

To appear in:

Daller, M. H. (ed.) (in press). The measurement of bilingual proficiency. *International Journal of Bilingualism (special issue 2011)*

Michael H. Daller, Cemal Yıldız, Nivja H. de Jong, Seda Kan and Ragıp Başbağlıⁱⁱⁱ

Language dominance in Turkish German bilinguals: methodological aspects of measurements in structurally different languages¹

0 Abstract

The purpose of this study is to establish measures of language dominance in bilinguals who speak structurally different languages, in our case German and Turkish, with tools that are based on fluency and oral proficiency. A “balanced” bilingual with equal proficiency in two (or more) languages is hardly ever found (e.g. Grosjean 1982, Olsson and Sullivan 2005) but the identification of the dominant language is a huge methodological problem especially in studies of structurally different languages (see Daller, van Hout and Treffers-Daller 2003).

The participants in the present study are a group of Turkish-German bilinguals who grew up in Germany and returned to Turkey during their school career, the so-called “returnees” (n = 60), and a group of Turkish secondary school students who grew up in Turkey and learned German as an L2, the so-called control group (n = 56). We firstly establish the language dominance of the two groups with a C-test. We then use oral picture descriptions in both languages to measure a variety of fluency measures, both manually and using scripts written in “Praat” (Boersma and Weenink 2007). On the basis of these scores we are able to develop measures of fluency that correlate highly with the C-test scores and have a highly predictive value in a logistic regression in the prediction of group membership (returnee or member of the control group). We conclude that this corroborates the validity of the measures. Overall we conclude that it is possible to develop measures of language dominance based on fluency and overall oral proficiency.

¹ We would like to thank Heather Hilton, Alex Housen and Brian Richards for their valuable comments on earlier versions of this article.

1. Introduction

Many immigrants to Western countries bring with them languages that are structurally different from the languages of the host countries. This leads to bilingual settings where Indo-European and Non-Indo-European languages are used by bilinguals on a daily basis side by side. Examples for these language combinations are Dutch/ Arabic, Turkish/German or English/Mandarin and many others. It is necessary to be able to compare the proficiency of these bilinguals for example to get an informed judgement on language dominance for educational decisions. Simple judgements about the dominance pattern between the two or more languages of a bilingual might be valid or not but what is needed are objective measures that can be replicated and make it possible to compare the findings of different studies (see also the discussion on the need for objective measures in the paper by Treffers-Daller, this special issue). There is, however, a methodological problem when comparing test scores from structurally different languages since the unit of counting might not be the same. Turkish, for example, is an agglutinating language in which fewer but longer words are used to convey the same information compared with German (see Schwenk 1988: 23). Therefore, a Turkish text normally has fewer words than a German translation and the number of words as the unit of counting in a measure might therefore be problematic. Various tools have been developed to allow comparisons between different languages (Daller 1999, Daller, van Hout, Treffers-Daller 2003; Treffers-Daller, Özsoy and van Hout 2007; Daller, Treffers-Daller and Furman in press) but to our knowledge no such measure is based on the analysis of oral fluency.

Fluency and related aspects of oral proficiency are good candidates for the analysis of language dominance for theoretical and for practical reasons. From a theoretical viewpoint fluency can give insight into the level of automaticity in a given language and therefore insights into the structure of underlying processing of language (see Segalowitz 2003 and the discussion below). An analysis of aspects of fluency of the two languages of a bilingual has therefore the potential to reveal structural differences between the proficiency in both languages and can therefore be a good indicator of language dominance. From a practical point of view, the measurement of fluency might be the right step forward as more and more automated measures are under development and time consuming transcriptions might not be necessary in the near future (De Jong and Wempe 2009). This will enable researchers to carry out studies with much larger samples than are currently typical and will certainly further our insight into bilingual proficiency.

In the present paper we first give an overview of the concepts of bilingual dominance and fluency. The methodological section follows, after which we present the analysis of three types of measures, the C-test, manual fluency measures and automated fluency measures. We also develop two measures of language dominance that are

despite (or because of?) their simplicity valid tools in the analysis of language dominance in bilinguals. The latter measures are based on spoken language and could therefore also be used in bilingual settings where the participant cannot read or write in one of their languages.

1.1 Language Dominance

Most researchers agree that a balanced bilingual with equal proficiency in both languages is hardly found in real life and that there is almost always a dominant language (e.g. Grosjean 1982, Olsson and Sullivan 2005). This is because bilinguals usually use their languages in different domains and children are usually more exposed to one language, usually the language of the dominant environment, and they need to use this language, e.g. in school, more than the other. Grosjean (1997: 165) coined the notion complementary principle to describe the typical bilingual who uses two or more languages in different domains for different purposes in daily life and will therefore develop a dominant and a less dominant language in each of the different domains. This is also true for children who are exposed to two languages simultaneously from a very early age (see Bosch & Sebastián-Gales 2001:73).

Whilst there is agreement that one language is usually dominant, there is no agreement on the definition of dominance and this concept has been described in various ways using the expressions such as “bilingual preference” (Dodson 1985, Caldas & Caron-Caldas 2000) or “language complexity” (Döpke 1992). Yip and Matthews (2006) make the distinction between dominance and proficiency. The first concept is related to the underlying knowledge of a language whereas the latter is related to language use. But even if we accept this distinction, it is clear that we can only measure language use in a given situation and draw from it conclusions about the underlying dominance structure of the languages involved. Therefore, in the present study we measure proficiency in both languages and try to draw conclusions about the underlying dominance patterns. The situation becomes even more complex if we accept that bilinguals can have a dominant language in one communicative context and a different one in another (see Caldas & Caron-Caldas 2000: 366) and that dominance can shift over time between the languages. For these reasons time variables have to be part of the research design to trace possible changes in shift.

Various measures have been used to operationalize proficiency in two languages in order to establish dominance. Yip and Matthews (2006) use the mean length of utterance (MLU) but one has to bear in mind that their study focuses on dominance in young children. MLU was originally developed for the description of language development in younger children (see Bowen 1998) and the interpretation

of this measure for L1 development of older children (48 months and older) is difficult (see Klee and Fitzgerald 1985). We therefore conclude that this is not a suitable measure for the present study because our participants are young adults.

Others (e.g. Bullock, Toriba, Gonzáles and Dolola 2006) focus on phonetic features such as the voice onset time (VOT). They find that VOT values of the non-dominant converge towards the values of the dominant language while the VOT of the dominant language are less influenced by the non-dominant language. Although Bullock et al. (2006) do not attempt to measure dominance, their findings could be used to determine dominance on the basis of VOT values. Other studies use measures of vocabulary knowledge to establish dominance (Paradis, Tremblay and Crago 2008). They use standardized vocabulary tests to establish language dominance in French-English bilingual children, parental questionnaires and tests for the acquisition of inflectional morphology in English and found high correlation between all measures involved.

Most of the studies discussed so far have a focus on young children. Much more research is necessary to establish dominance patterns in adults (see also the study by Treffers-Daller in this issue). One of the challenges of the present study is to compare measures in languages that are typologically different. German is a West-Germanic language and shares many properties with English. Turkish, on the other hand, belongs to the Altaic language family and is often seen as a prototype of the agglutinating language. The structural differences might make it difficult to compare scores between these two languages as the unit of counting might be different (e.g. three words in German might be expressed by only one word in Turkish).

1.2 Fluency

The definition of fluency is difficult and varies across different contexts (see Chambers 1997, Hilton 2008). For the layperson somebody is fluent when s/he speaks a language at a native or near-native level. In this case fluency becomes interchangeable with oral proficiency. In communicative language teaching fluency is quite often opposed to accuracy and is seen as a feature of performance rather than knowledge. For researchers in second language acquisition fluency is, however, only one component of oral proficiency. There seems to be general agreement that automaticity is closely related to the concept of fluency. According to Schmidt (1992: 358) “fluency in speech production is automatic procedural skill”. Segalowitz (2003) also relates the concepts of fluency and automaticity in performance and points out that it is not the mere speed of speech that characterizes automatic processes but that automaticity “refers to a significant change in the way processing is carried out”

(2003: 313). The restructuring of underlying processes during second language acquisition may, however, lead to a higher speech rate in more advanced learners. Wood (2001) sees automaticity and formulaic competence as the two main factors for fluency. These two aspects are certainly related and a fluent speaker can buy time during speech production through the use of prefabricated units as Pawley and Syder (1983) stated more than 25 years ago. They show that only “a minority of spoken clauses are entirely novel creations” (1983: 205) and that fluent speakers rely on memorized sentences and on lexical sentence stems, units that are mainly prefabricated.

Tavakoli and Skehan (2005) make an attempt to define fluency by introducing the concepts of breakdown fluency (pauses), repair fluency (false starts and repetitions) and speed (speech rate). Although a high speech rate is not necessarily an indication of fluent speech, the measurement of fluency usually includes time-related aspects of performance. Given the difficulties involved in defining the concept of fluency, Chambers (1997) sees temporal measures as a “useful anchorage”. These temporal measures can include “speech rate” as defined as syllables per second which comprises “articulation rate” (actual speech) and “pause time”. Wood (2001) lists a variety of temporal variables that are all relevant for fluency such as speed, pauses, hesitations and fillers. Early studies on fluency (e.g. Goldman-Eisler 1961) already focus on the importance of hesitation pauses for the measurement of fluency. Various measures are used by Goldman-Eisler, including the ratio of pause time to number of words in seconds, the words produced per pause or the length of word sequences uttered without break. Breaks were defined in Goldman-Eisler’s study as gaps in speech of more than .25 seconds. An overview of measures of fluency can be found in Hilton (2008). She reports several “baselines” for speech rate, for the English native speaker a rate of 2 – 3 words per second and a total amount of pauses of one-third of the performance time seems to be the norm. With reference to Chambers (1997) Hilton points out that the articulation rate (speed of speech) as such is not necessarily an indication of fluency rather than an appropriate amount of pauses at the right place. However, an extremely low speech rate (less than 50 words per minute) is seen as clinical disfluency. Many measures of fluency are time-related and use syllables as unit of counting. It should be noted that there are two ways to calculate speech rate, and that these calculations reflect different concepts. The first way refers to speaking rate in the sense that the number of syllables or words per time unit is calculated including pauses. This measure in effect encompasses a measure of pausing together with a measure of speech rate. The second measure can be referred to as articulation rate which is calculated measuring syllables, words, or even phones per time unit excluding pauses. Articulation rate reflects tempo of speech between pauses, rather than a combination of pausing behaviour and speech rate together. Most studies investigating L2 fluency

have used speech rate including pauses which makes it difficult to disentangle pausing behaviour from actual articulation rate. There are also measures that use the number of words as the unit of counting rather than syllables. O'Brian, Segalowitz, Freed and Collentine (2007) use next to measures of "fluidity" (mean run length in words between pauses, defined as silence longer than 400 ms, mean run containing no filled pauses, and longest run without filled or unfilled pauses) simply the number of words that a person uses in a two minute interview. They found this quite simple measure a useful tool for the definition of general oral ability. This means that the very basic measure "words per minute" is also a good candidate for the measurement of fluency. For this measure a baseline given by Levelt (1989: 22) is that: "Speech is normally produced at a rate of about two to three words per second". This is of course only an approximation and does not take into account norms of speech rate that vary between languages. However, one has to bear in mind that purely quantitative measures are not the only way to measure fluency. The appropriateness of pausing was mentioned earlier. Nevertheless, it is beyond the scope of the present study to include this aspect and we restrict ourselves to the measurement of quantitative, mainly time-related, aspects of fluency in the present study.

We conclude that there is a general agreement on the usefulness of measures for fluency but that the exact operationalization varies from study to study. The main focus of previous studies is on fluency in a second language and L1 speakers are generally seen as fluent by definition with the exception of clinical cases. Studies that use time-related measures in both languages of bilinguals are very rare, and the novelty of the present study lies in the fact that attempt is made to use measures of fluency to reveal underlying bilingual dominance patterns. Therefore, we analyse picture descriptions in both languages (see section 2.2). We define pauses as no speech for at least .2 seconds and base various measures of fluency on this for our automated analysis (see section 3.3). In the manual analysis (section 3.2) we use the total number of words per picture description and the number of words per second as measure of overall oral proficiency. In addition to the raw scores we use adjusted scores to take the structural differences between the two languages into account.

1.3. Language proficiency and social background

A further focus of this article is on the biographical background of the bilingual participants and to what extent biographical variables can predict language dominance and bilingual proficiency. There seems to be a clear link between education and professional background of parents and the language proficiency of their children in general. Wells (1985) found that the father's education correlates

significantly with early language development in pre-school years and that the professional background of the mother and the father are predictors for language development as well. Wells (1986) found high correlations for the family background of 10 year olds and their school achievement including L1 proficiency. Likewise Hart and Risley (1995) found that socioeconomic status was a good predictor for the amount of time that parents talked to their children and for the complexity of the utterances that young children heard. The socioeconomic status of the parents correlated significantly with vocabulary growth of three year olds and the number of different words children used. Furthermore the size of children's vocabulary was highly related to the size of the parents' vocabulary. These findings may only be valid in the specific L1 setting where they were carried out and they might not be replicable in other settings. Furthermore, all these studies focused on monolingual young children as participants who were learning their first language. Data about the relation between the educational and professional background of parents and the bilingual proficiency of young adults are rare. We assume that the findings from monolingual L1 acquisition studies apply to the bilingual language acquisition of older learners as well.

1.4 Hypotheses and aims of the present study

The aim of the present study is to develop and to test measures that can reveal different language dominance patterns in bilinguals who speak structurally different languages. For this aim we will test the following hypotheses:

1.4.1 A control group of monolingual Turkish L2 learners of German will have much lower C-test scores for German when compared with German-Turkish bilinguals who grew up in Germany, but the monolingual Turkish group will have higher scores in Turkish.

1.4.2 The C-tests scores in Turkish and German will show differences in language dominance patterns between the two groups.

1.4.3 The different language acquisition history of the two groups will have an influence on the proficiency levels in both languages.

1.4.4 The social background (occupation of the parents, especially of the father) will have an influence on the proficiency levels in both languages.

We also will investigate whether fluency indices (e.g. the ratio between fluency measures in both languages) will be a useful tool for the measurement of dominance in structurally different languages and whether automated fluency analyses will lead to similar results as precise manual measures (such as words per minute).

2. Method

2.1. Participants

All participants in the present study come from a so-called Anadolu school in Istanbul. The Anadolu schools in Turkey are secondary schools with a long tradition in teaching foreign languages and teaching other subjects in a foreign language as well. The school in the present study has a clear focus on German. Most students are monolingual Turkish speakers who grew up in Turkey and learn German as a foreign language at this school. The focus on German makes this school also attractive to Turkish children who grew up in Germany and went back to Turkey during their school career, the so-called “returnees”. The present study is based on a comparison between the monolingualⁱⁱⁱ Turkish speakers (the control group) and returnees from Germany (the bilingual group), all of whom were attending this school.

The bilinguals in the present study are children of Turkish immigrants to Germany who grew up in Germany speaking German and Turkish. They were either born in Germany or came there at a very young age (see Table 1 for more details). They attended schools in Germany and returned to Turkey at an average age of 15.6 years. In Germany the dominant language was German in many cases but after their return a shift towards Turkish took place (see Daller 1999; Daller, van Hout and Treffers-Daller 2003; Daller, Treffers-Daller and Furman, in press).

The students in the control group had an almost entirely monolingual Turkish upbringing and are L2 learners for German who hardly have any contact with German outside the classroom. A few of them visited German speaking countries as tourists, but if they did it was only for a few weeks. At the time of the data collection (beginning of class 10) the students who learn German as a foreign language had had about 400 hours of German. Therefore, both groups have completely different language backgrounds with regard to their acquisition and use of their languages. An overview of the biographical data for both groups is given in table 1.

Table 1 Biographical data for the two groups

group	n	Age in years	Age on arrival in Germany	Age at return to Turkey	Length of residence in Turkey (years)
Monolinguals (control group)	56	15.35 (SD 0.55)	n/a	n/a	always
Bilinguals (returnees)	60	16.58 (SD 1.13)	1.53 (SD 3.87)	15.58 (SD 1.20)	1.0 (SD 0.79)

2.2 Data collection

Two different types of data were collected in the present study: C-tests in both languages and narratives in both languages.

The C-test is generally seen as a measure of general language proficiency because it normally correlates highly with a variety of other tests that aim to measure sub skills such as reading and listening comprehension, writing or even oral proficiency (see Eckes and Grotjahn, 2006). It has been argued recently (Daller and Xue 2009) that the reason for these high correlations is the clear focus of the C-test on vocabulary knowledge. This makes the C-test a useful tool for the measurement of general language proficiency as vocabulary knowledge is an important component of language proficiency. It has been successfully administered in a variety of languages, including Turkish (see Daller, Treffers-Daller, Ünaldı-Ceylan and Yıldız 2002). A C-test usually consists of a number of texts with 20 gaps in each. From the second sentence onwards in each text every second half of every second word is deleted. Based on the results of pilot studies we constructed a German C-test with 7 texts (items) and a Turkish C-test with 6 texts, which means that there were 140 gaps in the German C-test and 120 in the Turkish one (one Turkish text had to be deleted after the pilot study because it did not contribute to the overall reliability of the test). We administered the Turkish and the German C-tests during class time at two different days. The test was “timed” in the sense that only one teaching hour was available for each test but most students finished the tests in a much shorter period so that time pressure is unlikely to play a role. The C-tests were marked by speakers of each language and exact scoring was applied, which means that only

one answer was accepted as correct and there was no weighting between different types of errors (see also Daller and Grotjahn 1999: 159).

The narratives consisted of descriptions of picture stories (father-and-son stories, Plauen 1986). In total there were four picture stories and we elicited descriptions of all four stories in both languages. For practical reasons two stories were put together and were described by each participant in succession. We attempted to make four recordings with each participant, two in Turkish and two in German, which was, unfortunately, not always possible for practical reasons. The two sets of picture stories are the “ball story and the bee story” (Ball/Bee) and the “bank story and the dog story” (Bank/Dog). They have been successfully used in previous studies with bilingual speakers (for an overview see Daller, Treffers-Daller and Furman, in press). The data collection took place in the schools during normal teaching hours. The Turkish data were elicited by a native speaker of Turkish^{iv}; the German data were elicited by a native speaker of German. The students were asked to produce a description for an imaginary third person. The interviewers were instructed to intervene as little as possible and to limit their talk to an introduction at the beginning which was not part of the recording. The participants had as much time as they wanted to look at the picture stories before they started with their description. In order to compensate for memory effects we did not ask the students to describe the same stories in both languages on the same day. Half of the students were asked to describe the stories first in German and two days later in Turkish. The other half did the picture description in the reverse order with Turkish first. The descriptions were then transcribed by native speakers of the two languages. Contributions by the interviewers were cut out of the recordings where technically possible and were not transcribed. Due to various factors, such as absence of students on specific days we could not collect complete data sets for all students. For this reason “n” is different in some of the computations carried out. In addition we administered a questionnaire to the returnees about the schools attended in Germany, their social background, migration history and the languages used at home. A shorter version of this questionnaire was also administered to the control group.

2.3 Data analysis

The picture descriptions were transcribed and analysed in two different ways for language dominance and fluency. In a first, manual analysis, we counted the number of words per second over the total run of the recordings including pauses using the programme “Cool Edit 2002”. In a second analysis, we computed automated measures using scripts written in the software programme Praat (Boersma &

Weenink 2007) for various aspects of fluency (for a discussion on automated syllable detection with Praat see De Jong and Wempe 2009). We first report on the manual analysis and then compare it with the results for our automated measures.

3. Results

3.1 The C-test scores

As a first measure of language dominance we analysed the C-test scores. One can argue that the C-test scores between the two languages are not directly comparable because of the different language structures but there is no methodological problem to compare the C-test scores of the two groups within each language. Both C-tests turned out to be highly reliable. Cronbach's alpha is .944 for the German test (7 items) and .928 for the Turkish test (6 items). A value of .6 or above for Cronbach's alpha is acceptable for scientific purposes but for decisions about individuals .9 or above is required (Meuffels 1992: 146/7). A value for Cronbach's alpha higher than .9 means that the C-tests in the present study fulfil the highest possible criterion for the reliability of a test. In total 116 participants filled in the C-tests, 60 returnees and 56 members of the control group. The results as a percentage for the two groups are given in Figure 1 and Figure 2:

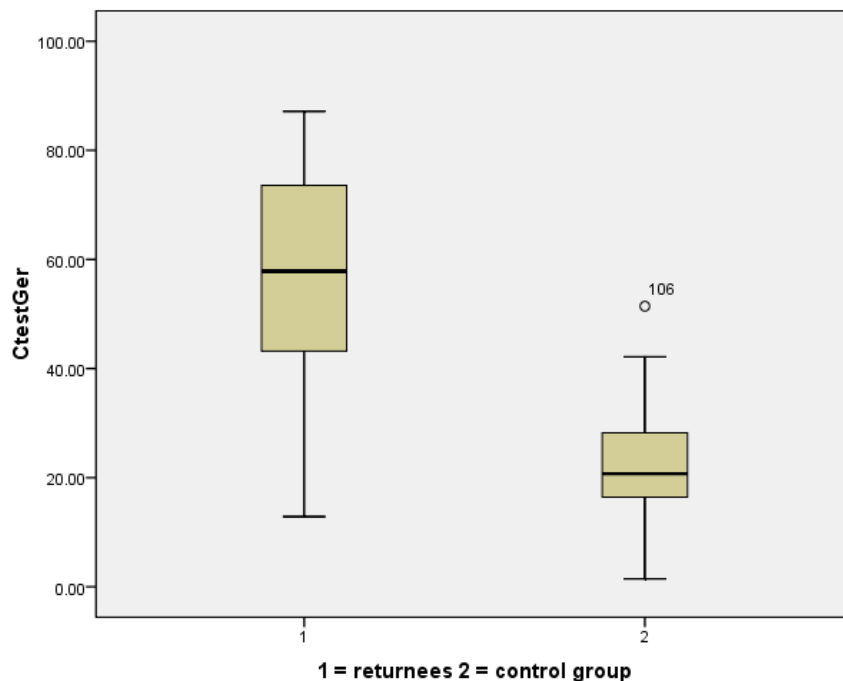


Figure 1 C-test scores for German (in percent)

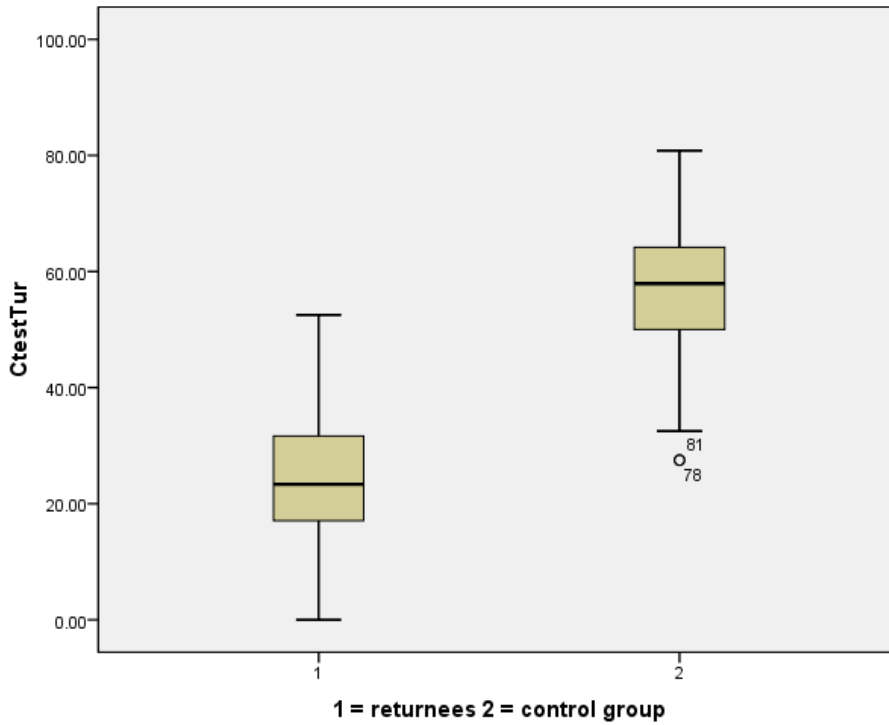


Figure 2 C-test scores for Turkish (in percent)

Table 1 gives the mean and the standard deviation for both tests:

Table 1 C-test scores in percent

		N	Mean	St.Dev.
German C-test	returnees	60	58.14	18.31
	control group	56	22.26	9.61
Turkish C-test	returnees	60	24.64	11.28
	control group	65	56.67	11.37

The C-test scores for each language show almost a mirror image. The returnees obtain scores in German that are roughly at the level of the Turkish scores for the control group. On the other hand the Turkish scores of the returnees are as low as the German scores for the control group for whom German is an L2. The differences between the groups are both significant (t-test for German: $t = 13.342$, $df = 90.519$, $p < .001$, equal variance not assumed; for Turkish: $t = 15.223$, $df = 114$, $p < .001$, equal

variance assumed). The control group is clearly dominant in Turkish which is in line with the expectations. The returnees are more dominant in German when compared with the control group. The C-test results make it likely that for the returnees German is still the dominant language. However, this can be different for individual members of this group and the C-test results as such do not provide a measure of absolute dominance for individuals as any cut-off point from which one language would be seen as being dominant would be arbitrary because the scores are not directly comparable between languages. Nevertheless, it is clear that the groups as a whole have a different language dominance profile which is important for the purpose of the present study.

For the returnees ($n = 59^v$) there are significant correlations between the C-test scores in German and age on arrival in Germany ($r = -.371$, $p < .001$), years spent in Germany ($r = .311$, $p < .05$) and school years in Germany ($r = .382$, $p < .01$). This is in line with the expectations. Those who came later to Germany or spent less time there have lower C-test scores. It is, however, surprising that no significant correlation between biographical data (years in Turkey after return) and the C-test scores in Turkish can be found for the returnees ($r = .239$, $p = .071$, two-sided, $n = 58$). There are also no significant differences between the C-test scores of female and male returnees, neither for Turkish nor for German.

By contrast, female students of the control group have higher scores than male students in the C-tests in both languages and the differences are significant (Turkish C-test: $t = 2.904$, $df = 53$, $p < .01$; German C-test: $t = 2.271$, $df = 53$, $p = .027$). The questionnaire with the data for the social background included questions about the professional background of the parents. These were ranked into three categories: manual, administrative, professional. For the control group there is a significant correlation between the test scores in German and the occupation of the mother ($\rho = .281$, $p = .039$, $n = 54$) but not with the occupation of the father. Students with mothers of a higher professional background achieve higher scores in the German C-test for .

As mentioned earlier there are clear indications that the C-test format taps mainly into vocabulary knowledge. To investigate this further we correlated the C-test scores for the returnees with the number of words that were produced in the picture descriptions. For German there was a significant correlation of $r = .569$ (Ball/Bee story, $p = .022$, $n = 16$) and of $r = .419$ (Dog/Bank story, $p = .047$, $n = 23$) but there was no significant correlation for Turkish for this group nor was there a significant correlation between the Turkish C-test scores of the control group and the word length of their descriptions. The focus on vocabulary of the C-test format can therefore only be established in the present study for German.

3.2 Manual measures of fluency

3.2.1 Turkish picture descriptions

The picture descriptions were first analysed manually with regard to “words per second” in Turkish. As noted above, this measure would be difficult to compare between Turkish and German as both languages are structurally quite different but a comparison within one language is possible. Figure 3 shows the results for the returnees and the control group in Turkish (Ball/Bee story).

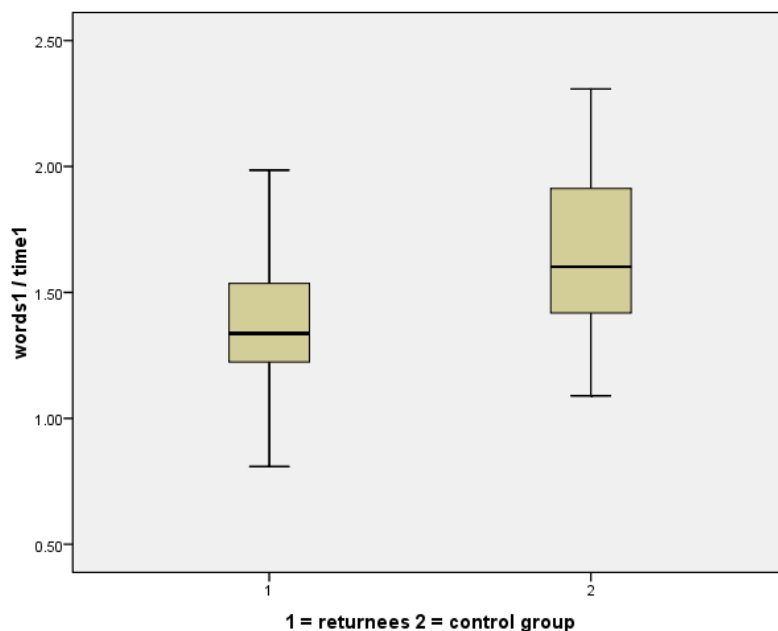


Figure 3 words per second in Turkish descriptions

The control group (n = 37) produce more words per second than the returnees (n = 34). The difference between the two groups is significant (t = 4.138, df = 69, p < .001).

For the returnees there is a positive significant correlation between the “words per second” and the professional background of the mother (rho = .410, p = .020, n = 32)^{vi}. For this group there is no correlation between the father’s background and the

variable “words per second”. No other biographical data, such as age on arrival in Germany or years spent in Turkey after return showed any significant correlation with fluency as measured in “words per second”. The difference in the total number of words produced by the two groups is also not significant.

3.2.2 Comparisons between Turkish and German

The comparison in Figure 3 is based on picture descriptions only in Turkish. A comparison between the languages is difficult because, as noted above, the agglutinating structure of Turkish leads to longer words than in German and therefore to fewer words when the same content is expressed. Turkish translations from German have therefore normally fewer but longer word tokens than the original and German translations from Turkish have normally more word tokens. We collected 9 parallel texts from the Internet that were translations from one language into the other. A Turkish-German bilingual judge confirmed that the content of the parallel texts was exactly the same. The Turkish texts were on average shorter by a ratio of .8957. This means that the number of tokens in the Turkish texts was nearly 10% lower than in the German parallel texts. For 13 returnees we had picture descriptions of the same stories (Ball/Bee) in both languages. We adjusted the number of tokens in German according to the ratio mentioned above. The results are shown in Figure 4.

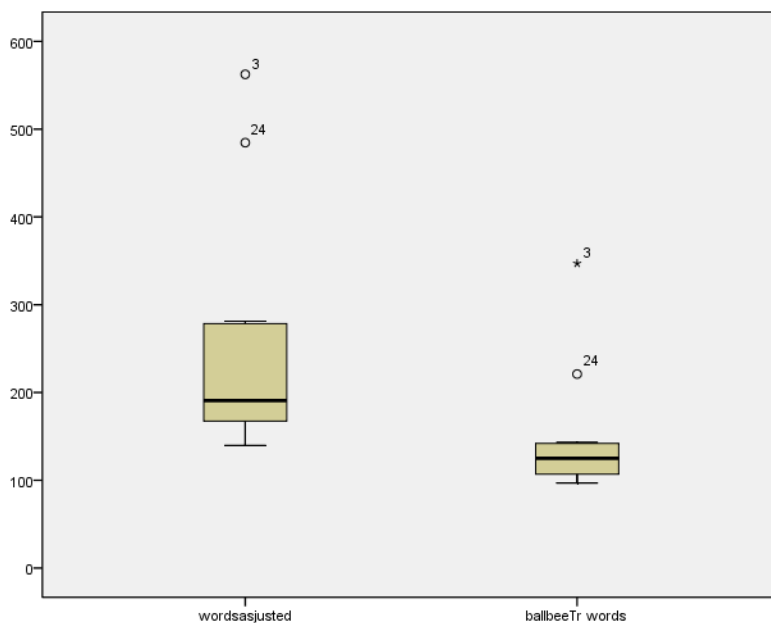


Figure 4 German words adjusted and Turkish tokens in the Ball and Bee story narrated in both languages by 13 returnees.

It is clear that the same persons use many more word tokens in German to describe these stories than they do in Turkish. The differences between the length of the descriptions in the two languages are significant (paired t-test, $t = 5.102$, $df = 12$, $p < .001$) and so is the correlation between the text length in the two languages ($r = .892$, $p < .001$, $n = 13$). This means that speakers who produce longer texts in one language also produce longer texts in the other. It is also clear for this group that there is a systematic tendency to produce shorter texts in Turkish even when compared with the adjusted text length in German. Interestingly the two outliers in Figure 4 follow exactly this pattern. This is further evidence that German is the dominant language for this group.

3.2.3. Measures of language dominance

The relation between the text lengths in both languages is an indication of the bilingual dominance of the speakers. We therefore computed a ratio between the number of words in German (adjusted) and in Turkish. The higher this ratio is the higher the dominance towards German might be. This is confirmed by the significant correlation of this ratio with the German C-test scores ($r = .594$, $p = .032$, $n = 13$). It also correlates significantly negatively with age on arrival in Germany ($r = -.628$, $p = .022$, $n = 13$), and positively with years in Germany ($r = .567$, $p = .043$, $n = 13$) and school years in Germany ($r = .569$, $p = .042$, $n = 13$). The only other significant correlation of this ratio is with age ($r = -.583$, $p = .037$, $n = 13$). Overall, these results confirm that the ratio of text length of the descriptions in each language is a good indicator of language dominance and that the language contact in the early years and at school still determines the direction of this dominance.

The findings so far are based on a small subsample of the bilingual group ($n = 13$) since it is only for this subsample that data are available for the same stories in both languages. To get further insight into the measurement of language dominance we need to compare the two groups. The data from the control group are important to contrast them with the returnees. The control group is clearly dominant in Turkish and they are only at beginner's or lower intermediate level in German. A comparison between the returnees and the control group is only possible on the basis of a different set of descriptions, the "Dog/Bank" stories in German and the "Ball/Bee" stories in Turkish. Table 2 shows the text length for these descriptions of the control group and returnees (raw scores).

Table 2 text length of the picture description in both languages

Group	German stories	Turkish stories
	Mean and (Std.Dev)	Mean and (Std.Dev)
Returnees (n = 15)	312.07 (175.0)	148 (66.51)
Control group (n = 20)	72.15 (59.31)	119.05 (39.87)

The returnees produce significantly more words in German than the control group ($t = 5.732$, $df = 33$, $p < .001$) but for Turkish the differences between the groups are not significant. More important than the actual scores (number of words) for the determination of language dominance, is the relationship between the scores in the two languages. The difference between the scores for the returnees leads to a value for Eta^2 of .499 which is much higher than the value for the control group ($\text{Eta}^2 = .125$). In order to find a measure of language dominance we first computed the difference between the scores in each language. First we computed this difference based on the raw scores in both languages (difference 1), then we computed it based on the adjusted scores in German and the raw scores in Turkish (difference 2). The value for these measures is clearly different between the two groups as shown in table 3.

Table 3 Scores for the measure “difference” with raw and adjusted scores

<i>measure</i>	<i>Group</i>	<i>Mean</i>	<i>Std.Dev.</i>
<i>Difference 1 (based on raw scores)</i>	<i>Returnees (n = 15)</i>	<i>163.87</i>	<i>119.00</i>
	<i>Control group (n = 20)^{vii}</i>	<i>-36.90</i>	<i>79.42</i>
<i>Difference 2 (based on adjusted scores for German)</i>	<i>Returnees (n = 15)</i>	<i>131.32</i>	<i>101.43</i>
	<i>Control group (n = 20)</i>	<i>-44.43</i>	<i>74.04</i>

For both measures the control group have a negative difference which means that their Turkish descriptions are longer and the returnees have a positive difference which means that their German descriptions are longer. In both cases the difference between the two groups is significant (difference 1: $t = 5.987$, $df = 33$, $p < .001$, difference 2: $t = 5.932$, $df = 33$, $p < .001$). This suggests that these measures are valid indices of language dominance. A negative difference indicates Turkish dominance and a positive difference indicates German dominance. To support the validity of this measure we computed the correlation between the measures and the C-tests scores in both languages. Table 4 shows the correlations between the C-test scores and the difference between the text length based on the raw scores and the adjusted scores in German and the scores in Turkish.

Table 4 correlations (adjusted scores in brackets, $n = 35$ where not stated otherwise)

	“difference”	German C-test	Turkish C-test
“difference”	-	.694** (.690**)	-.635** (-.633**)
German C-test		-	-.620** ($n = 115$) ^{viii}

** . Correlation is significant at the 0.01 level (2-tailed).

The correlations show that a high difference is significantly positively related to high scores in the German C-test and to low scores in the Turkish test. The variable “difference” is therefore an indication of language dominance.

An alternative to the “difference” measure is to compute a ratio between the scores in both languages. A ratio based on adjusted scores and on raw scores would lead to the correlations as the scores are all adjusted by the same proportion. We therefore computed the ratio only based on the raw scores and compared them with the C-test results. The results are shown in table 5.

Table 5 correlations (n = 35)^{ix}

	Ratio	German C-test	Turkish C-test
Ratio	-	.631**	-.656** (-.633**)
German C-test		-	-.689**

** . Correlation is significant at the 0.01 level (2-tailed).

There is a significant positive correlation between the ratio and the C-test scores in German and a significant negative correlation between the ratio and the C-test scores in Turkish. This means that a ratio is a good index of language dominance as well. As mentioned, the advantage of a ratio is that there is no difference between adjusted and raw scores when used in a correlation and it is therefore a useful tool when comparing dominance in structurally different languages.

3.2.4 The validity of the dominance measures

The validity of these two measures for the determination of language dominance can be supported further with an analysis based on a logistic regression. This statistical procedure predicts group membership of individuals (a categorical variable) on the basis of a continuous or categorical variable. If the measures discussed so far are valid measures for language dominance then they must be able to predict group membership of individuals. In our case the difference between text lengths or the ratio between text length in Turkish and German should allow us to determine to a certain degree whether a person is a returnee or a member of the control group. Since there is a high intercorrelation between the two predictor variables (*difference* and *ratio*) logistic regression was carried out for each variable separately based on the unadjusted scores. A logistic regression (method: enter) for a model with a constant and the variable "difference" leads to a model with an $R^2 = .739$ (Nagelkerke). The variable "difference" makes a significant contribution to that model at $p < .003$. The classification table for this model shows to what extent the allocation of group membership was correct.

Table 6 Prediction of group membership based on the variable “difference”

Observed	Predicted		Percentage
	Returnees	Control group	
Returnees (n = 15)	14	1	93.3
Control group (n = 20)	2	18	90.0
Overall percentage correctly predicted			91.4

This means that 91.4 % of the classifications based on the logistic regression with “difference” as predictor variable were correct. In 14 out of 15 cases returnees were classified correctly and in 18 out of 20 cases control group were classified correctly.

The validity of the measure “difference” has been investigated further with a discriminant analysis and a cross validation with a leave-one-out classification. With this method the group membership of each case (participant) is classified with a discriminant function based on all cases except the given case. We used “difference” as the independent and group membership as the dependent variable. The results are significant (Wilks’ lambda = .479, $\text{Chi}^2 = 23.896$, $\text{df} = 1$, $p < .001$) and 88.6 % of the cross-validated cases are correctly classified. The only difference between this analysis and table 6 is that 13 instead of 14 returnees are classified correctly.

We also carried out a logistic regression (method: enter) for a model with a constant and the variable “ratio” which leads to a model with a slightly lower $R^2 = .612$ (Nagelkerke). Again the variable “ratio” makes a significant contribution at a slightly lower p-value of $p < .001$. The classification table for this model shows a slightly lower percentage of correctly predicted group membership.

Table 7 Prediction of group membership based on the variable “ratio”

Observed	Predicted		Percentage
	Returnees	Control group	
Returnees (n = 15)	13	2	86.7
Control group (n = 20)	3	17	85.0
Overall percentage correctly predicted			85.7

A discriminant analysis for the measure “ratio” leads to significant results (Wilks’ lambda = .499, $\chi^2 = 22.598$, $df = 1$, $p < .001$) and 88.6 % of the cases cross-validated with the leave-one-out method are correctly classified. The only difference between this analysis and the one presented in Table 7 is that 14 instead of 13 returnees are classified correctly.

Overall, we conclude that the manually computed measures based on the length of the picture descriptions provide a useful and valid tool to predict group membership. Further research would be needed to decide which of the manual measures, the difference or the ratio between text lengths, is more useful.

3.3 Automated analysis of fluency

We calculated fluency measures using scripts written in the software programme “Praat”. To measure speech rate, we used the method explained in De Jong and Wempe (2009). They describe a script that automatically detects syllables in running speech in order to measure speech rate without the need of a transcription. The script takes sound files as input and detects syllables using intensity (loudness) and voicedness, because syllable nuclei are (usually) voiced peaks in intensity. In a validation study, they report a correlation between manually measured speech rate and automatically measured speech rate of .88 for sample of speech by speakers of Dutch as a second language. We used this script to measure speech rate and another script in Praat to detect pauses automatically (a modified version of which is now available under the button “To TextGrid (silences)” in Praat itself). Using these two scripts we computed the following measures:

- The total length speaking time excluding pauses (totrun)
- The total number of pauses (Npause)
- The total length of pausing time (totpause)
- The total length of performance (tot = totrun + totpause)
- The mean length of utterance in seconds between two pauses (Mlutt)
- The "phonation time ratio" which is the ratio between the total length of speaking time without pauses and the total length of performance
- The total number of syllables (Nsyl)
- Speech rate 1: the number of syllables per second over the total length of speaking time excluding pauses
- Speech rate 2: the number of syllables per second over the total length of performance including pauses.

First we analysed the Turkish descriptions (Ball/Bee) for the returnees only. Most recordings were of relatively poor quality due to the fact that the recordings had to be taken in a normal school setting with a lot of background noises, which is of course a major problem when automatically detecting pauses and syllables in sound files. Moreover, in such an automated manner it is impossible to distinguish between the speech of the participant and that of the interviewers. In most cases, the experimenter's speech was recorded on the same channel of the sound file. We cut out the contribution of the interviewers manually where possible. Due to noisiness of the sound files, in two cases zero pauses were reported for the Turkish of the returnees (Ball/Bee story) and in one case for their German (Dog/Bank story) which a manual analysis of the data did not confirm. These cases were excluded from the analysis. Most measures of fluency show significant intercorrelations in the expected direction, e.g. the mean length of utterance between pauses is negatively correlated with the number of pauses ($r = -.509$, $p = .005$, $n = 29$) which means that those who produce more pauses also have shorter utterances between pauses. The Turkish C-test results are not significantly correlated with the number of pauses or measures that are related to the length of performance, e.g. the number of syllables or the total length of speaking time. However, the Turkish C-test scores show a significant correlation with "speech rate 2", that is the number of syllables per second over the total length of performance ($r = .602$, $p = .001$, $n = 26$). This supports the assumption that automaticity and fast and efficient retrieval of vocabulary are relevant for the successful completion of C-tests (see the discussion on the C-test in section 3.1) Overall, the results show that speech rate in Turkish is an indication of a higher proficiency in this language. This finding supports the results of the manual analysis (see section 3.2. Figure 3).

We then analysed the German (Dog/Bank story) of the returnees with the automated measures. The German C-test scores of the returnees correlate significantly with measures that are related to length of performance, such as total length of speech

without pauses ($r = .483$, $p < .05$, $n = 22$), mean length of utterance in seconds between pauses ($r = .555$, $p < .01$, $n = 22$) and total number of syllables ($r = .489$, $p < .05$, $n = 22$). This is in line with the findings of the manual analysis where longer texts (measured in tokens) are an indication of higher proficiency for German but not for Turkish (see section 3.2.3). In contrast with the Turkish of the returnees there is no significant correlation with the C-test scores and speech rate measures. This means that for German the length of the descriptions (measures in time or tokens) is an indicator and for Turkish the speech rate is an indicator of higher proficiency for the returnees.

An analysis of the Turkish picture descriptions of the control group (Ball/Bee story) shows that speech rate is not related to the C-test scores in Turkish. Interestingly, the pausing seems to be an indicator of higher proficiency here. There are significant correlations between the C-test scores and pausing, such as the total length of pauses ($r = .419$, $p < .05$, $n = 35$) the phonation time ratio ($r = -.438$, $p = < .001$, $n = 35$) and the mean length of utterance between pauses ($r = -.422$, $p < .01$, $n = 35$). Note that these correlations are in the opposite direction than expected: many pauses, a low percentage of speaking time over total time (phonation time ratio) and shorter utterances between pauses are related to higher scores on the Turkish C-test for the control group. It is difficult to speculate about possible explanations for these results. One possible explanation that comes to mind has to do with extraversion. Previous research has shown that extraverts tend to speak faster with shorter silent pauses (e.g., Ramsay, 1968). Eysenck and Eysenck (1985) revealed that where extraverts tend toward faster performance compared to introverts, they also tend to perform less accurately compared to introverts. We can speculate that the speakers in the control group in our study that were pausing less in the narratives also tended to be less accurate on the C-test. This would be in line with the previous research on extraversion.

A comparison between the two groups based on the automated measures for the Turkish Ball/Bee story is given in Table 8.

Table 8

Automated measures for Turkish

measure	Returnees (n = 31)	Monolinguals (n = 35)	t-test	Significance
Total speaking time	72.8 (SD 35.4)	53.1 (SD 20.6)	t = 2.727 df 46.977	<.01
Number of pauses	35.7 (SD 26.2)	22.7 (SD 9.0)	t = 2.633 df 36.282	< .05
Total length of pausing time	28.2 (SD 21.2)	17.8 (SD 7.2)	t = 2.605 df 36.178	< .05
Total length of performance	101.1 (SD 52.6)	70.9 (SD 21.8)	t = 2.976 df 39.056	< .01
Mean length of utterance between two pauses	2.55 (SD 1.48)	2.76 (2.34)	t = .437 df 64	ns
Phonation time ratio	.745 (SD .1)	.741 (SD .1)	t = .201 df 64	ns
Total number of syllables	271.9 (SD 145.0)	204.4 (SD 79.4)	t = 2.303 df 45.242	< .05
Speech rate 1	3.67 (SD .52)	3.88 (SD .42)	t = 1.902 df 64	p = .062
Speech rate 2	2.70 (SD .30)	2.86 (SD .41)	t = 1.852 df 64	p = .069

The measure “mean length of utterance in seconds between pauses” and the measure “phonetimeratio” (ratio of total length of speech without pauses to total length of performance) are not significant.

All measures that are related the amount of text and pauses produced (number of pauses, total length of text of speech excluding pauses and number of syllables) show higher values for the returnees than for the monolinguals. The returnees make

more pauses but also speak longer even when pauses are excluded. However, the control group speak faster than the returnees and the differences between the two groups can be classified as a Trend (Speech rate 1: $t = 1.902$, $df = 64$, $p = .062$, Speech rate 2: $t = 1.852$, $df = 64$, $p = .069$; both tests two-sided). A possible interpretation is that the picture descriptions in Turkish take less effort for the control group as they are faster and that the returnees make more effort as they produce longer texts but also make more pauses.

We also wanted to explore whether the differences shown in table 8 can be used to predict group membership with a loglinear regression. This can, however, only be done on the basis of the Turkish data as no automated measures for the monolinguals in German were available for technical reasons. Nevertheless, this will give further insight into how measures of performances in Turkish alone can distinguish between the two groups. It is not possible to include all the variables into this computation as there are significant correlations between some of the measures (e.g. speech rate 1 correlates with speech rate 2; $r = .446$, $n = 66$, $p < .001$) and the problem of multicollinearity would arise. We therefore decided to include only three measures: "Total speaking time", "number of pauses" and "speech rate 1". These measure tap into three different aspects of fluency, such as the length of performance, pauses, and speed of performance. This decision for exactly these three measures was made on the basis of the t-value that each of the three measures obtained when compared with similar measures as shown in table 8. We decided to include those measures that lead to a higher t-value than related measures. We used the "backwards" method as recommended for explorative studies where there are two or more variables but no information about the predictive value of each (Field 2005: 227).

This leads to a model with $R^2 = .239$ (Nagelkerke) and only two predictor variables in the second step: "number of syllables" and "speech rate 1". The variable "total speaking time" is excluded from this model. 68.2 % of group membership was correctly predicted with the two variables remaining in the model as can be seen from classification table 9.

Table 9 Classification table for group membership^x

Observed	Predicted		Percentage
	Returnees	Control group	
Returnees (n = 31)	20	11	64.5
Control group (n = 35)	10	25	71.4
Overall percentage correctly predicted			68.2

Detailed information about the contribution of the variables can be seen in the list given in appendix 1. The variable “total speaking time” has been excluded in step 2 as not being significant. Not surprisingly the percentage of correctly predicted group membership is lower than in Table 7 since only one language was involved in this computation. Nevertheless almost 70 % can be predicted correctly based on the Turkish data alone.

A discriminant analysis with the two remaining variables leads to significant results (Wilks’s lambda = .823, $\chi^2 = 12.276$, df. = 2, p = .002) and 63.6 % of the cases cross-validated with the leave-one-out method are correctly classified.

4. Conclusions

The two groups have a clearly different language dominance profile, which is in line with the expectation given the different language acquisition history of the groups. Hypothesis 1.4.1 predicts that the returnees have significantly higher scores in the German C-test than the control group (L2 learners) but that the scores of the returnees are lower in Turkish than the scores of the control group. This hypothesis is clearly corroborated by the findings, with significant differences for all comparisons in the predicted direction. As stated earlier the C-test scores in the two languages are not directly comparable due to structural differences between Turkish and German and because the C-tests are based on different texts and therefore we cannot be entirely certain if higher C-test scores for one language than for the other

constitutes evidence for language dominance. If, for example, the C-test in Turkish were more difficult than the German C-test, a group with equal proficiency in both languages would yield lower scores in Turkish. A direct comparison between C-test scores in the two languages is therefore not possible. However, we can compare C-test scores *within* each language. When we compare the German C-test scores of the returnees with the German scores of the control group it is clear that the returnees are much more proficient in German; a similar comparison for Turkish shows that the returnees are much less proficient in Turkish than the control group. Both groups show almost a mirror image in their scores with one stronger and one weaker language. This confirms hypothesis 1.4.2 which assumes that the C-test scores can be used to predict differences in dominance patterns. Hypothesis 1.4.3 predicts a relation between biographical data and the proficiency levels in both languages. The C-test scores in German of the returnees reflect their language acquisition history and are related to biographical data such as date of arrival in Germany and school years there but the C-test scores in Turkish reflect biographical data such as years after return to a much lesser extent. The fact that there is only a very weak correlation between the Turkish scores and the time spent in Turkey after return is an indication that the returnees' language dominance profiles did not significantly change in the period after return to Turkey although individuals may well have made significant progress.

The specific language proficiency of the returnees is characterized by a very large difference between their proficiency in German and in Turkish. This is revealed by the two manual measures, the number of words per second and the total number of words in their Turkish descriptions when compared to the control group. Two variables of language dominance could be developed on the basis of the text lengths in both languages: the *difference* between the text lengths (tokens) of the stories in both languages and the *ratio* of the text lengths (tokens). The higher the values for these two variables are the more prominent is the proficiency in German in comparison with Turkish. As mentioned above it should be borne in mind that this is only a relative dominance between the two languages. Every cut-off point in absolute terms where a person would be judged as more dominant in one language than in the other would be arbitrary. However, both measures are valid tools to determine the specific form of language dominance of an individual or group in comparison with another individual or group. The two measures correlate significantly with the C-test scores in the predicted direction, positively with the German scores and negatively with the Turkish scores. We also tried to adjust the scores in German since structural differences between the two languages lead to systematically shorter texts in Turkish (as measured in token) due to the agglutinating structure of this language. However, this type of adjustment turned out to be not necessary for the validity of our measures. To investigate this validity further we used the variables *distance* and

ratio in a logistic regression to predict group membership of the individual participants (being a returnee or member of the control group). As the logistic regression demonstrates both variables are excellent predictors of group membership. This is an argument for the validity of the measures and we argue that this is not a circular conclusion as a prediction is made based on linguistic measures to predict biographical data (group membership). The validity of the measures is supported further with discriminant analyses and cross validations with the method “leave-one-out classification”. We therefore conclude that it is possible to develop indices of language dominance based on fluency measures (words per second) or measures of general oral proficiency (total number of words). These indices have a highly predictive power for group membership. We further analysed the Turkish data with automated measures of fluency. It turns out that the returnees produce longer texts and make more pauses but that the monolinguals are faster. A possible interpretation is that the picture descriptions in Turkish take less effort for the monolinguals and that the returnees make more effort to describe these pictures. This is in line with the manual analysis where speech rate in Turkish is one of the variables that distinguish the returnees from the monolinguals in Turkish (see Figure 3). The fact that the returnees produce longer texts in absolute terms in Turkish than the monolinguals is in line with the manual analysis (see Table 2). Again with the automated analysis the speech rate is higher for the monolinguals than for the returnees in Turkish. Both variables together allow the prediction of group membership based on the automated measures with an accuracy of almost 70%. At this point we cannot decide whether manual measures or automated measures are more powerful in the prediction of dominance.

Hypothesis 1.4.4 predicts that the family background, especially the occupation of the father, has an influence on the proficiency of the students. This hypothesis could be partially corroborated in the present study. The variable “words per second” in Turkish is significantly related to the mother’s profession for returnees and control group. For the control group in addition the C-test scores in German are related to the professional background of the mother. This is at first surprising since earlier studies found that it is the professional background of the father that has an influence on the language proficiency of the children. However, one has to bear in mind that this is only a relatively small study.

Overall, we conclude that it is possible to create measures of language dominance in bilinguals who speak structurally different languages. A mathematical adjustment of the scores in one language might mitigate differences between the two language scores but turned out not to be necessary and the measures are still valid tools. Although some of these measures are simply based on the number of tokens that are produced by a person, they have a strong predictive power for the language dominance of this individual. Manual and automated measures are both promising

candidates as measures of language dominance. If further studies corroborate the predictive power of automated measures, they will clearly be preferable as with fewer resources larger samples can be analysed. We therefore need further studies that replicate the present analysis, preferable with different language sets.

5. Bibliography

BOERSMA, P. and WEENINK, D. (2007). Praat (Version 4.5.25) [Software]. Latest version available from www.praat.org.

BOSCH, L. and SEBASTIÁN-GALLÉS, N. (2001). Early differentiation in bilingual infants. In Cenoz, J. and Genesee, F. (Eds.) *Trends in Bilingual Acquisition*, 71-93, Amsterdam: Clintons.

BOWEN, C. (1998). Brown's Stages: The development of morphology and syntax, Retrieved from <http://www.speech-language-therapy.com/BrownsStages.htm> on (24/10/09).

BULLOCK B.E., TORIBA, A.J., GONZÁLES V. and DOLOLA, A. (2006). Language Dominance and Performance Outcomes in Bilingual Pronunciation. In Grantham O'Brian, M., Shea C. and Archibald, J. (eds.) *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA)*. Cascadia Proceedings Project.

CALDAS, S. and CARON-CALDAS, S. (2000). The influence of family, school, and community on bilingual preference: Results from a Louisiana/Quebec case study. *Applied Psycholinguistics*, 21, 365 - 381.

CHAMBERS, F. (1997). What do we mean by fluency? *System*, 25 (4), 535 - 544.

DALLER, H. (1999). *Migration und Mehrsprachigkeit*. Frankfurt a. M.: Peter Lang.

DALLER, H. and GROTJAHN, R. (1999). The Language Proficiency of Turkish Returnees from Germany: An Empirical Investigation of Academic and Everyday Language Proficiency. *Language, Culture and Curriculum*, 12 (2), 156 – 172.

DALLER, H., VAN HOUT, R. and TREFFERS-DALLER, J. (2003). Lexical richness in spontaneous speech of bilinguals. *Applied Linguistics*, 24 (2), 197 - 222.

DALLER H., TREFFERS-DALLER J., ÜNALDI-CEYLAN A. and YILDIZ C. (2002). The development of a Turkish C-Test. In Coleman, Grotjahn and Raatz (eds.) *University Language Testing and the C-Test*, pp. 187 - 199. Bochum: AKS-Verlag .

DALLER, M. H. and XUE J. (2009). English Proficiency and Academic Success: A Study of Chinese Students in UK Higher Education. In Richards B., Daller M.H., Malvern D, Meara P., Milton, J. and Treffers-Daller, J. (eds.) *Vocabulary studies in first and second language acquisition: The Interface Between Theory and Application*, pp. 179 - 193. Houndmills: Palgrave Macmillan.

DALLER, M.H., TREFFERS-DALLER, J. and FURMAN, R. (In press). Conceptual transfer in narratives by Turkish-German bilinguals: the description of motion events in Turkish and German. In Jarvis S. (ed.). *Bilingualism, Language and Cognition* (CUP).

DE JONG, N. H. and WEMPE, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behaviour Research Methods*, 41 (2), 385 – 390.

DODSON, C.J. (1985). Second language acquisition and bilingual development: a theoretical framework. *Journal of Multilingual and Multicultural Development*, 6 (5), 325-346.

DÖPKE, S. (1992). *One Parent, One Language: A Interactional Approach*. Amsterdam: John Clintons.

ECKES, T. and GROTJAHN, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, Vol 23 (3), 290 -325.

EYSENCK, H.J., AND EYSENCK, M.W. (1985). *Personality and individual differences. A natural science approach*. New York: Plenum Press.

FIELD, A. (2005). *Discovering Statistics Using SPSS*. London: Sage.

GOLDMAN-EISLER, F. (1961). The distribution of pause duration in speech. In Teddington (ed.). *Language and Speech* (Vol 4), 232-237.

GROSJEAN, F. (1982). *Life with Two Languages*. Harvard University Press.

GROSJEAN, F. (1997). The bilingual individual. *Interpreting: International Journal of Research and Practice in Interpreting*, (1/ 2), 163 -187.

HART B. and RISLEY T. (1995). Meaningful differences in the Everyday Experience of Young American Children. Brookes Publishing Co.

HILTON, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal*, 36 (2), 136 – 166.

- KLEE, T. and FITZGERALD, M.D. (1985). The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language*, 12, 251-69.
- LEVELT, W.J.M. (1989). *Speaking. From Intention to Articulation*. Cambridge (M.): MIT Press.
- MEUFFELS, B. (1992). *Methods and Techniques of Empirical Research*. SICSAT: Amsterdam.
- O'BRIAN, I., SEGALOWITZ, N., FREED, B. and COLLENTINE, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition*, 29 (4), 557-581.
- OLSSON, Å. and SULLIVAN, K.P.H. (2005). Provoking Dominance Shift in a Bilingual Swedish-American English 4-Year-Old Child. In Cohen, J., McAlister, K., Rolstad, K. and MacSwan, J. (eds.). *Proceedings of the 4th International Symposium on Bilingualism*, pp. 1750 - 1764. Somerville, MA: Cascidilla Press.
- PARADIS, J., TREMBLAY, A. and CRAGO, M. (2008). Bilingual Children's Acquisition of English Inflection: The Role of Language Dominance and Task Type. In H. Chan, H., Jacob, H. and Kapia, E. (eds.), *Boston University Conference on Language Development 32 Proceedings*, pp. 378-389. Somerville, MA: Cascadilla Press.
- PAWLEY, A. and SYDER, F. H. (1983). Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In Richards, J.C. and Schmidt, R.W. (eds.) *Language and Communication*, pp. 191-226. London: Longman.
- PLAUEN, E.O. (= Erich Ohser). (1986). *Vater und Sohn* (Band 2). Ravensburger Taschenbuchverlag.
- PRAAT URL: <http://www.fon.hum.uva.nl/praat/>
- RAMSAY, R. W. (1968). Speech patterns and personality. *Language and Speech*, 11(1), 54-63.
- SCHMIDT, R. (1992). Psychological mechanisms underlying second language fluency. *SSLA* 14, 357-385.
- SCHWENK, H. (1988). *Das Sprachvermögen zweisprachiger türkischer Schüler*. Tübingen: Gunter Narr Verlag.
- SEGALOWITZ, N. (2003). Automaticity and Second Languages. In Doughty, C. and Long, M.H. (eds.) . *The Handbook of Second Language Acquisition*, pp. 382 – 408. Oxford: Blackwell.

TAVAKOLI, P. and SKEHAN, P. (2005). Strategic planning, task structure and performance testing. In. Ellis, R. (ed.). *Planning and task performance in a second language*, pp. 239 – 276. Amsterdam: Benjamins.

TREFFERS-DALLER, J., ÖZSOY, S. and VAN HOUT, R. (2007). (In)Complete Acquisition of Turkish Among Turkish-German Bilinguals in Germany and Turkey: An Analysis of Complex Embeddings in Narratives. *International Journal of Bilingual Education and Bilingualism*, 10 (3), 248 – 276.

WELLS, G. (1985). *Language development in the pre-school years*. Cambridge University Press.

WELLS, G. (1986). *The Meaning Makers. Children Learning Language and Using Language to learn*. London: Hodder and Stoughton.

WOOD, D. (2001). In Search of Fluency: What Is It and How Can We Teach It? *The Canadian Modern Language Review/ La Revue canadienne des langues vivantes*, 57 (4), 573 – 589.

YIP, V. and MATTHEWS, S. (2006). Assessing Language Dominance in Bilingual Acquisition: A Case for Mean Length Utterance Differentials. *Language Assessment Quarterly*, 3 (2), 97 – 116.

6. Appendices

6.1. Appendix 1

Loglinear regression with automated measures (see table 9)

Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	Total speaking time	.067	.089	.572	1	.450	1.069
	Number of syllables	-.026	.024	1.147	1	.284	.975
	Speech rate 1	2.585	1.464	3.118	1	.077	13.267
	Constant	-7.749	5.404	2.056	1	.152	.000
Step 2 ^a	Number of syllables	-.008	.003	6.836	1	.009	.992
	Speech rate 1	1.640	.652	6.322	1	.012	5.156
	Constant	-4.194	2.287	3.362	1	.067	.015

ⁱ We thank Prof Sumru Özsoy from the Bosphorus University for her support and the Turkish Ministry of Education for its permission to carry out the present study.

ⁱⁱ We are grateful for the support of the head masters and teachers of the two Anadolu school for their support. We wish to give a special thank to the students of this schools for their willingness to participate in the data collection.

ⁱⁱⁱ The term “bilingual” and “monolingual” is used in the literature in various ways. We decided to call the returnees “bilingual” because they learned both of their languages in a natural setting from a very early age. The control group is termed “monolingual” because they learned their first language Turkish in a natural setting from birth but they have contact with German in a classroom setting.

^{iv} The main data collection in Turkish was carried out by Seda Kan.

^v Here and at further computations n is lower than 60 for the returnees due to missing values.

^{vi} The differences in the number of participants per analysis is due to missing values.

^{vii} N is 22 for the control group but two participants (person 101 and 119) were excluded from these computations because of incomplete data sets.

^{viii} N is much higher for the comparison between the C-tests as only a subset of the total sample did the picture descriptions.

^{ix} Strictly speaking data from a ratio are not on an interval level and it could be argued that a non-parametric measure of association would be more appropriate. We double checked with a Spearman correlation and yielded almost to the same co-efficient.

^x We also computed a loglinear regression with the variables “total speaking time”, “number of syllables”, “speech rate 1” and all possible interactions between them. The final model lead to a slightly higher R^2 of .251 (Nagelkerke) but the percentage of correctly predicted group memberships was exactly the same (68.2%). The variables remaining in this model were “number of syllables” and “interaction between number of syllables and speech rate 1”. We do not report this in detail here as it would lead to almost the same conclusions as with the model presented in the text.