition may need to be reassessed. As was noted above, a range of studies have administered this task alongside other elicitation techniques, such as narratives and interviews. A metastudy of these investigations, using the methods described here, may be a valuable step forward in this context. In summary, the answer to RQ1a is that attriters are indeed outperformed by controls on the VFT, but that effect sizes are weak and that performance on this task does not contribute much to the profile of an attriter vs. a control.

Free speech

Where free speech was concerned, we first observed that type-token frequency based measures failed to distinguish between attriters and controls. There were some differences related to the distribution of items across the interview, indicating that for some of the attriters

there were longer intervals intervening between items of the same type than in the data from the controls. This is possibly related to the nature of the interview data, where attriters maybe more often went off on 'tangents' related to the emigration experience and also produced longer speech samples overall. No such distributional effects were found in the film retelling data, probably due to the fact that the sequence of events was predetermined here by the stimulus, and lexical selection was thus more constrained. In other words, in answer to RQ1b, it is likely that such distributional differences between populations are the outcome of the type (and length) of data produced and not an indication of changes in lexical access due to language attrition.

Word frequency was assessed on the basis of two corpora: one that consisted only of the data collected in the present study (that is, only of texts of a very similar nature) and one that measured lexical frequency in a large corpus of texts comprising 5.4 billion words from a wide range of written and spoken sources (the COSMAS II corpus collected by the *Institut für Deutsche Sprache*). In order to assess the use of high-, medium- and low-frequency items, we divided all lexical lemmas into five frequency bands, each representing 20% of all tokens. The group comparisons showed an overuse of the high-frequency and an underuse of the low-frequency items among the attriters. It was shown that for the interview data, the overuse affected the items of the very highest frequency, while in the film retelling, it was the second frequency band that was overused. Again, this is probably due to the nature of the stimuli: as was pointed out above, the highest frequency band in the film retelling data contained a large number of items that figured prominently in the film sequence and were therefore necessary for all speakers.

We also found an overuse in both attriting populations of the mid-frequency band 3 in the interview data. This is probably related to the fact that very low-frequency items are dispreferred by these speakers, who then come to rely more on the mid-range. However, it

should be pointed out again that effect sizes here were very modest indeed. It is certainly not the case that attriters stop using low-frequency items overall: every single speaker used at least a certain number of words from this category, as is illustrated in the histograms in Supplementary Figures 5a/b.

The interpretation of the frequencies that were derived from the COSMAS II corpus were far less straightforward. This indicates that it may be difficult to rely on overall lexical frequencies, and preferable to use a corpus for reference purposes that is very similar in type to the data under investigation. This finding is helpful for researchers working on less well-documented languages for which large corpora may not be available, as it suggests that the corpus at hand may yield the most reliable results (provided that it is large enough). In answer to RQ1c, the analyses presented here have thus established that attriters do indeed develop a preference for somewhat more common and more high-frequent lexical items within the context specified by a certain task, but that this preference does not appear to be reflected when lexical frequency is based on overall language use from a wide range of text types and tasks.

RQ1d focussed on the important but complex issue of whether attrition is determined more by a reduction in lexical accessibility, due to infrequent use of the L1, or to problems related to inhibiting the more frequently used L2. It was predicted that speakers of an L2 that was very closely related to the L1 would have fewer problems related to activation due to underuse, since activation levels degrade over time, but would be maintained due to the use of similar items in the L2. Such speakers would, on the other hand, suffer more from problems related to inhibition than bilinguals with a less similar L2, since the close neighbourhood of items in L1 and L2 would make inhibiting the competing items more costly..Two findings stand out in this respect:

1. The attriters in the Netherlands differ significantly from the controls with respect to evenness in the interview data, while there is no difference between the Canadians and the controls. It is possible that this effect is linked to some extent to the similarity between the two languages, which requires the Dutch L2ers to inhibit their L1 more strongly in their daily lives when interacting in their L2. Potentially, this inhibition effort makes the L1 less available, so that those items whose activation is boosted by semantic similarity to the topic at hand are preferred.
2. The Dutch attriters overuse cognates while the Canadian attriters do not. This finding may again relate to the proportion of the lexicon that is shared between a speaker's L1 and L2, and the assumption proposed by e.g. de Groot, Dannenburg and van Hell (1994) that cognates share the same semantic representation (recall that more than two thirds of the lexical items used in the present corpus were cognate between Dutch and German, as compared to less than a third of the English-German pairs). L2 speakers of English have more reason to distrust forms that look the same, and may therefore opt for a different word if they have the choice.

The lack of clearer results in this respect may to some extent be due to the fact that, although English is less closely related to German, it still belongs to the same language family. Further studies invoking more distant languages may be able to shed more light on this issue.

Diagnosing attrition

We then attempted to combine the measures that were derived from the data collected for the present study in order to determine which combination of factors would produce the most reliable model that would discriminate attriters vs. non-attriters. We entered the measures from the verbal fluency task, the measures concerning lexical diversity and sophistication plus a number of fluency measures from the film retelling task into the analysis, which yielded a

model comprising eleven factors. What was most immediately striking was that both the formal task and the film retelling task did not contribute to the final model (with the exception of the speech rate from the film retelling task).

The model revealed that only four of the 54 measures included in the analysis together accounted for more than 90% of the overall variance. These measures pertained mainly to the diversity and distribution of types across the interview data, and also included the use of items in the mid-range frequency band (which, as was mentioned above, was overused by the attriters). A second set of measures, accounting for a further 9% of the variance, distinguished the Dutch attriters form the other two groups and contained mainly measures pertaining to word frequency in the interview data, in particular to the high- and low-frequency lexical items.

Where the use of different tasks and various measures of lexical diversity in language attrition research are concerned (RQs 1 and 2), the findings from the DA suggest that the predominantly formal tasks and even elicited data that have provided the data for most previous investigations of attrition are less powerful and yield less valid comparisons of attriters and controls than true spontaneous speech. Similarly, the relatively easily derived measures that are based on type-token frequencies appear to give less reliable insights into the attritional process than measures that assess the frequency of lemmas, either in the corpus at hand or (where available) in larger corpora of the language. It also appears important not only to look at the presence/absence of items in the data but also at their distribution (evenness) across the entire discourse, which also may have changed in the attritional process.

The impact of extralinguistic variables

While the group comparisons showed robustly significant differences between attriters and controls and the model yielded by the DA was quite reliable in assigning individuals to the

different populations (the classification was accurate in 85% of the cases), the analyses did leave us with the question of why some of the attriters in the individual populations had stronger attrition effects than others. In an attempt to gain insight into the questions formulated under RQ3, we attempted to investigate the impact of external factors pertaining to the length of time spent in an L2 environment, the use of the L1 in different situations and settings, attitudes towards the native language and culture, as well as L2 proficiency on these functions. In order to do so, linear regression analyses with these factors as predictors were conducted on the outcome variables that had most strongly differentiated attriters and controls (the number of items produced on the VFT, the use of high-frequency items in interview and film retelling, Dispersion in the interview and the two functions that had been calculated by the DA). The only significant impact that these analyses were able to find, however, concerned the formal task, where speakers who used the L1 in their workplace and who were more proficient in the L2 were shown to be more productive. Other than that, these analyses produced no significant results and left us no nearer to answering the puzzling question of why some individuals may be more susceptible to attrition than others. Based on the data investigated here, length of residence, frequency of L1 use, attitudes or L2 proficiency (as addressed by RQs 3a and 3b) do not appear to have an impact on individual variation in lexical access.

Conclusion

This study has provided an in-depth examination of lexical L1 attrition in relation to the likelihood that attriters will experience decreased lexical accessibility as a consequence of their reduced use of the L1, and as a result of the limited number of contexts in which they use it. We investigated lexical diversity and distribution, lexical sophistication, and verbal fluency, and analyzed whether generalizable and predictable differences could be found

between Germans who have remained in Germany, Germans living in Canada, and Germans living in The Netherlands.

Lexical diversity in this study has been examined as a matter of multidimensional compositional complexity that extends beyond the relationship between types and tokens to other dimensions of word choice, including the degree to which tokens are distributed evenly across different types, the distances with which repetitions of the same type are dispersed, and the relative rarity of the words that are used in a sample of speech. The last of these also overlaps with the construct of lexical sophistication, which we have examined carefully in relation to the percentage of words in each participant's speech samples that can be found in each of the lexical frequency bands in the present data as well as in the COSMAS II native German corpus.

Overall, we found that the attriters do indeed differ significantly from each other and from the German controls on a number of measures relating to lexical diversity and distribution, lexical sophistication, and verbal fluency. In most cases, though, the differences in lexical diversity are quite subtle and cannot be detected with measures that are based only on type and token frequencies. The significance levels and effect sizes also differ by task, such that significant effects for lexical diversity are found only in the interview. Importantly, this is the task where the largest differences were found between groups in terms of number of words produced, so there is a strong possibility that the lexical diversity differences between groups in this task were to some extent affected by sample size. Although we used sample size as a covariate, we believe that future studies will benefit by using these same measures of lexical diversity applied to equally-sized subsamples in order to further reduce the potential effects of sample size.

Regarding rarity and lexical sophistication, we recognize that the frequency rank of a word in a corpus is not a reliable measure of the perceived sophistication of a word or of

whether it has been used appropriately in its context. The findings reported here do nevertheless reveal that the three groups of participants did not use all of the same words, and that there are significant and predictable differences between the words they used and the patterns with which they used them. Future work that explicates these qualitative differences has the potential to substantially further our understanding of the nature of lexical L1 attrition and the mechanisms, such as lexical accessibility, through which it occurs.

Lastly, the finding that language use in the interview was the most strongly predictive task in classifying a speaker as an attriter or a non-attriter (all but one of the eleven factors retained in the Discriminant Analysis originated from this task) underscores the fact that attrition affects that skill that is most characteristic of what native speakers know how to do: use language in free speech. This finding is important in light of the fact that most attrition studies, in particular those assessing lexical attrition, typically use controlled tasks. To what extent such tasks actually assess attrition, or whether they may be detecting a weakening of metalinguistic skills or knowledge instead, remains to be seen.

References:

Ammerlaan, T. (1996). You get a bit wobbly. Exploring bilingual lexical retrieval processes in the context of first language attrition. Ph.D. dissertation, Katholieke Universiteit Nijmegen.

Andersen, R. W. (1982). Determining the linguistic attributes of language attrition. In R. D. Lambert & B. F. Freed (eds.), *The loss of language skills,* pp. 83-118. Rowley, MA: Newbury House.

Berthele, R. (2011). On abduction in receptive multilingualism. Evidence from cognate guessing tasks. *Applied Linguistics Review, 2,* 191-220.

Bialystok, E. (2005). Consequences of bilingualism for cognitive development. In J.F. Kroll & A. M. B. de Groot (eds), *Handbook of bilingualism*, pp. 417-432. Oxford: Oxford University Press.

Cherciov, M. (2011). Between Attrition and Acquisition: the Case of Romanian in Immigrant Contexts. PhD dissertation, University of Toronto.

de Leeuw, E., Schmid, M.S. & Mennen, I. (2010). Perception of foreign accent in native speech *Bilingualism: Language and Cognition*, 13 (1), 33-40.

Dijkstra, T. (2005). Bilingual visual word recognition and lexical access. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 179-201). Oxford: Oxford University Press.

Dostert, S. (2009). Multilingualism, L1 attrition and the concept of 'native speaker'. PhD dissertation, Heinrich-Heine Universität Düsseldorf.

Green, D. W. (1986). Control, activation and resource: A framework and a model for the control of speech in bilinguals. *Brain and Language* 27: 210-223.

Green, D. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition* 1 (1), 67-81.

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and Discriminant Analysis*. Hoboken, NJ: Wiley.

Hulsen, M. (2000). *Language loss and language processing: Three generations of Dutch migrants in New Zealand.* Ph.D. dissertation. University of Nijmegen, the Netherlands.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19,* 57-84.

Jarvis, S. (2009). Lexical transfer. In A. Pavlenko (Ed.), *The bilingual mental lexicon* (pp. 99-124). Bristol, UK: Multilingual Matters.

Jarvis, S. (2012). Lexical challenges in the intersection of applied linguistics and ANLP. In C. Boonthum-Denecke, P.M. McCarthy, & T. Lamkin (Eds.), *Cross-disciplinary advances in applied natural language processing: Issues and approaches* (pp. 50-72). Hershey, PA: IGI Global.

Jarvis, S. (2013a). Capturing the diversity in lexical diversity. *Language Learning, 63* (supplement 1), 87-106.

Jarvis, S. (2013b). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13-44). Amsterdam: Benjamins.

Jarvis, S., & Daller, M. (Eds.) (2013). *Vocabulary knowledge: Human ratings and automated measures*. Amsterdam: Benjamins.

Keijzer, M. (2007). First Language Attrition: A Crosslinguistic Investigation of Jakobson's Regression Hypothesis. PhD dissertation, Vrije Universiteit Amsterdam.

Köpke, B. & Nespoulous, J-L. (2001). First language attrition in production skills and metalinguistic abilities in German-English and German-French bilinguals. In T. Ammerlaan, M. Hulsen, H. Strating & K. Yağmur (eds.), *Sociolinguistic and Psycholinguistic Perspectives on Maintenance and Loss of Minority Languages*, pp. 221-234. Münster: Waxmann.

Köpke, B., & Schmid, M. S. (2004) First language attrition: The next phase. In M. S. Schmid, B. Köpke, M. Keijzer, M. & L. Weilemar (eds.), *First language attrition: Interdisciplinary perspectives on methodological issues*, pp. 1-43. Amsterdam: John Benjamins.

Linck, J.A., Kroll, J.F., & Sunderman, G. (2009). Losing access to the native language while immersed in a second language: Evidence for the role of inhibition in second language learning. *Psychological Science, 20(12)*: 1507-1515.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mägiste, E. (1979). The competing language systems of the multilingual: a developmental study of decoding and encoding processes. *Journal of Verbal Learning and Verbal Behavior*, *18*, 79-89.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488.

McCarthy, P., & Jarvis, S. (2013). From intrinsic to extrinsic issues of lexical diversity assessment: An ecological validation study. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 45-78). Amsterdam: Benjamins.

Montrul, S. (2008). *Incomplete Acquisition in Bilingualism. Re-examining the age factor*. Amsterdam/Philadelphia: John Benjamins.

Opitz, C. (2011). First language Attrition and Second Language Acquisition in a Second-Language Environment. PhD dissertation, Trinity College Dublin.

Paradis, M. (1993). Linguistic, psycholinguistic, and neurolinguistic aspects of interference in bilingual speakers: the activation threshold hypothesis. *International Journal of Psycholinguistics*, 9 (2), 133-145.

Paradis, M. (2004). *A Neurolinguistic Theory of Bilingualism*. Amsterdam: John Benjamins.

Paradis, M. (2007). L1 attrition features predicted by a neurolinguistic theory of bilingualism. In B. Köpke, M. S. Schmid, M. Keijzer & S. Dostert, (eds), *Language Attrition: Theoretical perspectives*, pp. 121-134. Amsterdam: John Benjamins.

Perdue, C. (1993). *Adult language acquisition: Cross-linguistic perspectives*. Cambridge: CUP.

Pielou, E. C. (1969). *An introduction to mathematical ecology*. Hoboken, NJ: Wiley.

Schmid, M. S. 2007. The role of L1 use for L1 attrition In M. S. Schmid, B. Köpke, M. Keijzer & L. Weilemar (eds.), *First Language Attrition: Interdisciplinary perspectives on methodological issues*, pp. 135-153. Amsterdam/Philadelphia: John Benjamins.

Schmid, M.S. (2011). *Language Attrition*. Cambridge University Press.

Schmid, M.S. & Beers Fägersten, K. (2010) Fluency and language attrition. *Language Learning* 60 (4), 753–791.

Schmid, M.S. & Köpke, B. (2007). Bilingualism and attrition. In: B. Köpke, M.S. Schmid, M. Keijzer & S. Dostert (es) *Language attrition: Theoretical perspectives*, pp. 1-8. Amsterdam/Philadelphia: John Benjamins.

Schmid, M.S. & Köpke, B. (2008). L1 attrition and the mental lexicon. In: A. Pavlenko (ed) *The bilingual mental lexicon: Interdisciplinary approaches*, pp. 209-238. Clevedon: Multilingual Matters.

Schmid, M.S. & Dusseldorp, E. (2010). Quantitative analyses in a multivariate study of language attrition. *Second Language Research* 26 (1), 125-160.

Schmid, M. S., Verspoor, M. & MacWhinney, B. (2011). Coding and extracting data. In M. Verspoor, W. Lowie & K. de Bot (eds), *A dynamic approach to second language development: methods and techniques*, pp 39-54. Amsterdam/Philadelphia: John Benjamins.

Segalowitz, N. (1991). Does Advanced Skill in a Second Language Reduce Automaticity in the First Language? *Language Learning* 41 (1), 59-83.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Sorace, A. (2005). Selective optionality in language development. In L. Cornips & K. P. Corrigan (eds.), *Syntax and variation. Reconciling the biological and the social*, p. 55-80. Amsterdam: John Benjamins.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd edition). New York: Harper Collins.

Tsimpli, I. (2007). First language attrition from a minimalist perspective: Interface vulnerability and processing effects. In B. Köpke, M. S. Schmid, M Keijzer & S. Dostert. (eds.), *Language Attrition: Theoretical perspectives,* pp. 83-98. Amsterdam: John Benjamins.

Varga, Z. 2012. First language attrition and maintenance among Hungarian speakers in Denmark. PhD dissertation, Århus University.

Waas, M. (1996). *Language Attrition Downunder*. Frankfurt: Peter Lang.

Yağmur, K. (1997). *First Language Attrition Among Turkish Speakers in Sydney*. Tilburg: Tilburg University Press.

Yılmaz, G. & Schmid, M. S. (submitted). L1 lexical accessibility among Turkish speakers in the Netherlands.

Figure caption

Fig. 1: Average productivity in each 10-second segment of the two verbal fluency task across populations