# The Myth of the Digital Native? Analysing Language Use of Different Generations on Facebook

**Jennifer-Carmen Frey, Aivars Glaznieks**
Eurac Research, Institute for Applied Linguistics, Bolzano, Italy
JenniferCarmen.Frey@eurac.edu, Aivars.Glaznieks@eurac.edu

**Abstract**

Digital Natives, i.e. people who grew up in a digital world, are said to be different to their counterparts, digital immigrants, regarding their communication habits and use of digital services. In this paper, we investigate the linguistic behavior of digital natives compared to digital immigrants in a sociolinguistically annotated corpus of personal Facebook texts using methods from corpus linguistics, computational sociolinguistics and data mining. The texts are data donations from the profiles of 133 users of various ages from the northern Italian province of South Tyrol. In order to investigate if and how digital natives differ from older generations with respect to language choice, variety choice and the use of style markers, we use three analysis methods: (1) we disclose and compare central tendencies of the two groups in a quantitative analysis, (2) we train text classifiers to distinguish both groups automatically and compare prediction results, and (3) we investigate a ranking of features. The two groups differ in particular in their use of language varieties. However, taking into account the user's first language, their choice of language and use of CMC-specific style markers also differ significantly.

**Keywords:** Facebook, CMC, youth language, sociolinguistics

## 1. Introduction

In 2001, Prensky published an essay on the distinctiveness of post- and pre-digitalization generations (Prensky, 2001), which he named *digital natives* and *digital immigrants*, respectively. The digital natives, i.e. people who were born in an already digital era and hence grew up with computers and other digital devices, were said to be different to their older counterparts, the digital immigrants, with regard to communication habits and their use of digital services, for example. Since then, several studies from domains like sociology and pedagogy have investigated his claim, trying to figure out if and how both generations differ (e.g. Palfrey and Gasser 2013, Kennedy et al. 2008, Bennett et al. 2008, Helsper and Eynon 2010). However, there is a lack of empirical linguistic investigations of such "generational" differences due to the unavailability of socio-linguistically annotated data that could represent such differences. Without doubt, age is a relevant category in computer-mediated communication (CMC) and its impact on writing has been further acknowledged in recent studies (Hilte et al., 2016; Glaznieks and Glück, Forthc; Peersman et al., 2016; Verheijen, 2017). However, we are not aware of any linguistic study investigating Prensky's note on post- and pre-digitalization generations. In this paper, we used the DiDi Corpus of South Tyrolean CMC (Frey et al., 2016) to investigate linguistic differences in the writings of digital natives and digital immigrants. We will focus our analysis on three characteristics of the investigated texts: (a) the writer's choice of language, (b) his/her choice of language variety and (c) the use of style markers that are specific to CMC.

We will start with a brief overview of the data used for this analysis (section 2.) followed by a detailed description of our approach and the methodology used (section 3.). In section 4., we report on the results obtained with regard to the two groups and summarize them in section 5.

## 2. Data: The DiDi Corpus

The data we used for our investigation is a corpus of Facebook texts published on the personal Facebook accounts of 133 voluntary data donors from South Tyrol. The so-called DiDi Corpus (Frey et al., 2016) is a multilingual corpus that contains in total around 40,000 texts (~11,000 status updates, ~6,500 comments and ~23,000 chat messages) from German and Italian native speakers and provides socio-demographic metadata such as gender, first language, education and age (collected via a questionnaire that was filled in by the data donors) for each text. The data donors were recruited via a Facebook application following the necessary privacy restraints and obligations (cf. Frey et al. 2014).

For the analysis described in this paper, we used three types of information on language use provided in the corpus:

**Languages:** The corpus provides language labels for each text that are based on a semi-automatic annotation[1]. The labels state the predominant language of the text, ignoring any kind of code-switching. The main languages in the corpus are German (58.7%), Italian (20.9%) and English (9.5%). Texts exclusively composed of non-language elements such as emoticons or hyperlinks are labeled as "non-language" texts.

**Varieties:** The corpus provides variety labels for all German-tagged texts. The variety labels are: dialect (contains dialect-specific lexical items and/or a high ratio of non-standard spellings), non-dialect (no dialect-specific items, a very low amount of non-standard spellings) or an undefinable variety (text too short to classify or contains mixed spellings).

**CMC style markers:** The corpus provides labels for style markers frequently named in the literature on CMC (Crystal, 2001; Vandergriff, 2013; Darics, 2013; Androutsopoulos, 2011), namely acronyms, emoticons, emojis, hashtags, hyperlinks, @mentions and iterations of graphemes. As CMC style markers are provided on token level, we will use the total number of style markers (and the number per subcategory) normalized for text length for our investigation.

With reference to Palfrey and Gasser (2013) and Bennet

[1] For further details see: Frey et al. (2016).

(2008), we split our data donors into two groups: people born from 1980 onwards (i.e. digital natives) and people born before 1980 (i.e. digital immigrants). Accordingly, 42% of the writers were classified as digital natives and 58% as digital immigrants. While digital natives and immigrants are almost equally represented in terms of writers, immigrants produced significantly more texts (66% of all texts compared to 34% written by digital natives). Table 1 gives an overview of available profiles and texts for both groups.

| | profiles | texts | mean | sd |
|---|---|---|---|---|
| Digital Natives | 56 | 13,529 | 242 | 439.2 |
| Digital Immigrants | 77 | 26,296 | 342 | 516.0 |

Table 1: Overview of profiles and texts in the DiDi Corpus

## 3. Methodology

We explored three strategies for our analysis of the use of languages, varieties and CMC style markers by digital natives vs. immigrants.

First, we conducted a manual statistical analysis and compared measures of central tendencies for the investigated features for both groups. We used the Mann-Whitney U test and Student's t-test to check the statistical significance (.95 confidence level) of the averaged differences.

Secondly, we applied a data mining approach comparing prediction performances of different text classifiers using machine learning. In particular, we based our research on other studies in author profiling, computational sociolinguistics (Nguyen et al., 2016) and age prediction, in which machine learning is used to predict author characteristics on the basis of their texts (Rosenthal and McKeown, 2011; Nguyen et al., 2013; Schler et al., 2006). We trained a number of text classifiers to distinguish digital natives and digital immigrants on the basis of our selected features. Then we evaluated accuracy and F-measures using 10-fold cross validation (CV) in order to validate the classifier's ability to learn underlying relations in the data. Although more sophisticated methods like neural networks would probably provide better prediction results, we used a decision tree algorithm (J48 implementation of WEKA (Witten et al., 2016)) to build our classifiers, because we were rather interested in the interpretation of the models than in reaching high accuracies.

Finally, we used a feature ranking method to check for the most informative features as it is frequently carried out in computational sociolinguistics (e.g. Simaki et al.2016, Vajjala 2017).

## 4. Results

In the following section we report the results of the three approaches described above.

### 4.1. Comparing central tendencies

Since the majority of the users in the DiDi Corpus stated German as their L1, we only used texts from L1 German users for our statistical analyses to remove potential interactions (e.g. regarding L1-dependent language choice). Furthermore, we excluded all users who wrote less than 10 texts in order to account for data skewness. The analysed subset thus contained 29,808 texts from 90 users. Table 2 shows the calculated measures of central tendencies for both groups for each feature and the corresponding p-values of the significance tests.[2]

| Feature | Natives | Immigrants | p |
|---|---|---|---|
| German | 70.83% | 83.33% | 0.1 |
| Italian | 1.09% | 5.66% | 0.003 |
| English | 9.01% | 2.08% | 1e-04 |
| non-lang. | 13.71% | 5.72% | 3e-05 |
| dialect | 41.94% | 10.91% | 5e-06 |
| non-dialect | 15.38% | 43.07% | 1e-05 |
| CMC (tokens per text) | 1.205 | 0.762 | 9e-05 |

Table 2: Comparison of central tendencies

**Languages:** After calculating the proportion of each language per user, we used median values to aggregate over both groups and performed a two-tailed Mann-Whitney U test ($\alpha = 0.05$) to test if the differences are statistically significant. The results show significant differences for the use of English, Italian and non-language texts between digital natives and digital immigrants (see Table 2). While there is no significant difference with regard to the use of German, the natives use significantly more English, produce more non-language texts and use less Italian than the digital immigrants.

**Varieties:** Per user, we compared the percentages of dialect and non-dialect texts of all German-tagged texts (in total 20,337 of 29,808 texts of the subset) averaged for both groups. As averages were not distributed normally, we used the median to average the percentages for the two groups. The results show a significant difference in the use of varieties of German between digital natives and digital immigrants. Digital natives wrote significantly more dialectal texts than immigrants when writing in German (Table 2).

**CMC style markers:** We calculated the average number of CMC style markers per text for each user and compared mean values for digital natives and immigrants (as the values were normally distributed). A two-tailed Student's t-test showed a significant difference between digital natives and digital immigrants. As can be seen in Table 2, natives used more CMC style markers (1.21 per text) on average than immigrants (0.76 per text).

### 4.2. Comparing prediction results

In our second approach, we trained a number of decision tree classifiers to label texts automatically on the basis of the provided features, instead of meticulously sampling our data and analysing aspect per aspect individually. We compared the results for classifiers with different feature combinations and controlled the effects of class imbalance and first language as a confounding factor using both the whole data set as well as the subset for training.

[2]Percentage values are median proportions per user of the group, CMC style markers represent the users' average amount of CMC-specific tokens per text, aggregated for the group.

The classification performance of our classifier, trained with all three feature categories (language, variety and number of CMC style markers) on the whole data set, proved to be significantly above the baseline (71.2% accuracy compared to 66.03%, which would be achieved when always assigning the majority class).
Table 3 shows the classification results for the different feature categories and combinations of categories.[3] When investigating each feature category individually, we found that only variety choice gave prediction results that were significantly above the baseline. However, when accounting for the interaction between a users' first language and his/her language choice by using only the L1 German subset, we could also achieve performances above the baseline with the language feature category. The number of CMC style markers, when used exclusively, did not achieve any performance improvement to the baseline. However, in combination with other features, CMC style markers contribute to the overall classification result.

| Feature | Whole corpus | | L1 German subset | |
|---|---|---|---|---|
| | Acc. | F-Score | Acc. | F-Score |
| CMC | 0.661 | 0.53 | 0.572 | 0.42 |
| Language | 0.660 | 0.53 | 0.592* | 0.51 |
| Variety | 0.704* | 0.67 | 0.675* | 0.68 |
| CMC + Lang. | 0.667* | 0.55 | 0.598* | 0.53 |
| CMC + Variety | 0.706* | 0.68 | 0.674* | 0.67 |
| Lang. + Variety | 0.703* | 0.67 | 0.695* | 0.70 |
| All | 0.712* | 0.69 | 0.700* | 0.70 |
| Baseline | 0.660 | 0.53 | 0.572* | 0.42 |

Table 3: CV results for different feature combinations

### 4.3. Feature ranking

Table 4 shows a feature ranking based on the information gain metric. According to the ranking, the use of Italian, the use of the South Tyrolean dialect and the use of the non-dialect variety in German texts are the highest ranked and thus the most informative features to distinguish digital natives from digital immigrants in the DiDi Corpus.

| Rank | Feature | InfoGain |
|---|---|---|
| 1 | Lang_IT | 0.077 |
| 2 | Var_dialect | 0.052 |
| 3 | Var_non-dialect | 0.026 |
| 4 | Lang_DE | 0.022 |
| 5 | Lang_non-lang. | 0.008 |
| 6 | CMC | 0.003 |
| 7 | Lang_EN | 0.0004 |

Table 4: Information Gain ranking

## 5. Conclusion

In this paper, we approached the distinction of digital natives and digital immigrants using three different methods, a) calculating central tendencies for both groups and testing for statistical significance, b) training a text classifier to apply a data mining strategy based on machine learning and c) calculating the most informative features by applying a feature ranking method.
The results of this study show that the investigated features of language choice, variety choice and the use of CMC style markers have proven informative for the distinction of texts written by digital natives and digital immigrants in the DiDi Corpus.
The compared measures of central tendencies showed statistically significant differences between digital natives and digital immigrants for all investigated features. The digital natives used more English as well as more dialectal writings. They also used significantly more CMC style markers, but less Italian.
The data mining approach based on text classification with decision trees similarly showed relations between the choice of both language and variety, the use of CMC style markers and the categorization of the writer as digital native or digital immigrant.[4]
In the manual investigation, all features were analysed individually using a well-defined subset. The machine learning approach provided further possibilities to test feature combinations as well as to test and rank more fine-grained features. However, the data mining approach was also sensitive to the interaction between users' first language and their language choice. When using individual feature categories for training on the whole data set, language features could not achieve performance above the baseline. This shows us that, for this approach too, methods should not be used without critical reflection, especially when relatively small data sets are used.
Furthermore, we saw that variety choice was the most important feature for the automatic text classification to discriminate between both groups. However, investigating the features individually, the use of Italian as an L1 German speaker, the use of the South Tyrolean dialect in German texts and the use of a non-dialect variety were the most important features for text classification.
The relevance of these features is also reflected in the results of the information gain calculation which ranked the use of Italian as most informative feature, followed by the use of the dialect and non-dialect variety in German texts.
The results support the general impression that South Tyrolean writers from the younger generation are more open to using different global and local varieties in CMC. In addition, they are more open to various writing styles, comprising non-language texts and texts with a high amount of CMC style markers. However, whether this originates from being a digital native or from belonging to different social groups with different communication habits cannot be answered with our data. The fact that older generations composed more texts in Italian than the younger generation (their second language with a high local and national value)

[3] Values are weighted averages for 10-fold CV. Values with asterisk are significantly higher than a baseline accuracy achieved when always assigning the majority class.

[4] Although the performance of the trained text classifiers was not particularly high (around 71%), we still accept this result as an indication to answer our linguistic research question, as we were interested in the inherent structure of the data and not in the prediction of age groups.

might also hint at societal changes (in the region or in general) in which younger people are internationally connected and English becomes more and more important.

## 6. Outlook

In future work, we plan to extend this research in two directions. First, by questioning the split of the age groups at the year 1980. For this, we want to compare different splits of age groups based on the numerical age, as well as taking into consideration alternative age concepts based on digital media experience (cf. Glaznieks and Stemle 2014). Second, methodologically, by using more sophisticated models for the statistical analysis (mixed-effects models to consider random effects) and extended feature sets for the classification approach (e.g. phenomena of multilingualism, shallow features like word or character n-grams).

## 7. References


Androutsopoulos, J. (2011). Language change and digital media: a review of conceptions and evidence. *Standard Languages and Language Standards in a Changing Europe*, pages 145–159.

Bennett, S., Maton, K., and Kervin, L. (2008). The 'digital natives' debate: A critical review of the evidence. *British journal of educational technology*, 39(5):775–786.

Crystal, D. (2001). *Language and the Internet*. Cambridge University Press, Cambridge.

Darics, E. (2013). Non-verbal signalling in digital discourse: The case of letter repetition. *Discourse, Context and Media*, 2(3):141–148.

Frey, J.-C., Stemle, E. W., and Glaznieks, A. (2014). Collecting language data of non-public social media profiles. In Gertrud Faaß et al., editors, *Workshop Proceedings of the 12th Edition of the KONVENS Conference (2014)*, pages 11–15, Hildesheim, Germany, oct. Universitatsverlag Hildesheim, Germany.

Frey, J.-C., Glaznieks, A., and Stemle, E. W. (2016). The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts. In Anna Corazza, et al., editors, *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016), 5-6 December 2016, Napoli*, pages 157–161, Torino. Academia University Press.

Glaznieks, A. and Glück, A. (Forthc.). From the Valleys to the World Wide Web: Non-Standard Spellings on Social Network Sites. In Egon W. Stemle et al., editors, *Postvolume Monograph of the 5th CMC-Corpora Conference*. Clermont Auvergne University Publishing House, Clermont Auvergne.

Glaznieks, A. and Stemle, E. W. (2014). Challenges of building a CMC corpus for analyzing writer's style by age: The DiDi project. *Journal for Language Technology and Computational Linguistics (JLCL)*, 29(2):31–57, dec.

Helsper, E. J. and Eynon, R. (2010). Digital natives: where is the evidence? *British educational research journal*, 36(3):503–520.

Hilte, L., Vandekerckhove, R., and Daelemans, W. (2016). Expressiveness in Flemish Online Teenage Talk: A Corpus-Based Analysis of Social and Medium-Related Linguistic Variation. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana, Slovenia*, pages 30–33.

Kennedy, G. E., Judd, T. S., Churchward, A., Gray, K., and Krause, K.-L. (2008). First year students' experiences with technology: Are they really digital natives? *Australasian journal of educational technology*, 24(1):108–122.

Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. (2013). 'How old do you think I am?': A study of language and age in Twitter. In *Proceedings of the seventh international AAAI conference on weblogs and social media, 8-11 July 2013, Cambridge, Massachusetts, USA*, pages 439–448.

Nguyen, D., Dogruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593.

Palfrey, J. G. and Gasser, U. (2013). *Born digital: Understanding the first generation of digital natives*. Basic Books.

Peersman, C., Daelemans, W., Vandekerckhove, R., Vandekerckhove, B., and Van Vaerenbergh, L. (2016). The effects of age, gender and region on non-standard linguistic variation in online social networks. *arXiv preprint arXiv:1601.02431*.

Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the horizon*, 9(5):1–6.

Rosenthal, S. and McKeown, K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2006). Effects of Age and Gender on Blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Simaki, V., Mporas, I., and Megalooikonomou, V. (2016). Evaluation and sociolinguistic analysis of text features for gender and age identification. *American Journal of Engineering and Applied Sciences*, 9(4):868–876.

Vajjala, S. (2017). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 18:1–27.

Vandergriff, I. (2013). Emotive communication online: A contextual analysis of computer-mediated communication (CMC) cues. *Journal of Pragmatics*, 51.

Verheijen, L. (2017). WhatsApp with social media slang?: Youth language use in Dutch written computer-mediated communication. In Darja Fišer et al., editors, *Investigating Computer-Mediated Communication. Corpus-Based Approaches to Language in the Digital World*, pages 72–101. Ljubljana University Press, Ljubljana.

Witten, I. H., Frank, E., Hall, M., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.