Amsterdam University Press

TAAL & TONGVAL 75.1 (2023) 29-57
HTTPS://DOI.ORG/10.5117/TET2023.1.003.GILL

Regional variation, internal change and language contact in Luxembourgish: results from an app-based language survey[1]

Peter Gilles
University of Luxembourg

1 Aims of the study

Like many other small languages in Europe, Luxembourgish is set in a specific multilingual situation, which resulted in intricate patterns of language variation. Language contact with German and French is clearly one of the main factors behind language variation. Furthermore, Luxembourgish is characterized by regional variation within the country as well as internal changes, both related to the mainly spoken status of Luxembourgish. These processes are further complicated by the ongoing language standardization in the written domain. The research presented in this article utilises a large-scale, crowd-sourcing data collection approach and several case studies of linguistic variables to ensure a broad overview. A novel smartphone application was developed for the purposes of data collection allowing variable linguistic phenomena to be elicited in a coherent way. With this big data approach, we were able to collect over 300.000 audio speech samples from over 3700 speakers, which has permitted us to analyze variation on the phonetic, morphological, syntactic and lexical level on a hitherto unprecedented quantitative level. The aims of this long-term project are thus to document spoken Luxembourgish and its variation and to develop a new kind of linguistic atlas in which variation is not only illustrated as a geographical phenomenon but also correlated with several social and demographic factors. Data analyses will then provide a comprehensive picture of language variation and general trends in Luxembourgish.

This article is structured as follows: Section 2 briefly describes the language situation in Luxembourg. The design of the smartphone application *Schnëssen* and an overview of the dataset are then provided in section 3.

© Peter Gilles
This is an open access article distributed under the terms of the CC BY-NC-ND 4.0 license
https://creativecommons.org/licenses/by-nc-nd/4.0/

Guest (guest)
IP: 158.64.79.7

Section 4 presents selected case studies of linguistic variables covering regional variation, internal variation and variation caused by language contact.

2 Language in Luxembourg and language variation of Luxembourgish

Luxembourgish is the national language of the Grand-Duchy of Luxembourg. It is a mostly spoken language variety that originates from a Moselle-Franconian dialect and, therefore, shares several linguistic features with this dialect group in Germany. However, due to national independence (since 1815/1839) and an increasingly positive attitude toward the language, Luxembourgish ('Lëtzebuergesch') has gained more and more independence from these German dialects and developed into an Ausbau language. Located on the Germanic-Romance language border, Luxembourg is a multilingual area, which was characterized as bilingual (German and French) until the mid-20th century. Since then and since the recognition of Luxembourgish as an independent language in its own right, the situation is best described as trilingualism (Horner & Weber 2008; Fehlen 2009). Today, Luxembourg is also a country with high rates of immigration. In fact, roughly 47% of the residents have non-Luxembourgish nationalities. The largest group of immigrants originate from Portugal (see Statec 2021). Luxembourgish is the first language of 53% of the population (i.e. approximately 336,000 of the total population of 634,700 in 2021). It is also the second (or third) language of approximately 15% of the population (Fehlen et al. 2013). German and French are learnt as foreign languages (although to a high level) in school, whereas Luxembourgish is only taught at a rudimentary level in 10th grade (one hour per week with a focus on literature and culture). French, and, increasingly, English are the languages of the workplace and in most of the public life. The Luxembourgish society is thus characterized by a high degree of social and individual multilingualism, making it necessary to use several languages on any given day depending on the situation and the person being spoken to. As such, it is possible to use Luxembourgish as a spoken language unrestrictedly in all domains of private and public life. It is, for example, the (only) spoken language in parliament. As a written language, it is being used increasingly in private and public documents. This is mainly due to the increase in informal text genres in digital-based media.

Although it is safe to say that there is no social differentiation in the usage of Luxembourgish as such, i.e. as opposed to the usage of German or

Guest (guest)
IP: 158.64.79.7

French, there is a great deal of language variation in Luxembourgish itself deriving from regional variation, internal variation and language contact. The main purpose of the present study is to document and analyze this language variation on a broad scale. To date research on linguistic variables is limited to anecdotal evidence or outdated studies. One primary task of the *Schnëssen* survey is thus to collect the variants for as many variables as possible for the first time and correlate them in a subsequent step with social factors to estimate the dynamics of language variation and change.

3 A Smartphone application to elicit and document language variation

The chosen approach for the survey relies on crowd-sourcing by using a smartphone application to record the audio and social data of the participants. In recent years, several similar applications have been developed to document various language varieties on a large scale, e.g. Hilton & Leemann 2021 in the Netherlands; Leemann et al. (2016) and Hasse, Bachmann & Glaser (2021) in Switzerland; see Bettinson & Bird (2017) for a more generic approach. Figure 1 shows some screenshots of the *Schnëssen* app used in this study (the name is derived from the Luxembourgish verb *schnëssen* 'to chat'). Participants were prompted to donate spoken language data for survey items that were specifically designed to elicit certain linguistic variables. These survey items contain phenomena from all linguistic levels, mostly from phonetics/phonology, morphology and the lexicon, but, to a lesser extent, also from syntax and pragmatics (for technical design details, see Entringer et al. 2021).

Three types of survey items were used: an image naming task, a translation task and a reading task. While the image naming task is intended to elicit one certain lexical concept (and also to bring some variety in the tasks for the user), the translation task is based on sentences, which are constructed specifically to elicit variants for several phenomena at the same time. Participants had to deliver their responses orally through the microphone of their smartphone. The following two items for a translation task illustrate this design for a German and a French sentence (Table 1, Table 2). As can be seen below, there are often several potential Luxembourgish variants for nearly every word of the sentence. Most of the variation phenomena come from phonetics (*eng/en* 'a', *huet/hatt* 'has (3PSg)', *schw*[*aːʀ*]*z/schw*[*a*]*z/schw*[*ats*] 'black', *Kleeder/Kléider* 'dresses', *bescht/best* 'best'), morphology (plural suffix variation *Witzer/Witzen* 'jokes (Pl)'), superlative variation

Guest (guest)
IP: 158.64.79.7

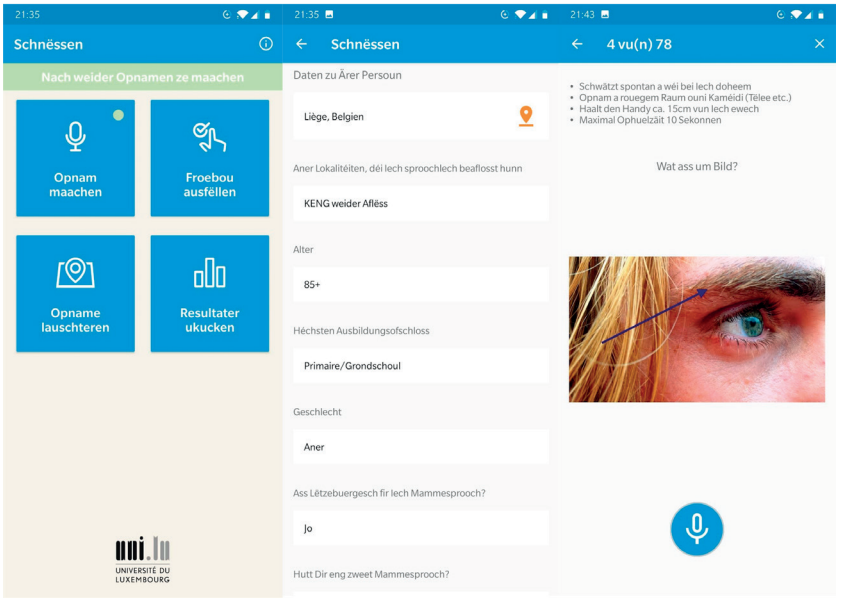


Figure 1. Screenshots of the smartphone application *Schnëssen* illustrating the main screen (left), the form to enter social information about the participant (middle) and an example for a recording item (right)

(*bescht/beschten*), verb morphology (past participle *gebitzt/gebutt* 'sewn') or lexical variation (*dacks/oft* 'often').

Most of the translation tasks are based on German input. French is used specifically to prevent the interference of a potential German variant into Luxembourgish. The intention of the sentence in Table 2 was (among other aspects) to capture the variation of the personal pronoun 'she', where *hatt* and *si* are both possible. Using a German sentence for this translation would potentially interfere with this variation as the German *sie* is phonetically identical to one of the Luxembourgish variants (*si*). Whereas the French input *elle* avoids any potential interference.

Despite the disadvantage of this method drawing on largely decontextualized language that is being used in a non-conversational setting (in contrast to spontaneous recordings or interviews), the advantage is that it provides a corpus of comparable data, which opens multiple avenues for quantitative analysis. This elicitation method, which is based on written/visual stimuli, resembles the approach used in traditional dialectological questionnaires. However, instead of relying on participants' written responses or those of a dialectological fieldworker, we receive participants'

Guest (guest)
IP: 158.64.79.7

Table 1. Example of a German sentence for the translation task. The second row lists some possible Luxembourgish variants for the respective words in the German sentence.

Eine	**Frau**	**hat**	**am**	**Freitag**	**die**	**schwarzen**	**Kleider**	**genäht.**
Eng En	Fra Frau Fr[ɔ:]	huet hatt	de e um Ø	Freideg Fregdig Freddeg Freiden	déi d' di	schw[a:ʁ]z schw[a:]z schw[ɑts]	Kleeder Kléider Klegder	gebitzt gebutt gebout gebikst genéit
'a	woman	has	on	Friday	the	black	dresses	sewn'

Table 2. Example of a French sentence for the translation task. The second row lists some possible Luxembourgish variants for the respective words in the French sentence.

Ta	soeur	est	géniale;	elle	raconte	souvent	les	meilleures	blagues!
Deng Dein	Schwëster Sëschter Schwester	ass	[ʒ]enial [g]enial	hatt si se et 't	erzielt verzallt zielt	dacks oft	déi di	bescht beschte(n) beste(n)	Witzer Witzen
'your	sister	is	great;	she	tells	often	the	best	jokes'

actual oral speech production, which reflects more closely their everyday speech habits. As already mentioned, it cannot be excluded entirely that the language of the input stimulus can influence the response. However, as German and, especially, French are considered to be different languages from Luxembourgish, the linguistic distance between the languages is believed to reduce this immediate influence. Thus, the authenticity of data is not considerably diminished by this factor. The strongest influence of the source language might be expected for certain syntactic constructions, which are in parts similar in German and Luxembourgish as well as for certain lexical items. Especially for phonetic variables, however, the data of this corpus are believed to be rather reliable.

Data collection with the app started in February 2018. The generally high interest in every aspect of the Luxembourgish language among the public made it relatively easy to motivate a large number of people to participate. Coverage on TV, Radio and in news media also boosted participant numbers considerably. Setting up a dedicated Facebook page also helped in raising

Guest (guest)
IP: 158.64.79.7

Table 3. Number of recordings captured in the Schnëssen survey as of September 2022

Round	Recordings	Average recordings per survey item	Survey items
Round 1	126512	1265	100
Round 2	82995	847	98
Round 3	31924	431	74
Round 4	22405	400	56
Round 5	23873	341	70
Round 6	5833	167	35
Round 7	8993	141	64
Round 8	1970	141	14
Grand Total	304505	596	511

and maintaining awareness of the app in the public sphere. From the outset, the app was conceptualized as part of a sustainable, long-term project. In fact, approximately four months after the launch, a set of new items were added to the app. By implementing these 'rounds' of survey tasks we were able to attract both returning and new participants. This approach, which distinguishes our app from comparable ones, allowed us to include more and more linguistic variables in the survey, contributing to a constant increase in the size of the corpus. For the app user, all former rounds still remain accessible. It takes 20 to 30 minutes to complete one round and participants can interrupt their recording sessions at any time and continue later. Between 2018 and 2022, eight rounds were made available (among them one dedicated to the language phenomena related to COVID-19 and the pandemic). Table 3 provides an overview of the audio recordings collected so far: for the eight rounds, a total of 511 survey items were published on the app. Participants provided over 300,000 recordings in response to these items. On average, every survey item was recorded by 596 participants. The decreasing numbers starting with round 3 are due to a decrease in the public interest after the initial hype, but every round was still attracting more than 100 participants. The *Schnëssen* app will be maintained as a research and data collection tool in the future and is currently also being used for data collection for Master and PhD students' projects at the University of Luxembourg.

Of the 511 survey items, more than half were sentence translation tasks and each of these sentences contained several variation phenomena. It can be estimated that the entire dataset contains over 2000 tokens targeting specific variation phenomena.

Guest (guest)
IP: 158.64.79.7

The corpus compiled in this study can be regarded as the largest structured database for spoken Luxembourgish to date, which offers rather new perspectives on the research on language variation and change in Luxembourg. Considering only participants that have recorded at least five items, the corpus is based on approximately 3,700 persons. Every participant provided a comprehensive set of social information which was later used to compute correlations with language use. Social information included the following:

- Location where the participant grew up (selected on a geographical map)
- Age group (divided into ≤ 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65+)
- Gender (male/female/other)
- Education (4 levels)
- Luxembourgish as first language (yes/no)
- Other first languages
- Language competency in French (scale 1 to 7)
- Language competency in German (scale 1 to 7)
- Language competency in Luxembourgish (if Luxembourgish was not the first language, scale 1 to 7)
- Self-reported influences of other Luxembourgish dialects regions or languages

Figure 2 shows the distribution of the sample across the six age groups. As one of the aims of the study is to analyze language variations within different age groups, it is necessary to have sufficient participants in each. It can be seen that half of the participants are under 34 years of age and that there are fewer and fewer participants in each of the older age groups. The oldest age group 65+ represents just 5.7 % of the sample and is thus the smallest group, but still consists of 214 individuals. This distribution comes as no surprise because older people tend to participate less often in app-based surveys.

Nevertheless, the overall size of the sample can be considered as big data, both compared to similar surveys, and also in relation to the total population of Luxembourg. Based on the assumption that 300,000 to 350,000 residents are competent in Luxembourgish, the *Schnëssen* corpus of more than 3700 participants represents around 1 % of this total population (cf. Fehlen & Heinz 2016, Entringer et al. 2021 for more details on speaker numbers). Regarding the regional distribution, Figure 3 shows the percentage of participants by commune. As can be seen, this ranges between 0.5 % and 1.5 % of the population of each commune. This approximately equal distribution of participants across the entire country is crucial because it allows reliable linguistic mapping of the dynamics of regional variation.

Guest (guest)
IP: 158.64.79.7

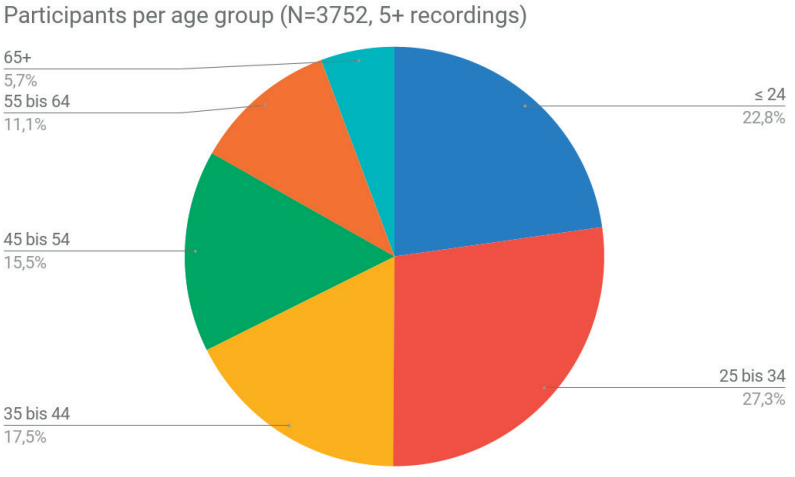


Figure 2. Number of participants per age group who provided more than five recordings in the Schnëssen app survey

Participants' audio files and the associated social data are stored in a database which is continuously updated when new data is entered into the app. The duration of the audio recordings varies from one second, e.g. for image descriptions eliciting a single word, to up to several seconds for longer sentences. The data for every item is then transferred to online spreadsheets, where one row represents one recording. These tables are used for annotating and analyzing the data, as well as for the auditory transcription, which has been conducted manually by the research team.

The quantitative distribution of the variants of a variable and their various correlations with the social information of the participants is visualized and published online in the 'Atlas of linguistic variation of Luxembourgish' ('Variatiounsatlas vum Lëtzebuergeschen', https://infolux.uni.lu/variatiounsatlas, (Gilles 2021a). Every phenomenon is presented in the atlas with various maps to illustrate the regional distribution of all individual variants as well as a summary map to indicate the most frequent variant per locality (see Figure 4). If a variation phenomenon was already discussed in the older linguistic atlas of Luxembourgish (LSA 1963), this map is displayed alongside the map of the present situation, allowing any changes in the dialect landscape to be estimated. Note that in contrast to traditional dialect atlases, the present atlas uses polygon maps rather than symbol maps. As the data available per locality consists of many observations and a mix of variants, a symbol map was not

Guest (guest)
IP: 158.64.79.7

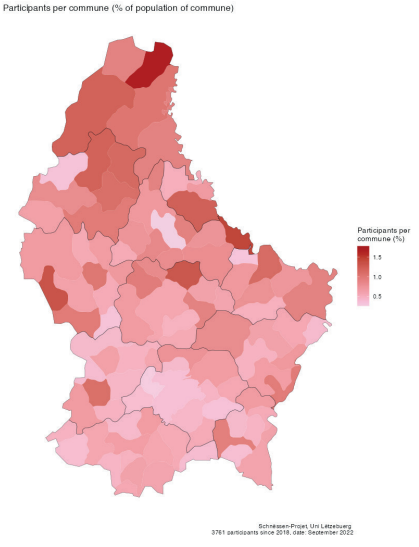


Figure 3. Participants per commune in the Schnëssen survey: Percentage of participants in relation to the population per commune (N = 3761 participants)

fit for purpose. Instead, the shades of the colors indicate the frequency of a variant in each locality. On closer observation, it is obvious that the isoglosses which were used in the old atlas to describe (pseudo-)homogeneous dialect areas do not show up as clearly on the new maps. Instead, in this more realistic view, the gradual overlap of variants in a space becomes visible.

Alongside the maps, every variation phenomenon is further described by correlations with various social characteristics of the participants, i.e. age, gender, education, language competencies in French and German and stays abroad. In addition, the participants' variants are also correlated with the socio-demographic factors of the place where they grew up, i.e. the degree of urbanization, the population per km², a socio-economic index and the percentage of non-Luxembourgish residents. Examples for these correlations will be presented in the following sections. Finally, the audio of every observation is available for listening. At present (summer 2022), some 700 variation phenomena are presented in the online atlas, each based on 200 to 1700 observations, separated into four linguistic categories (phonetics, lexicon, grammar and language contact). The atlas and its associated data analysis pipeline have been developed entirely on the open-source platform R, using mainly the packages shiny, dplyr and ggplot2 (R Core Team 2022). To ensure long-term availability, it is foreseen to publish the source code on a GitHub repository at https://github.com/PeterGilles/Variatiounsatlas.

Guest (guest)
IP: 158.64.79.7

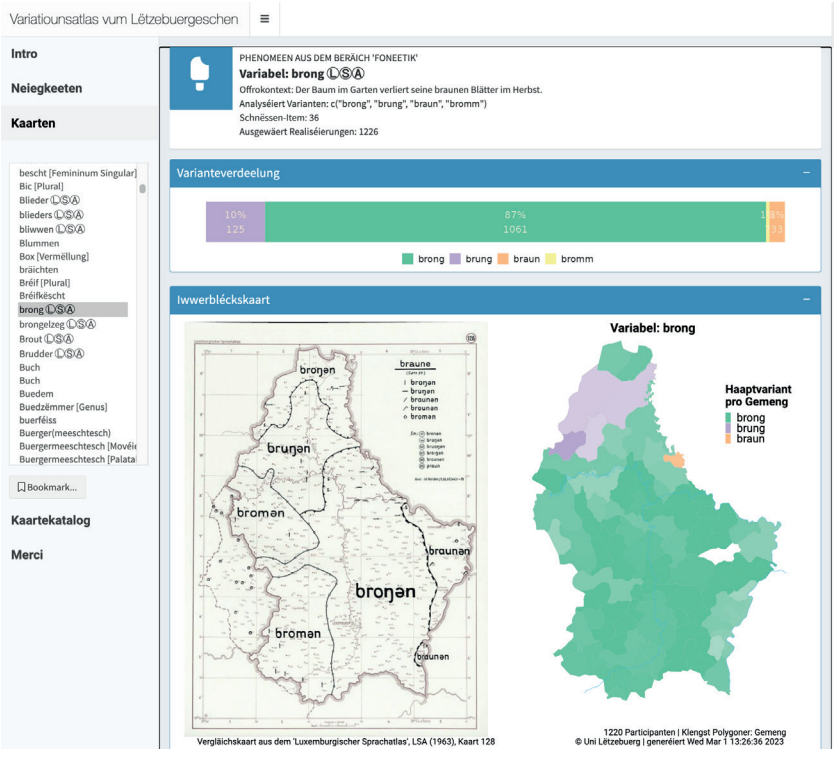


Figure 4. Layout of the 'Atlas of linguistic variation of Luxembourgish': Selection of the variable (left), key information about the variable (top), display of various maps (right)

4 Insights into the language variation of Luxembourgish

This section presents ongoing trends in the language variation of Luxembourgish for selected variables: regional variation (4.1), internal change (4.2) and language contact (4.3). These results are combined with methodological discussions on how this extended dataset can be used for studies in social dialectology and variationist linguistics. The *Schnëssen* dataset has already been used in the following studies: Entringer (2022) on morphological variation, Martin (2019) on the socio-pragmatics of neutral pronouns for female persons, Gilles (2019) for the ongoing sound merger of the consonants [ʃ] and [ɕ] and Gilles (2021b) for ongoing chain-shifts in the vowel system.

4.1 Regional variation

The language history of Luxembourgish is tightly linked to the development of dialect variation within its territory. For a long time, the classification

Guest (guest)
IP: 158.64.79.7

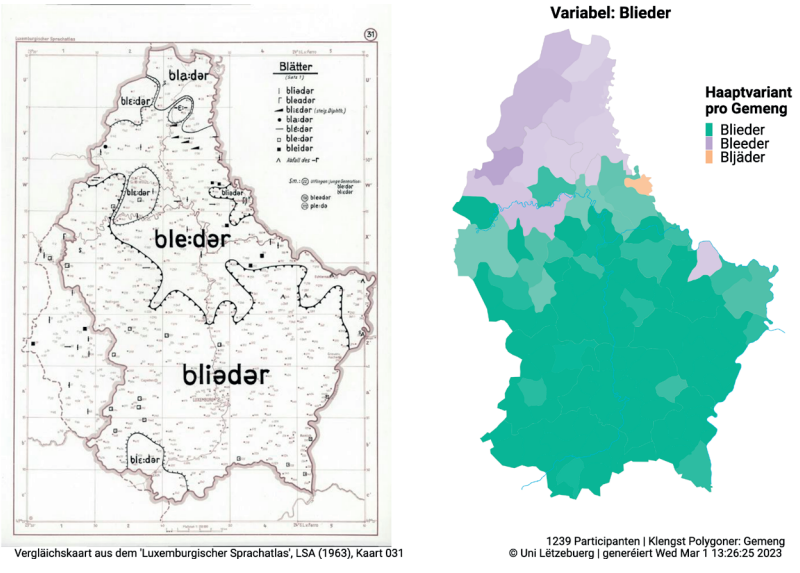
and description of Luxembourgish dialects within the wider context of the Moselle Franconian dialect (and within the German dialects in general) was the main topic of research (cf. Bruch 1954). Especially from the 1950s onwards, the Luxembourg-internal dialect situation came into focus and was characterized by a leveling of dialects, in which dialects in southern, eastern, northern, and western regions were converging towards the variety at the center of the country (cf. Gilles 2006). In this particular case of dialect leveling, the central variety developed into the standard variety of Luxembourgish (cf. Gilles 2006). It is important to note that this central region is also the location of the capital, Luxembourg City, with its important economic, political, and cultural infrastructure.

Thanks to the *Schnëssen* dataset it is now possible to analyze the current status of regional variation, which is most noticeable in the domain of phonetics. The dataset also contains nearly all structural features that were also analyzed as part of the older linguistic atlas of Luxembourgish (LSA 1963). Due to the sheer size of the new dataset, it is now possible to estimate the degree and the quality of dialect leveling by comparison with the old maps.

The first example in Figure 5 concerns the phonetic variation in the word *Blieder* 'leaves', for which the historical data on the left report two major variants, [bleːdər] and [bliədər] for the north and the south respectively. The far north also shows [blaːdər] and [bliɛdər] in smaller areas. On the right, the polygon map for the present situation shows the frequent variant per locality/polygon, where the intensity of the color indicates the frequency itself. Thus, the more intense the color, the more dominant is the respective variant; the lighter the color, the more the locality is characterized by a mix of variants, which in turn is a potential indicator of ongoing change. Compared to the historical situation, some remarkable shifts have occurred: the variant with the largest regional spread, *Blieder* (marked in green), continues to spread north, pushing the *Bleeder* (purple) area further upwards. In the north itself, *Bleeder* is still the most common form and it has suppressed the former [blaːdər] and [bliɛdər] variants nearly entirely. The reducing number of variants in the north means that *Bleeder* is now becoming the dominant variant in that region.

While Figure 5 gives the general overview of the most frequent variants per locality, the exact distribution of the individual variants remains unclear on this type of map. By offering separate maps per variant, this individual distribution becomes much clearer. Figure 6 shows how the *Blieder* variant is gradually moving into the north, where it is now present in nearly all localities. Indeed, the gradual nature of the diffusion is visible in the decreasing intensity of the color.

Guest (guest)
IP: 158.64.79.7



Vergläichskaart aus dem 'Luxemburgischer Sprachatlas', LSA (1963), Kaart 031

Figure 5. Regional distribution of phonetic variants of Blieder 'leaves'. Historical situation in LSA (1963) (left) and the situation in 2022 (right). The color indicates the most frequent variant as per polygon (= commune). The intensity of the color stands for the relative frequency itself.

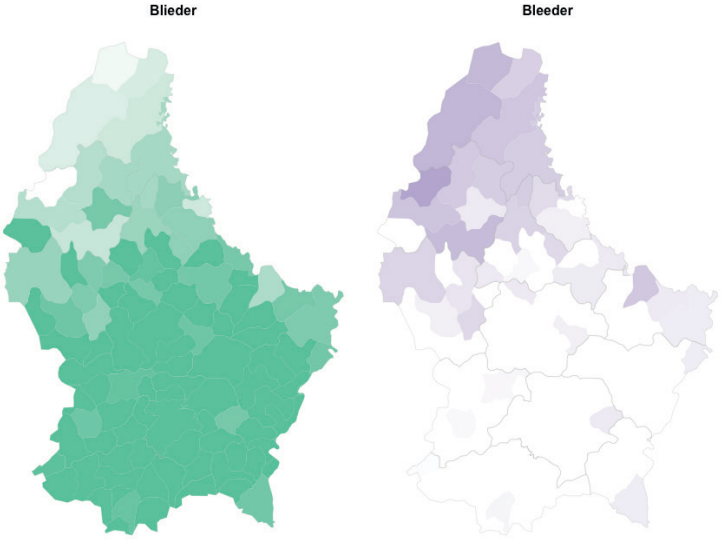


Figure 6. Individual maps for the regional distribution of the phonetic variants of *Blieder* 'leaves'

Guest (guest)
IP: 158.64.79.7

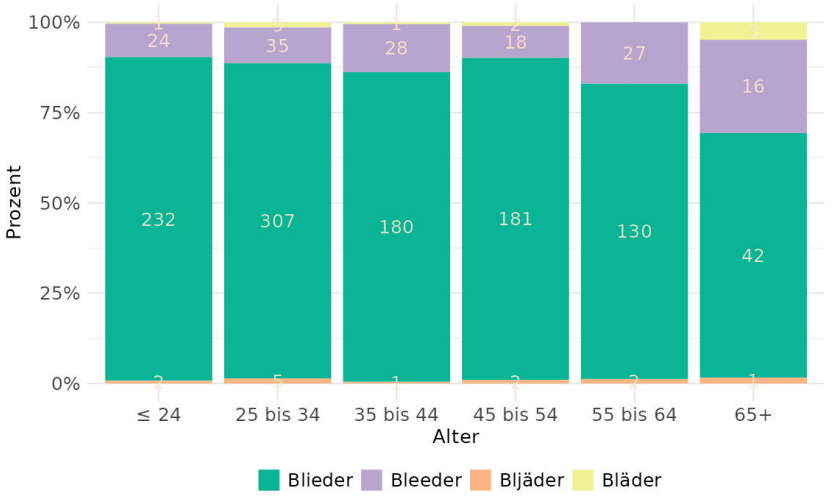


Figure 7. Age distribution of the phonetic variants for *Blieder* 'leaves'

The ongoing loss of the *Bleeder* variant as well as other minor ones like *Bljäder* and *Bläder* is corroborated through correlation with the socio-demographic factor of age. Figure 7 shows the distribution of variants for each of the six age groups. A considerable proportion of the oldest participant group (65+) retain the *Bleeder* variant, which is then increasingly replaced in the younger age groups by the now dominant *Blieder* variant. As is probably to be expected, the pattern of dialect variants receding holds true in the apparent-time, i.e. older speakers tend to keep more regional variants, whereas younger speakers tend to import the new, viz. standard variants into the region. The advantage of the present dataset is that this loss can be traced in great detail across the six age groups. According to figure 7, it can be argued that the most profound loss is taking place between the two oldest age groups and that the loss is progressively less significant as age decreases.

The next example demonstrates a case of considerable dialect stability. The verb 'to mow' occurs in three phonetic variants, *méien* [meiən] (standard variant), *méinen* [meinən] (with a hiatus-bridging nasal) and a monophthongal *mien* [miən]. The distribution maps in Figure 8 display three regions spanning from west to east. It can be seen that the two variants with the widest distribution still occupy the same area today. The more detailed maps in Figure 9 show that *méien* is entering the *méinen* region in the south-west and *méinen is* protruding into the *méien* area in the north-east,

Guest (guest)
IP: 158.64.79.7

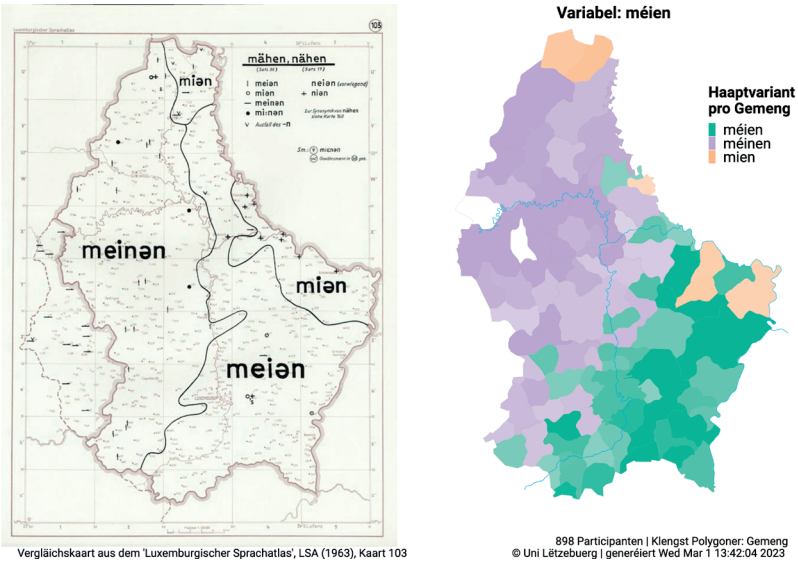


Figure 8. Regional distribution of the phonetic variants of méien 'to mow'. Historical situation in LSA (1963) (left) and the situation today (right).

but overall the historical situation remains intact. The eastern variant *mien* is receding in favor of *méien* despite retaining a strong presence in the far north. This example is remarkable insofar as the standard, central, variant *méien* has not been spreading significantly, with the competing variant *méinen* still covering a large area.

Without going into further detail about the ongoing dialect leveling and dialect stability, the findings so far clearly demonstrate that former dialects in the north, east and south are subject to continuous leveling in favor of the central variety. Both examples provide evidence for dialect leveling and dialect loss. Although the example of *méien 'to mow' does show that* the former distribution of variants can still be observed, albeit on a reduced scale, in general younger speakers are tending to abandon the former dialect variants in favour of the modern standard. The size of the *Schnëssen* dataset offers the opportunity to trace this leveling in great detail concerning both regional distribution and the correlation with sociolinguistic aspects such as age.

Guest (guest)
IP: 158.64.79.7

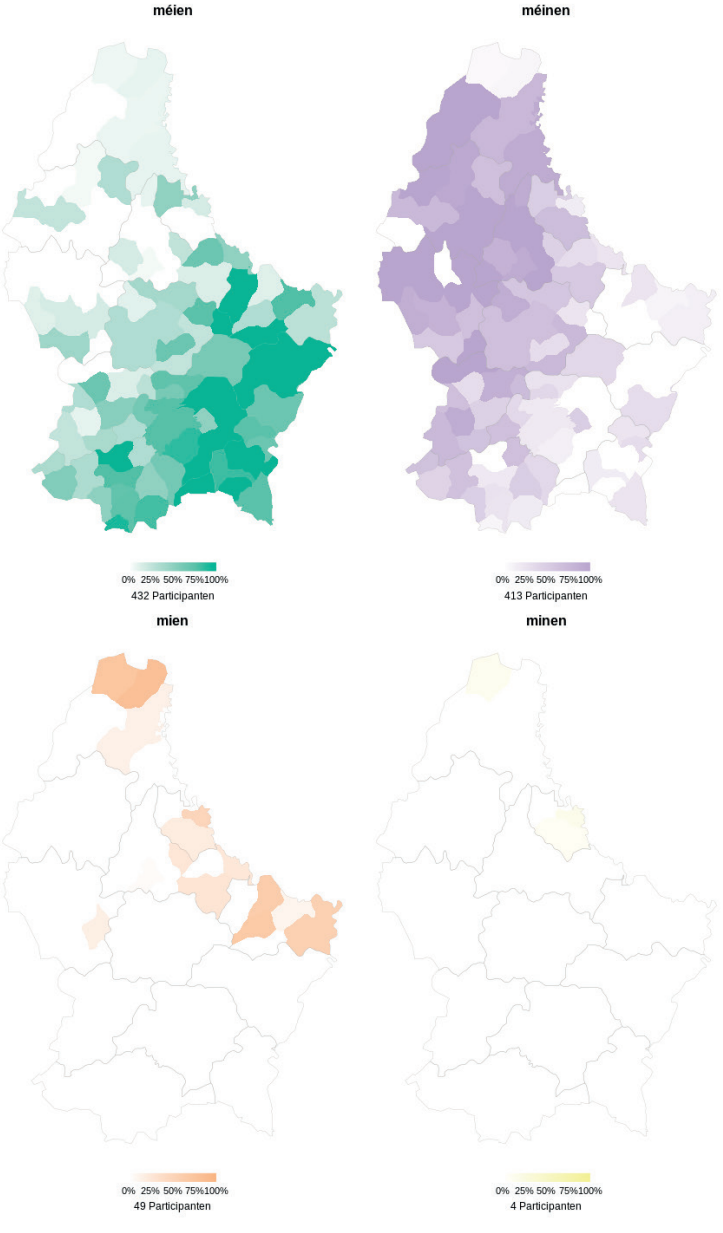


Figure 9. Individual maps for the regional distribution of the phonetic variants of méien 'to mow'

Guest (guest)
IP: 158.64.79.7

4.2 Internal change

This section is devoted to variation phenomena that have neither a regional basis nor are due to language contact. They refer to phenomena of morphology and syntax and are indicators of the internal restructuring of the grammar of Luxembourgish.

As for morphology, numerous nouns are undergoing a change in their plural marker. The nouns of Luxembourgish have lost their case marker and the only remaining suffixes indicate plural. The most common plural markers are *-en, -er, -0* (combined with umlaut and consonantal changes; cf. Nübling 2006) and their distribution is largely governed by morphological (gender) and phonological (rhyme complexity, syllable stress) factors. Among these suffixes *-en* is the most common and least distributionally restricted, and is also used as a default suffix for the plural of loan nouns. There is a particular group of nouns that is variable with regard to the suffix, i.e. alternating between *-en* and *-er* (e.g. in the singular/plural pairs *Bic* > *Bicken/Bicker* 'pen(s)', *Rendez-vous* > *Rendez-vousen/Rendez-vouser* 'encounter(s)', *Apparat* > *Apparaten/Apparater* 'device(s)'). With the *Schnëssen* dataset, it is now possible to trace this ongoing change in the speech community on a broad scale and discover how socio-demographic factors are influencing this change. As one of many examples, Figure 10 presents the variant distribution of the noun *Bus* 'bus' for 1237 speakers. For the 65+ age group *Bussen* is clearly the main variant, which decreases in usage gradually as speakers become progressively younger with the age group ≤ 24 preferring *Busser*. This variant pattern is thus characterized by a nearly complete reversal of the distribution and it shows how the suffix *-er* is taking over for these groups of nouns. Thanks to the size of the dataset, it is now possible to show convincingly that the increase of the *-er* suffix is indeed a constant one that is progressing from age group to age group.

Similar patterns were obtained for further nouns from this group (e.g. *Witz* > *Witzen/Witzer* 'joke', *Exercice* > *Exercicen/Exercicer* 'exercise'). These changes themselves are part of an extensive restructuring of the plural marking system (cf. Entringer 2022: 22 for an extensive discussion).

The next phenomenon concerns the use of personal pronouns to designate females. While females are usually referenced with the pronoun *si* 'she' (and the related grammatical forms), there exists a special pronoun, *hatt* (weak form: *et*) 'it', especially for younger women when being addressed by their first names. The pronoun *hatt,* originally derived from a neutral pronoun, is used in complex socio-pragmatic contexts, not only in Luxembourgish, but also in neighboring dialects in Germany and the Netherlands (Nübling, Busley & Drenda 2013; Nübling 2015). In Luxembourgish the neuter pronoun

Guest (guest)
IP: 158.64.79.7

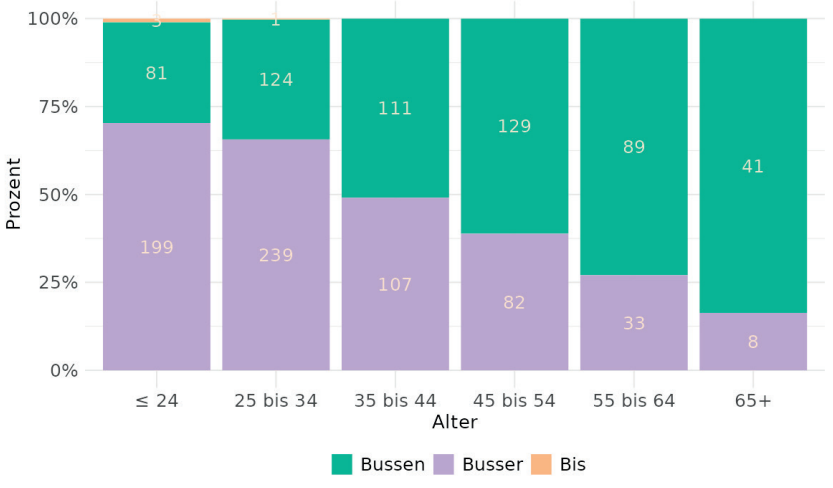


Figure 10. Age distribution for the morphologic variants of the plural of the noun *Bus* 'bus'

hatt is used categorically for all females that are also being addressed by their first names, e.g. *d'Marie* triggers the neuter pronoun *hatt*. Variation arises when it comes to the pronominalization of female family members. Traditionally, female nouns like *Schwëster* 'sister', *Cousine* 'cousin (fem.)' or *Frëndin* 'friend (fem.)/girlfriend' along with many others triggered the female personal pronoun *si* 'she'. However, the use of the neuter pronoun *hatt* for these inherently female nouns has been increasing recently. In order to analyze this variation, the survey included several items that elicit the pronouns in these variable cases. One of them was the French sentence *Ta soeur est géniale, elle raconte souvent les meilleures blagues!* 'Your sister is amazing, she often tells the best jokes!', which was intended to elicit the pronoun referring to the noun *Schwëster* 'sister'. When inspecting the distribution across age groups in Figure 11 (N = 1519), it is obvious that the neuter form *hatt* is dominant among all age groups and thus creating a conflict of grammatical gender between the female noun *Schwëster* and the referring neuter personal pronoun *hatt*, i.e. *Schwëster* <- *hatt*. The original, gender-congruent constellation *Schwëster* <- *si* is only observable in any considerable frequency in the oldest age group (35 %) and drops off as groups decrease, with female pronouns being practically absent in the youngest two age groups.

Similar results have been obtained for the nouns *Frëndin* 'friend (fem.)/girlfriend', *Cousine* 'cousin (fem.)' and *Sekretärin* 'secretary': In all cases the

Guest (guest)
IP: 158.64.79.7

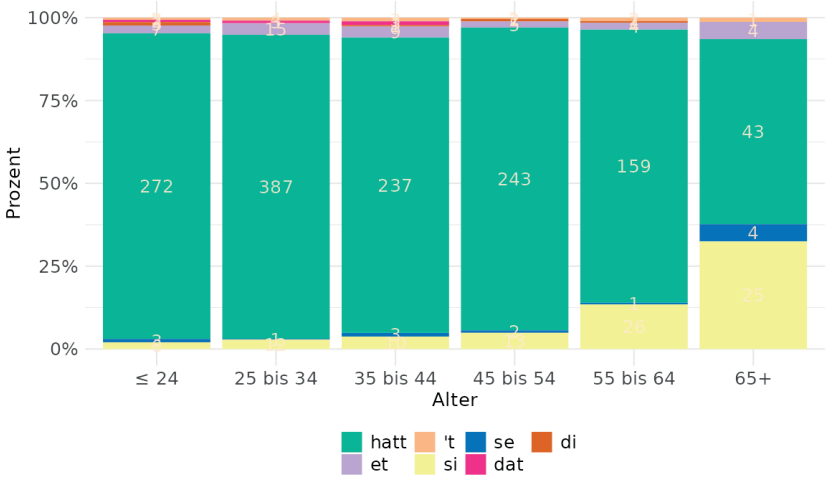


Figure 11. Age distribution of the variants of the pronouns to refer to the noun *Schwëster* 'sister' (N = 1519)

neuter personal pronoun *hatt* is used increasingly by younger speakers (see Martin 2019 for further details). This demonstrates a profound change in the grammar of personal pronouns, the detailed propagation of which through the age groups can be witnessed in the dataset.

4.3 Language contact

This final case study utilizes the dataset to explore the influence of language contact on the lexical level. The widespread incorporation of German and French words into the lexicon is a long recognized feature of Luxembourgish (Southworth 1954; Conrad 2017). These loans can coexist in the lexicon of either the individual speaker or the speech community alongside the Luxembourgish term. Speakers thus have a lexical choice of using, for example, either the German loan *Strand* or the French loan *plage* 'strand/beach'. The choice of either variant is governed by sociolinguistic factors. In the case of 'strand/beach' the most relevant factors are age, education, level of competence in French and the degree of urbanity of the residence of the speaker. The next section will discuss some of these factors. Firstly, in terms of age, it can be seen from Figure 12 that the prevalence of the French word for beach, *Plage* decreases with age. In fact, the oldest speakers clearly favor *Plage*, while for the youngest speakers, the German form *Strand* is

Guest (guest)
IP: 158.64.79.7

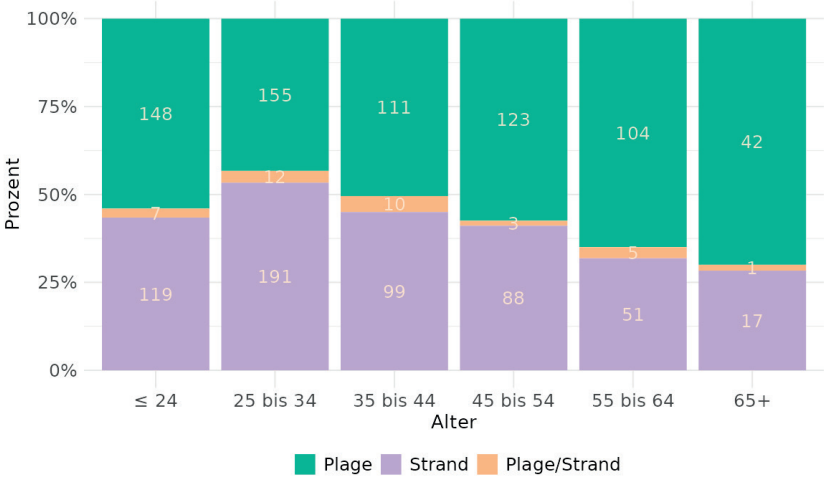


Figure 12. Lexical variation of 'strand/beach' across six age groups; 1166 participants

nearly as frequent as *Plage* (N = 1166), leading to a situation with heightened variability.

The correlation with the competence of French is also quite revealing. The higher the competence, the higher the proportion of speakers who use *Plage,* indicating that language competence does indeed correlate with language use (Figure 13). A further correlation is to be found with the degree of the urbanity of the residence of the speaker (Figure 14). The more urbanized the residence of the speaker, the more *Plage* can be found (and vice-versa).

For this specific lexical variable, it is obvious how sociolinguistic and demographic factors are shaping the pattern of variation. In the following extended case study the factors determining the choice of a French loan will now be investigated on a more systematic and statistical level. This study is based on 28 lexical variables consisting of at least two, sometimes more, variants, where one is a French loan word and the other a Germanic word, e.g. *Arbitter/Schidsrichter* 'referee', *Goût/ Geschmaach* 'taste' (see Table 4 for the entire list of the variables). The respective variants are all considered synonyms. For these variables, the *Schnëssen* corpus contains 18.922 observations coming from 2740 participants. Each participant has therefore provided approximately seven lexical variants.

Guest (guest)
IP: 158.64.79.7

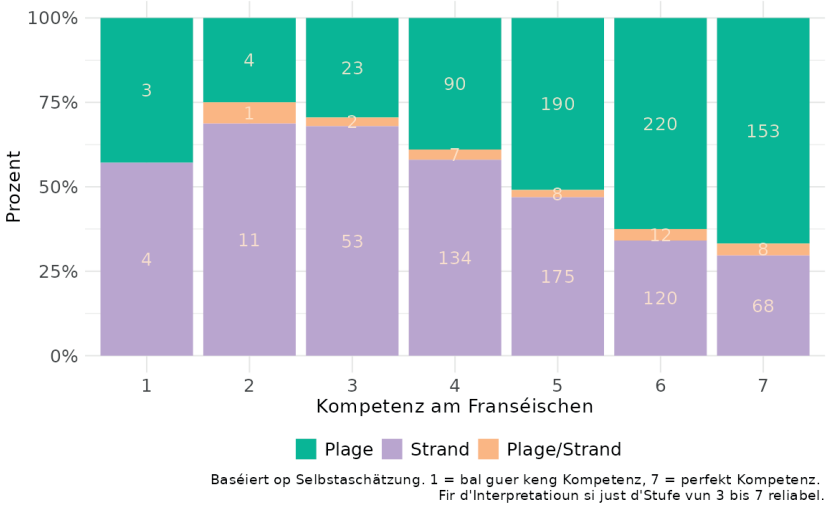


Figure 13. Lexical variation of 'strand/beach' correlated with the French competence of the participant (1 = hardly any competence, 7 = near-native competence; 1166 participants)

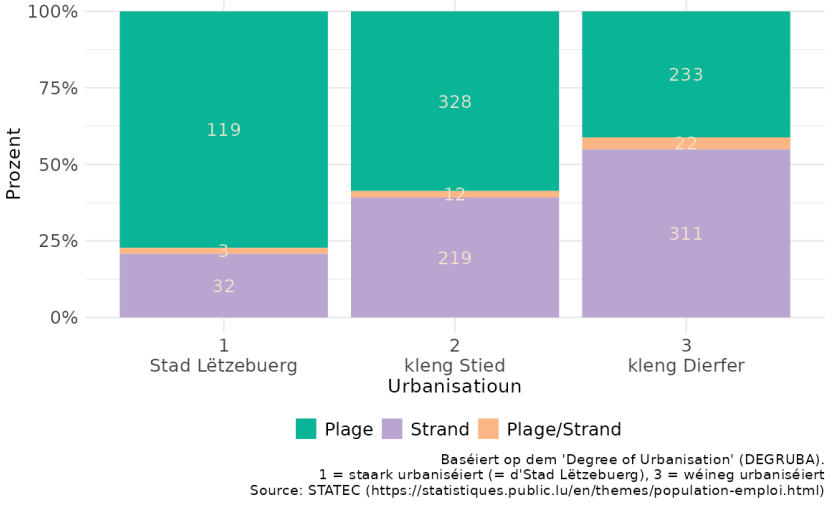


Figure 14. Lexical variation of 'strand/beach' correlated with the degree of urbanity of the residence of the participant (1 = highest urbanization, i.e. the capital Luxembourg City, 2 = smaller cities, 3 = villages; 1166 participants)

Guest (guest)
IP: 158.64.79.7

Table 4. Pairs of lexical variables (N = 28) containing at least one French loan word (in boldface)

Variable	Variants	
Arbitter	**Arbitter**, Schidsrichter	'referee'
Bëbee	Puppelchen, **Bëbee**, **Beebee**	'baby'
Bréifkëscht	**Bréifboîte**, Bréifkëscht	'letter box'
Chantier	Baustell, **Chantier**, **Schantjen**	'construction site'
Couvre-feu	Ausgangsspär, **Couvre-feu**	'curfew'
Decisiounen	Entscheedungen, **Decisiounen**	'decisions'
Déierendokter	Déierendokter, Véidokter, **Veterinaire**	'vet'
Dëschelduch	Dëschelduch, Dëschdecken, **Dëschnapp**, **Napp**, **Toile cirée**	'tablecloth'
Drucker	Drucker, Printer, **Imprimante**	'printer'
Exercice	**Exercice**, (Haus-)aufgab	'exercise'
Fernsee	Fernsee, **Tëlee**, Fernseeër, **Televisioun**	'TV'
Gefaangen	Gefaangen, **Prisonéier**	'prisoner'
Geschmaach	**Goût**, Geschmaach	'taste'
H	Ha, **Hasch**	pronunciation of letter 'H'
Homeoffice	**Teletravail**, Homeoffice/Homeworking	'home office'
Impfstoff	**Vaccin**, Impfstoff	'vaccine'
Kannapee	**Kannapee**, Kusch	'sofa'
Klinick	**Klinick**, Spidol	'hospital'
Pharen	**Pharen**, Grouss Luuchten	'full beam'
Plage	**Plage**, Strand	'strand/beach'
Poubelle	**Poubelle**, Dreckseemer, Dreckskëscht, **Drecksbac**	'trash can
Schwämm	Schwämm, **Piscine**	'swimming pool'
Suen	**Suen**, Geld	'money'
Telecom-mande	Fernbedienung, **[te**:**le**:**]commande**	'remote control'
Wallis	Koffer, **Wallis**	'suitcase'
Wartesall	Wartesall, **Salle d'attente**	'waiting room'
Y	**[i**:**'ɡræk]**, Ypsilon	pronunciation of letter 'Y'
Zoppeläffel	**(Zoppe)Louche**, (Zoppe)Läffel	'soup ladle'

As for the statistical method, correspondence regression analysis was applied using the R package "corregp" (Plevoets 2020). This special type of correspondence analysis is employed in lectometry, where structural distances between lects (language varieties, registers or languages) and linguistic variants are mapped in a multi-dimensional space (see Speelman,

Guest (guest)
IP: 158.64.79.7

Grondelaers & Geeraerts 2003; Ghyselen 2016). This rather descriptive method is intended to model the degree of association between response variables (= the lexical variables with their variants) and the predictor variables (= socio-demographic factors). The resulting association factors can explain the degree of association in the data. Usually, the first two association factors are the most important ones and explain the largest share of the association in the data (see Plevoets 2020: 153). These two association factors can then be mapped in a biplot for inspection of the interdependence between the response variables and the predictor variables. Within the two dimensions of the biplot, the proximity of data points indicates the strength of their association: The closer the data points lie together, the more often they co-occur, i.e. they share common characteristics with the predictor variables – and vice-versa, the more distant the data points are, the less properties they have in common. Note that in correspondence regression, all variables are treated as groups making it possible to calculate the distances between the variants of all variables in a single step. Following the proposal in Plevoets (2020), it is also possible to plot the confidence intervals of the predictor variables themselves as ellipses along with the data points of the response variables. This offers the researcher the opportunity to compare and analyze the proximity or distance of variants and the socio-demographic factors directly. When plotting the two highest association factors in a biplot, these axes do not have a predefined meaning, but are instead the abstract dimensions of the correspondence regression, similar to factor analysis or multidimensional scaling. However, as Plevoets (2020) points out, the axes can be interpreted as so-called 'latent variables' which represent underlying factors that are structuring the dataset.

For this study of lexical choice, two correspondence regression analyses were conducted for a dataset of 28 lexical variables, each based on a combination of two socio-demographic predictor variables. The first analysis is based on the social factors 'age' (recoded as 'young', 'middle-aged' and 'old') and 'education' and their interaction.[2] The eigen-values for this correlation indicate that the first association factor explains 51% and the second 21% of the total variation. Thus, 72 % of the total variation is already explained by the first two association factors. An ANOVA on the eigen-values shows that both predictors as well as their interaction are significant.[3] In the biplot in Figure 15 all 67 variants of the 28 variables are mapped with the respective distances in a two-dimensional space. Color-coding helps to distinguish Germanic (blue) from French variants (green). The placement of the variants of a variable then allows us to determine the distance or proximity of the two in this space. Thus, for example, the French variant *Plage* is found at the

Guest (guest)
IP: 158.64.79.7

bottom in the lower right quadrant, while its Germanic counterpart is to be found high in the top-left quadrant. This suggests that these two variants are used by quite different populations of speakers, which are in this case characterized by the socio-demographic predictors 'age' and 'education'. The same is true for *Drucker* (left, middle) and the French *Imprimante* 'printer' (right, middle) or the Germanic *Geschmaach* (upper half, middle and the French *Goût* 'taste' (bottom, middle). In this type of visual representation, the great distances between the Germanic and French variants of a variable are immediately obvious. From the color-coding, it is clear that most French variants are in the lower half, whereas most German variants are mostly in the (left quadrant of the) upper half. Therefore, it can be concluded that the vertical dimension represents the latent variable of 'French preference', with lower values indicating a preference for French variants and higher values indicating a dispreference for French variants (thus, a preference for Germanic forms).

This single 'latent variable', though, is not sufficient to capture the entire distribution of the two groups of variants—for example, there are also French variants present in the top-right quadrant. A second, superimposed latent variable has to be assumed, which seems to be related to the predictor variable 'age'. As can be seen by the orange confidence interval ellipses for 'young', 'middle-aged' and 'old', following the horizontal axis closely from left to right. The variants on the left, then, are mostly used by younger speakers with older speakers on the right-hand side. Furthermore, the vertical axis is associated with the second predictor variable, i.e. 'education', whereby lower education ('technical school') is in the top half and higher education ('classical secondary school', 'university') is in the lower half. A lower educational level is thus associated primarily with Germanic variants, whereas higher education is clearly associated with the French variants.

From this first correspondence regression analysis, it can be concluded that French and Germanic variants form more or less separate clusters in the space created by the socio-demographic predictors 'age' and 'education'. At least one latent variable can be established steering the preference/disinclination for French loans, which is in turn linked to 'education'.

The following second correspondence regression is calculated on the predictor variables 'gender' (two levels) and 'competence_french'.[4] The latter refers to the self-reported competence level of the participant, recoded here in three levels. The total amount of explained variation amounts to 82 % (54 % for the first factor, 28 % for the second). The biplot in Figure 16 offers a somewhat different, but still consistent picture, with the groups of variants split along the horizontal axis: Germanic variants are located more on the

Guest (guest)
IP: 158.64.79.7

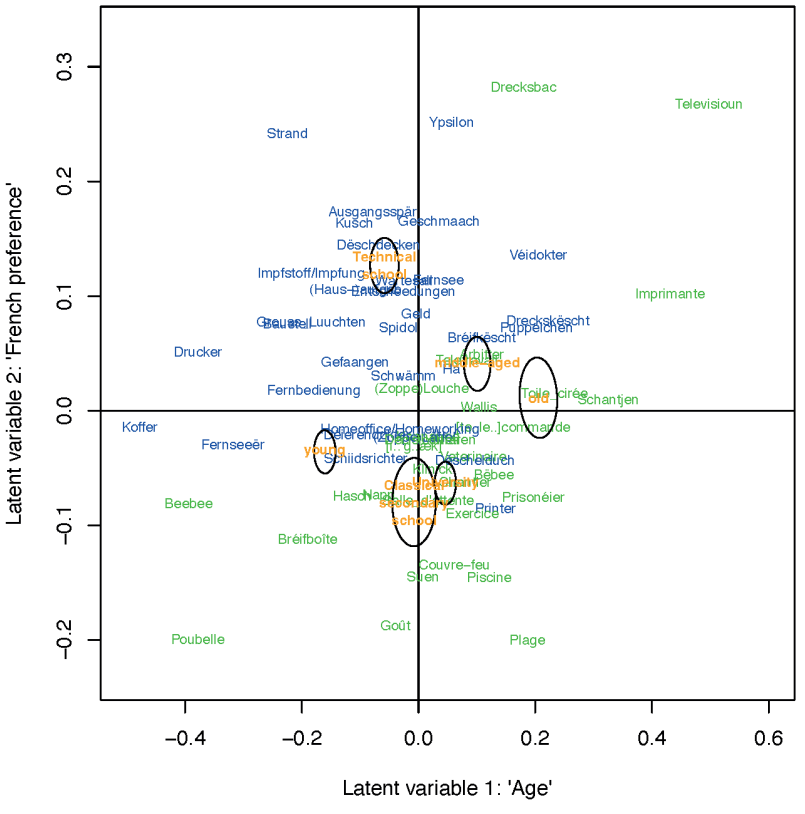


Figure 15. Biplot of the correspondence regression for the socio-demographic predictor variables 'age' * 'education'. Latent variable 1 explains 51 %, latent variable 2 21% of the total variation. Germanic variants (blue) and French variants (green) overlaid by the confidence intervals of the two predictor variables (orange). 'Young', 'middle-aged', 'old' for the predictor 'age' and 'technical school', 'classical secondary school' and 'university' for the predictor 'education'.

left and the French ones on the right. Thus, the choice of a lexical variant is largely based on the latent variable on the horizontal axis, i.e. it is clearly associated with the preference/disinclination for French loans. The distances between pairs of variants and the overall configuration are somewhat different compared to Figure 15, but the general clustering of 'green' and 'blue' variants remains unchanged. Unsurprisingly, the horizontal axis also strongly correlates with the predictor variable 'competence_french': That is, the levels 'French low', 'French average' and 'French high' are horizontally aligned from left to right, indicating that a low competence in French is also associated with the preferred use of Germanic loans (e.g. *Strand, Baustell*

Guest (guest)
IP: 158.64.79.7

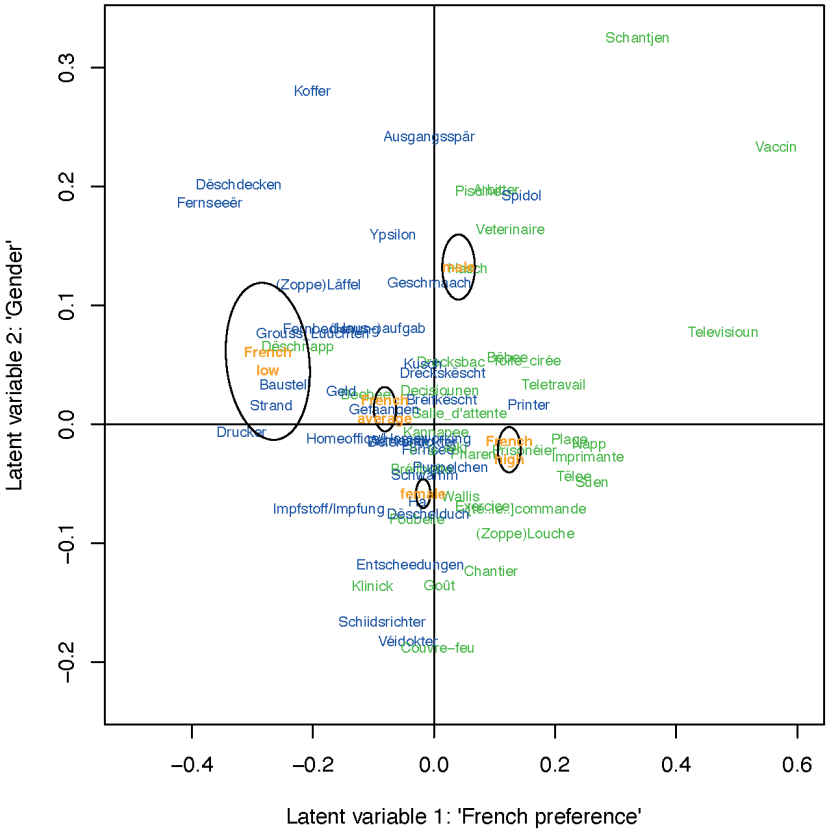


Figure 16. Biplot of the correspondence regression for the socio-demographic predictor variables 'gender' * 'competence_french'. Latent variable 1 explains 53 %, latent variable 2 28 % of the total variation. Germanic variants (blue) and French variants (green) overlaid by the confidence intervals of the two predictor variables (orange): 'French low', 'French average', 'French low' for 'the predictor 'competence_french' and 'female' and 'male' for the predictor 'gender'.

'construction site', *Grouss Luuchten* 'full beam') and high competence comes along with increased French loans (e.g. *Prisonéier* 'prisoner', *Pharen* 'full beam', *Napp* 'tablecloth').

The vertical dimension seems to be of less importantance for the distribution of Germanic and French variants, as it accounts for only 28 % of the total variation. However, as the orange ellipses for 'male' and 'female' indicate, this latent variable is related to the gender of the speaker. The (Germanic and French) variants for men tend to be in the top half and those for women tend to be located in the bottom half. For example, men would prefer the

Guest (guest)
IP: 158.64.79.7

German *Koffer,* while the French *Wallis* 'suitcase' is used more by women. A similar tendency can be observed for *Veterinaire* (for men) vs. *Déierendokter* 'vet' (for women). While it is obvious that gender is not steering the choice of French variants on a general level, it is nevertheless influencing lexical choice for certain words and has thus to be considered an overlapping factor.

It was the purpose of this extended case study to demonstrate how a substantial subset of the *Schnëssen* corpus could be employed to explore lexical choices for over 2700 speakers. By applying correspondence regression, the association distances of the variants have been calculated and visualized. The choice of French variants is governed by a complex interplay of age, educational level and competency in French. These findings are in line with the results from Conrad (2017) for phonological variables and Conrad (in prep.) for the lexical domain of football language. By developing an index to measure general 'language preference' ('sprachliche Orientierung') in a multilingual setting, the author can show how a general preference for either German or French also leads to an increased use of German or French words.

5 Conclusion

As a small and largely spoken language that is contained within an intricate multilingual setting, Luxembourgish is subject to extended language variation and change on all linguistic levels. The development of the *Schnëssen* smartphone application using crowd-sourcing techniques allows language variation and change to be addressed in a big data perspective for the first time. Using this app it was possible to compile a corpus of over 3700 speakers from all age groups providing over 300,000 individual recordings for 511 survey items. Large parts of this corpus have already been annotated manually and analyzed. The case studies presented here provide some insight into recent trends in the development of Luxembourgish in terms of regional variation, internal changes and language contact. The high number of speakers in the sample permitted a fine-grained subdivision into six age groups which in turn allowed variations and changes to be traced in a highly detailed manner, which was not possible before. As for regional variations, this new data set leads to a much more complete and dynamic picture of the ongoing process of dialect leveling. In terms of data presentation, the quantitative results have been published as an atlas of language variation in Luxembourgish, where linguistic variants are geographically mapped and also correlated with several social and demographic factors. As a research

Guest (guest)
IP: 158.64.79.7

tool for students and researchers, but also as an insightful platform for laypeople, language learners and language enthusiasts, the atlas offers ample opportunity to further analyze language variation and change in Luxembourgish in a systematic way.

Bibliography

Bettinson, Mat & Steven Bird. 2017. Developing a suite of mobile applications for collaborative language documentation. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 156–164. https://doi.org/10.18653/v1/W17-0121.

Bruch, Robert. 1954. *Das Luxemburgische Im Westfränkischen Kreis*. Luxemburg.

Conrad, François. 2017. *Variation durch Sprachkontakt* (Luxemburg-Studien / Études Luxembourgeoises). Vol. 14. Frankfurt / New York: Peter Lang.

Conrad, François. in prep. ‚Sprachliche Orientierung' als Interpretationsfaktor (in) einer komplexen, mehrsprachigen Sprachlandschaft: Eine erweiterte Perspektive auf das Verständnis sprachkontaktinduzierter Variation in Luxemburg.

Entringer, Nathalie. 2022. *Vun iwwerfëlltene Bussen bis bei déi beschte Witzer. Morphologische Variation im Luxemburgischen – eine variations- und perzeptionslinguistische Studie*. Luxembourg: University of Luxembourg PhD Dissertation. http://hdl.handle.net/10993/50110.

Entringer, Nathalie, Peter Gilles, Sara Martin & Christoph Purschke. 2021. Schnëssen. Surveying language dynamics in Luxembourgish with a mobile research app. *Linguistics Vanguard* 7(s1). 20190031. https://doi.org/10/gh6d34.

Fehlen, Fernand. 2009. *BaleineBis : une enquête sur un marché linguistique multilingue en profonde mutation. Luxemburgs Sprachenmarkt im Wandel*. Luxembourg: SESOPI Centre intercommaunitaire.

Fehlen, Fernand & Andreas Heinz. 2016. *Die Luxemburger Mehrsprachigkeit. Ergebnisse einer Volkszählung*. Bielefeld: Transcript.

Fehlen, Fernand, Andreas Heinz, François Peltier & Germain Thill. 2013. *Les langues parlées au travail, à l'école et/ou à la maison / Umgangssprachen*. Luxembourg: STATEC.

Ghyselen, Anne-Sophie. 2016. *Verticale structuur en dynamiek van het gesproken Nederlands in Vlaanderen: een empirische studie in Ieper, Gent en Antwerpen*. Gent: University of Gent PhD dissertation. http://hdl.handle.net/1854/LU-8055169.

Gilles, Peter. 2006. Dialektausgleich im Luxemburgischen. In Claudine Moulin & Damaris Nübling (eds.), *Perspektiven einer linguistischen Luxemburgistik. Studien zu Diachronie und Synchronie* (Germanistische Bibliothek 25), 1–27. Heidelberg: Winter.

Guest (guest)
IP: 158.64.79.7

Gilles, Peter. 2019. Using crowd-sourced data to analyse the ongoing merger of [ɕ] and [ʃ] in Luxembourgish. In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, 1590–1594. Canberra, Australia: Australasian Speech Science and Technology Association Inc. https://assta.org/proceedings/ICPhS2019/papers/ICPhS_1639.pdf.

Gilles, Peter. 2021a. Variatiounsatlas vum Lëtzebuergeschen. R/Shiny. Luxembourg: University of Luxembourg. https://petergill.shinyapps.io/variatiounsatlas.

Gilles, Peter. 2021b. Ongoing sound changes in the Luxembourgish vowel system. Poster at *Phonetik und Phonologie 17*, Frankfurt, 29.-30.9.2021, https://www.linguistik-in-frankfurt.de/17-pp-in-frankfurt-gu/.

Hasse, Anja, Sandro Bachmann & Elvira Glaser. 2021. Gschmöis–Crowdsourcing grammatical data of Swiss German. *Linguistics Vanguard* 7(s1) 20190026. https://doi.org/10.1515/lingvan-2019-0026.

Hilton, Nanna Haug & Adrian Leemann. 2021. Using smartphones to collect linguistic data. *Linguistics Vanguard* 7(s1) 20200132. https://doi.org/10.1515/lingvan-2020-0132.

Horner, Kristine & Jean-Jacques Weber. 2008. The language situation in Luxembourg. *Current Issues in Language Planning* 9(1). 69–128.

LSA = Bruch, Robert. 1963. *Luxemburgischer Sprachatlas. Laut- und Formenatlas von Robert Bruch. Für den Druck vorbereitet von Jan Goossens*. Marburg: Elwert.

Leemann, Adrian, Marie-José Kolly, Ross Purves, David Britain & Elvira Glaser. 2016. Crowdsourcing Language Change with Smartphone Applications. *PLoS ONE* 11(1). e0143060. https://doi.org/10.1371/journal.pone.0143060.

Martin, Sara. 2019. Hatt or si? Neuter and feminine gender assignment in reference to female persons in Luxembourgish. *STUF – Language Typology and Universals* 72(4). 573–601. https://doi.org/10.1515/stuf-2019-0022.

Nübling, Damaris. 2006. Zur Entstehung und Struktur ungebändigter Allomorphie: Pluralbildungsverfahren im Luxemburgischen. In Claudine Moulin & Damaris Nübling (eds.), *Perspektiven einer linguistischen Luxemburgistik. Studien zu Diachronie und Synchronie* (Germanistische Bibliothek 25), 107–125. Heidelberg: Winter.

Nübling, Damaris. 2015. Between feminine and neuter, between semantic and pragmatic gender: hybrid names in German dialects and in Luxembourgish. In Jürg Fleischer, Elisabeth Rieken & Paul Widmer (eds.), *Agreement from a Diachronic Perspective*, 235–265. Berlin, München, Boston: De Gruyter.

Nübling, Damaris, Simone Busley & Juliane Drenda. 2013. Dat Anna und s Eva – Neutrale Frauenrufnamen in deutschen Dialekten und im Luxemburgischen zwischen pragmatischer und semantischer Genuszuweisung. *Zeitschrift für Dialektologie und Linguistik* 80(2). 152–196.

Guest (guest)
IP: 158.64.79.7

Plevoets, Koen. 2020. Lectometry and Latent Variables: A Model for Underlying Determinants of (Normative) Choices in Written and Audiovisual Translations. *Zeitschrift für Dialektologie und Linguistik* 87(2). 144-172. https://doi.org/10.25162/zdl-2020-0006.

R Core Team. 2022. *R: A language and environment for statistical computing*. Manual. Vienna, Austria. https://www.R-project.org/.

Southworth, F. C. 1954. French elements in the vocabulary of the Luxemburg dialect. *Bulletin liguistique et ethnologique* (Section de Linguistique, de Folklore et de Toponymie/Institut Grand-Ducal) 2. 1–20.

Speelman, Dirk, Stefan Grondelaers & Dirk Geeraerts. 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities*. Springer 37(3). 317–337. https://doi.org/10.1023/A:1025019216574.

Statec. 2021. *Le Luxembourg en chiffres 2021*. Luxembourg. http://statistiques.public.lu/fr/publications/series/luxembourg-en-chiffres/2021/luxembourg-en-chiffres/index.html.

Notes

1. For many I valuable comments I would like to thank Caroline Döhmer and Anne Breitbarth.
2. The corresponding formula to run the function of the package 'corregp' in R is: corregp(variant ~ age * education, data=corr_data, part="variable", b=3000).
3. ANOVA Table (Type III Tests)

	X^2	Lower	Upper
age	463.56941	385.63762	546.10566
education	204.52730	152.45393	261.53016
age.education	44.57588	26.75012	76.02506

4. The corresponding formula to run the function of the package 'corregp' in R is: corregp(variant ~ gender * competence_french, data=corr_data, part="variable", b=3000).

Guest (guest)
IP: 158.64.79.7