Celebrating and quantifying the linguistic diversity of the UBC student community

MOLLY BABEL, LINE LLOY, ZARA KHALAJI PIRBALUTI,
RACHEL SOO & LEXIA SUITE
University of British Columbia

1 Introduction

While the University of British Columbia (UBC) does not survey the student body's language background, the undergraduate student body is a diverse lot.[1] The goal of this short paper is to provide a first-pass description of the linguistic diversity in the student population at UBC.

That UBC would be linguistically diverse is unsurprising given the surrounding speech community. Metro Vancouver boasts high levels of linguistic diversity itself. For example, while English is the dominant societal language of both the university and Vancouver, only 51.2% of Metro Vancouver residents are mother tongue speakers of English (ISO 639-3: eng). French (ISO 639-3: fra) is not widely spoken as a mother tongue in Metro Vancouver, with less than 2% speaking French as their mother tongue. Over fifty different mother tongue languages are spoken by the non-English and non-French mother tongue speakers, according to the most recent census (Statistics Canada, 2023).

In our characterization of the linguistic diversity of UBC students, we consider various aspects of the language experience and quantify language patterns from several angles. This multi-pronged approach is in recognition that bilingualism — or, more broadly, multilingualism — is a challenging, if not impossible, construct to quantify (Marian and Hayakawa, 2021). Moreover, we highlight that any measure of bilingualism is a continuum and not a categorical variable (Luk and Bialystok, 2013). A common instrument used to describe individuals' multilingual experiences is the Language Experience and Proficiency Questionnaire (LEAP-Q, Marian et al., 2007). The LEAP-Q probes participants' lan-

[1] Thank you to the UBC community for sharing your language background with us. Thanks to Khia A. Johnson and Khushi Nilesh Patil for their contributions to the projects from which these data originate. We thank Hotze Rullmann for being such a wonderful teacher and colleague! We are lucky to have you in our lives.

guage history, use, attitudes, and self-rated proficiency, providing data that can be quantitatively or qualitatively described.

Using responses on the LEAP-Q, Gullifer and Titone (2020) recently introduced a measure called *language entropy* that quantifies the predictability of an individual's language use in different contexts. In the quantification of language entropy, a monolingual individual would have a score of 0 in any context; there is no doubt about the language that will be used, as the individual is monolingual. A bilingual individual who uses both of her languages equally in a given environment would have an entropy value of 1, indicating that it is unpredictable which of the two languages would be used. The maximum entropy value increases with the number of languages spoken, but, regardless of the number of languages spoken, a low language entropy value indicates that it is highly predictable what language that individual would use in a given context and a high language entropy value indicates unpredictability in language use. Gullifer and Titone (2020) characterize these types of language use associated with low and high language entropy as *compartmentalized* and *integrated*, respectively, pointing to the ways in which an individuals' multiple languages are used in varying social contexts. As a kind of validation of this interpretation, language entropy is positively correlated with language mixing and switching practices (Kałamała et al., 2022), though it appears to be independent from cognitive processing measures like proactive control (Wagner et al., 2023; Gullifer and Titone, 2021).

The goal of this paper is to provide a description of the multilingualism of UBC students. Because we intend for this paper to be broadly readable, we avoid quantitative analyses and, instead, provide qualitative descriptions of the patterns.

2 Methodology

2.1 Participants

1026 UBC students completed the LEAP-Q. Ten individuals did not report their month and year of birth. The mean participant age was 22 (SD = 3.7). As this is a rather contracted age range, we do not discuss age further.[2]

[2] We note changes in language use over time may be an interesting and meaningful dimension to consider, should the data allow.

2.2 Materials and procedures

The LEAP-Q was administered on Qualtrics. For the subset of data from Suite et al. (2023), this instrument was presented after a short vocabulary assessment in a survey that followed completion of a sentence transcription task. For the subset from Lloy et al. (2024), the LEAP-Q was completed in a multilingual survey that also included the Bilingual Language Profile (Gertken et al., 2014) and the Bilingual Code Switching Questionnaire (Rodriguez-Fornells et al., 2012). In both projects, the LEAP-Q was completed by participants online in a location of their choosing.

3 Results

3.1 What type of multilingual?

Figure 1 presents two panels that broadly summarize the type of multilingual speakers in the UBC speech community. On the left, Panel A is a histogram of the number of individuals who report experience with different numbers of languages. The mode of this distribution is 3, indicating the most common situation is to have experience with three languages. Bilingual and quadrilingual experiences are the next most likely language backgrounds. It is more common for UBC students to have experience with five languages than to be monolingual.

An important distinction in the bilingual (or multilingual) experience is whether an individual acquired their first two languages simultaneously or sequentially. Sequential bilinguals who learn a second language much after their first often, but not always, exhibit linguistic patterns distinct from simultaneous bilinguals. To determine whether UBC students are simultaneous or sequential bilinguals, the two lowest reported ages of acquisition for individuals with experience with more than two languages were compared. The difference in these values is reported on the x-axis in the right panel of Figure 1. There is a large spike at 0, indicating that the mode is for individuals to be simultaneous bilinguals; there is no difference in the ages at which individuals begin acquiring their first two languages. A second clear peak in the data occurs before the onset of schooling. As most participants report age 0 as the onset of acquisition of their first language, this second peak in early childhood may suggest that many individuals begin acquiring a second language in an early childcare

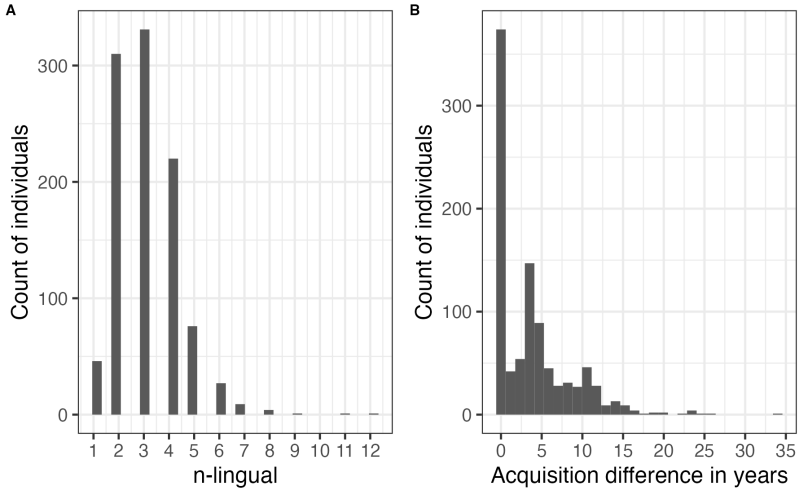

Figure 1: (A): A histogram of how multilingual participants are. The vertical axis shows the counts of individuals for each *n*-lingual bin on the horizontal axis. Trilinguals are the most common type of multilingual. (B): A histogram of the difference in the ages at which individuals acquire their first two languages. An acquisition difference of 0 represents simultaneous bilinguals, for whom there is no difference in age of acquisition of their first two languages.

or educational setting either due to entrance in a language immersion program or an introduction to English in daycare or preschool. English, the societally dominant language in the Lower Mainland, is then introduced at this point after having familial experience with another language.

Our calculation of language entropy provides separate values for speaking/signing[3], exposure, and reading, as individuals can vary in how often they produce a language, how often they are exposed to language, and how often they read a language. These varied experiences are observed in the panels in Figure 2, which shows speaking by exposure entropy, and Figure 3, which shows speaking and exposure entropy by reading entropy. Because of an interest in characterizing different calculations of entropy, particularly speaking and exposure entropy, we present these data in scatterplots that show pairwise correlations and histograms along the top and right sides of the figures.

[3] We use the term 'speaking' for any kind of oral or signed language production.

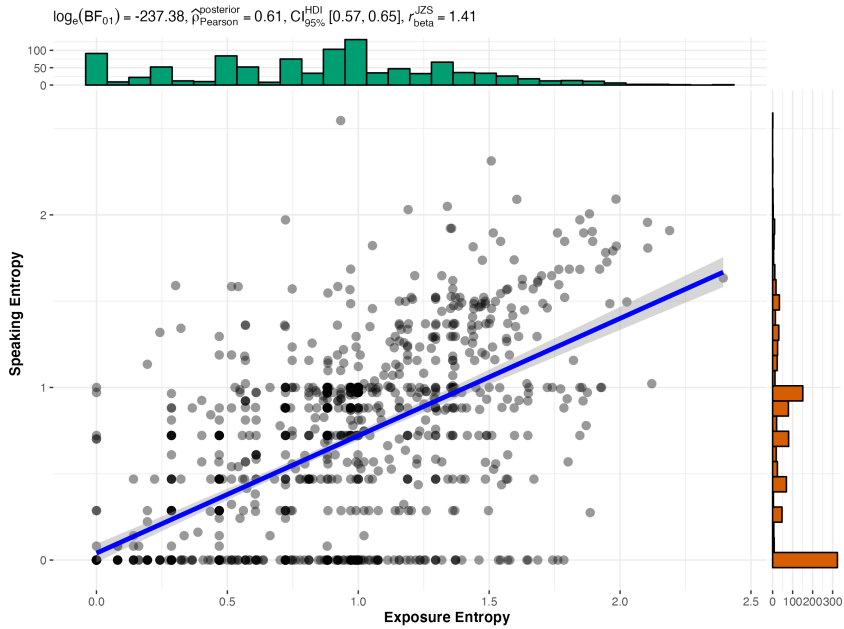

Figure 2: Speaking Entropy (vertical axis) by Exposure Entropy (horizontal axis). Histograms for both variables are on the opposing axes.

Each pairwise comparison demonstrates a positive correlation. This suggests that individuals with high entropy for speaking/signing, exposure, or reading are also more likely to have high entropy values for any of these dimensions. So, while we see from the histograms in these figures that, for example, there is a more prominent low entropy peak for reading and speaking than exposure, the overall pattern in these values is that more integrated language use in one domain is associated with more integrated language use in another domain. However, the strength of the relationship is the strongest for speaking and exposure entropy, suggesting that reading is a more distinct mode.

3.2 What languages are represented?

Having established that UBC students have experience with multiple languages, let us identify what those languages are.

Participants reported speaking 104 distinct languages. This language

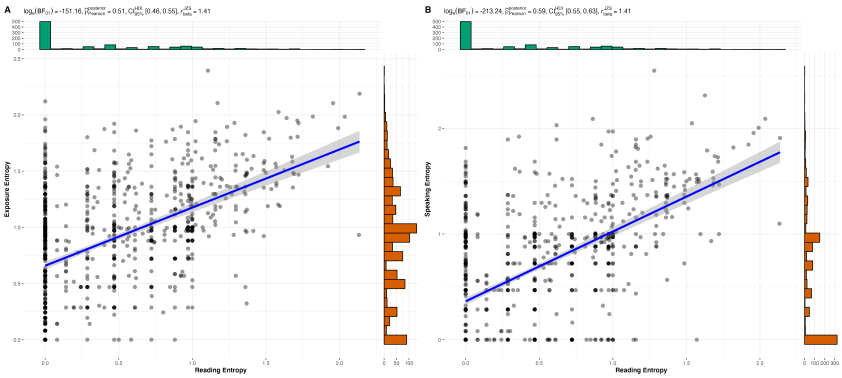

Figure 3: (A) Exposure Entropy (vertical axis) by Reading Entropy (horizontal axis). (B) Speaking Entropy (vertical axis) by Reading Entropy (horizontal axis). In both panels, the histograms for both variables are on the opposing axes.

diversity was attenuated in participants' reports of their most dominant language; there were 24 languages reported as participants' most dominant. The linguistic diversity increased for individuals' second (59 reported languages) and third most dominant languages (54 reported languages). The 15 most commonly spoken languages in each of these groups — all languages, and the most, second, and third dominant languages — are reported in Table 1. English dominates the column for all reported languages and dominant languages. This is unsurprising given that the language of instruction at UBC is generally English. French is the most often reported second and third most dominant language, but only 4 individuals report French as their most dominant language. This presumably is a UBC manifestation of the mother tongue census data in BC, which reports that less than 2% of BC residents are mother tongue speakers of French.

With the exception of English, the number of speakers reporting a particular language as a second most dominant language outnumber the count of individuals who report that same language as a most dominant language. This asymmetry is likely indicative of the rich diversity in home languages in our domestic student population. The home language environment is ultimately usurped by English dominance due to the soci-

etal dominance of English.

The languages reported for the top three dominant languages and all languages are presented visually in word clouds in Figure 4. Interactive versions of these four figures are available for download here.[4] English has been removed from these visualizations since its size inhibits the readability of the other languages.

Meriting special mention are the four First Nations languages reported in our sample: Anishinaabemowin (ISO 639-3: oji/ojg), Chinuk Wawa (IISO 639-3: chn), həńq̓əmiń̓əḿ (ISO 639-3: hur), and Nehiyawewin (ISO 639-3: crk). We celebrate their presence and hope to see an increase in the number of First Nations languages spoken by our student body in future years.

Table 1: The 15 most frequently reported languages and their reported counts, ordered by frequency. The columns present the languages most commonly reported overall (first column), and those most commonly identified as the dominant language (second column), as the second most dominant language (third column), and as the third most dominant language (fourth column).

All languages	**Dominant Language**	**2nd Dominant Lg**	**3rd Dominant Lg**
English (1025)	English (841)	French (190)	French (210)
French (525)	Mandarin (73)	English (165)	Mandarin (94)
Mandarin (348)	Cantonese (36)	Mandarin (131)	Spanish (78)
Spanish (210)	Korean (23)	Cantonese (103)	Japanese (50)
Cantonese (184)	Japanese (8)	Spanish (59)	Cantonese (31)
Japanese (151)	Russian (6)	Korean (41)	Hindi (31)
Korean (129)	French (4)	Punjabi (38)	English (19)
Hindi (66)	Hindi (4)	Tagalog (34)	Korean (19)
Punjabi (57)	Spanish (4)	Japanese (31)	German (17)
German (54)	Farsi (3)	Hindi (23)	ASL (13)
Tagalog (51)	Punjabi (3)	Arabic (20)	Italian (12)
Arabic (38)	Tagalog (3)	German (15)	Tagalog (10)
Russian (30)	Turkish (3)	Portuguese (12)	Punjabi (9)
Italian (26)	Arabic (2)	Vietnamese (11)	Hokkien (8)
ASL (23)	Bahasa Indonesian (2)	Farsi (9)	Russian (6)

[4] On the interactive html files, moving one's cursor over the language name shows the number of individuals who reported that language.

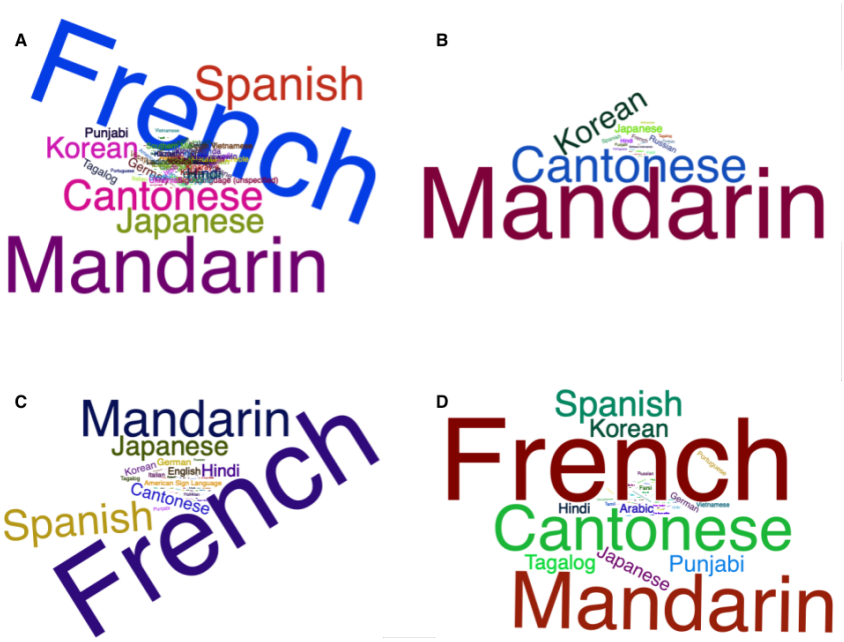

Figure 4: Language frequency word clouds. Word clouds, from right to left, top to bottom, visualizing the frequency of (A) all languages in the data set, (B) most dominant languages, (C) second most dominant languages, (D) third most dominant languages. English is excluded from all clouds.

3.3 Self-ratings

Individuals provided self-proficiency ratings for speaking and understanding in each of their languages. These data are provided in Figure 5 for up to the six most dominant languages. Self-ratings for speaking and understanding are at ceiling for Language 1 (individuals' most dominant language) and gradually lower as the language becomes less dominant.[5] The second and third most dominant languages demonstrate an interest-

[5] The individual data points presented as circles in these boxplots represent responses that are aberrant with respect to the general response distribution. In some cases, these data points are due to a likely misreading of the survey; individuals were asked to enter in their languages in the order of dominance, but some entered their languages in the order of acquisition.

ing asymmetry in language use: Individuals report higher proficiency in understanding than speaking. This difference does not exist for the most dominant languages or fourth most dominant languages and above.

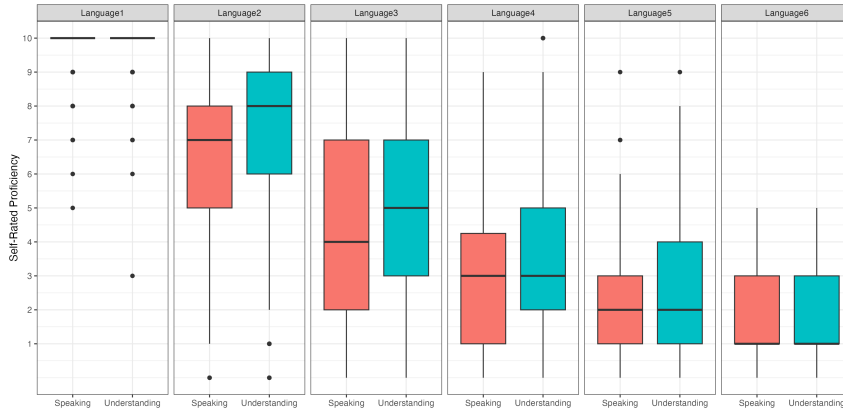

Figure 5: Boxplot visualization of the self-ratings for individuals' top six most dominant languages.

4 Conclusion

This short paper represents a first attempt at describing the linguistic diversity in the student population at UBC. Using questionnaire data collected from over 1000 UBC students (Suite et al., 2023; Lloy et al., 2024), we provide a qualitative description of the types of multilinguals, their various language entropy scores, and the languages they speak.

While we should celebrate the linguistic diversity of our UBC students, this multilingual profile is not unique. The majority population in the world is multilingual (Grosjean, 2021). At the same time, monolingual speakers are often placed on a pedestal, as though their linguistic competence and performance is more authentic than that of a multilingual speaker (Cheng et al., 2021). In celebrating the linguistic variation of UBC students, we also showcase the opportunity to innovate discipline-moving research questions that improve our theory and understanding of linguistic knowledge, behaviour, and processes.

References

Cheng, L. S., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., and Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology*, 12:715843.

Gertken, L. M., Amengual, M., and Birdsong, D. (2014). Assessing language dominance with the Bilingual Language Profile. In Leclercq, P., Edmonds, A., and Hilton, H., editors, *Measuring L2 proficiency: Perspectives from SLA*, pages 208–225. Multilingual Matters, Bristol, UK.

Grosjean, F. (2021). *Life as a bilingual: Knowing and using two or more languages*. Cambridge University Press, Cambridge.

Gullifer, J. W. and Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and Cognition*, 23(2):283–294.

Gullifer, J. W. and Titone, D. (2021). Engaging proactive control: Influences of diverse language experiences using insights from machine learning. *Journal of Experimental Psychology: General*, 150(3):414.

Kałamała, P., Senderecka, M., and Wodniecka, Z. (2022). On the multidimensionality of bilingualism and the unique role of language use. *Bilingualism: Language and Cognition*, 25(3):471–483.

Lloy, L., Patil, K. N., Johnson, K. A., and Babel, M. (2024). Language-general versus language-specific processes in bilingual voice learning. *under review*, pages 1–45.

Luk, G. and Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology*, 25(5):605–621.

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, 50(4):940–967.

Marian, V. and Hayakawa, S. (2021). Measuring bilingualism: The quest for a "bilingualism quotient". *Applied Psycholinguistics*, 42(2):527–548.

Rodriguez-Fornells, A., Krämer, U. M., Lorenzo-Seva, U., Festman, J., and Münte, T. F. (2012). Self-assessment of individual differences in language switching. *Frontiers in Psychology*, 2:388.

Statistics Canada (2023). (table) Census Profile. 2021 Census of Population. Statistics Canada Catalogue no. 98-316-X2021001. Ottawa.

Suite, L., Freiwirth, G., and Babel, M. (2023). Receptive vocabulary predicts multilinguals' recognition skills in adverse listening conditions. *The Journal of the Acoustical Society of America*, 154(6):3916–3930.

Wagner, D., Bekas, K., and Bialystok, E. (2023). Does language entropy shape cognitive performance? A tale of two cities. *Bilingualism: Language and Cognition*, pages 1–11.