10 Challenges for the Development of Linked Open Data for Research in Multilingualism

María Blume, Isabelle Barrière, Cristina Dye, and Carissa Kang

Introduction

The study of multilinguals is fundamental for linguistic research, since multilinguals constitute the majority of the world population as well as a growing proportion of the population in many countries (McCabe et al. 2013; Gambino, Acosta, and Grieco 2014; Special Eurobarometer 386, 2012). We use the term *multilingual* to refer to speakers who know more than one language to a variable extent, regardless of when they learned those languages (thus encompassing simultaneous and sequential bilinguals, as well as second-language speakers/learners and heritage speakers).

The multilingual brain is dealing with more than one linguistic system, and thus theories of language structure and cognitive models of language development or processing must account for language use, processing, and acquisition by all people who know more than one language. The language abilities of multilinguals change throughout their lifetime, so our data need to capture differences in a person's ability through time, including language attrition when or if it occurs. Studies on bilingualism, multilingualism, second-language acquisition, and language attrition have grown exponentially in the last decades, and their data need to be accessible and comparable so that all the research community can benefit from it.

With these facts in mind, we discuss three major issues related to research with multilingual populations:

- Requirements for conducting research with multilingual populations
- Challenges for the development of Linguistic Linked Open Data (LLOD) in the field of multilingual acquisition
- Capacities and needs of any primary research tool that would allow us to achieve the vision of LLOD

Requirements for Conducting Research with Multilingual Populations

Several methodological issues arise in doing research with human participants in the field of linguistics (Blume and Lust 2017). However, working with multilingual participants creates additional challenges.

Complexity Inherent in Multilingual Population

Most people become multilingual because their circumstances force them to do so. These different circumstances can be summarized as follows[1] (Austin, Blume, and Sánchez 2015, 39):

- Individuals who are multilingual from birth either speaking two languages at home or one at home and one outside the home
- Early multilinguals who start learning a second language sometime after birth but still during childhood, typically speaking one language at home and one outside the home
- Individuals who learned a second language in adulthood and speak it mostly for work-related activities
- Individuals who, as a result of immigration, must learn a second language to survive in the new country, or who spoke a minority language in their own country but must learn the dominant language in their new country

Even within groups, multilingual speakers differ greatly in many respects. They may range from monolingual speakers having limited exposure to a second language (e.g., a few hours per week in a classroom), to more fully multilingual speakers (e.g., people who learned both languages simultaneously in childhood, using both frequently in everyday life across various situations). A speaker's proficiency may also change across different contexts (Fishman 1965) and throughout the speaker's lifetime (more so than that of a monolingual speaker), requiring assessments across various situations and at multiple points in their language development.

Determining the nature of a participant's multilingualism is fundamental for research since it has effects on such important areas as further linguistic development, cognition, and literacy.

Challenges for Research Posed by Population Complexity

Because many complex factors account for a multilingual person's language profile, it is often challenging to select individuals for study who have only some specific characteristics that a researcher wishes to compare, or to form groups of speakers of similar characteristics that one can then compare to different groups (in the same study or across studies). Detailed metadata must be carefully collected and documented to allow for such comparisons.

Those who are even considered to be possible participants change across studies. Depending on the type of research and how researchers define multilingualism, types of

participants who are recruited may differ considerably. Some studies may call a participant *multilingual*, for example, only if he or she is a simultaneous multilingual—someone who learned two or more languages from birth with only a few days of difference between the beginning of exposure to each language (De Houwer 2009). Others will count students in their first stages of classroom-only exposure to a second language as being *multilingual*.

This may be largely attributed to the fact that there is no single definition in the field nor any clear set of criteria for deciding who is a multilingual speaker (Hamers and Blanc 1989; Grosjean 2010; Mackey 2012), arising from the complexity of the multilingualism phenomenon. Criteria used to characterize multilingual speakers include psychological ones (such as the degree of competence in one language versus another; Lambert 1955); the domains of competence (spoken and/or written production, oral comprehension and/or reading abilities; Bialystok 2007); and sociological ones, such as the contexts of use of a language and whether it was acquired in a naturalistic context or a formal setting (Fishman 1965). To complicate the matters further, terms commonly used to classify speakers in the literature, such as *balanced*, *dominant, native,* or *beginner*, refer to different concepts and are related to the different types of criteria, making comparison across studies less direct (Flege, MacKay, and Piske 2002; Genesee 1989; Genesee, Nicoladis, and Paradis 1995; Hamers and Blanc 1989).

Although some criteria undoubtedly exist in our field, not all relevant factors are systematically taken into account (for example, speakers are classified according to age of acquisition, but patterns of use may not be considered). At other times, speakers are carefully selected, though the criteria for selection are not completely or clearly detailed in publications (Grosjean 2008, 2011; Thomas 1994). It may not be realistic to expect all researchers to agree on the exact definitions of terms or to have them list in their research articles every last criterion used for classifying speakers, largely because of space limitations. However, the value of each study data can be incremented if researchers make this detailed information available online, so that other researchers can decide whether the population studied fits the profile they are looking for, either for further research with the same data or for comparison with other data.

To be able to compare groups of speakers, researchers need to control several potentially confounding factors in order to conclude that a speaker's multilingualism modulates, for example, the use of a particular linguistic structure or leads to a proposed cognitive difference. Two such factors are the context of acquisition and the type or level of multilingualism involved.

Context of Acquisition

To be able to establish the context of acquisition of an individual's languages, a researcher needs to have information on the speaker's language history—such as "Which languages has the participant acquired?" or "When and how were the languages acquired?" Age of acquisition is a good predictor of further language proficiency, with people who acquire a second language early usually outperforming speakers who acquired the language later in

terms of linguistic abilities. The status of the language in the speaker's society is also important. Speakers tend to use and maintain languages that the majority of the population speaks and that their societies consider important more than they use and maintain minority languages, often because of a lack of educational resources and opportunities to use the language in daily life. The relationship between a speaker's languages (e.g., How closely related are they? Which aspects of the language systems are similar and which are not?) should be also taken into account. A language that is closely related to a person's native language may be easier to acquire for a second language speaker than a more distantly related one (Grosjean 2008, 2011).

This information, and a participant's biographical data (such as sex, age, and socioeconomic status), make it possible to begin to assemble a language profile for the participant.

Type/Level of Acquisition

Information is also needed on the speaker's knowledge and use of each of his or her languages. This information is relevant to research for the reasons listed here, among others:

- *Language proficiency in the four skills (speaking, comprehension, reading, and writing) in each language:* Speakers may be similar in their comprehension skills but quite different in their expressive skills; some highly competent speakers may even be illiterate, and literacy has been shown to affect language processing.
- *Function of languages:* Which languages are used for what purposes? In what context and to what extent is each language used? Some speakers may have an extremely developed home-related lexicon in a language but not an academic one, or they may be able to have conversations about certain topics but not others. This may affect their performance on certain linguistic tests or their self-perception as multilingual speakers.
- *Language stability:* Are one or several languages still being acquired? Has a certain level of language stability been reached? In the past, wrong conclusions on the cognitive or linguistic capacities of multilingual speakers have often been reached when not taking into account that the subjects were incipient language learners of the language used for testing them.
- *Language modes:* This refers to the duration and frequency spent by the participant in both monolingual and multilingual modes. The mode may affect performance, especially in processing tasks. A speaker with less code-switching experience (i.e., alternating between more than one language) may provide very different answers to a study searching for syntactic or pragmatic constraints on code-switching than would a more experienced one.

Most studies gather information on language proficiency. However, language proficiency is not always understood or operationalized in the same way, and different instruments are frequently used to measure it. For example, some studies measure language competence

(i.e., knowledge of the grammatical rules of a language), while others assess proficiency or communicative competence (i.e., the knowledge and ability to use language in socially acceptable ways, including grammatical, sociolinguistic, strategic, and discourse competence; Canale and Swain 1980; Canale 1983). Researchers are now more aware of this difference, and studies today tend to be more precise on their definition of competence.

To enable comparisons across multilingual speakers and groups, researchers need data on how their level of multilingualism was determined, including whether competence or proficiency were studied, the specific measures and tests used in assessing them, the task modality (e.g., comprehension or production), and the linguistic domain tested (e.g., vocabulary, grammar, pronunciation).

Sometimes speakers' proficiency is never directly measured—for instance, studies with L2 (second-language) learners are frequently conducted in formal settings (universities and schools) and course level is often used as a proxy measure for the speaker's proficiency (Thomas 1994). The problem with this approach is that courses that are officially at the same level (say, intermediate) may not actually be equally demanding at different institutions or across languages in the same institution.

When studies do gather independent data, questionnaires are frequently used. The questionnaires vary across labs in terms of length and type of information asked. Some are very short (approximately 10 questions), while others are much longer.[2] Although shorter questionnaires may be more practical, it is sometimes challenging to tell whether the results of a given study will generalize to other groups of multilingual speakers without detailed information.[3] Moreover, not all questionnaires of similar length ask the exact same questions about the speaker's acquisition, proficiency, and use.

Parental questionnaires have long been used as a measure of child language development (Gutierrez-Clellen and Kreiter 2003; Squires, Bricker, and Potter 1997; Thordaardottir and Weismer 1996), a recent study found that a more precise estimate of grammar can be achieved by adding a direct observation measure to the child's evaluation. In the study by Lust et al. (2014), two Korean-dominant children who were four years of age with Korean as their L1 (first language) and English as their L2 were assessed through a questionnaire and also an elicited imitation task. The parental reports and general linguistic histories predicted similar proficiency for the two children. However, in the experimental task, one child demonstrated a more developed level of grammar in his production in both of his languages than the other one. Thus, children who seem to be very similar according to parental reports can differ tremendously on their performance in experimental tasks both in the L2 and in the L1.[4]

While studies sometimes use standardized instruments to assess the development of linguistic abilities of the speaker, most such instruments exist strictly in English or only in a few well-studied languages (although a collection of instruments for research on second language acquisition can now be found through IRIS).[5] New instruments (or translations of existing instruments) that are reliable have proven difficult to create (e.g., Esquinca,

Yaden, and Rueda 2005; Gathercole 2010; Paradis, Emmerzael, and Duncan 2010; Peña 2007), and it can take years to validate and norm them (Alcock et al. 2015). Some instruments measure only some aspects of linguistic knowledge—for example, vocabulary (e.g., Peabody Picture Vocabulary Test, Dunn and Dunn 2007). Most are normed on the basis of monolingual speakers (Espinosa and García 2012; Barrière 2014) and, as is well-known, bilinguals are not two monolinguals in one person and therefore cannot be compared directly to monolinguals (Grosjean 1989; Barac et al. 2014; and Sánchez 2015, among others).

Awareness of these differences among speakers and acknowledgment of the importance of having detailed information on their evaluation or classification has grown with the development of the field. This awareness is leading researchers to use more than one method to select and evaluate participants, as well as to collect more-careful metadata on each one. This is good for the field but it increments the amount of data and metadata that we professionals need to collect, store, and share.

Additional challenges may arise when researchers work with less-studied languages, in multilingual areas. It may often be the case that at least one of these less-studied languages is acquired by children and used in contexts where they constitute a minority language (Baker, van den Bogaerde, and Woll 2008). For instance, in New York City, 50% of children use a language other than English at home, including Haitian Creole, Yiddish, African Languages, Tagalog, Urdu, or Gujarati, and many others (García, Zakharia, and Otcu 2013, 13).

Even when both languages are well documented in adults, they may be less so for children. For example, while the use of both Spanish and English by Spanish-speaking adults in New York City has been documented (e.g., Otheguy and Zentella 2012 and references therein), little is known about the contextual factors that affect the acquisition of both languages in multilingual children. Barrière et al. (2015) investigated the acquisition of subject–verb agreement markers in English and Spanish by low socioeconomic status (SES) children of Mexican descent with Spanish as an L1: Their speakers were homogeneous with respect to the variety of Spanish they were acquiring, ensuring that the effects of bilingual acquisition were not confounded with dialectal variation in Spanish that impacts the speed and pattern of acquisition of Spanish grammatical inflections (e.g., Miller and Schmitt 2010). It was, however, difficult to determine the characteristics of the variety of English (such as Mainstream American English versus Chicano English or African American English or other Caribbean English) spoken by each participant. That determination was needed because different language varieties exhibit different norms regarding the third-person singular marker, and also because monolingual English-speaking children enrolled in the same preschools as their bilingual or trilingual colleagues perform differently on experimental tasks. That difference arises depending on the variety of English they are acquiring: Only preschoolers who are acquiring Mainstream American English (but not those acquiring other varieties, such as African American English or Jamaican

English) show evidence of comprehension in a video matching task that requires the exclusive use of the third-person singular–*s* to determine number of participants (examples of stimuli: *the boy skip**s*** versus *the boy**s** skip**ø***; Barrière et al. 2016).

The challenge of determining participants' language variety is significantly exacerbated when the languages to which the children are exposed to have not been well documented. This is the case of the Hasidic Yiddish-speaking community—a rapidly increasing population in two areas of Brooklyn—whose members speak varieties that come from three distinct areas in Eastern Europe that are now in contact both with one another and with English (Barrière 2010).

Studies conducted on multilinguals also frequently gather information on the attitudes that such speakers and their communities have about the languages they speak, attitudes that are relevant for explaining language dominance. Language preference has been shown to contribute to children's developing language abilities (Armon-Lotem et al. 2014). Some studies require more specific information; for example, Kang, Martohardjono, and Lust (unpublished manuscript) asked participants to self-rate the frequency of their daily language-mixing, the extent of their multilingualism, and even their attitudes toward code-switching, so as to investigate how code-switching attitudes and habits relate to code-switching fluency. Although language preference and code-switching behavior may affect multilingual development, they are rarely included in participant profiles.

All the previous examples illustrate how multilingual research requires extensive and detailed metadata to be gathered from each participant, which then need to be made accessible and searchable. The main issue is that more variability occurs among multilingual speakers' proficiencies than among those of monolingual speakers, and therefore researchers need to be able to describe multilingual participants in precise ways that are both meaningful and consistent across the field. These extensive data then must be documented and shared so that they benefit the wider research community.

Development of Linguistic Linked Open Data (LLOD)

Metadata

All the aspects of conducting research with multilingual populations discussed in the section "Requirements for Conducting Research with Multilingual Populations" point to the necessity of gathering extensive metadata on each participant before even testing them on the particular linguistic aspect of interest—metadata that are more extensive than for monolinguals. These metadata need to include not only the biographical and language context data mentioned above but also the specific measures used to classify the speakers' language abilities. Furthermore, multiple measures may be associated with each participant, since his or her abilities may change with age or development.

Most important, all these metadata must be linked to the particular data of the participant being studied.

Advantages of Accessible Extensive Metadata

Published studies should provide as much information as possible about their participants, the criteria used to classify multilinguals in various groups, and the language assessment tools used in the study; however, this is not always possible owing to length constraints. Having this information available online, then, would greatly facilitate research and calibration across studies.

Since it is often difficult to identify participants with a shared profile, studies of multilingual populations usually have small sample sizes. A tool that allows researchers to conduct meta-analysis studies (e.g., combining data collected from studies that employed a given task, or studies that focused on the development of a particular grammatical element) would certainly be advantageous, yet such analyses can only be properly conducted if we have access to exhaustive metadata for all studies.

Challenges

This extensive metadata documentation is now partially possible through some online tools (e.g., the DTA tool,[6] the Language Archive,[7] the Open Science Framework [OSF][8]), although the metadata, while available, are not always searchable automatically for less technically proficient researchers and the tools used to create them are often incompatible.

Gathering such detailed and often-personal data has the advantage of allowing us researchers to build an accurate linguistic profile of a multilingual speaker, but this brings with it the challenge of protecting the individual's identity, especially since multilingual speakers may come from minority and at-risk populations.

Data Challenges

In many cases, metadata and primary data either are not online or are not searchable; for example, the Electronic World Atlas of Varieties of English (EWAVE),[9] classifies varieties of English according to whether it is an L1 or L2 for the speakers, yet it provides no metadata on the informants. Many studies of multilingualism, for example, gather data and metadata through questionnaires. Although the results of a given study may be available online, the questionnaires themselves often are not, and at best they are attached as PDF forms to participants' metadata. This situation creates difficulties for comparison, calibration, and replication of studies.

Data Markup Challenges

Cross-Linguistic Differences

The main problem that multilingual data present is precisely that of being *multilingual*. Structures require an additional level of coding, indicating which language they belong to (in those cases where the researcher can even confidently decide the language). While this may be easy to do for independent words or one-language utterances, it can be more chal-

lenging in multi-language utterances and in utterances where words themselves contain morphemes from more than one language.

Enabling the cross-linguistic analyses needed to compare a multilingual speaker's two or more languages requires a rich markup capacity. Coding systems for the two or more languages need to be available for the researcher, and specific coding conventions may also need to be created, depending on the languages involved, since some phenomena common in the speech of a number of multilingual communities may be rare or nonexistent in others. For example, when analyzing the imitation of relative clauses in three languages (Flynn and Lust 1981; Foley 1996; Somashekar 1999), coding was tailored to the similarities of these structures across languages: lexically headed versus free relative, type of *wh*-word heading the relative clause, and the similarities of the expected response to the stimuli across languages, whether the subjects' imitation had matched the target or not. The coding also had to reflect the differences across languages, that is, their language-specific characteristics, for example, information of the relativized position was needed in French but not in English or Tulu; specific morphemes appear in Tulu but not in the other two languages (see Blume et al. in this volume, for a detailed explanation). The data complexity here is not only morphological complexity; it is relational complexity—that is, relation of discrete parts of the child's form to other discrete parts, and relation of each to the parts of the stimulus form.

Language-Variety Differences

Research with multilingual populations frequently involves working with better-known Indo-European languages, as well as lesser-studied languages such as Haitian Creole, Yiddish, and Quechua. This type of research, just as do cross-linguistic studies, needs researchers to include in addition detailed and calibrated information on the language variety, so that cross-linguistic development can be compared.

Language Switching

Multilingual populations may also switch back and forth between languages in a single transcript or within utterances (i.e., code-switching/mixing data). For example, in an experimental study attempting to measure adult code-switching, Kang, Martohardjono, and Lust (forthcoming) asked English-Chinese multilinguals to switch back and forth between their two languages. Participants were given various topics to talk about for two minutes each and were instructed to switch from one language to another upon hearing a beep. These beeps occurred every 30 seconds. Markup was developed to identify the languages at multiple levels (e.g., lexical, morphological, syntactic), in order to examine the types of switches made (e.g., do participants switch faster when they switch functional items, such as discourse connectors or content words?). This requires any coding tool either to switch easily between the markups appropriate for each language or to allow for several coding fields in each screen.

Figure 10.1
Markup created in the Data Transcription and Analysis Tool (DTA).

Working with code-switching data may imply the need to code for elements linked to language processing. For example, this experimental study focused on both fluency (defined as the time taken to switch from one language into the other after the beep) and productivity (defined as the number of words produced within two minutes), besides the types of switches. Figure 10.1 shows some of this markup created in the Data Transcription and Analysis Tool (DTA).

Multimodal Data Markup

Another set of issues pertains to the modality in which languages are expressed as well as the status and information of the language(s) under investigation. While many studies have focused on the acquisition and use of two *spoken* languages, individuals who acquire more than one sign language and those who acquire *both* spoken and signed languages are also multilingual. The transcription and analysis of sign languages present specific challenges: They do not benefit from standard orthography, and no notation system for them is currently standard (Baker, van den Bogaerde, and Woll 2008).[10] The simultaneous use of different channels of speech production—the hands and the face—complicate the accurate representation of the different components of the utterance and may have modality-specific effects in the context of interactions (Morgan, Barrière, and Woll 2006). With respect to multilingual children's acquisition of both a spoken and a sign language, research shows that "Deaf children in such a multilingual situation often produce utter-

ances in which both the manual and vocal channels are used simultaneously" (Baker, van den Bogaerde, and Woll 2008, 20). The meanings expressed through each distinct channel may be separate or may combine, in which case transcribing the two independently from each other may not provide an accurate meaning of the full proposition (Baker, van den Bogaerde, and Woll 2008). Ultimately, data will need to be shared across researchers who work with both spoken and signed languages.

Experimental Data

As we saw in the case of the code-switching study, experimental data, for multilinguals as well as for monolinguals, require specific markup, depending on the method used. Given the current variation regarding both designs of experiments and coding systems by research teams, one needs to be able to calibrate results of different extensive markup systems indicating, for instance, the type of response (e.g., looking, pointing, moving props and toys, speaking), the timing of exposure to relevant stimuli (e.g., the point at which a child hears verbal stimuli when presented with visual stimuli in a picture- or video-matching task), and the data source (total looking time versus first long gaze in an Intermodal Preferential Looking Paradigm).

Linking Data to Metadata

As we hope to have shown in our discussion of the complexities of multilingual data, any study of language development or use must link data to rich metadata; for example, the code-switching study above looked at how attitudes toward code-switching and frequency of code-switching influenced its productivity and its fluency. Having each participant's metadata on hand in the same database is, therefore, critical for several reasons.

Design of Any Primary Research Tool Appropriate to Achieve the Vision of LLOD

It is obvious for the linguistic community working on multilingual acquisition and use that sharing data in an LLOD approach is essential to the progress of the field, since it enables us to replicate studies[11] and make full use of or reanalyze data that already exists. As we have shown, sharing Open Data would be most advantageous in terms of increasing sample sizes, allowing the identification of comparable populations, and allowing for meta-analyses.

Being able to share these data requires us to (1) standardize assessment tools as well as questionnaires, (2) capture metadata and data in efficient ways and in a design that is informed by past research, (3) link across projects and datasets, (4) allow for the capacity to query fields and relations among fields, and (5) at the same time allow for enough flexibility to capture the large diversity and richness of multilingual data.

Below, we discuss the capabilities of the Data Transcription and Analysis Tool (DTA)—but only briefly, since this tool is discussed in more detail in Blume et al. (this volume)—as an example of what is entailed in transforming any primary research tool to allow for the LLOD vision in multilingualism. The DTA tool is a primary research web application created mainly for the study of monolingual and multilingual language acquisition; it features a powerful relational database that handles both experimental and naturalistic data.

The DTA tool structures both the metadata documentation and the data creation process. It allows researchers to use built-in labels or to create project-specific labels (*codings*) to code their data, which in turn enables them to perform multiple types of analyses on their own data as well as to link data across projects.

A tool such as the DTA tool achieves requirements 2, 3, and 4 above, thus enabling researchers to share experimental (and natural speech) data so that people with varying types of expertise can reuse and repurpose them. Since the metadata and markup are so clear and specific, it becomes easy for new researchers to find all the details of a study in one place and then use that information to critique, reanalyze, and, if desired, repurpose the data.

However, the data creation process still requires many hours of dedicated and detailed work by individual researchers, since little is automated. With large sets of data, this process can take many years, so collaboration would be welcomed with other tools that have already achieved some level of automation or more efficient ways to speed up data creation (e.g., the CHILDES' CLAN[12] system or the LENA system[13]).

In terms of requirement number 4, although the tool is extremely flexible, dealing with the type of data we have described above entails some adjustments—some easier than others, but all possible. For example, capturing multimodal data would require us to display videos in the coding screen and not merely on the transcription screen. This is easily achieved and it would benefit all forms of language coding. Creating specific codes for sign language is now possible, but linking video and transcript/code is very time-consuming on the system currently available. Another challenge is that of language switch. At this point, there is no efficient way to tag the language of every word in an utterance. While this can be achieved by breaking the utterance word by word and tagging each word, this clearly could be better resolved by some automated process that may be available elsewhere.

To achieve Open Data, any tool needs to be able to speak to other tools and databases, and this bring us back to our first and major challenge. Having data that are really comparable across projects will never be achieved until we solve the standardization issues on metadata collection and presentation in requirement 1.

In sum, having an LLOD perspective and then acquiring and using any primary research tool that would aid researchers to achieve linking of their data in the study of multilingualism would require a cyberinfrastructure to support collaborative cross-linguistic research, calibration of complex multilingual markup systems, and the capacity to store, link, and search through extensive metadata.

Acknowledgments

Funding for Dr. Barrière was provided by NSF, USA/BCS#1251828 and 1251707 awarded to I. Barrière and G. Legendre; ESRC, UK; PSC-CUNY. Dr. Barrière would also like to acknowledge her collaborators: Katsiaryna Aharodnik (Graduate Center City University of New York, USA), Jennifer Culbertson (University of Edinburgh, Scotland), Guetjens (Prince) Fleurio (ENARTS, Port-au-Prince, Haiti), Nayeli Gonzalez-Gomez (Oxford University, Brooks, UK), Lisa Hsin (Harvard University, Boston, USA), Blandine Joseph (Long Island University, Brooklyn, USA), Sarah Kresh (Graduate Center City University of New York, USA), Géraldine Legendre (Johns Hopkins University, Baltimore, USA), Gary Morgan (City University, London, UK), Thierry Nazzi (University of Paris V & Centre National de la Recherche Scientifique, France), Bencie Woll (University College, London, UK), and Erin Zaroukian (US Army Research Laboratory, USA).

The authors would like to thank their colleagues at the VCLA for their input and discussion of these matters.

Notes

1. Multilingual speakers can be classified in many different ways. This classification intends to summarize and simplify on major life circumstances that may determine the speaker's level of competence and use.

2. For an example of an extensive questionnaire (78 questions, 42 pages long), see Blume and Lust's (2017). supplemental site: http://pubs.apa.org/books/supp/blume/?_ga=1.998898.2130472459.1479745044.

3. Multiple independently created questionnaires are available. One important task would be to compare them and decide which questions truly help researchers classify speakers so that a standard "minimal level" questionnaire can be created that also enables independent researchers to add questions as needed for their particular studies.

4. Pease-Álvarez, Hakuta, and Bayley (1996) also found discrepancies between children's linguistic abilities and their linguistic history.

5. https://www.iris-database.org/iris/app/home/index.

6. https://webdta.clal.cornell.edu/.

7. https://tla.mpi.nl/.

8. https://osf.io/.

9. http://ewave-atlas.org/.

10. ASL SignBank! is now being developed at the University of Connecticut by Diane Lillo-Martin and the members of the Sign Linguistics & Language Acquisition Lab.

11. This is being done for psychological studies in the Estimating the Reproducibility of Psychological Science project (https://osf.io/ezcuj/wiki/home/) and for second language acquisition by the Effects of Attention to Form on Second Language Comprehension: A Multi-Site Replication Study (https://osf.io/tvuer/), both hosted inside OSF.

12. http://dali.talkbank.org/clan/.

13. https://www.lena.org.

References

Alcock, K. J., K. Rimba, P. Holding, P. Kitsao-Wekulo, A. Abubakar, and C. R. J. C. Newton. 2015. "Developmental Inventories Using Illiterate Parents as Informants: Communicative Development Inventory (CDI) Adaptation for Two Kenyan Languages." *Journal of Child Language* 42 (4): 763–785.

Armon-Lotem, Sharon, Susan Joffe, Hadar Abutbul-Oz, Carmit Altman, and Joel Walters, J. 2014. "Language Exposure, Ethnolinguistic Identity and Attitudes in the Acquisition of Hebrew as a Second Language among Bilingual Preschool Children from Russian and English-Speaking Backgrounds." In *Input and Experience in Bilingual Development*, edited by Theres Grüter and Johanne Paradis, 77–98. Philadelphia: John Benjamins.

Austin, Jennifer B., María Blume, and Liliana Sánchez. 2015. *Bilingualism in the Spanish-Speaking World*. New York: Cambridge University Press.

Baker, Anne, Beppie van den Bogaerde, and Bencie Woll. 2008. "Methods and Procedures in Sign Language Acquisition Studies." In *Sign Language Acquisition,* edited by Anne Baker and Bencie Woll, 1–50. Philadelphia: John Benjamins.

Barac, Raluca, Ellen Bialystok, Dina C. Castro, and Marta Sanchez. 2014. "The Cognitive Development of Young Dual Language Learners: A Critical Review." *Early Childhood Research Quarterly* 29:699–714.

Barrière, Isabelle. 2010. "The Vitality of Yiddish among Hasidic Infants and Toddlers in a Low SES Preschool in Brooklyn." In *Yiddish—a Jewish National Language at 100*. Proceedings of Czernowitz Yiddish Language 2008 International Centenary Conference, edited by Wolf Moskovich, 170–196. Jerusalem-Kyiv, Hebrew University of Jerusalem.

Barrière, Isabelle. 2014. "Assessment of Language Abilities." In *Encyclopedia of Language Development*, edited by Patricia J. Brooks and Vera Kempe, 21–25. Sage.

Barrière, Isabelle, Sarah Kresh, Victoria Fay, Erika Lanham, Claribel Polanco, Stephanie Rauber, Jenice Robertson, et al. 2015. "The Contribution of Language-Specific Characteristics to Spanish-English Bilingual Preschoolers' Comprehension of Subject–Verb Agreement." Paper presented at the Tenth International Symposium on Bilingualism, Rutgers University, New Brunswick, NJ, May 20–24.

Barrière, Isabelle, Sarah Kresh, Katsiaryna Aharodnik, Géraldine Legendre, and Thierry Nazzi. 2016. "The Comprehension of 3rd Person Subject–Verb Agreement by Low SES NYC English-Speaking Preschoolers Acquiring Different Varieties of English: A Mutidimensional Approach." Paper presented at the Third Formal Ways of Analyzing Variation Workshop, Graduate Center, CUNY, New York, May 18–19.

Bialystok, Ellen. 2007. "Acquisition of Literacy in Bilingual Children: A Framework for Research." *Language Learning* 57 (Suppl. 1): 45–77.

Blume, María, and Barbara C. Lust. 2017. *Research Methods in Language Acquisition: Principles, Procedures, and Practices.* Washington, DC: American Psychological Association and De Gruyter Mouton.

Canale, Michael. 1983. "From Communicative Competence to Communicative Language Pedagogy." In *Language and Communication*, edited by Jack C. Richards and Richard W. Schmidt, 2–27. London: Longman.

Canale, Michael, and Merrill Swain. 1980. "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing." *Applied Linguistics* 1:1–47.

De Houwer, Annick. 2009. *Bilingual First Language Acquisition*. Bristol/Buffalo/Toronto: Multilingual Matters.

Dunn, Lloyd M., and Douglas M. Dunn. 2007. *Peabody Picture Vocabulary Test*, 4th ed. (PPVT-4). Pearson Education.

Espinosa, Linda M., and Eugene García. 2012. "Developmental Assessments of Young Dual Language Learners with a Focus on Kindergarten Entry Assessments: Implications for State Policies." Working paper # 1, Center for Early Care and Education Research-Dual Language Learners (CECER-DLL). Chapel Hill: University of North Carolina, Frank Porter Graham Child Development Institute, 1–16.

Esquinca, Alberto, David Yaden, and Robert Rueda. 2005. "Current Language Proficiency Tests and Their Implications for Preschool English Language Learners." In *Proceedings of the Fourth International Symposium on Bilingualism*, edited by James Cohen, Kara T. McAlister, Kellie Rolstad, and Jeff MacSwan, 674–680. Somerville, MA: Cascadilla Press.

Fishman, Joshua A. 1965. "Who Speaks What Language to Whom and When?" *La Linguistique* 2:67–68.

Flege, James Emile, Ian R. A. MacKay, and Thorsten Piske. 2002. "Assessing Bilingual Dominance." *Applied Psycholinguistics* 23:567–598

Flynn, Suzanne, and Barbara C. Lust. 1981. "Acquisition of Relative Clauses: Developmental Changes in Their Heads." In *Cornell Working Papers in Linguistics*, 2 (Spring), edited by Wayne Harbert and Julia Herschensohn, 33–45. Ithaca, NY: Department of Modern Languages and Linguistics, Cornell University.

Foley, Claire. 1996. "Knowledge of the Syntax of Operators in the Initial State: The Acquisition of Relative Clauses in French and English." PhD diss., Cornell University.

Gambino, Christine P., Yesenia D. Acosta, and Elizabeth M. Grieco. 2014. "English-Speaking Ability of the Foreign-Born Population in the United States: 2012." American Community Survey Reports. U.S. Census Bureau. Accessed November 10, 2017. https://www.census.gov/library/publications/2014/acs/acs-26.html.

García, Ofelia, Zeena Zakharia, and Bahra Otcu. 2013. *Bilingual Community Education for American Children: Beyond Heritage Languages in a Global City*. Bristol, UK: Multilingual Matters.

Gathercole, Virginia C. M. 2010. "Bilingual Children: Language and Assessment Issues for Educators." In *International Handbook of Psychology in Education,* edited by Karen Littleton, Claire Wood, and Judith Kleine Staarman, 713–748. Bingley, UK: Emerald Group.

Genesee, Fred. 1989. "Early Bilingual Development: One Language or Two." *Journal of Child Language* 16:161–179. Reproduced in *The Bilingualism Reader*, edited by Li Wei, 327–343. London: Routledge.

Genesee, Fred, Elena Nicoladis, and Johanne Paradis. 1995. "Language Differentiation in Early Bilingual Development." *Journal of Child Language* 22:611–631.

Grosjean, François. 1989. "Neurolinguists, Beware! The Bilingual Is Not Two Monolinguals in One Person." *Brain and Language* 36 (1): 3–15.

Grosjean, François. 2008, 2011. *Studying Bilinguals*. Oxford: Oxford University Press.

Grosjean, François. 2010. *Bilingual: Life and Reality*. Cambridge: Harvard University Press.

Gutierrez-Clellen, Vera, and Jacqueline Kreiter. 2003. "Understanding Child Bilingual Acquisition Using Parent and Teacher Reports." *Applied Psycholinguistics* 24:267–288.

Hamers, Josiane F., and Michel H. A. Blanc. 1989. *Bilinguality and Bilingualism*. Cambridge: Cambridge University Press.

Kang, Carissa, Gita Martohardjono, and Barbara C. Lust. (unpublished manuscript). "Underlying Cognitive Mechanism for Code-switching Differs across Bilinguals."

Lambert, Wallace E. 1955. "Measurement of the Linguistic Dominance in Bilinguals." *Journal of Abnormal and Social Psychology* 50:197–200.

Lust, Barbara C., Suzanne Flynn, María Blume, Seong Won Park, Carissa Kang, Sujin Yang, and Ah-Young Kim. 2014. "Assessing Child Bilingualism: Direct Assessment of Bilingual Syntax Amends Caretaker Report." *International Journal of Bilingualism* 20 (2): 153–172.

Mackey, William. 2012. "Bilingualism in North America." In *Handbook of Bilingualism and Multilingualism*, edited by Tej K. Bhatia and William C. Ritchie, 707–724. Oxford: Blackwell.

McCabe, Alyssa, Catherine S. Tamis-LeMonda, Mark H. Bornstein, Carolyn Brockmeyer Cates, Roberta Golinkoff, Alison Wishard Guerra, Kathy Hirsh-Pasek, et al. 2013. "Multilingual Children beyond Myths and towards Best Practices." *Social Policy Report* 27 (4): 1–37.

Miller, Karen, and Cristina Schmitt. 2010. "Effects of Variable Input in the Acquisition of Plural in Two Dialects of Spanish." *Lingua* 120 (5): 1178–1193.

Morgan, Gary, Isabelle Barrière, and Bencie Woll. 2006. "The Influence of Typology and Modality in the Acquisition of Verbal Agreement in British Sign Language." *First Language* 26 (1): 19–43.

Otheguy, Ricardo, and Ana Celia Zentella. 2012. *Spanish in New York: Language Contact, Dialect Leveling and Structural Continuity*. Oxford: Oxford University Press.

Paradis, Johanne, Kristyn Emmerzael, and Tamara Sorenson Duncan. 2010. "Assessment of English Language Learners: Using Parent Report on First Language Development." *Journal of Communication Disorders* 43:474–497.

Pease-Álvarez, Lucinda, Kenji Hakuta, and Robert Bayley. 1996. "Spanish Proficiency and Language Use in California's Mexicano Community." *Southwest Journal of Linguistics* 15:137–51.

Peña, Elisabeth D. 2007. "Lost in Translation: Methodological Considerations in Cross-Cultural Research." *Child Development* 78 (4): 1255–1264.

Sánchez, Liliana. 2015. "Crosslinguistic Influences, Functional Interference, Feature Reassembly and Functional Convergence in Quechua–Spanish Bilingualism." In *The Acquisition of Spanish as a Second Language: Data from Understudied Language Pairings*, edited by S. Perpiñan and T. Judy, 19–48. Amsterdam: John Benjamins.

Somashekar, Shamitha. 1999. "Developmental Trends in the Acquisition of Relative Clauses: Cross-Linguistic Experimental Study of Tulu." PhD diss., Cornell University.

Special Eurobarometer 386. 2012. "Europeans and Their Languages Report." European Commission, Brussels. Accessed November 10, 2017. ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_386_en.pdf.

Squires, Jane, Diane Bricker, and LaWanda Potter. 1997. "Revision of Parent-Completed Developmental Screening Tool: Ages and Stages Questionnaire." *Journal of Pediatric Psychology* 22:313–328.

Thomas, Margaret. 1994. "Assessment of L2 Proficiency in Second Language Acquisition Research." *Language Learning* 44 (2): 307–336.

Thordaardottir, Elin, and Susan Ellis Weismer. 1996. "Language Assessment via Parent Report: Development of Screening Instrument for Icelandic Children." *First Language* 16:265–285.

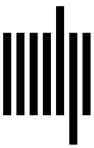
Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences

Edited by: Antonio Pareja-Lora, María Blume, Barbara C. Lust, Christian Chiarcos

Citation:

Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences
Edited by: Antonio Pareja-Lora, María Blume, Barbara C. Lust, Christian Chiarcos
DOI: 10.7551/mitpress/10990.001.0001
ISBN (electronic): 9780262357210
Publisher: The MIT Press
Published: 2020

The open access edition of this book was made possible by generous funding and support from Knowledge Unlatched



The MIT Press

The Open Access edition of this book was published with generous support from the National Science Foundation (grant number BCS-1463196), Pontificia Universidad Católica del Perú, and Knowledge Unlatched.

This book was set in Times New Roman by Westchester Publishing Services. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Pareja-Lora, Antonio, editor. | Blume, María, editor. | Lust, Barbara C., 1941– editor. | Chiarcos, Christian, editor.
Title: Development of linguistic linked open data resources for collaborative data-intensive research in the language sciences / edited by Antonio Pareja-Lora, María Blume, Barbara C. Lust, and Christian Chiarcos.
Description: Cambridge : MIT Press, 2019. | Includes bibliographical references and index.
Identifiers: LCCN 2019019588 | ISBN 9780262536257 (paperback)
Subjects: LCSH: Language and languages--Study and teaching. | Language and languages--Research. | Linked data.
Classification: LCC P53 .D398 2019 | DDC 025.06/4--dc23
LC record available at https://lccn.loc.gov/2019019588

10 9 8 7 6 5 4 3 2 1