Research Methods in Applied Linguistics 4 (2025) 100170

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Categorising speakers' language background: Theoretical assumptions and methodological challenges for learner corpus research

Olga Lopopolo [a,b,*], Arianna Bienati [a,c], Jennifer-Carmen Frey [a], Aivars Glaznieks [a], Stefania Spina [b]

[a] Institute for Applied Linguistics, Eurac Research Bolzano, Italy
[b] Dipartimento di Lingua, letteratura e arti italiane nel mondo, University for Foreigners of Perugia, Italy
[c] Dipartimento di Educazione e Scienze Umane, University of Modena and Reggio Emilia, Italy

ARTICLE INFO

Keywords:
Learner corpora
Metadata
Multilingualism
Dominant language constellation
Cluster analysis

ABSTRACT

In this article, we investigate how speakers can be categorised based on their language background in the field of Learner Corpus Research (LCR). Specifically, we discuss three key aspects: first, the theoretical assumptions and methodological choices made in learner corpus design, second the integration of a holistic perspective for speaker categorisation in LCR and third the consequences that different categorisations might have on study outcomes. Through a comprehensive review of corpora used in the field, we identify the most common terms, definitions and criteria of categorisation used to describe a speaker's language background. Focusing on the most central metadata encoding language backgrounds, the *L1* metadata, we inspect different operationalisations made and scrutinise the theoretical assumptions underlying them. Drawing on research on plurilingualism, we propose a holistic view of speaker's language background for Learner Corpus Research, combining various aspects of speaker's language use by methods inspired from the Dominant Language Constellation framework. We apply this methodology to re-evaluate the language categorisation system in LEONIDE, a multilingual corpus of Italian, German and English texts from secondary school students of diverse language backgrounds. We use the same corpus to evaluate the consequences of using different categorisations of the students on the outcome of possible linguistic studies. Despite a generally high overlap between study results across categorisations, we observe that variables combining multiple aspects of the speakers' language backgrounds seem to explain group differences for more of the linguistic features investigated.

Introduction

Any exploration of the social world is guided by interpretative frameworks that sort social reality into classes or categories, which function as a powerful cognitive tool to extend previous knowledge and make inferences (Medin & Coley, 1998). According to Hackert

Declarations of interest: none

[*] Corresponding author at: Olga Lopopolo, Viale Druso 1, 39100 Bolzano, Italy.
E-mail address: olga.lopopolo@eurac.edu (O. Lopopolo).

https://doi.org/10.1016/j.rmal.2024.100170
Received 15 February 2024; Received in revised form 14 November 2024; Accepted 17 November 2024
Available online 25 November 2024

(2012: 35), it is through discourse, in its Foucauldian sense, that these categories are constructed within scientific disciplines to meaningfully represent specific concepts. Through discourse, the object of study is represented as if it was 'really there', i.e., it can be named, defined and set off from other phenomena, thus contributing to shape a specific social reality (Hackert, 2012: 36). The way in which research discursively constructs an object of study produces what has been defined a system of statements, which emerges from the contingent socio-historical conditions and is continuously negotiated as social realities change and new conceptualisations gain momentum.

A concrete example of this discursive practice is the categorisation of speakers according to their language backgrounds. Although such classification often passes unnoticed and is rarely questioned, it involves a complex intersection of factors — language, culture and identity — each of which is itself shaped by disciplinary discourses within Applied Linguistics.

In our study, we investigate the categorisation of speakers' language backgrounds in Learner Corpus Research (henceforth, LCR), considering the conceptual foundations and terminology adopted in the field. We begin with an overview of the different conceptualisations and related terminology about a speaker's language background in Applied Linguistics, before we continue with reviewing terminology, definitions and criteria for categorisation used in corpus descriptions in LCR. Afterwards, we propose an alternative to current prevalent ways of categorising speakers, which is grounded in holistic accounts of speakers' experiences with languages. This alternative approach is tested together with more traditional categorisation systems in a study that explores the potential consequences of different categorisations on study results. We conclude by considering the importance of standardised metadata and comprehensive documentation to enhance transparency and comparability across corpus designs, as well as by discussing the possible use cases of a methodology that holistically integrates language background variables to capture the linguistic practices of speakers in diverse societies.

Unveiling the discourse behind the categorisation of speakers' language background

In research on language acquisition (e.g. Berthele & Udry, 2021), various aspects of the acquisitional history of individuals (amount of exposure and use, context and manner of acquisition, affective factors, etc.) and their interplay have been found to contribute to differences in how the speaker processes, acquires, masters and experiences languages. The way in which research has dealt with what we refer to the *categorisation of language background*, i.e. how individuals' histories with languages in terms of use, proficiency, exposure and identification has been categorised, is based on various aspects such as time of acquisition, level of proficiency, amount and domains of use and self-identification. The perspective taken in Second Language Acquisition (SLA) is usually to look at a speaker's language background in terms of time-relation (Hammarberg, 2014). Two major time-related criteria are important for the field of SLA: the first is the chronological order of acquisition that identifies one *L1* or *primary language* as the language that is encountered first and thus develops as the original system. The second is based on cognitive maturity and considers as *native languages* (NL) all languages which a speaker has been exposed to during a specific stage of maturational development (the critical period hypothesis, Penfield & Roberts, 1959; Lenneberg, 1967). The speaker is called a *native speaker* (NS) of those specific varieties. The age span for which one can speak of a native language-type of acquisition, however, is still debated (cf. Paradis, 2004 and Meisel, 2011 for two different positions on this matter).

The two categorisations not only bring forward different descriptions of the same linguistic system, but also various terms (e.g. *L1, primary language, native language* or *mother tongue*) that are used to describe similar and yet different concepts. This can be seen with the terms *native speaker* and *mother tongue*, which have been criticised and problematised as ambiguous and misleading by many scholars from historical, political, linguistic, economic, and educational standpoints (Paikeday, 1985; Davies, 2013; Holliday, 2006; Hackert, 2012; Copland et al., 2016, Cheng et al., 2021). For categorisations based on the proficiency of a speaker, the native speaker is conceptualised as the "ideal speaker-listener" (Chomsky, 1965). From this perspective, native speakers, having achieved full mastery of a particular language, are therefore able to provide authoritative judgments about grammaticality and are used as a benchmark for language learners or those who are considered otherwise "non-native speakers" (NNS) (Abrahamsson & Hyltenstam, 2009). However, the traditional dichotomy of native vs. non-native speaker has faced criticism from numerous scholars in SLA. The critique mainly stems from the "comparative fallacy" (Bley-Vroman, 1983), which involves equating "identity with idealised native speaker production as a definition of success", consequently making it "difficult to avoid seeing the learner's interlanguage as anything but deficient" (Larsen-Freeman, 2014: 217) and propounding the idea of a superiority of one group over the other (Dewaele, 2018: 239). Therefore, alternative terms, such as *native speaker* vs. *L2 user* (Cook, 2002) or *L1 user* vs. *LX user* (Dewaele, 2018) have been proposed. A similar criticism concerns categorisations based on the aspect of identity. A self-declared "native language" can, in fact, stand as a symbol of group affiliation and identity (Gal & Irvine, 1995). This, however, mixes linguistic categories with ideologies on heritage and belonging that could be problematic for certain types of studies that do not focus on this aspect. The same is the case for the term *mother tongue* that stands somewhere in between categorisations based on age and those based on domains of use. While in some cases the term relates to the linguistic environment of initial exposure (typically the family), in other cases a *mother tongue* emphasises a biological component that justifies the affiliation to a certain ethnic group. To this respect, Bonfiglio (2010; 2013) has traced the genesis of the metaphors of maternality used in the discourse about language, discussing how they represent a manifestation of specific ideologies such as nationalism. Nationalism, in fact, is tightly bound to the early modern period where national characteristics like language were connected to geography to construct "myths of congenital community" (Bonfiglio 2013: 54).

Finally, when employing any of the concepts and related terminology discussed above – such as using the label *L1* in an age-related categorisation to indicate the language of first exposure (chronologically) as compared to *L2, L3* and so on – one also refers to specific models of languages that are employed for their potential descriptive or explanatory power. The practice of counting languages as discrete entities or "linguistic solitudes" (Cummins, 2008) is a common approach also used in studies on bilingualism, where the concept of parallel monolingualism, i.e. considering bilinguals' languages as independent systems, is one of the most accepted and reproduced according to Melo-Pfeifer (2015). In contrast, there are also approaches that question the usefulness of identifying and counting languages, advocating a perspective on languages as integrated systems in speakers' repertoires (Blommaert & Backus, 2013). In the wake of the so-called multilingual turn (May, 2014), new approaches have been developed (cf. Hufeisen, 2018), including the Factor Model (Hufeisen, 2010), the Dynamic Systems Theory model of multilingualism (Herdina & Jessner, 2002), the Plurilingual Didactic Monitor Model (Meissner, 2004) and the Dominant Language Constellation (DLC, Aronin, 2019; Aronin & Moccozet, 2021). These models of multilingualism acknowledge that the linguistic repertoire of a multilingual individual results in an integrated set of resources that are in constant mutual interaction and development. They draw on ideas such as multicompetence (Cook, 1992), translanguaging (Garcia & Li Wei, 2014) or languages seen as part of a dynamic and complex system (cf. Larsen--Freeman, 2002). In all these cases, often sharing the assumption that multilingualism is a normal human condition (Puig-Mayenco et al., 2018: 2), speakers are seen holistically rather than through categorical labels of language.

In the next section, we will investigate how the various theoretical approaches to a speaker's language background are visible in corpus designs, by reviewing the terms, definitions and criteria used to categorise study participants in LCR.

Terms, definitions and criteria for categorisation in LCR

The information on a speaker's language background is of utmost importance in the field of LCR. It is used to describe both those who are considered learners as well as those who learners are compared to and thus helps to understand the studied population and classify subsets of the data for comparison. This section provides a review on how information on a speaker's language background is collected and organised in corpora used in LCR. For the review, we define frequently used corpora as all corpora that are listed in the widely cited and used *UCLouvain list of learner corpora around the world* (Centre for English Corpus Linguistics, 2019) and featured in at least two different studies presented at the biannual LCR conferences over the past twelve years (2011–2022) .[1]

Using an R script that searched for all occurrences of either the full name or the corpus acronym in the Book of Abstracts of the past six editions of the conference, a list of 48 corpora corresponding to these criteria was retrieved.[2] We successively reviewed corpus description papers, handbooks, corpus websites and any other type of documentation retrievable. For three corpora, no corpus description or any other kind of documentation was available and these were excluded from the review. For corpora for which various forms of documentation were available (corpus description paper, handbooks, websites, etc.) we cross-checked the information available, by also looking at corpus interfaces, where freely accessible. For most of the corpora, we found the related documentation in English. However, since one of the foci of our review is a critical discussion of terminology, the 7 corpus descriptions that are not available in English were excluded from the review to avoid methodological problems arising from translation.

From the remaining 38 corpora in our sample, we can draw the following considerations. Information on a speaker's language background is either set a priori by sampling decisions or solicited from study participants via questionnaires and stored as metadata. Reflecting the interest in different factors that might influence a person's acquisition and mastery of a language, some of the reviewed corpora offer a variety of information about the linguistic profiles of the participants. Among those, we found information about the languages spoken by the parents, home and school language(s), as well as other known languages. However, the language background of a speaker was not always so richly encoded. Usually, only few metadata variables that refer to the speaker's language background were recorded, often focusing on the aspects that reflect the research aims for which the corpus was collected. Almost all corpora reviewed (33/38) provided a metadata variable that is supposed to be the most central variable encoding language background and is usually referred to as L1 (cf. Granger, 2002). This variable has also been proposed as one of the "core metadata" for LCR (cf. Paquot et al., 2023) that should be documented for every corpus in order to establish best practices in the field.

The significance of this information as a core metadata variable arises from two assumptions. The first of these is that the L1 of a speaker influences their process of acquisition of another language (e.g., due to transfer, Osborne, 2015). Methodologically, it is common in LCR to compare groups of 'learners' with different L1s to test whether a certain linguistic feature is specific to a certain group of speakers because of their L1 (cf. Gilquin, 2001; Jarvis, 2010). The second assumption postulates that linguistic systems acquired later are acquired differently than the systems acquired first. This assumption is at the foundation of studies that focus on comparing 'learners' with a reference group of 'native speakers' (Callies, 2015: 38). When the corpus compilation project envisages a reference corpus to be provided with the learner corpus, the L1 metadata plays a crucial role in distinguishing those participants who will be part of the reference group from those who will be part of the learner group (e.g., FALKO, Lüdeling et al., 2008; LAS2, Ivaska, 2014).

[1] The use of these criteria requires two disclaimers. First, the Louvain list includes corpora collected from student populations in English-dominant countries, such as BAWE (Heuboeck et al., 2010) and MICUSP (Ädel & Römer, 2012). These corpora are primarily used as reference data, representing the language use of proficient speakers in the target language, regardless of their L1. Naturally, they also include data from heritage speakers or international students whose declared L1 differs from English. While these participants are not typically considered 'learners' in the conventional sense, they may still be classified as such based solely on their recorded L1. Given our neutral stance on the criteria that define a learner corpus, we included these corpora in our review simply because they are part of the Louvain list. Second, since the LCR conferences are predominantly hosted by European institutions and attended by European scholars, our inclusion criteria may introduce a bias towards corpora compiled in Western(ised) countries, primarily representing WEIRD (Western, Educated, Industrialised, Rich, and Democratic) populations.

[2] All supplementary materials (including the scripts and spreadsheets used for the literature review and the case study) are accessible at https://gitlab.inf.unibz.it/commul/lca/rmal2024_speaker-categorization.

For this central metadata, we systematically collected the following information:

- original description of the metadata from the main reference source(s);
- terminology adopted in the corpus documentation to refer to the L1 metadata and other related terminology;
- definition (explicitly stated or inferred) of the L1 metadata;
- criterion (explicitly stated or inferred) according to which study participants are categorised.

Five of the 38 reviewed corpora do not provide the L1 as metadata in their corpus interfaces, for different reasons. In LOCCLI (Spina & Siyanova-Chanturia, 2018), NICT JLE (Izumi et al., 2005) and SPLLOC (Dominguez, 2010) the L1 is not provided as metadata since it is set a priori by sampling decisions. In CELI, in contrast, it was not possible to derive it from the base data (language certifications), "as candidates [were] asked to report only their nationality" (Spina et al., 2023: 463). MICUSP (Ädel & Römer, 2012) is an interesting case, where the L1 might have been collected, yet it was decided to just report on the "native speaker status" of the speakers. In fact, in the corpus interface[3] there is just one variable related to the speakers' language background, namely *nativeness*. Even though the L1 variable is not provided as metadata, these corpora still use criteria to categorise (or select) the participants and terminology to describe their language backgrounds.

Terminology used to refer to the L1 metadata

The terminology specifically used to describe the L1 metadata in the corpora reviewed can be traced back to five main terms: *L1, native language, mother tongue, first language* and *dominant language*. The terms, together with the number of corpora in which they are employed, are reported in Table 1.

Of the five main terms used to refer to the L1 metadata, some of these display a range of variants either combining two terms, as in *mother tongue (L1)*, adding specifications, as in *mother tongue background* or signalling plurals, such as *first languages* or *native languages*. The plural variants are particularly interesting with a view on multilingualism. Variants such as *native language(s)* in the SweLL corpus (Volodina et al., 2016) explicitly acknowledge the existence of multilingual participants, in that they record "both unique L1 and combinations" (Volodina, 2021). Singular variants, in contrast, indicate that the existence of multiple L1s per participant was either not foreseen or implicitly integrated without making it visible. Interestingly enough, some corpora that use the singular variant for the L1 metadata combine this terminology with plural forms for additional metadata on e.g. other languages known. In LONGDALE (Meunier, 2016), for instance, the singular term *L1* is paired with the plural *L2s*. In some other corpora, e.g., VESPA (Paquot et al., 2022) and InterFra,[4] the terms employed to refer to the participants' L1 metadata are in singular form, yet they encode both single languages as well as combinations of two or more languages. This information is not explicitly stated in corpus manuals or websites but was retrieved by consulting the questionnaire used for metadata elicitation in the case of VESPA, and by exploring the possible values of the L1 variable in the case of InterFra.

While the terms used to describe the L1 metadata usually refer to different languages in the participants' language backgrounds, the reviewed corpus descriptions also often explicate how this variable has been employed to categorise individuals into groups, opposing terms such as *native speaker* vs. *non-native speaker* (vs. *near-native speaker*) or *native speaker* vs. *learner*.

Definitions given for the terms employed

Of the 33 corpora providing the L1 as metadata, only five give a definition of the term L1. The corpora, original descriptions, definitions and terminologies are cross tabulated in Table 2.

The small number of corpora that provide definitions for the L1 variable does not allow us to draw any generalising claim about term-definition pairs. However, it is possible to comment on the type of definitions that were proposed and the problems of univocally defining what this metadata encodes. Starting from the self-declaration of the L1 (BAWE, KOLIPSI, TRAWL), it is evident that this operationalisation leaves considerable ambiguity as to which exact phrasing led to the responses collected by the questionnaire and how this phrasing was potentially influencing the interpretation of the respondents. With regards to the two descriptions that convey a definition summarised under the concept 'home language' (LEONIDE, VESPA), it is important to highlight how not only the context (the family setting) plays a role in defining the variable, but also the time-relation, implied by the verbs *raised* and *brought up*, is present. This definition thus seems to stand in between categorisations based on age and those based on domains of use.

Apart from these few cases, the prevalent tendency (n=28) in corpus descriptions reviewed is to leave terms employed undefined. In those cases, one can infer from vague statements that person-related metadata have been collected, but neither a definition of the terminology used nor an explanation of the operationalisation of the variable is given. For instance, in the ICCI corpus description paper it is stated that "each learner was asked to provide at least the following information: […] iii. Mother tongue; iv. Gender; v. Grade (age)" (Tono & Díez-Bedmar, 2014: 170), but no other information is available regarding what was meant by *mother tongue*. This suggests that this term is treated as a straightforward, objective variable (such as chronological age or school year) for which no definition is needed.

[3] https://elicorpora.info/main, last accessed May 16, 2024.

[4] Information retrieved from https://www.su.se/romklass/interfra/brief-description-of-corpora, (G) Table secondary school students, last accessed May 5, 2024.

Table 1
Terminology used to refer to the L1 metadata in the 38 corpora reviewed.

Name of the L1 metadata	N. of corpora
L1	16
Native language	7
Mother tongue	5
First language	4
Dominant language	1
(L1 not provided as metadata)	5
Total	38

Table 2
Original descriptions of the L1 metadata and the explicitly stated or inferred definitions per corpus.

Corpus	Original description of the variable	Definition	Terminology
BAWE	"first language as indicated by student" (Heuboeck et al., 2010:14)	Self-perception	First language
KOLIPSI	"Metadata regarding the learner's L1 (author_L1) is based on what authors perceive as their first language(s)." (Glaznieks et al., 2023)	Self-perception	L1
LEONIDE	"Forty pupils were raised with a language other than German or Italian (17 of which attended German schools, 23 Italian schools) as L1. In addition, 36 pupils came from a multilingual household in which at least one of the three target languages was spoken (German, Italian or English)." (Glaznieks et al., 2022:104)	Home language	L1
TRAWL	"All the students who agreed to contribute to the corpus were asked to fill in a questionnaire on language knowledge and use. Around 15 % of the participants listed other L1s than Norwegian. Searches in all sub-corpora can be filtered by the L1s listed by students." (Dirdal et al., 2022:199)	Self-perception	L1
VESPA	"This is the language in which you were brought up. If you were raised in two languages at home, please enter them both, separated by a semi-colon (e.g. English; French)."	Home language	Native language

It is to be noted that, in another point of the paper, the L1 seems to be defined as self-reported languages: "The term L1 is reserved for the pupils' first language(s), as indicated by themselves" (Glaznieks et al., 2022: 98). Yet, having the chance to further inspect the unpublished questionnaires, we discarded this possibility, as there is no question that directly asks the students about their L1(s).
The passage reported is preceded by this sentence: "Our use of the terms L1, L2 and L3 above to refer to Norwegian, English and French/German/Spanish, respectively, is based on the order and level at which these languages are taught in Norwegian schools" (Dirdal et al., 2022: 199). However, inspecting the corpus interface we discarded the possibility that L1 would be defined as the language first taught (by order and level) in a certain school system, because the 'Student's L1' variable in TRAWL allows for values different than the languages taught in the Norwegian school system.
Quote from https://cdn.uclouvain.be/public/Exports/%20reddot/adri/documents/VESPA_learner_profile.pdf, last accessed May 6, 2024.

Criteria for categorisation of speakers by their background

In addition to the five types of terms and their variants which have been used to label the L1 variable, additional terminology was used in corpus descriptions to name speaker categories based on the L1 metadata. Fig. 1 illustrates the intricate relationship between terms, definitions and speakers' categorisations found in the corpora reviewed. The graph was obtained by first creating a term-definition matrix that contains, for each term, how many times it connects with a certain definition. Then the *igraph* library (Csardi & Nepusz, 2005) was used to render the image.

The terms (black labels) and speaker categories (grey labels) are organised and linked to three different nodes (in blue), which serve as definitions (or lack thereof) of the terms. The bigger the node size, the more frequent the terms and the larger the number of corpora that uses that definition for the L1 metadata. The frequency of occurrence of form-meaning pairs is represented by the thickness of the lines connecting each term to its definition. Solid lines connect the terms used to name the L1 variable to their definitions, whereas the dashed lines identify the additional terminology (synonyms and speaker categories) found in the corpus descriptions.

Very central is the distinction between *native* and *non-native speaker*, as well as the term *learner*. In most cases these categories of *learners* and *natives* are based entirely on the L1 metadata and its definition, which might differ between corpora. In two cases, however, proficiency functions as an additional criterion to categorise speakers. In the MICASE corpus, for example, *near-native speakers* are differentiated from *non-native speakers* by their "native-like fluency and grammatical proficiency" (Simpson et al., 1999: 8). Similarly, in the UEA corpus from the FLLOC project, English L1 recordings used as a reference for French Erasmus students' spoken productions include recordings from two non-native students "with native-like proficiency in English" .[5]

Very concrete challenges might arise with corpus designs that allow for multiple languages in the L1 metadata. Multilinguals might be filtered out from the sample as they make "the control of the L1 variable difficult" (cf. CAES, Rojo & Palacios, 2016:14) or because they blur the borders between what are usually considered learners and reference data (cf. SPLLOC, Dominguez, 2010), having extensive contact to the target language. While this approach simplifies the categorisation process, it significantly reduces the sample sizes for areas with high linguistic diversity and perpetuates a *monolingual bias*, i.e., the notion that "the prototypical human is seen as

[5] Quote from http://www.flloc.soton.ac.uk/uea/learners.html, last accessed February 10, 2024.

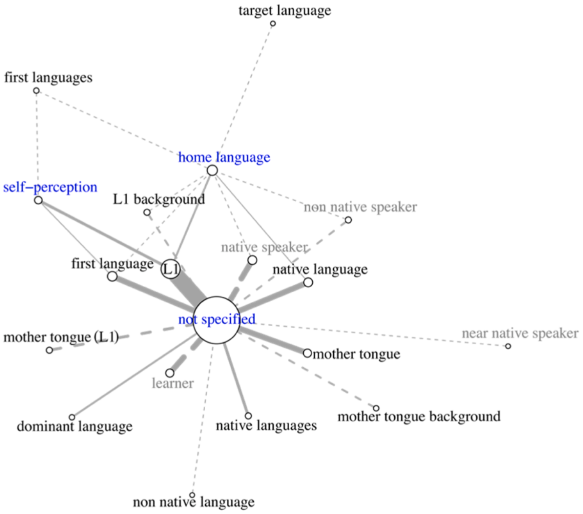


Fig. 1. Bipartite Network Graph Showing the Relationships Between Terms, Definitions and Speakers' Categorisations in the Reviewed Corpora.

having only one language" (Barrat, 2018: 1).

A case study on speaker categorisation

As shown in our review of corpora, in the field of LCR it is common practice to operationalise language background in discrete categories, documented in person-related metadata. This methodological choice implicitly aligns with assumptions inherent in formal models of language acquisition, which assume the identifiability and boundedness of languages and aspects of language experience within an individual's language repertoire (cf. Berthele, 2021). In our review we focused specifically on one segment of a speaker's language background, the L1 metadata, which consistently emerges as core metadata across nearly all reviewed corpora. We found that multiple terms are employed for the underlying concept encoded in the metadata, often without clear definitions. Even when definitions are provided, there is no consensus on a single conceptualisation of the metadata. As a consequence, the criteria for creating groups of speakers with different backgrounds vary, sometimes incorporating additional criteria. Different categorisations of the same language background might however impact the results obtained when analysing differences between groups, and thus make direct comparisons between studies impossible and the combination of data from various corpora methodologically spurious.

It is equally relevant to observe that none of the reviewed corpora seems to align with holistic models of language acquisition. Holistic models such as Complex Dynamic Systems (Larsen-Freeman, 2002) and Dominant Language Constellations (Aronin, 2019; Aronin & Moccozet, 2021) highlight the dynamic interaction not only among the language subsystems as integrated (instead of separated) ones, but also as socially embedded systems that make sense only in their concrete practices. This aspect is clearly connected to the methodological dilemma posed by multilingual speakers that we found for some of the corpora reviewed. As discussed, this aspect often leads corpus compilers a) to assume that the L1 is a single, homogeneous language associated with each speaker, resulting in few corpus projects that allow for multiple L1s; and b) to exclude or simplify multilingual speakers because their language backgrounds are challenging to capture with a single metadata variable. As much as multilingual speakers may have seemed exceptional in past learner corpus compilation projects, shifting the perspective towards viewing language as a practice rather than a set of different segments helps us craft new methodologies that, while accommodating the majority of cases, might also include those typically excluded. In fact, viewing languages as activities embedded in social and cultural contexts prompts the need for the field to consider holistic models of language acquisition for the operationalisation of metadata about speakers' language backgrounds.

We address these problems in a case study by a) proposing a viable methodology to integrate a holistic model of language acquisition into the field of LCR, namely Dominant Language Constellation (Aronin, 2019; Aronin & Moccozet, 2021), and b) testing whether choosing between formal- vs holistic-derived categorisation to determine speakers' language backgrounds might impact the study results. Our aim is thus to address the following research questions:

1) How can we integrate a holistic perspective on speakers' language backgrounds into the field of LCR?
2) What are the consequences that different categorisations of the speakers' language backgrounds can have on the results of typical linguistic studies in LCR?

Data

To answer both research questions, we selected LEONIDE (Glaznieks et al., 2022), a longitudinal trilingual learner corpus for the languages Italian, German and English. It consists of more than 2500 texts written by 163 students from German and Italian schools[6] in the northern Italian Autonomous Province of Bolzano – Alto Adige (South Tyrol). The data were collected within the *One School, Many Languages* project (Stopfner & Engel, 2019) over the period of three years, i.e., the full cycle of lower secondary schools in Italy (grades 6–8; age of students 11–14 years). Assessment tasks were repeated in each project year with a different set of prompts to elicit narratives and argumentative essays in each language.[7] The collection of the corpus data was accompanied by a detailed questionnaire, aiming to document the full picture of students' language background. While the published corpus only provides some selected information on the language background, we were able to use the original questionnaire data for the purpose of this study. The relevant questionnaire items inquired which languages the students use with whom (e.g., family members and friends) and how often on a 4-point Likert scale (1 never, 2 sometimes, 3 often, 4 always).

Based on the information about the students' language use with their parents, a person-related metadata variable named *author_L1* was created, referring to the languages that students declared to *always* use with their parents. The pre-specified languages were integrated into the values of the *author_L1* metadata, i.e., *IT* (for Italian), *DE* (for German), *EN* (for English), while all other languages that appeared as the students' family languages were gathered in one value, i.e. *OTHER*. As students could come from plurilingual family backgrounds, combinations of different languages are possible in the *author_L1* metadata. In addition, information about the school main language of instruction (*school_language*) is provided in the corpus. Table 3 shows the combination of the *school_language* and *author_L1* metadata, the two person-related language background metadata variables encoded in LEONIDE for all students.

The first observation emerging from Table 3 is that the largest groups of students are those whose language of instruction (either German or Italian) aligns with the language used at home: more specifically, 40 students attending schools having German as main language of instruction report exclusively using German with their parents, and the same for 43 students attending Italian-speaking schools who use only Italian with their families. There are, however, several students whose home language differs from their language of instruction, meaning that the language environment in which they are immersed at school does not reflect the same language environment of the family. For example, when looking at students attending German schools, we can observe that 17 students use a different language (classified as "OTHER") and 8 students use combinations of languages in which German is not present (IT-OTHER n=1; IT-EN n=1; EN-OTHER n=1; IT n=3; EN-OTHER n=1; IT-EN-OTHER-OTHER n=1). The same scenario occurs with students attending Italian schools, in which 24 of them do not make use of Italian at home but use another language or a combination of other languages in which Italian is not present (OTHER n=22; DE n=1; OTHER-OTHER n=1). In addition, Table 3 shows that some students make use of multiple languages at home beside their language of instruction, thus showing a high degree of multilingualism in the family context (for example, for students attending Italian schools, combinations comprising Italian are IT-OTHER n=8; DE-IT n=3; DE-IT-OTHER n=1; IT-EN n=1; IT-EN-OTHER n=1). This shows the fact that, depending on individual constellations, students not only make use of multiple languages at home or combinations of languages that might or might not comprise the language in which they are mainly instructed at school, but that their use of languages highly depends on the contexts of use, which do not necessarily overlap between school and family practices.

In what follows, we first propose a method to define language background holistically and categorise speaker's language background based on various aspects of language use reported by the study participants in the questionnaire data, as compared to other prevalent definitions that are based on single aspects of language use, such as language of instruction, family language or order of onset. Then we perform an empirical study, testing the effect of different categorisations based on either the language background-related metadata available in LEONIDE (language of instruction and language(s) used at home) or the aforementioned holistic view on language background on the outcomes of a series of linguistic studies.

A holistic perspective on speakers' language background

As we have seen, the metadata variable *author_L1* in the current version of LEONIDE provides only a partial view of participants' repertoires, in view of the considerable variation recorded by the questionnaire in the possible combinations of languages spoken in the family context. Yet context is a crucial element if we consider the fact that speakers make use of their languages not only in the family context but also in social interaction.

As already mentioned, it is possible to view speakers holistically rather than through categorical labels of language. Among the different holistic models of multilingualism, Dominant Language Constellation (DLC) (Aronin, 2019; Aronin & Moccozet, 2021) is the framework chosen in the present investigation to portray students' most expedient language uses in the configuration of DLCs. The DLC model operationalises speakers' language backgrounds "considering the whole set of languages as units, rather than focusing, one by one, on the specific languages used by given individuals or groups" (Aronin & Singleton, 2012: 69). A DLC is a group of an individual's most important languages, functioning as an entire unit that enables them to meet all needs in a multilingual environment (Aronin, 2006). The approach thus shifts the focus from the investigation of separate languages to the exploration of their constellations (Aronin & Jessner, 2014: 64). The integration of different contexts of practice in the configuration of the DLCs has been employed to the repertoires of multilingual school children in Norway in Storto et al. (2023) using visual models. The study effectively demonstrates the utility of such models in illustrating the linguistic repertoires across different contexts, including family, school, holidays, friends,

[6] The local schooling system offers schools with Italian as the language of instruction (and German taught as L2), schools with German as the language of instruction (and Italian taught as L2), and, in the two Ladin valleys, schools with Italian and German as evenly distributed languages of instruction (and Ladin as subject). The data in LEONIDE were collected only in Italian and German schools.

[7] The full range of prompts can be accessed on the Eurac Research Learner Corpus Portal PORTA: https://www.porta.eurac.edu/lci/leonide/.

Table 3
Grouping of LEONIDE students by language(s) spoken with their parents (author_L1) and school main language of instruction (school_language).

School main language of instruction (*school_language*)	**Language(s) spoken with parents** (*author_L1*)	n	%
German	German (DE)	40	25 %
	Other (OTHER)	17	10 %
	German-Italian (DE-IT)	14	9 %
	Italian (IT)	3	2 %
	German-Other (DE-OTHER)	3	2 %
	German-Italian-Other (DE-IT-OTHER)	1	1 %
	Italian-Other (IT-OTHER)	1	1 %
	Italian-English (IT-EN)	1	1 %
	Italian-English-Other-Other (IT-EN-OTHER-OTHER)	1	1 %
	English-Other (EN-OTHER)	1	1 %
Italian	Italian (IT)	43	26 %
	Other (OTHER)	22	13 %
	Italian-Other (IT-OTHER)	8	5 %
	German-Italian (DE-IT)	3	2 %
	German-Italian-Other (DE-IT-OTHER)	1	1 %
	German (DE)	1	1 %
	Other-Other (OTHER-OTHER)	1	1 %
	Italian-English (IT-EN)	1	1 %
	Italian-English-Other (IT-EN-OTHER)	1	1 %
Total		163	100 %

and media. For the South Tyrolean context, Colombo & Stopfner (2018) discuss various prototypical language constellations identified in the *One School, Many Languages* project from which the data in LEONIDE stems. In a case-study-based analysis, they demonstrated that multiple constellations are active in the South Tyrolean environment when considering different contexts of use, such as family and school. Three main profiles based on the language practices of students emerge: a) a predominantly German or Italian-speaking monolingual setting, b) a balanced plurilingual setting in which both languages are spoken with the parents with approximately equal frequency and c) an unbalanced plurilingual setting, in which students make use of several languages with different tendencies towards a certain set of languages and frequency of use.

Recategorising speakers' language backgrounds

We quantitatively replicate the concept of dominant language constellations based on the students' language use across various contexts of practice, as reflected in their questionnaire responses. Specifically, we employ the TwoStep Cluster Analysis technique (Norusis, 2011) for its suitability in handling complex datasets and multiple variables. The aim is to identify similar groups, or clusters, of individuals based on their answers to specific statements. This analysis mirrors the idea of DLC by providing a clustered representation of prototypical use of languages in different contexts of LEONIDE participants.

We modelled the students' DLCs based on three contexts of language use declared in the questionnaire: the family, friends and school contexts. The questionnaire statements considered as variables are listed in Table 4. The answers to these statements indicate the frequency of use of three pre-specified languages (Italian, German and English) with the possibility of adding other languages, on a 4-point Likert scale ranging from 1 (*never*) to 4 (*always*).

For the family context, we not only include information about the language(s) spoken with parents (a, b), as was done to operationalise the L1 metadata in LEONIDE, but also consider responses about language use with siblings (c), taking into consideration that students may use different language(s) with their parents and their siblings. Regarding the school context (e, f, g), students indicated whether they attended kindergarten, primary and lower-secondary school in German, Italian or another language of instruction. For both the family and school context, we aggregate the respective questionnaire items calculating the median of the responses.[8] For the friends context, we focus on the single item (d), recording the frequencies of use of languages spoken with friends.

The TwoStep Cluster Analysis was conducted using the Statistical Packages for the Social Sciences (SPSS). This method offers various advantages over other clustering techniques: (1) While methods such as k-means clustering require numerical variables, the TwoStep Cluster Analysis algorithm standardises all variables, making it suitable for categorical variables as used in this study (Norusis, 2011). (2) Using TwoStep Cluster Analysis, the available sample size of 163 students allows us to retrieve reliable and valid clusters based on several key categorisation variables which are considered independent. (3) The algorithm does not require a predefined number of clusters, which would have been difficult to set a priori given the high variability of students' use of multiple languages mentioned in the questionnaires but determines the number of clusters automatically. (4) With TwoStep Cluster Analysis, it

[8] The median was chosen as a measure of central tendency for both the family and school context because it helps to summarise the prevailing language category across different school cycles, allowing for a meaningful aggregation of the data. For the school context, for example, where the data were categorical (i.e., Italian, German, or other language of instruction for each school cycle), the median was computed by selecting the language of instruction (Italian, German or other) across the three school cycles (kindergarten, primary and lower-secondary school) (questionnaire items e, f, g) for each student.

Table 4
Contexts of language practice and questionnaire items considered.

Context of language practice	Questionnaire statements
Family	a. My mother speaks with me …
	a. My father speaks with me …
	a. My siblings speak with me….
Friends	a. My friends speak with me…
School	a. Which kindergarten did you attend?
	a. Which primary school did you attend?
	a. Which lower-secondary school do you attend?

is possible to examine the importance of individual variables in a cluster using a Chi-Squared test and to identify if the item was relevant for the best solution.

To construct a valid cluster model, we first verified that the likelihood distance measure among the variables considered is independent and that each categorical variable has a multinomial distribution. Subsequently, we observe the silhouette measure of cohesion and separation, which measures the relationship of the variables within and between clusters (Norusis, 2011) to ensure that both within-cluster distance and between-cluster distance are non-zero, indicating sufficient variation between variables, allowing us to accept the model as valid. The silhouette measure of cohesion and separation for our final model showcased a “fair separation” distance (ibid.) between clusters (0.3). Table 5 displays the presence of four main clusters for each context, each characterised by distinct language use patterns that enable grouping the 163 students in LEONIDE.

In the *Dominant German* cluster (31.9 %), the exclusive use of German in the family context holds the highest predictor importance (1), with 100 % of the students declaring consistent use of German at home. The same applies to the school context, with all students declaring to have attended all grades at schools with German as the language of instruction. Similarly, 95.2 % of the students in this cluster report always using German with friends. The absence of English and other languages emerges as a predictor across all contexts, albeit with lower predictor importance.

In the *Dominant Italian* cluster (27.6 %), the lack of German in the family context has the highest predictor importance (1), with 100 % of the students in this cluster indicating never using German at home. Additionally, the non-use of German with friends influences cluster affiliation, with a predictor importance of 0.44. Furthermore, 88.6 % of students in this cluster report always using Italian at home and 100 % of them always using Italian with their friends. All students in this cluster attended schools with Italian as the language of instruction. The use of English and other languages exhibits lower predictive power, contributing to cluster affiliation only when never used in any of the contexts.

In the *Dominant Heritage Language + Italian* cluster (22.7 %), 74.1 % of the students always use ‘other’ languages besides German, Italian and English in the family context; 70.4 % of them never use German within the family setting, while 44.4 % prefer using Italian in this context. This pattern may be attributed to the cluster’s predominantly Italian-medium school attendance (40.7 %), which carries moderate predictor importance (0.44). The remaining students attended schools in German or in other languages outside South Tyrol. In the friends context, language use is more varied, with a prevalent tendency to always use Italian in most interactions (77.8 %), whereas German and English are rarely used (respectively, 44.4 % and 68.6 % of students indicate they never use them).

Most students (78.9 %) in the smallest cluster, *Dominant Italian + German* (17.8 %), declare that they always use Italian, while often using German in the family setting. Similarly, in interactions with friends, most students indicate consistent use of German (63.2 %) and Italian (84.2 %). Additionally, most students in this cluster received German-medium instruction in all school grades (68.4 %).

Comparing alternative categorisations

The cluster analysis condensed the original values of the LEONIDE *author_L1* variable into four distinct clusters as shown in Fig. 2. Smaller groups with values *OTHER-OTHER, IT-EN-OTHER*, and *EN-OTHER* were fully incorporated into the *Heritage Language + Italian* cluster, whereas the *IT-EN* group was clustered with the *Dominant Italian* cluster. Conversely, some of the larger groups were split into different clusters. While German-only speakers (*DE*) are entirely integrated into the *Dominant German* cluster, some students who reported using exclusively Italian with their parents (*IT*) are now grouped into the *Dominant Italian + German* cluster. The former Italian-German bilingual group (*DE-IT*) (17) was completely recategorised into the *Dominant Italian + German* cluster. This cluster also includes students with the *author_L1* values *DE-IT-OTHER* and even some with *OTHER* and *IT*, who, upon considering their daily language practices, exhibit enough similar profiles to the declared Italian-German bilingual group. Consequently, this cluster encompasses students previously categorised as monolingual speakers in LEONIDE, totalling now 29 students with the new additions, thus allowing for a richer contextualisation of their language practices. The group labelled as *OTHER* in LEONIDE initially consisted of students who reported speaking languages other than German, Italian, and English with their parents. While a substantial portion of these students (25) finds placement within the *Dominant Heritage Language + Italian* cluster, others are dispersed across multiple clusters such as *Dominant German* (8), *Dominant Italian* (3) and *Dominant Italian + German* (4).

The identification of these new clusters provides some interesting observations. Firstly, the clusters replicated the idea of dominant language constellations in a way that we are now able to identify prototypical language profiles of LEONIDE participants on the basis of shared practices of language use. Compared to the previous grouping offered in the *author_L1* metadata, speakers shift among the clusters in different ways. This is attested for example by the addition of a second dominant bilingual cluster (*Dominant Heritage Language + Italian*) next to the Italian-German bilinguals one, as well as by the numerous students previously categorised as

Table 5
Results of the cluster analysis showing clusters, contexts and predictors.

Cluster name and size	Dominant German *n* = 52 (31.9 %)	Dominant Italian *n* = 45 (27.6 %)	Dominant Heritage Language + Italian *n* = 37 (22.7 %)	Dominant Italian + German *n* = 29 (17.8 %)
Family				
German	always (100 %) (1)	never (100 %) (1)	never (70.4 %) (1)	often (78.9 %) (1)
Italian	never (64.3 %) (0.56)	always (88.6 %) (0.56)	sometimes (44.4 %) (0.56)	always (78.9 %) (0.56)
English	never (85.7 %) (0.06)	never (85.7 %) (0.06)	never (74.1 %) (0.06)	never (47.4 %) (0.06)
Other	never (90.5 %) (0.06)	Never (74.3 %) (0.36)	Always (74.1 %) (0.36)	never (47.4 %) (0.36)
Friends				
Italian	never (54.8 %) (0.38)	always (100 %) (0.38)	always (77.8 %) (0.38)	always (84.2 %) (0.38)
German	always (95.2 %) (0.44)	never (42.9 %) (0.44)	never (44.4 %) (0.44)	always (63.2 %) (0.44)
English	never (88.1 %) (0.05)	never (68.6 %) (0.05)	never (55.6 %) (0.05)	never (63.2 %) (0.05)
School (language of instruction)	German (100 %) (0.43)	Italian (100 %) (0.43)	Italian (40.7 %) (0.43)	German (68.4 %) (0.43)

As far as the clusters are concerned, the *n* refers to the absolute number of participants belonging to a certain cluster with the respective percentage in brackets. For each language belonging to a context of use (family, school and friends), the relative importance of the predictor to the overall clustering solution is indicated on a scale from 0 (least important) to 1 (most important), and the respective percentage of cases for each of the response categories.

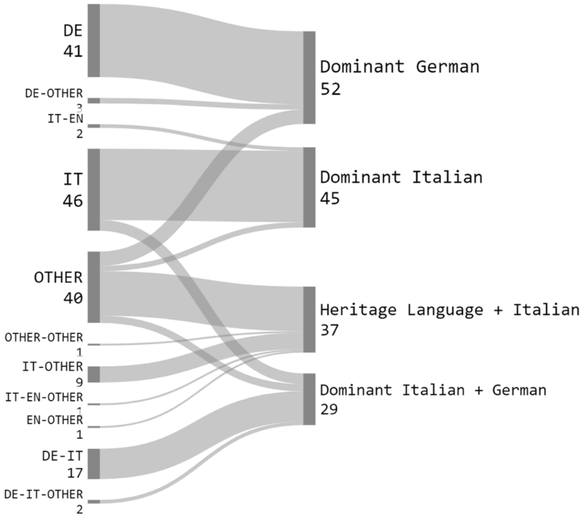


Fig. 2. Comparison between author_L1 in LEONIDE and the clusters using a Sankey Diagram.

monolingual speakers who now shifted to one of these two clusters.

Secondly, the cluster analysis has shown that, when considering other contexts of language use, students initially classified under *OTHER* due to the presence of heritage languages in the family context use the majority languages (either Italian or German or both) in their daily interactions with peers and within the school environment. These languages could potentially become their dominant languages, making them indistinguishable from informants the literature usually refers to as ‘native speakers’. This evidence supports existing research on heritage language speakers, which suggests regarding them as part of the native language continuum (Wiese et al., 2022; Rothman & Treffers-Daller, 2014; Rothman et al., 2023). Although this perspective may diminish the documentation of the existing linguistic diversity in the corpus, it would also support decisions in corpus creation that prioritise the privacy of students with very specific heritage language backgrounds that might otherwise not be guaranteed if languages are specified in detail. It is imperative at this point to explore the implications of this recategorisation in conjunction with the previous possible categorisations available in LEONIDE to assess its impact on study results.

An exploration of potential consequences of different categorisations for linguistic inquiry

In order to estimate the potential effect of categorising individuals depending on different aspect of language background, we performed an analysis that explores a number of simulated linguistic studies based on automatically extracted linguistic features gathered from both German and Italian texts in LEONIDE. We use the Common Text Analysis Platform (CTAP, see https://all4ling.eurac.edu/ctapWebApp/) to extract linguistic features for the texts of both languages. The tool, originally implemented for English text analysis (Chen & Meurers, 2016), allows the extraction of linguistic features for Italian, German and English texts (cf. Okinina et al., 2020) based on the same underlying processing pipeline, using the Unstructured Information Management (UIMA) framework (Ferrucci et al., 2004). The features provided by the tool span descriptive statistics on text, sentence and word length, lexical and syntactic complexity features, lexical sophistication features, PoS density features, cohesion features and raw frequencies of various other linguistic elements. Of 297 different features retrievable from the system, we selected the subset of 73 features that were available for both German and Italian texts.[9] Texts with less than 30 characters were excluded from the dataset as they could not be analysed with the tool. In total, our dataset contains 1658 texts. As an example, we display one of the LEONIDE texts (DE_op_1_55 × 31A01_100) in Fig. 3 and comment few features extracted from the text with CTAP. A (normalised) translation in English of the text is provided in the footnote.[10]

The text is 110 tokens long (CTAP feature: Number of Tokens) and tokens are, on average, 5.26 characters long (feature: Mean Token Length in Letters). Sentences are on average 12.22 tokens long (feature: Mean Sentence Length in Tokens) and each sentence has on average 0.44 dependent clauses (feature: Dependent Clauses per Sentence), highlighting the very simple syntax used in the text. The lexical diversity, measured with RTTR (Root Type-Token Ratio) is 6.58, which is slightly above the average of the German subcomponent of the corpus (RTTR mean = 5.66).

In our analysis, we compared different aspects of a student's language background and their effect on the outcome of linguistic studies having as the dependent variable one of the 73 linguistic features extracted with CTAP. In particular, we investigated whether grouping students depending on the presence or absence of a certain target language in a) their educational environment, b) their family environment or c) their dominant language constellation yields different results, when testing for group differences with linear mixed effects models observing both statistical significance and effect size. Instead of being interested in the differences between assumed "learners" and assumed "natives" for individual features, our analysis rather aims to explore whether some features are more influenced by certain aspects of a speaker's language background than by others. This would be in line with other more qualitative studies analysing texts of multilingual speakers (e.g., Cenoz & Gorter, 2011; Kobayashi & Rinnert, 2013). Furthermore, we want to investigate how holistic models of classifying language background relate to these results. Thus, we create three different categorisations that serve as alternatives for building groups and identifying group differences. In an exploratory fashion, we investigate group differences for each linguistic feature, using in turn one of the categorisations stated above. To disassociate the categorisation from the target language of the text, we define for each text:

a) whether the (text) target language is the students' school language (i.e. categorisation based on language of instruction);
b) whether the (text) target language is one of the family languages (as encoded in the *author_L1* variable);
c) whether the (text) target language is the only dominant language, the dominant language next to a heritage language, the dominant language next to another official language of the region or not a dominant language in their language constellations at all (i.e. categorisation based on DLC).

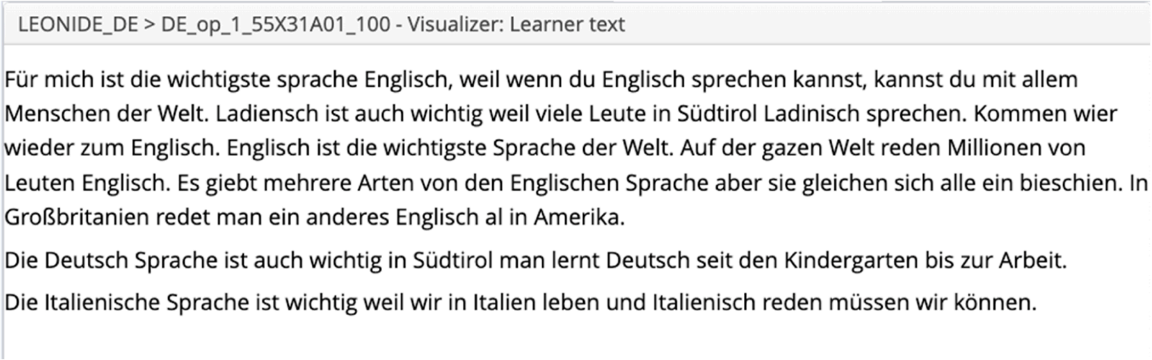
We used linear mixed effects models (Gries, 2021; Link & Cunnings, 2015) to identify significant group differences while accounting for individual variation of students who contributed more than one text to the sample. We established the significance of the predictor variable (i.e. the categorisation) at a confidence level of 0.95 (i.e., alpha 0.05) with a log-likelihood ratio test (Chi-Squared test), comparing two nested linear mixed effects regression models: (i) a null model that only accounts for individual variation using the student's ID as a random factor and (ii) a model that additionally contains the categorisation in question as a main factor. From every specified model, we also extracted he marginal R-squared, namely the variance explained by the fixed effect predictor (i.e., the categorisation) relative to the total variance in the response.

All analyses were performed in R, using the *lme4* package version 1.1–35.1 (Bates et al., 2015) for building the linear mixed effects models and the *anova* function for model comparison. Since this exploratory analysis involves running a high number of tests (73 features x 3 analyses with different versions of the predictor variable = 219 tests), we adjusted the p-values obtained implementing the False Discovery Rate with the Benjamini-Hochberg method (Benjamini & Hochberg, 1995).[11]

[9] The full list of features can be found in the supplementary materials provided at: https://gitlab.inf.unibz.it/commul/lca/rmal2024_speaker-categorization.

[10] "For me, the most important language is English, because if you can speak English, you can talk to anyone in the world. Ladin is also important because many people in South Tyrol speak Ladin. Let's come back to English. English is the most important language in the world. Millions of people all over the world speak English. There are several types of English, but they are all a bit alike. English is spoken differently in the UK than in America. The German language is also important in South Tyrol, where German is learnt from kindergarten to work. The Italian language is important because we live in Italy and we have to be able to speak Italian."

[11] The final tables with p-values and R-squared used for analysis are available in the supplementary materials at: https://gitlab.inf.unibz.it/commul/lca/rmal2024_speaker-categorization.



LEONIDE_DE > DE_op_1_55X31A01_100 - Visualizer: Learner text

Für mich ist die wichtigste sprache Englisch, weil wenn du Englisch sprechen kannst, kannst du mit allem Menschen der Welt. Ladiensch ist auch wichtig weil viele Leute in Südtirol Ladinisch sprechen. Kommen wier wieder zum Englisch. Englisch ist die wichtigste Sprache der Welt. Auf der gazen Welt reden Millionen von Leuten Englisch. Es giebt mehrere Arten von den Englischen Sprache aber sie gleichen sich alle ein bieschien. In Großbritanien redet man ein anderes Englisch al in Amerika.

Die Deutsch Sprache ist auch wichtig in Südtirol man lernt Deutsch seit den Kindergarten bis zur Arbeit.

Die Italienische Sprache ist wichtig weil wir in Italien leben und Italienisch reden müssen wir können.

Fig. 3. The DE_op_1_55 × 31A01_100 text from the German subsection of the LEONIDE corpus.

Differences in the explanatory power of the investigated categorisations

Our analysis quantifies and examines the linguistic features for which there are differences in the explanatory power across the three categorisations (family language, school language and DLC clusters). We do so by visualising the features where certain categorisations fail to reach significance (Fig. 4) and by plotting the distribution of R-squared values (Fig. 5). We also compare the predictive power of each categorisation, identifying which categorisation explains the most features and, where discrepancies occur, assessing the unique contributions of each categorisation.

For most of the linguistic features investigated (60/73), the log-likelihood ratio test yielded a significant result for all three alternative categorisations, indicating that adding any of the categorisations as a predictor to a null model improves the fit of the model. For only one feature – the number of coordinating conjunctions per token (label: CoordinatingConjunction in Fig. 4) – none of the categorisations was found to be a significant predictor. For the remaining 12 features, only some of the alternative categorisations were found to be significant. These discrepancies are highlighted in Fig. 4.

For example, the categorisation based on family language(s) would not be considered a significant predictor for the PoS density of adjectives in a text (label: Adjective), while the categorisation based on school language and DLC clusters would do so.

Fig. 4 also clearly shows that the language background categorisation based on the holistic model of the DLC was the categorisation showing the lowest number of non-significant results, with only one feature for which DLC clusters were not a significant predictor at alpha 0.05. In contrast, categorisations based on home language and school language could not always be found to be predictive for the linguistic features investigated. In addition, the features for which home language and school language categorisations were not significant predictors only partly overlap.

Comparing the number of successful study outcomes (i.e. the categorisation being predictive according to a log-likelihood ratio test) for the categorisations in a pairwise fashion with a McNemar test showed no significant difference between *school_language* and *author_L1* or between *school_language* and the DLC clusters, but did show a difference between *author_L1* and the clusters based on the DLC (McNemar's chi-squared = 7.1111, df = 1, p-value = 0.007661).

Finally, looking at the difference in effect sizes for the three categorisations, Fig. 5 summarises the marginal R-squared obtained for the fixed effects in the models for all three categorisations.

The categorisation based on the DLC clusters had the highest average R-squared, while the categorisation based on the family languages had the lowest. For all categorisations, lexical diversity measures showed the highest effect sizes (R-squared between 0.14 and 0.28). Although the categorisation based on family languages generally showed lower effect sizes than those based on the school language and DLC, it yielded higher effect sizes for lexical diversity features, suggesting a potential positive effect of exposure to the target language at home on vocabulary richness. Other features showing clear differences between the variance explained by one categorisation as compared to other categorisations included the number of sentences and the number of subordinating conjunctions. The number of sentences was better explained by the categorisation based on languages used with the parents at home (R-squared: 0.039 home; 0.029 school). This finding could be related to a higher fluency of writers that frequently use the language at home and therefore are able to produce more sentences in this timed task setting. In contrast, the number of subordinating conjunctions was better explained by the categorisation based on the language of instruction (R-squared: 0.048 school; 0.022 home), which is also not surprising, considering that educational contexts might be more formal, needing more complex sentence structures than everyday language used at home.

Discussion

We started with the hypothesis that using different criteria for categorising speakers would have an impact on the outcome of linguistic studies. The analysis revealed that this is indeed the case for one sixth of the features investigated. When adopting a confidence level of 0.95, 12 of the simulated linguistic studies have different outcomes depending on the criteria used for categorisation. When looking just at the significance of the log-likelihood ratio tests, the holistic categorisation implemented via the DLC clusters was

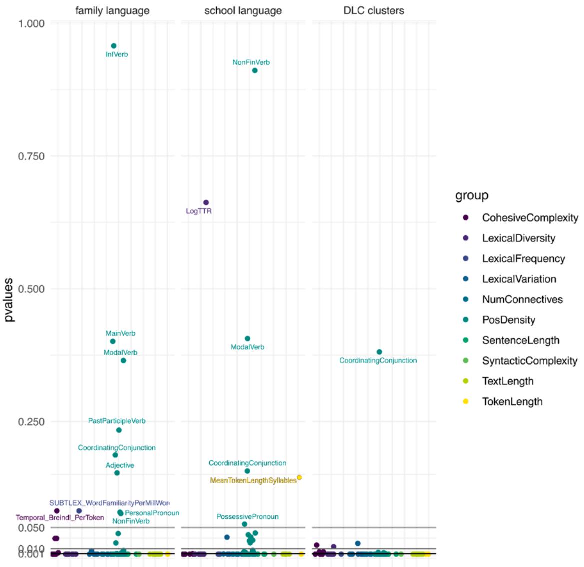


Fig. 4. Features for which the three tested categorisations did not have a significant effect on the variance explained by a mixed effects model. This graph highlights features for a p-value above 0.05.[12]

[12] The area between 0.05 and 0.001 indicates the existence of additional features not being significant, if a more conservative confidence level for establishing predictor significance would be chosen.

predictive for most of the features. This means that this categorisation significantly explained the variance observed in the data for features, where single aspects of language background did not do so. This suggests that for our heterogeneous dataset, with a substantial number of students whose educational language environment differs from their family language environment, the DLC was able to create rather homogeneous speaker groups by clustering speakers over various aspects of language use. While this might not be necessary for datasets with a more homogeneous distribution of language backgrounds (e.g., all participants use the same language across all domains of use), it is worthwhile considering when creating and/or analysing datasets from increasingly diverse linguistic environments in Europe and beyond.

When considering the variance explained by the different categorisations through effect sizes, we were able to identify features

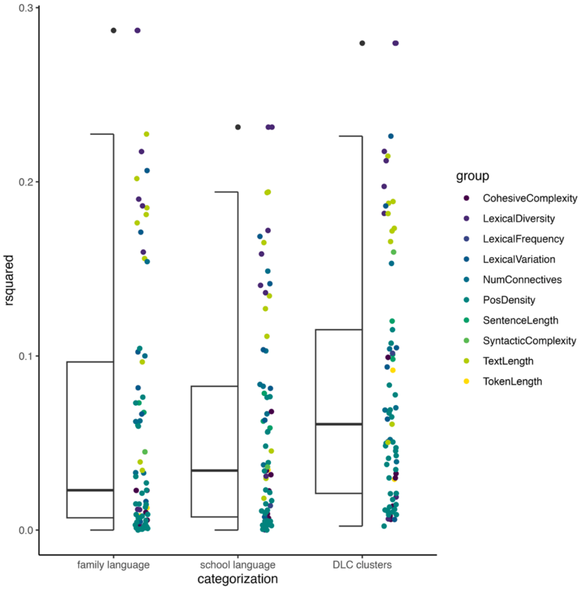


Fig. 5. Summary of marginal R-squared for the three categorisations of language background.

which are more strongly related to exposure in educational contexts vs. those more strongly related to language use in the family environment, besides those that are better explained by a more holistic view of the language practices of the speakers. Our results therefore suggest that an increased transparency on how language background is operationalised could allow us to zoom into individual aspects of language use with more clarity. Unambiguously defining *a priori* the language background and its operationalisation (at least with one metadata identifying the L1 but preferably documenting various language background variables side by side which can then be combined), would allow corpus users to focus on aspects of their interest and readers of corpus analyses to judge the relevance of the results for their own research, facilitating the progress of the field through cumulative research (e.g., Larsson et al., 2024).

From a methodological point of view, our analysis aims at illustrating the potential effect of using p-values to decide on study outcomes. While the challenges in using, interpreting and relying on p-values are well-known in the literature (cf. Wasserstein et al., 2019; Lakens, 2021), it is still common practice in LCR to test whether (usually one of few) linguistic features can be explained by certain characteristics of a writer using mixed effects models and testing for statistical significance of predictor variables, as we did in RQ2. Accompanying the interpretation of p-values with effect sizes (the variance explained by the fixed effects of the models) was key to better understand the differences across categorisations over the tested features.

Concluding remarks

The conceptual and empirical work sketched out in this contribution connects different ideas that have shaped practices of categorising speakers based on their language backgrounds. We attempted to reconstruct the way these practices have been put in place in the field, shedding light on the implications of the use of certain categorisations and related terms. The aim has been to fill a gap in the methodologies used for speaker categorisation in LCR, where it is usually a common practice to categorise speakers' language backgrounds in the form of metadata that capture different aspects of speakers' experience with languages. Of these metadata, we focused on the L1 metadata variable among the various person-related metadata, because of its central role in encoding language background across nearly all the corpora selected for our review.

Our review has revealed several key observations about the way in which this metadata variable is termed, described and operationalised. We encountered challenges related to accessing corpus documentation: in fact, some corpora lacked comprehensive and coherent documentation, with information being fragmented across disparate sources. Missing documentation or discrepancies in the information published in different sources pose notable challenges for comparability of results and interoperability of resources (see also Hashimoto & Nelson, 2024). Our review showed that formalist models of language dominate the field of LCR. These models suggest that, in the repertoires of speakers, languages are categorical bounded entities that can be ordered and labelled. Consequently, languages are deemed as countable (as it is the case for the L1 metadata variable) and language-related aspects (like proficiency in the L2) are documented in separate metadata. Holistic models that, instead, regard language learning as a process that should be studied and documented as a whole system considering the interactional practices of the speakers, are largely overlooked in LCR.

To demonstrate how holistic models can be introduced in LCR, we used information about speakers' language practices documented in LEONIDE and developed a new categorisation based on the principles of the holistic model of Dominant Language Constellations. Our investigation not only confirmed that integrating holistic models in LCR is feasible, but also demonstrated that this approach reveals a clearer picture about speakers' multi-faceted language practices in their everyday life. Additionally, adopting a holistic perspective comes with practical advantages: it avoids the need to impose predefined grouping criteria, as the clusters derived from participants ordinal or continuous responses can serve directly as categorical variables in planned analyses.

While offering different aspects of a language background as metadata (e.g., age of onset, exposure, context in which a language is spoken, etc.) as proposed by Paquot et al. (2024) would be ideal to allow corpus users to decide *a posteriori* which aspect (or set of aspects) they want to use, this comes with the drawback of eliciting detailed language profiles, which might not always be possible when participants' anonymity should be preserved (cf. Dirdal et al., 2022: 119). Offering categorisations adopting a holistic view on language background that integrates various aspects into one variable could be a viable alternative, balancing both usability and privacy concerns.

However, while our new categorisation provides a more comprehensive methodology to describe participants having similar patterns of language use in different contexts and seemed to explain more variance in our case study, it remains another form of categorisation, albeit a more multi-faceted one. In line with other scholars (Luk and Bialystock, 2013), future research should go beyond the transformation of speakers' language background information into a static categorical variable. Instead, operationalising it in terms of *gradable category membership* across several dimensions, along which category membership may vary, as suggested by Berthele (2021: 110) among others, seems to be a promising alternative. This approach aligns conceptually better with the idea that language backgrounds are dynamic within the individual, and that relevant language experience factors contribute to determine varying degree of membership of speakers in multiple language-related categories.

CRediT authorship contribution statement

Olga Lopopolo: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Arianna Bienati:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jennifer-Carmen Frey:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Aivars Glaznieks:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Stefania Spina:** Writing – review &

editing, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the reviewers and the editors for their invaluable feedback, as well as Verena Platzgummer for her insightful suggestions on previous stages of this work, and Rossella Viggiano for her technical support with the cluster analysis. The authors also thank the Department of Innovation, Research University and Museums of the Autonomous Province of Bozen/Bolzano for covering the Open Access publication costs.

References

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning, 59*(2), 249–306. https://doi.org/10.1111/j.1467-9922.2009.00507.x

Ädel, A., & Römer, U. (2012). Research on advanced student writing across disciplines and levels: Introducing the Michigan Corpus of Upper-level Student Papers. *International Journal of Corpus Linguistics, 17*(1), 3–34. https://doi.org/10.1075/ijcl.17.1.01ade

Aronin, L. (2006). Dominant language constellations: An approach to multilingualism studies. In M. Ó. Laoire (Ed.), *Multilingualism in educational settings* (pp. 140–159). Schneider Publications.

Aronin, L. (2019). Dominant language constellation as a method of research. In E. Vetter & U. Jessner (Eds), *International research on multilingualism: breaking with the monolingual perspective* (pp. 13–26). Springer.

Aronin, L., & Jessner, U. (2014). Methodology in Bi- and Multilingual Studies: From simplification to complexity. *AILA Review, 27*(1), 56–79.

Aronin, L., & Moccozet, L. (2021). Dominant language constellations: Towards online computer- assisted modelling. *International Journal of Multilingualism, 20*(3), 1067–1087. https://doi.org/10.1080/14790718.2021.1941975

Aronin, L., & Singleton, D. (2012). *Multilingualism.* John Benjamins.

Barratt, L. (2018). Monolingual Bias. *The tesol encyclopedia of english language teaching* (pp. 1–7). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118784235.eelt0024

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*(1), 289–300.

Berthele, R. (2021). The Extraordinary Ordinary: *Re*-engineering Multilingualism as a Natural Category. *Language Learning, 71*(S1), 80–120.

Berthele, R., & Udry, I. (Eds.). (2021). *Individual differences in early language learning: the role of language aptitude, cognition, and motivation.* Language Science Press.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: the case of systematicity. *Language Learning, 33*(1), 1–17. https://doi.org/10.1111/j.1467-1770.1983.tb00983.x

Blommaert, J., & Backus, A. (2013). Superdiverse repertoires and the individual. In I. de Saint-Georges & J.-J. Weber (Eds.), *Multilingualism and multimodality: current challenges for educational studies* (pp. 11–32). Sense Publishers.

Bonfiglio, T.P. (2010). *Mother Tongues and Nations. The Invention of the Native Speaker*. De Gruyter Mouton. https://doi.org/10.1515/9781934078266.

Bonfiglio, T. P. (2013). Inventing the Native Speaker. *Critical Multilingualism Studies, 1*(2), 29–58.

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), The cambridge handbook of learner corpus research (pp. 35–56). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.003.

Centre for English Corpus Linguistics. (2019). *Learner corpora around the world.* https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

Cenoz, J., & Gorter, D. (2011). Focus on multilingualism: A study of trilingual writing. *The Modern Language Journal, 95*(3), 356–369.

Chen, X., & Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC* (pp. 113–119).

Cheng, L. S. P., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.715843

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Colombo, S., & Stopfner, M. (2018). Alte und neue Formen der Mehrsprachigkeit in Südtirol. In P. Mauser & M. Dannerer (Eds.), *Formen der mehrsprachigkeit: sprachen und varietäten in sekundären und tertiären bildungskontexten* (pp. 123–142). Stauffenburg.

Cook, V. (1992). Evidence for multicompetence. *Language Learning, 42*(4), 557–591.

Cook, V. (2002). Portraits of the L2 User. *Multilingual Matters*.

Copland, F., Mann, S., & Garton, S. (2016). Introduction: positions, experiences and reflections on the native speaker issue. In F. Copland, S. Garton, & S. Mann (Eds.), *LETs and NESTs: voices, views and vignettes* (pp. 5–19). British Council.

Csardi, G., & Nepusz, T. (2005). The Igraph software package for complex network research. *InterJournal, Complex Systems*, 1695.

Cummins, J. (2008). Teaching for transfer: Challenging the two solitudes assumption in bilingual education. In N. H. Hornberger (Ed.), *Encyclopedia of language and education* (pp. 1528–1538). Springer. https://doi.org/10.1007/978-0-387-30424-3_116.

Davies, A. (2013). *Native speakers and native users: loss and gain*. Cambridge University Press.

Dewaele, J. M. (2018). Why the dichotomy 'L1 versus LX User' is Better than 'Native Versus Non-native Speaker. *Applied Linguistics, 38*(2), 236–240. https://doi.org/10.1093/applin/amw055

Dirdal, H., Hasund, I. K., Danbolt Drange, E.-M., Thue Vold, E., & Berg, E. M. (2022). Design and construction of the Tracking Written Learner Language (TRAWL) Corpus: A longitudinal and multilingual young learner corpus. *Nordic Journal of Language Teaching and Learning, 10*(2), 115–135. https://doi.org/10.46364/njltl.v10i2.1005

Dominguez, L. (2010). *Spanish learner language oral corpora (SPLLOC)*. http://www.splloc.soton.ac.uk/.

Ferrucci, D., & Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering, 10*(3-4), 327–348.

Gal, S., & Irvine, J. T. (1995). The boundaries of languages and disciplines: How ideologies construct difference. *Social Research, 62*(4), 967–1001.

Garcia, O., & Wei, L. (2014). Language, languaging and bilingualism. Eds.. In O. García, & L. Wei (Eds.), *Translanguaging. Language, bilingualism and education* (pp. 5–18). Palgrave Macmillan

Gilquin, G. (2001). The integrated contrastive model: Spicing up your data. *Languages in Contrast, 3*(1), 95–123. https://doi.org/10.1075/lic.3.1.05 gil

Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L., & Nicolas, L. (2022). Leonide: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research, 8*(1), 97–120. https://doi.org/10.1075/ijlcr.21004.gla

Glaznieks, A., Frey, J.-C., Abel, A., Nicolas, L., & Vettori, C. (2023). The Kolipsi corpus Family: Resources for learner corpus research in Italian and German. *Italian Journal of Computational Linguistics, 9*(2). https://doi.org/10.13092/lo.127.11087

Granger, S. (2002). A Bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.6.04 gra.

Gries, S. Th (2021). (Generalized linear) Mixed-effects modeling: A learner corpus example. *Language Learning, 71*(3), 757–798.

Hackert, S. (2012). *The emergence of the English native speaker. A chapter in nineteenth-century linguistic thought*. De Gruyter Mouton. https://doi.org/10.1515/9781614511052

Hammarberg, B. (2014). 1. Problems in defining the concepts of L1, L2 and L3. In A. Otwinowska, & G. De Angelis (Eds.), *Teaching and learning in multilingual contexts: sociolinguistic and educational perspectives* (pp. 3–18). Multilingual Matters. https://doi.org/10.21832/9781783091263-003.

Hashimoto, B., & Nelson, K. (2024). Recent trends in corpus design and reporting: A methodological synthesis. *Research in Corpus Linguistics, 12*(1), 59–88. https://doi.org/10.32714/ricl.12.01.03

Heuboeck, A., Holmes, J., & Nesi, H. (2010). *The BAWE Corpus Manual. Coventry University*. https://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/.

Herdina, P., & Jessner, U. (2002). *A dynamic model of multilingualism: Perspectives of change in psycholinguistics*. Multilingual Matters.

Holliday, A. (2006). Native speakerism. *English Language Teaching, 60*(4), 385–387.

Hufeisen, B. (2010). Theoretische Fundierung multiplen Sprachenlernens–Faktorenmodell 2.0. *Jahrbuch Deutsch als Fremdsprache, 36*, 200–207.

Hufeisen, B. (2018). Models of multilingual competence. In A. Bonnet, & P. Siemund (Eds.), *Foreign language education in multilingual classrooms* (pp. 173–189). John Benjamins.

Ivaska, I. (2014). The Corpus of Advanced Learner Finnish (LAS2): Database and toolkit to study academic learner Finnish. *Apples-Journal of Applied Language Studies, 8*(3).

Izumi, E., Uchimoto, K., & Isahara, H. (2005). Error annotation for corpus of Japanese learner english. In *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC-2005)*. IJCNLP 2005 https://aclanthology.org/I05-6009.

Jarvis, S. (2010). Comparison-based and detection-based approaches to transfer research. *EUROSLA Yearbook, 10*(1), 169–192. https://doi.org/10.1075/eurosla.10.10jar

Kobayashi, H., & Rinnert, C. (2013). L1/L2/L3 writing development: Longitudinal case study of a Japanese multicompetent writer. *Journal of Second Language Writing, 22*(1), 4–33.

Lakens, D. (2021). The practical alternative to the *p* Value is the correctly used *p* Value. *Perspectives on Psychological Science, 16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Larsen-Freeman, D. (2002). Language acquisition and language use from a chaos/complexity theory perspective. In C. J. Kramsch (Ed.), *Language acquisition and language socialization: ecological perspectives* (pp. 33–46). Continuum.

Larsen-Freeman, D. (2014). Another step to be taken - Rethinking the end point of the interlanguage continuum. In Z Han, & E. Tarone (Eds.), *Interlanguage. forty years later* (pp. 203–220). John Benjamins.

Larsson, T., Biber, D., & Hancock, G. R. (2024). On the role of cumulative knowledge building and specific hypotheses: The case of grammatical complexity. *Corpora, 19*(3).

Lenneberg, E. H. (1967). *Biological foundations of language*. Wiley.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(S1), 185–207.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K., & Walter, M. (2008). Das Lernerkorpus Falko. *Deutsch als Fremdsprache*. https://doi.org/10.37307/j.2198-2430.2008.02.02, 2/2008.

Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology, 25*(5), 605–621. https://doi.org/10.1080/20445911.2013.795574

May, S (Ed.). (2014). *The multilingual turn: implications for SLA, tesol and bilingual education*. Routledge/Taylor & Francis.

Medin, D. L., & Coley, J. D. (1998). Concepts and Categorisation. In J. Hochberg (Ed.), *Perception and cognition at century's end* (pp. 403–439). Academic Press. https://doi.org/10.1016/B978-012301160-2/50015-0.

Meisel, J. M. (2011). *First and second language acquisition*. Cambridge University Press.

Meissner, F.-J. (2004). Modelling plurilingual processing and language growth between intercomprehensive languages. In L. N. Zybatow (Ed.), *Translation in der globalen welt und neue wege in der Sprach- und Übersetzerausbildung* (pp. 31–57). Peter Lang.

Melo-Pfeifer, S. (2015). Multilingual awareness and heritage language education: Children's multimodal representations of their multilingualism. *Language Awareness, 24*(3), 197–215.

Meunier, F. (2016). Introduction to the LONGDALE Project. In E. Castello, K. Ackerley, & F. Coccetta (Eds.), *Studies in learner corpus linguistics* (pp. 123–126). Peter Lang.

Norusis, M. J. (2011). *IBM SPSS statistics 19 procedures companion*. Addison-Wesley.

Okinina, N., Frey, J. C., & Weiss, Z. (2020). CTAP for Italian: Integrating components for the analysis of Italian into a multilingual linguistic complexity analysis tool. In *Proceedings of the 12th Conference on Language Resources and Evaluation* (pp. 7123–7131). *LREC 2020*.

Osborne, J. (2015). Transfer and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), The cambridge handbook of learner corpus research (pp. 333–356). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.015.

Paikeday, T. (1985). *The native speaker is dead!* Paikeday Publishing.

Paquot, M., Larsson, T., Hasselgård, H., Ebeling, S. O., Meyere, D. D., Valentin, L., Laso, N. J., Verdaguer, I., & Vuuren, S.van (2022). The Varieties of English for Specific Purposes dAtabase (VESPA): Towards a multi-L1 and multi-register learner corpus of disciplinary writing. *Research in Corpus Linguistics, 10*(2). https://doi.org/10.32714/ricl.10.02.02. Article 2.

Paquot, M., König, A., Stemle, E., & Frey, J.-C. (2023). *Core metadata schema for learner corpor*a [data set]. *Open Data @UCLouvain*. https://doi.org/10.14428/DVN/4CDX3P

Paquot, M., König, A., Stemle, E., & Frey, J.-C. (2024). The core metadata schema for learner corpora: Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research. *International Journal of Learner Corpus Research, 10*(2), 280–300. https://doi.org/10.1075/ijlcr.24010.paq

Paradis, M. (2004). *A neurolinguistic theory of bilingualism*. John Benjamins.

Penfield, W., & Roberts, L. (1959). *Speech and brain-mechanisms*. Princeton University Press.

Puig-Mayenco, E., González Alonso, J., & Rothman, J. (2018). A systematic review of transfer studies in third language acquisition. *Second Language Research, 36*(1), 31–64. https://doi.org/10.1177/0267658318809147

Rojo, G., & Palacios, I. M. (2016). Learner Spanish on computer: The CAES 'Corpus de Aprendices de Español' project. In M. Alonso-Ramos (Ed.), *Spanish learner corpus research: current trends and future perspectives* (pp. 55–87). John Benjamins Publishing Company. https://doi.org/10.1075/scl.78.03roj.

Rothman, J., & Treffers-Daller, J. (2014). A prolegomenon to the construct of the native speaker: heritage speaker bilinguals are natives too! *Applied Linguistics, 35*(1), 93–98. https://doi.org/10.1093/applin/amt049

Rothman, J., Bayram, F., DeLuca, V., Di Pisa, G., Duñabeitia, J. A., Gharibi, K., Kolb, N., Kubota, M., & Wulff, S. (2023). Monolingual comparative normativity in bilingualism research is out of "control": Arguments and alternatives. *Applied Psycholinguistics, 44*(3), 316–329. https://doi.org/10.1017/S0142716422000315

Simpson, S. L., Briggs, J. O., & Swales, J. M. (1999). *The Michigan corpus of academic spoken English*. The Regents of the University of Michigan. https://ca.talkbank.org/access/0docs/MICASE.pdf.

Spina, S., & Siyanova-Chanturia, A. (2018). The longitudinal corpus of chinese learners of Italian (LOCCLI) [Poster]. In *13th Teaching and Language Corpora conference*. Cambridge.

Spina, S., Fioravanti, I., Forti, L., & Zanda, F. (2023). The CELI corpus: Design and linguistic annotation of a new online learner corpus. *Second Language Research, 40*(2), 457–477. https://doi.org/10.1177/02676583231176370

Stopfner, M., & Engel, D. (2019). Communicative competence in the context of increasing diversity in South Tyrolean schools. In E. Vetter, & U. Jessner (Eds.), *International research on multilingualism: breaking with the monolingual perspective* (pp. 59–80). Springer Nature.

Storto, A., Haukås, Å., & Tiurikova, I. (2023). Visualising the language practices of lower secondary students: outlines for practice-based models of multilingualism. *Applied Linguistics Review, 15*(5), 2035–2059. https://doi.org/10.1515/applirev-2022-0010

Tono, Y., & Díez-Bedmar, M. B. (2014). Focus on learner writing at the beginning and intermediate stages: The ICCI corpus. *International Journal of Corpus Linguistics, 19*(2), 163–177. https://doi.org/10.1075/ijcl.19.2.01ton

Volodina, E. (2021). ReadMe: SweLL-pilot collection. Swell-Release-V1. Retrieved December 19, 2023, from https://spraakbanken.github.io/swell-release-v1/Readme-SweLL-pilot.html.

Volodina, E., Pilán, I., Enström, I., Llozhi, L., Lundkvist, P., Sundberg, G., & Sandell, M. (2016). SweLL on the rise: Swedish learner language corpus for european reference level studies. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 206–212). European Language Resources Association (ELRA). https://aclanthology.org/L16-1031.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "$p < 0.05$. *The American Statistician, 73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., Martynova, M., Pashkova, T., Rizou, V., Schroeder, C., Shadrova, A., Szucsich, L., Tracy, R., Tsehaye, W., Zerbian, S., & Zuban, Y. (2022). Heritage speakers as part of the native language continuum. *Frontiers in Psychology, 12*, Article 717973. https://doi.org/10.3389/fpsyg.2021.717973