15

Issues about Analysing Multilingual Communication in Immigrant Contexts

A. SEZA DOĞRUÖZ

Introduction

Human mobility brings people with diverse backgrounds, languages, and cultures together and is a key feature of immigration. With the availability of computational tools and methods, it is possible to harness large-scale digital data and make predictions about diverse research topics related to immigration. However, analysing language use through text-based datasets in immigrant communities is a complex issue. It involves analyses of multiple languages as well as cultural and social factors. Ignoring this amalgam of factors operating in the immigrant context and just focusing on the automatic analysis of multilingual textual data from the data science perspective will lead to a partial understanding of the community dynamics that lacks useful insights for policy-making targeting these communities (cf. Doğruöz et al., 2021).

In order to avoid these pitfalls, this chapter aims to inform the reader about the (socio)linguistic issues operating in immigrant contexts, clarify the terminology about multilingualism, present findings of relevant studies, and make comparisons (e.g., research questions, types of datasets, methods of analyses) across academic fields (e.g., sociolinguistics and computational areas of research) to address the challenges and opportunities for interdisciplinary research in this area.

Clarification of terminology about language use in immigrant contexts

Meaningful analysis of language use is useful for understanding the dynamics of social interaction in immigrant communities. In order to address the relevant

Proceedings of the British Academy, **251**, 334–349, © The British Academy 2022.

research questions in this context, there is a need to clarify the terminology as follows.

Sociolinguistics investigates language in relation to society (Hudson, 1996). Bilingualism/multilingualism is commonly observed in immigrant contexts, and it is one of the sub-areas of Sociolinguistics. Duration, intensity of contact, and social factors (and their interaction with each other) determine the outcomes of linguistic contact between languages (Thomason, 2001). Initially, speakers of immigrant languages start borrowing lexical items from the host languages and become bilingual/multilingual at later stages of contact (Winford, 2003). However, host languages may also borrow lexical items (e.g., food-related words) from the immigrant languages as well (Winford, 2003). Immigrant community members may use diverse languages in their daily lives depending on the interlocutors and communication contexts. Linguistic competencies of the speakers, their attitudes toward languages (e.g., immigrant vs host language), generation of immigration, formality of communication environment, social rules and conventions may influence the communication patterns in multilingual contexts (Thomason, 2001; Winford, 2003). A heritage language refers to the language/s and/or dialect/s spoken by the immigrant community members. It is usually different than the language of the non-immigrant (i.e., local and/or host) community. Some scholars (e.g., Winford, 2003) associate immigrant languages with the minority groups and the host language with the dominant groups in society. Due to the variation in the backgrounds of the speakers, their immigration histories, minority status, and linguistic proficiencies, the definition of a heritage language and terminology (e.g., 'heritage', 'ancestral', 'immigrant') may vary across immigration contexts.

In the Canadian context, Cummins (1991) excluded English and French from the definition of heritage languages and included languages of the indigenous people and/or immigrants. Fishman (2001; as reported in King and Ennser-Kananen, 2012) defines heritage languages as ancestral languages, which may be indigenous (e.g., Navajo in the US), colonial (e.g., early immigrants of European descent in the US), or may have roots in immigration (e.g., Hmong, Mexican communities in the US). However, these definitions do not necessarily entail or guarantee that the heritage languages are spoken at home or in the community regularly. This view is in line with Kondo-Brown (2003), who stated that a heritage language can be related to the identification with a certain group rather than the linguistic proficiency. For example, new generations may not speak the heritage language of the immigrant community anymore, but they may still use the heritage language to identify themselves with their family roots.

Rothman (2009, p. 156) defined a heritage language as 'a language spoken at home or otherwise readily available to young children and crucially this language is not a dominant language of the larger (national) society'. Similar to Valdés (2001), Rothman expects heritage speakers to obtain some command of heritage language naturalistically along with the majority language spoken in the wider community. On the other hand, Aalberse and Muysken (2013) emphasised the link between language and cultural heritage of the speakers. More specifically, they highlighted the

heritage speakers who are familiar with the language/s spoken in their community but did not hear or speak these languages within their family environment while growing up (e.g., Amboy Malay is spoken mostly by first-generation Indonesian immigrants in the Netherlands but it is not necessarily used for raising children).

Analysing heritage languages of immigrant communities is one of the key areas of research in immigrant contexts. More specifically, linguistic structures (e.g., word order, morphology, vocabulary, phonology) are analysed either in terms of how they change through the influence of the host language or how they resist change and remain intact (maintained) despite the influence of the host language. Heritage languages of Spanish (Silva-Corvalán, 1994) and Russian (Polinsky, 2018a) immigrant communities in the US and Germany, the Turkish immigrant community in the Netherlands (Doğruöz and Backus, 2007, 2009) and Germany (Treffers-Daller et al., 2007), and the Chinese community in immigrant contexts (He, 2008) have already been investigated in terms of changing and maintained linguistic structures. Not all immigrant contexts are the same, and there are controversial findings about the changing vs maintained linguistic structures in relation to the social factors and duration of contact operating in the respective contexts. For example, word order in a heritage language is expected to change due to contact with the host language. However, Doğruöz and Backus (2007) revealed that the basic word order of the Turkish spoken in the Netherlands (heritage language) does not change significantly due to Dutch (host language) influence.

In addition to analysing linguistic changes in heritage languages, some researchers focus on analysing the mixed language communication in immigrant contexts. Example (1) illustrates language mixing between German and Turkish (in bold) at the phrase/sentence level (Karakoç and Herkenrath, 2019).

Example (1)
Hat sie dir was abgegeben? …**Öyle dedin.**
*Did she share anything with you? …**You said so.***

Although first-generation immigrants may or may not learn the language of the host community, the second and third generations (usually) grow up learning the language of the host community. In addition, some immigrant families speak more than one heritage language/dialect at home as well (e.g., French, Lingala, and Swahili spoken by the Zairian immigrant community in Belgium, as reported by Meeuwis and Blommaert, 1998). In these multilingual contexts, switching across languages is very common and it is not merely a linguistic phenomenon. It is influenced by social factors and contexts as well. For example, Ting et al. (2020) analysed multilingual communication in business negotiations in Malaysia, which hosts a major Chinese immigrant community. According to their findings, English was the preferred language of negotiation for more expensive products whereas Hokkien (a dialect of Chinese) and Malay (host community language) were used for cheaper products. In other words, being aware of the language dynamics of the

target group could also be useful for successful business interactions in the particular immigrant context.

Code-switching and language mixing in immigrant communities have been studied extensively across languages and contexts, e.g., Chinese–English switches in the UK by Wei and Milroy (1995) and Wei (2013), Spanish–Moroccan Arabic switches in Spain by Vicente (2007), Greek–English switches in Australia by Alvanoudi (2018), Turkish–Dutch switches in the Netherlands (Backus and van der Heijden, 2002; Papalexakis et al., 2014), Spanish–English switches within the Mexican community in Los Angeles (Silva-Corvalán, 1994) and the Puerto Rican community in New York (Poplack, 1980), and Persian–English switches in Canada (Samar and Meechan, 1998).

From a developmental point of view, it is not always easy to measure the linguistic development of children in immigrant settings while they acquire multiple languages simultaneously or sequentially (Håkansson et al., 2003). If their linguistic abilities are not measured properly, it may not be possible to decide whether or not there is a need for intervention (e.g., extra learning activities, treatment for speech disorders). This concern is also shared by Grimm and Schulz (2014), who drew attention to the need for more research about the misdiagnosis (e.g., under or overdiagnosis) concerning specific language impairment in children who acquire a second language. Their claim is directly relevant for children growing up with multiple languages in immigrant settings, since they may be misdiagnosed for language impairments due to their multilingual skills.

Learning the language of the host community is often considered to be useful for adult immigrants to lead an independent life and take part in the host society. However, adults usually need explicit instruction to learn foreign languages and there is a variation among members of the immigrant communities in terms of educational means, motivations, and resources to learn the language of the host community. According to van Tubergen and Wierenga (2011), age of immigration, proficiency in mother tongue, participation in voluntary activities in the host society, and years of education within the education system of the host community are factors that influence second language learning among male Turkish and Moroccan immigrants in Belgium.

In terms of gender differences, van der Slik et al. (2015) reported a study carried out among immigrants learning Dutch as a second language in the Netherlands. Analyses of the test scores in a language proficiency test (27,199 participants with 49 different language backgrounds) indicated that females perform better than males in spoken and written proficiency tests but there are no significant differences in reading and listening skills.

Janko et al. (2019) compared the level of education vs length of residence for acquisition of the host community language (Greek) by Albanian immigrants. Their findings indicate that the level of education has some influence on grammar-related tasks, whereas length of residence has more influence on the fluency in the host language. Mocciaro (2019) claimed that the literacy level of an immigrant community member in his/her native language has an influence on learning Italian as

a second language. Similarly, Ćatibušić et al. (2021) analysed the needs of Syrian refugees while learning English in Ireland and intercultural support mechanisms. Their findings highlight the importance of learning the language of the host community for the personal well-being, autonomy, and dignity of the participants with an immigrant background in addition to other well-known effects (e.g., employment prospects and social interaction). Similarly, Abou-Khalil et al. (2019) investigated the language learning needs of Syrian refugees who settle down in Lebanon and Germany through participatory design. Through interacting with the participants, the authors found that refugees have different needs (e.g., family reunion, finding jobs) and styles of learning foreign languages. Although they may have access to digital language learning tools and apps (e.g., mobile phones), they also express the need for social learning (e.g., practising German with native German speakers), guidance for self-motivation, discipline, and time-management skills.

Attitudes and policies regarding language use have also been studied in immigrant contexts. From a theoretical point of view, Chick and Hannagan-Lewis (2019) discussed the policies for Syrian refugees in the UK, and Gándara and Rumberger (2009) explained the educational policies toward immigrant students in the US. Following a survey-based method, Utych (2018) studied the influence of using dehumanising words towards immigrants, Zhang and Slaughter-Defoe (2009) investigated the attitudes toward heritage language use among Chinese immigrant families, Hopkins (2015) explained the link between the perception of immigrant community members and their accent in the host community languages, and Ibarraran et al. (2008) studied the perceptions and attitudes of students with immigrant backgrounds toward multilingual communities.

Analysing language use in immigrant contexts through spoken data

Traditionally, linguistic performances of immigrant community members are mostly evaluated through spoken data since their literacy skills and language proficiencies may not be adequate for eliciting written data. Spoken data could be elicited through controlled (e.g., task-based and experimental methods) or less controlled methods (e.g., informal and unstructured conversations). Polinsky (2018b) offered a detailed description of tasks and methods for assessing linguistic proficiencies of heritage speakers and eliciting production data from them. In addition to the scientific descriptions of such tasks, there are also practical issues and challenges about conducting linguistic research in immigrant settings.

In controlled experiments, there is a need for a homogeneous group of participants to measure the influence of social variables (e.g., age, gender, educational and linguistic background) on linguistic variation. However, it is often challenging to find a balanced set of speakers within the immigrant community with similar background characteristics and who are willing to take part in a scientific study. Therefore, researchers may be limited to relatively small sample sizes of participants for their experimental studies. When the study involves an experimental task,

participants are invited to the experimental location (e.g., a university lab) where the researchers have access to the technical equipment. This type of setup protects the participants from external distractions and enables them to focus solely on the experimental task. On the other hand, this safe option may also inhibit the participants from taking part in the study, since it means more time and effort for them to travel to the experimental location. In addition, an isolated experimental environment may also lead to divergences from the usual linguistic practices of the participants in their daily lives.

In the case of conversational data collection, the researcher is not limited to a certain location which could be convenient for the participants. However, s/he still needs to make sure that the recordings (audio and/or video) are of good quality without external distractions and noise (cf. Eppler and Codó, 2016). For example, multiple participants and interruptions during the conversation lead to difficulties in transcription and linguistic analyses later.

Similar to the controlled tasks, it is time-consuming to find and convince participants within the immigrant community to assist in conversational data collection. Unless the researcher is familiar with the community, s/he needs to invest time and create a network to gain the trust of the community members. Once that trust is established, the researcher allocates time and resources to meet with the participants from the immigrant community. Similar to studies in other research domains, the participants could change their minds and decide not to take part in the study as well. In these cases, the researcher needs to secure a large pool of participants who may replace each other in the case of a drop out. However, this practice could be challenging due to relatively small participant pools in immigrant settings.

Participants may vary and accommodate their linguistic performance depending on the context, the topic of conversation, their mood, or the conversation partners they interact with. It is not uncommon to encounter situations in multilingual settings where speakers adapt their linguistic performance in line with what they assume the researcher needs to hear (Thomason, 2007). As a result, the collected data may not reflect the actual language practices in the immigrant community.

In order to analyse the conversational spoken data, transcription is the first step. Although there are some automatic speech-to-text tools, they may not be available for the languages spoken in immigrant contexts. Even when they are available, transcribing conversational data involving multiple participants is rather hard for most automatic tools (Aoki et al., 2006). There is often a need for additional manual error correction. However, this additional correction could be more time-consuming than the manual transcription in the first place. The process of manual transcription is also time-consuming (approximately four hours of transcription for one hour of spoken data) and requires special training depending on the research goals and protocols of the study.

As an alternative to collecting spoken data from immigrant community members, some researchers conduct surveys to assess spoken language use to reach out to a wider participant pool in a shorter time. However, language use reported in surveys may not always mirror the findings in spoken data (Doğruöz and Gries, 2012).

In addition, education level, literacy rate, and linguistic competency (both in the heritage and host languages) among immigrant community members may influence their participation in surveys as well. For example, younger generations in immigrant communities often receive education and schooling in the language of the host community. Although they may speak and understand heritage languages spoken in the community, they may lack literacy skills in these languages (Benmamoun et al., 2013). Therefore, conducting surveys about heritage languages presents a challenge for these groups of speakers. In addition, there are also criticisms about the predictability of replies in surveys as well (Arnulf et al., 2014).

The discussion about the difficulties of finding participants and collecting and analysing spoken data brings us to the question of searching for alternative data sources for linguistic research in immigrant communities. Which other data sources are available? What are the pros and cons of such data sources? The next section explores the answers to these questions.

Analysing language use in online environments for immigrant communities

Qualitative methods of analysis

Androutsopoulos (2007) was one of the early researchers to investigate communication between users with a Persian immigrant background on a digital discussion platform in Germany. His linguistic analyses suggest a link between language and the topic choices of the users. More specifically, the users in his study preferred using Persian for topics related to the traditional culture and entertainment in comparison to other topics of discussion. However, the amount of data, the duration of data collection, and methods of analyses were not explained in detail.

On a similar topic, Androutsopoulos (2015) analysed the online communication among Facebook users in Germany with a Greek immigrant background. With the permission of these users, he collected data (90 pages) from their Facebook accounts over a period of one year and reported one week's communication of this dataset in his study. Although there are descriptions of language switches (Greek–German) within the same post and across posts (e.g., users replying to each other's messages), he does not mention any patterns regarding the reasons why and how these speakers switch across languages.

Dorleijn (2017) analysed the code-switching (Turkish–Dutch) on a dataset (17,000 words) obtained from a digital platform dominated by users with immigrant backgrounds in the Netherlands. Through qualitative methods of analyses, she analysed the complexity of code-switching through the interplay of different grammatical units (e.g., phonology, syntax, morphosyntax). However, it was not possible to make any generalisations due to the small-scale and ongoing data analyses.

In addition to linguistic analyses, there have also been studies investigating the language preferences of immigrant users and their reported experiences in online

environments through surveys. For example, Velázquez (2017) analysed how Spanish immigrant speakers in the US utilise Spanish with respect to their reported literacy, media consumption, and social media use (as a digital resource) through ethnographic interviews and an online survey. The analyses of her data revealed that Spanish was highly (83%) relevant for the reading and writing activities of the heritage speakers in daily life. In terms of social media use, 61% of immigrant speakers reported using Spanish frequently in their online communication. Similarly, 89% of the immigrant speakers reported using Spanish for texting messages. The lower use of Spanish (the heritage language) in social media is linked to the fact that the speakers with an immigrant background in this study attended US colleges where English was the dominant language.

The studies above investigated the language practices of immigrant speakers in online environments using qualitative methods of analyses. However, making comparisons across these studies is challenging. First of all, communication style, text size, and types of media for sharing (e.g., text, photos, videos) differ across social media platforms (e.g., Facebook and Twitter) in online environments. Users may select or limit the audience for their posts and accommodate their language accordingly (Nguyen et al., 2016). These inherent differences across online media services and platforms may influence the linguistic performance of immigrant language users and qualitative studies may not capture this variation systematically.

Second, there is no consensus on how to report the size and duration of the online data for immigrant language analyses in qualitative studies. For example, Androutsopoulos (2015) reported his data size in terms of page numbers, whereas Dorleijn (2017) reported hers in number of words. This difference in measurement of data size leads to difficulties for systematic comparisons across online immigrant communities and their language use practices.

Third, there are not always detailed explanations about the procedures and methods that are followed to extract and compile data from online environments for qualitative studies. Without these explanations, it is hard to carry out similar studies in the future.

Despite these drawbacks, qualitative studies provide insights and prepare the ground for new research questions in online environments using computational methods of analyses in larger and longitudinal datasets. The next subsection focuses on the computational approaches to studying language use and communication in immigrant contexts and digital environments.

Computational methods of analysis

Computational methods are useful for analysing language use on digital platforms with immense data sizes and variation in data sources. To start to analyse multilingual data there is a need to automatically identify the language of the digital post/message content. However, automatic language identification is a challenging task for multilingual digital posts. Examples (2) and (3) illustrate language switches

between Turkish and Dutch (in bold) as they were observed on an online discussion platform used by immigrant community members in the Netherlands (Nguyen and Doğruöz, 2013). Although Turkish and Dutch languages dominate the discussions on the platform, the use of English and/or Moroccan Arabic (another minority population in the Netherlands) is occasionally observed as well (Papalexakis and Doğruöz, 2015). Since these posts also contain creative use of social media language (e.g., unconventional spellings of Turkish characters), it is hard for automatic tools to make meaningful linguistic identification and segmentation.

Example (2)
Mijn dag kan niet stuk :) Cok guzel bir haber aldim.
This made my day:) I received good news.

Example (3)
Kahvalti **met vriendinnen by mij thuis.**
Breakfast with friends at my home.

Off-the shelf automatic methods for language identification and processing are usually developed after training on monolingual and standard written language data (Nguyen et al., 2016). These methods are useful for language identification of text-based documents (e.g., web pages). However, automatic language identification is more challenging for shorter texts (e.g., social media messages) and multilingual digital posts which contain multiple languages and unconventional forms. Nguyen and Doğruöz (2013) was one of the first to tackle this challenge for automatic language identification of multilingual digital posts written by users within an immigrant community. After the initial steps of cleaning the data (e.g., smileys, URLs, links, user names, images), annotation, and normalisation, they developed automatic language identification methods to recognise the language of digital posts at various levels (e.g., word, phrase, and/or post levels) with relatively high accuracy. Considering the difficulty of manual language identification for such large and online datasets, automatic methods save time in classifying the multilingual digital posts (e.g., per language) accurately and preparing the data for further research.

Social and linguistic factors influencing code-switching (i.e., language mixing) in multilingual communication have been described earlier in sociolinguistic studies (e.g., Thomason, 2001; Gardner-Chloros and Edwards, 2004). However, it has not been possible to predict when users/speakers switch across languages automatically. Using the linguistic and non-linguistic features from earlier studies, Papalexakis et al. (2014) were one of the first to develop computational methods to predict code-switching in multilingual communication of immigrant community members on an online platform.

Even when there is no trace of code-switching in communication, linguistic changes could still take place at the underlying linguistic levels in immigrant contexts. If the host language has an influence on these changes, they will not be observed in the non-immigrant variety. Although it may sound simple, it is not easy to detect these underlying linguistic differences between the varieties immediately.

For example, Turkish spoken by the immigrant community in the Netherlands is quite similar to Turkish spoken in Turkey in terms of basic word order (Doğruöz and Backus, 2007). However, there are ongoing changes in multi-word expressions that are translated literally from Dutch into Turkish (Doğruöz and Backus, 2009). Using these changing linguistic cues as features in computational experiments, it has been possible to automatically differentiate between the immigrant (e.g., Turkish spoken in the Netherlands) and non-immigrant varieties (e.g., Turkish spoken in Turkey) of the same language (Doğruöz and Nakov, 2014).

Language use in immigrant contexts has also been used for automatic author identification. From a historical and digital humanities perspective, Müller et al. (2020) trained neural networks for automatic author identification in texts written by the Czech immigrants to Berlin (Germany) in the 18th century.

In addition to focusing on language use by immigrant community members, computational methods are also used to analyse the language use toward immigrant communities and/or minorities. As a collective initiative, Basile et al. (2019) organised a shared task for automatic identification of hate speech against immigrants and women on Twitter in English and Spanish. Along similar lines, Capozzi et al. (2019) introduced a platform to detect and monitor hate speech against immigrants, Roma, and minorities in Italian social media. In a separate paper, Pontiki et al. (2020) automatically analysed verbal aggression and xenophobic attitudes about immigrants in Greek on Twitter.

From a computational social science perspective, Lapesa et al. (2020) introduced a manually annotated (e.g., claim identification, claim classification, actor identification, claim attributes, date, and polarity) dataset to follow the immigration debate on German media. Similarly, Fokkens et al. (2018) developed a tool to automatically detect stereotypes about Muslims in Dutch media in the Netherlands. Through combining questions from social sciences with textual data and computational methods of analyses, the outputs of these projects reflect the attitudes toward immigrants and minorities across contexts and societies.

Depending on the particular context and social factors, members of the immigrant communities are often expected to learn the language of the host community. In addition to the traditional methods of language learning in classroom settings, there are also studies about using computer-assisted tools and services for language learning for members of the immigrant communities. However, these tools should not be mistaken for commercial language learning products, which are not necessarily developed according to the needs and preferences of immigrant community members. In other words, the commercial success or wide availability of a digital language learning tool does not guarantee the desired results for language learning in immigrant communities. Instead, there is a need for research-based language learning tools that are specifically designed for the specific immigrant context, e.g., a computer-assisted vocabulary learning tool, SweLLex, for learners of Swedish with an immigrant background in Sweden (Volodina et al., 2016)).

As the studies in this section illustrate, the availability of computational tools to harness very large and longitudinal datasets makes online data attractive for analysing communication and language use among immigrant community members. However, analysing online communication data in immigrant contexts introduces new challenges from computational, linguistic, and social sciences perspectives as well.

First of all, not all members of the immigrant community may have access to the internet, as is widely assumed. For example, older generations of the immigrant communities often lack the resources and digital skills to use internet-based devices and services (Chen et al., 2020). In addition, purchasing smart devices and paying for internet-based services could be financially cumbersome for immigrant community members with restricted income. Therefore, analysing language use in online environments for immigrant communities faces challenges about representativeness.

Second, analysing large-scale communication data among immigrant community members in online environments has several linguistic challenges. In order to analyse the content of the multilingual communication, there is a need for an automatic language identification process (Nguyen and Doğruöz, 2013; Frey et al., 2019). However, creative language use in social media communication and language mixing lead to difficulties for automatic language identification processes. Although off-the shelf computational tools are available, these tools are usually developed through training on monolingual and standard written language data. If the training data is in the standard and written language, it may not capture the dynamic, creative, and ever-changing aspects of multilingual language used by immigrant community members in digital environments. In addition, most of these tools are developed for major languages (e.g., English) and may not be available for processing the less spoken and/or low resource languages in immigrant contexts.

Third, unless the data services are publicly accessible and/or immigrant community members provide their consent, it is not easy to collect social media data. Even in publicly accessible cases, there may be quota restrictions that may prohibit extensive research about language use in immigrant communities.

Finally, the background (e.g., age, education, gender) and socioeconomic status of the users are difficult to verify in digital environments (Nguyen et al., 2016). Although some background information about users may be available, these data are often incomplete and/or filled in randomly to hide the true identity of the users. Lack of personal information about the users prevents researchers form making links between language use and social factors to analyse communication in online communication for immigrant communities.

Discussion and conclusion

The goal of this study was to inform the reader about the scientific discussion around language use by and about immigrant communities through making comparisons across academic fields, research questions, datasets, and methodologies.

There is no unique remedy to avoid the pitfalls and carry out a 'perfect' study that reflects and captures multilingual language use with all its diversity in immigrant contexts. For example, qualitative studies investigating multilingual language use within immigrant communities usually focus on small datasets (e.g., the datasets described in the 'Qualitative methods of analysis' section) in depth. Although their results may not be generalisable across all users or contexts, the descriptions of the data or reported results could serve as starting points for hypothesis testing in larger and longitudinal datasets for research on immigrant languages using computational methods of analysis.

Online sources offer larger and longitudinal datasets for immigrant language research in comparison to traditional data collection and analysis methods. However, extracting adequate and representative datasets from online sources requires certain computational skills (e.g., programming skills, web crawling, database management systems), time, and expertise (e.g., cleaning and classifying data into meaningful categories). In addition, systematic analysis of these datasets requires expertise in computational methods (e.g., machine learning and natural language processing) which may not be the main focus of research training in many humanities and/or social sciences programmes. Similarly, data scientists may have the computational skills and tools to harness large-scale online data, but lack the theoretical background to interpret the interwoven linguistic and social factors operating in multilingual immigrant contexts. Therefore, there is a need for collaboration across academic fields for research on language use in immigrant communities.

Finally, big data is not very meaningful on its own. Only after extensive cleaning and restructuring can one look for patterns and analyse the data to make it more meaningful for research purposes. In order for multilingual textual data to be processed, there is a need for normalisation and standardisation stages; see also Box 8.1 in Kim et al. (2022), published as Chapter 8 in this volume. However, valuable information about multilingualism and language variation (e.g., unique linguistic styles associated with certain speakers and contexts) within immigrant communities could be lost during the cleaning and normalisation stages (Nguyen et al., 2016; Eisenstein, 2013).

References

Aalberse, S. and Muysken, P. (2013), 'Language contact in heritage languages in the Netherlands', *in* J. Duarte and I. Gogolin, eds, *Linguistic Superdiversity in Urban Areas: Research Approaches*, Vol. 2, John Benjamins Publishing, Amsterdam, 253–74.

Abou-Khalil, V., Helou, S., Flanagan, B., Pinkwart, N., and Ogata, H. (2019), 'Language learning tool for refugees: Identifying the language learning needs of Syrian refugees through participatory design', *Languages* 4(3), 71.

Alvanoudi, A. (2018), 'Language contact, borrowing and code switching: A case study of Australian Greek', *Journal of Greek Linguistics* 18(1), 3–44.

Androutsopoulos, J. (2007), 'Language choice and code-switching in German-based diasporic web forums', *in* B. Danet and S. C. Herring, eds, *The Multilingual Internet: Language, Culture, and Communication Online*, Oxford University Press, Oxford, 340–61.

Androutsopoulos, J. (2015), 'Networked multilingualism: Some language practices on Facebook and their implications', *International Journal of Bilingualism* 19(2), 185–205.

Aoki, P. M., Szymanski, M. H., Plurkowski, L., Thornton, J. D., Woodruff, A., and Yi, W. (2006), 'Where's the "party" in "multi-party"? Analyzing the structure of small-group sociable talk', *in Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work*, 393–402.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., and Bong, C. H. (2014), 'Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour', *PLOS One* 9(9), e106361.

Backus, A. and van der Heijden, H. (2002), 'Language mixing by young Turkish children in the Netherlands', *Psychology of Language and Communication* 6(1), 55–73.

Basile, V., et al. (2019), 'Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter', *in Proceedings of the13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, 54–63.

Benmamoun, E., Montrul, S., and Polinsky, M. (2013), 'Heritage languages and their speakers: Opportunities and challenges for linguistics', *Theoretical Linguistics* 39(3–4), 129–81.

Capozzi, A. T., et al. (2019), 'Computational linguistics against hate: Hate speech detection and visualization on social media in the "Contro L'Odio" project', *in Proceedings of the Sixth Italian Conference on Computational Linguistics, CLiC-it 2019*, 2481.

Ćatibušić, B., Gallagher, F., and Karazi, S. (2021), 'Syrian voices: An exploration of the language learning needs and integration supports for adult Syrian refugees in Ireland', *International Journal of Inclusive Education* 25(1), 22–39.

Chen, X., Östlund, B., and Frennert, S. (2020), 'Digital inclusion or digital divide for older immigrants? A scoping review', *in* Q. Gao and J. Zhou, eds, *Proceedings of the International Conference on Human–Computer Interaction*, Springer, New York, 176–90.

Chick, M. and Hannagan-Lewis, I. (2019), 'Language education for forced migrants: Governance and approach', *Languages* 4(3), 74.

Cummins, J. (1991), 'Introduction', *The Canadian Modern Review* 47(4), 601–5.

Doğruöz, A. S. and Backus, A. (2007), 'Postverbal elements in immigrant Turkish: Evidence of change?', *International Journal of Bilingualism* 11(2), 185–220.

Doğruöz, A. S. and Backus, A. (2009), 'Innovative constructions in Dutch Turkish: An assessment of ongoing contact-induced change', *Bilingualism: Language and Cognition* 12(1), 41–63.

Doğruöz, A. S. and Gries, S. T. (2012), 'Spread of on-going changes in an immigrant language', *Review of Cognitive Linguistics* (published under the auspices of the Spanish Cognitive Linguistics Association) 10(2), 401–26.

Doğruöz, A. S. and Nakov, P. (2014), 'Predicting dialect variation in immigrant contexts using light verb constructions', *in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1391–5.

Doğruöz, A. S., Sitaram, S., Bullock, B. E., and Toribio, A. J. (2021), 'A survey of code-switching: Linguistic and social perspectives for language technologies', *in Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, Association for Computational Linguistics, 1654–66.

Dorleijn, M. (2017), 'Is dense codeswitching complex?', *Language Sciences* 60, 11–25.

Eisenstein, J. (2013), 'What to do about bad language on the internet?', *in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, 359–69.

Eppler, E. D. and Codó, E. (2016), 'Challenges for language and identity researchers in the collection and transcription of spoken interaction', *in* S. Preece, ed., *The Routledge Handbook of Language and Identity*, Routledge, London, 304–19.

Fokkens, A., Ruigrok, N., Beukeboom, C., Sarah, G., and Van Atteveldt, W. (2018), 'Studying Muslim stereotyping through microportrait extraction', *in Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Frey, J.-C., Stemle, E. W., and Doğruöz, A. S. (2019), 'Comparison of automatic vs. manual language identification in multilingual social media texts', *in* C. R. Wigham and E. W. Stemle, eds, *Building Computer-Mediated Communication Corpora for Socio-Linguistic Analysis*, University of Clermont Publications, France, 47–69.

Gándara, P. and Rumberger, R. (2009), 'Immigration, language, and education: How does language policy structure opportunity?', *Teachers College Record* 111(3), 750–82.

Gardner-Chloros, P. and Edwards, M. (2004), 'Assumptions behind grammatical approaches to code-switching: When the blueprint is a red herring', *Transactions of the Philological Society* 102(1), 103–29.

Grimm, A. and Schulz, P. (2014), 'Specific language impairment and early second language acquisition: The risk of over- and underdiagnosis', *Child Indicators Research* 7(4), 821–41.

Håkansson, G., Salameh, E.-K., and Nettelbladt, U. (2003), 'Measuring language development in bilingual children: Swedish-Arabic children with and without language impairment', *Linguistics* 41(2), 255–88.

He, A. W. (2008), 'Chinese as a heritage language', *in* W. S.-Y. Wang and C. Sun, eds, *The Oxford Handbook of Chinese Linguistics*, Oxford University Press, New York, 578–90.

Hopkins, D. J. (2015), 'The upside of accents: Language, inter-group difference, and attitudes toward immigration', *British Journal of Political Science* 45(3), 531–57.

Hudson, R. A. (1996), *Sociolinguistics*, Cambridge University Press, Cambridge.

Ibarraran, A., Lasagabaster, D., and Sierra, J. M. (2008), 'Multilingualism and language attitudes: Local versus immigrant students' perceptions', *Language Awareness* 17(4), 326–41.

Janko, E., Dąbrowska, E., and Street, J. A. (2019), 'Education and input as predictors of second language attainment in naturalistic contexts', *Languages* 4(3), 70.

Karakoç, B. and Herkenrath, A. (2019), 'Understanding retold stories: The marking of unwitnessed events in bilingual Turkish', *Turkic Languages* 23(1), 81–121.

Kim, J., Pollacci, L., Rossetti, G., Sîrbu, A., Giannotti, F., and Pedreschi, D. (2022), 'Twitter data for migration studies', *in* A. A. Salah, E. E. Korkmaz, and T. Bircan, eds, *Data Science for Migration and Mobility*, Proceedings of the British Academy, British Academy / Oxford University Press, London.

King, K. A. and Ennser-Kananen, J. (2012), 'Heritage languages and language policy', *in* C. A. Chapelle, ed, *The Encyclopedia of Applied Linguistics*, Wiley, Chichester.

Kondo-Brown, K. (2003), 'Heritage language instruction for post-secondary students from immigrant backgrounds', *Heritage Language Journal* 1(1), 1–25.

Lapesa, G., Blessing, A., Blokker, N., Dayanık, E., Haunss, S., Kuhn, J., and Padó, S. (2020), 'DEbateNet-mig15: Tracing the 2015 immigration debate in Germany over time', *in Proceedings of the 12th Language Resources and Evaluation Conference*, 919–27.

Meeuwis, M. and Blommaert, J. (1998), 'A monolectal view of code-switching: Layered code-switching among Zairians in Belgium', *in* P. Auer, ed, *Code-Switching in Conversation: Language, Interaction and Identity*, Routledge, Abingdon, 76–98.

Mocciaro, E. (2019), 'Emerging constructions in the L2 Italian spoken by low literate migrants', *Languages* 4(4), 86.

Müller, K., Tikhonov, A., and Meyer, R. (2020), 'LiViTo: Linguistic and visual features tool for assisted analysis of historic manuscripts', *in Proceedings of the 12th Language Resources and Evaluation Conference*, 885–90.

Nguyen, D. and Doğruöz, A. S. (2013), 'Word-level language identification in online multilingual communication', *in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 857–62.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., and De Jong, F. (2016), 'Computational sociolinguistics: A survey', *Computational Linguistics* 42(3), 537–93.

Papalexakis, E. and Doğruöz, A. S. (2015), 'Understanding multilingual social networks in online immigrant communities', *in Proceedings of the 24th International Conference on World Wide Web*, 865–70.

Papalexakis, E., Nguyen, D., and Doğruöz, A. S. (2014), 'Predicting code-switching in multilingual communication for immigrant communities', *in Proceedings of the First Workshop on Computational Approaches to Code Switching (EMNLP 2014)*, Association for Computational Linguistics, 42–50.

Polinsky, M. (2018a), 'Bilingual children and adult heritage speakers: The range of comparison', *International Journal of Bilingualism* 22(5), 547–63.

Polinsky, M. (2018b), *Heritage Languages and their Speakers*, Cambridge University Press, Cambridge.

Pontiki, M., Gavriilidou, M., Gkoumas, D., and Piperidis, S. (2020), 'Verbal aggression as an indicator of xenophobic attitudes in Greek Twitter during and after the financial crisis', *in Proceedings of the Workshop about Language Resources for the SSH Cloud*, 19–26.

Poplack, S. (1980), 'Sometimes I'll start a sentence in Spanish y termino en espagñol: Toward a typology of code-switching', *Linguistics* 18, 581–618.

Rothman, J. (2009), 'Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages', *International Journal of Bilingualism* 13(2), 155–63.

Samar, R. G. and Meechan, M. (1998), 'The null theory of code-switching versus the nonce borrowing hypothesis: Testing the fit in Persian–English bilingual discourse', *International Journal of Bilingualism* 2(2), 203–19.

Silva-Corvalán, C. (1994), *Language Contact and Change: Spanish in Los Angeles*, Clarendon Press, Oxford.

Thomason, S. (2007), 'Language contact and deliberate change', *Journal of Language Contact* 1(1), 41–62.

Thomason, S. G. (2001), *Language Contact*, Edinburgh University Press, Edinburgh.

Ting, S.-H., Then, D. C.-O., and Ong, O. G.-B. (2020), 'Prestige of products and code-switching in retail encounters', *International Journal of Multilingualism* 17(2), 215–31.

Treffers-Daller, J., Özsoy, A. S., and Van Hout, R. (2007), '(In)-complete acquisition of Turkish among Turkish–German bilinguals in Germany and Turkey: An analysis of complex embeddings in narratives', *International Journal of Bilingual Education and Bilingualism* 10(3), 248–76.

Utych, S. M. (2018), 'How dehumanization influences attitudes toward immigrants', *Political Research Quarterly* 71(2), 440–52.

Valdés, G. (2001), 'Heritage language students: Profiles and possibilities', *in* J. Peyton, D. Ranard, and S. McGinnis, eds, *Heritage Languages in America: Preserving a National Resource*, Delta Systems Company Inc., McHenry, IL.

van der Slik, F. W., van Hout, R. W., and Schepens, J. J. (2015), 'The gender gap in second language acquisition: Gender differences in the acquisition of Dutch among immigrants from 88 countries with 49 mother tongues', *PLOS One* 10(11), e0142056.

van Tubergen, F. and Wierenga, M. (2011), 'The language acquisition of male immigrants in a multilingual destination: Turks and Moroccans in Belgium', *Journal of Ethnic and Migration Studies* 37(7), 1039–57.

Velázquez, I. (2017), 'Reported literacy, media consumption and social media use as measures of relevance of Spanish as a heritage language', *International Journal of Bilingualism* 21(1), 21–33.

Vicente, Á. (2007), *Two Cases of Moroccan Arabic in the Diaspora*, Routledge, Abingdon.

Volodina, E., Pilán, I., Llozhi, L., Degryse, B., and François, T. (2016), 'Swellex: Second language learners' productive vocabulary', *in Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, 76–84.

Wei, L. (2013), 'Codeswitching', *in* R. Bayley, R. Cameron, and C. Lucas, eds, *The Oxford Handbook of Sociolinguistics*, Oxford University Press, New York, 360–78.

Wei, L. and Milroy, L. (1995), 'Conversational code-switching in a Chinese community in Britain: A sequential analysis', *Journal of Pragmatics* 23(3), 281–99.

Winford, D. (2003), *An Introduction to Contact Linguistics*, Wiley-Blackwell, Oxford.

Zhang, D. and Slaughter-Defoe, D. T. (2009), 'Language attitudes and heritage language maintenance among Chinese immigrant families in the USA', *Language, Culture and Curriculum* 22(2), 77–93.