

Categorizing Speakers' Language Background

Theoretical Assumptions and Methodological Challenges for Learner Corpus Research

Olga Lopopolo^{1,2}, Arianna Bienati^{1,3}, Jennifer-Carmen Frey¹, Aivars Glaznieks¹,
Stefania Spina²

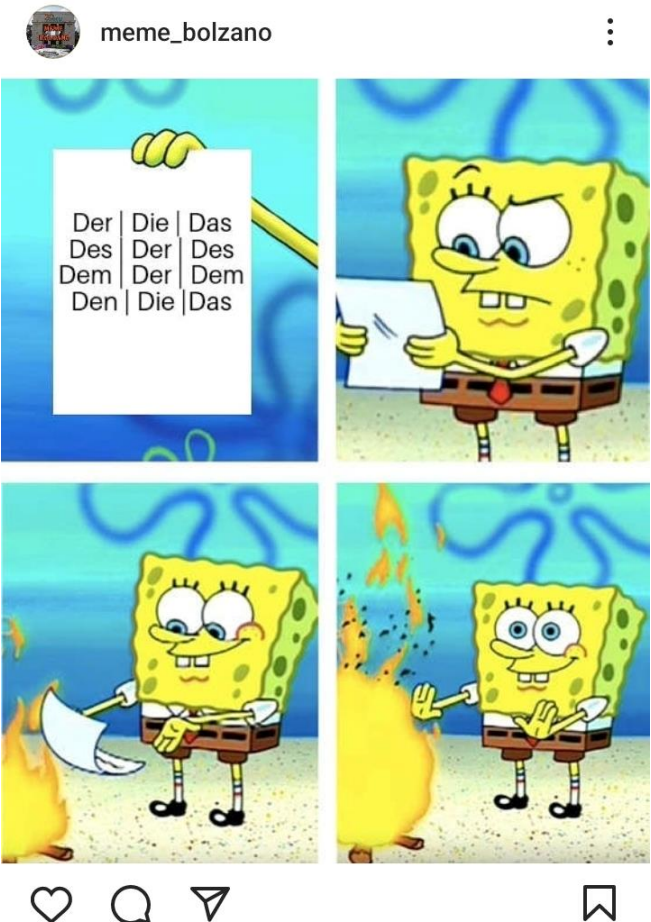
¹Eurac Research Bolzano

²University for Foreigners of Perugia

³University of Modena and Reggio Emilia

The multilingual province of Bolzano/Bozen in South Tyrol (Italy)

- **Autonomy** Statute of the region Trentino-South Tyrol (Constitutional Law, 10 November 1971).
- “**Language groups**”: Italian, German and Ladin speakers living in South Tyrol
- **Parity** between Italian, German and Ladin language (Art. 99) and **equality** of rights and representation within the political and public spheres
- Language is a **line of demarcation** for establishing ethnic identity
- “**Linguistic separatism**” (Carrozza, 1993)
- Three autonomous **education authorities** and school systems.
- **Membership declaration** of one of the 3 language groups at the moment of population census



A prominent 'multilingual' territory with multiple languages spoken seems to be 'monolingual' in the way in which public life is organized.

It is a site of tensions between different linguistic paradigms: a **multilingual paradigm** that is seen in the multilingual practices of individuals, and a **monolingual** one, for which the separation between languages and communities should be preserved.

What was the solution for South Tyrol?

Creation of a living environment where the three main language groups are **separated** ‘as much as possible’ (Pallaver 2014: 376), while the ‘principle of consociational democracy’ works to encourage cooperation between ‘the language groups’ elites’ (ibid.).

As a result, the relationships in particular between the German- and Italian-speaking communities are described as having “**sharp and bright boundaries** that are managed through power-sharing and “forced” cooperation” (Wisthaler 2015: 4).

Some scholars have seen a connection between the dominant ideology of **monolingualism** and its implications for social categorization in relation to language (and speakers)

(May 2012; Bauman & Briggs 2003; Ortega 2014; Pennycook & Otsuji 2015)

At its core, the **nation-state's promotion** of standardized monolingualism aligns with its imperative to consolidate a cohesive national identity.

Doing Learner Corpus Research in South Tyrol...

- ➔ Need for simplification of linguistic profiles to create meaningful groups, filter and sample observations
- ← Stay true to reality and shed light on a multilingual society, so far often over-simplified in corpus studies

Observe and discuss **'hidden' methodological decisions** regarding corpus design, sampling, analysis and reporting of results **with regard to speakers' language backgrounds in Learner Corpus Research (LCR)**

1. How does the field of LCR categorize speakers by their language background?
 - Which terminology, definitions and criteria for categorization are used?
2. Case study one: Potential consequences for analyses in LCR
3. Case study two: Integrating holistic views on multilingualism

The *L1* metadata as proxy for speakers' language background in Learner Corpus Research

- “**core metadata**” for LCR (Paquot et al. 2023)
- present in some form in practically all learner corpora
- plays a crucial role in distinguishing those participants that will be part of the reference group and those who will be part of the learner group
- relevant for comparisons across corpora of learners having different interlanguage varieties (Granger 2015)

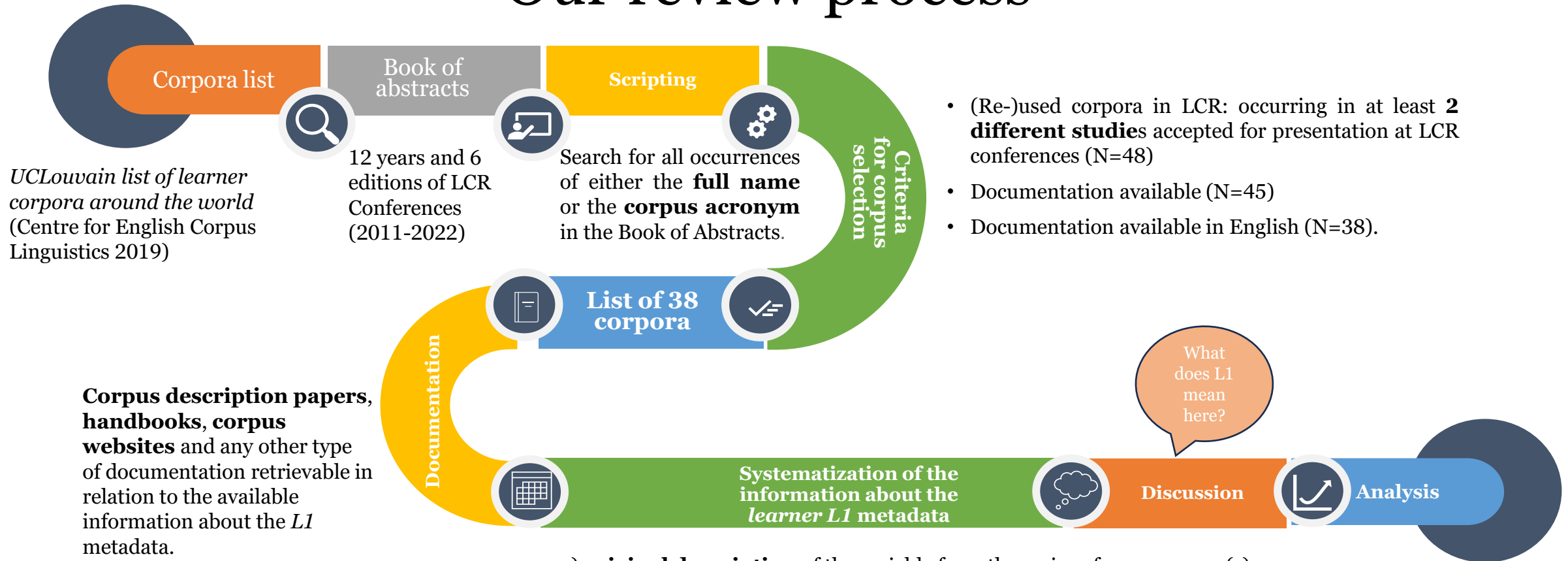
Other metadata related to language background observed sporadically but not systematically.

How is this metadata recorded in corpora?

- Which **terminology** is used to refer to L1 of people?
- Which **definitions** are given for the L1?
- Which **criteria** guided the **creation of groups/corpus samples**?
- Which **tools** have been used to assign L1 to texts?
- Which **other/additional variables** were used to describe the language background of people?

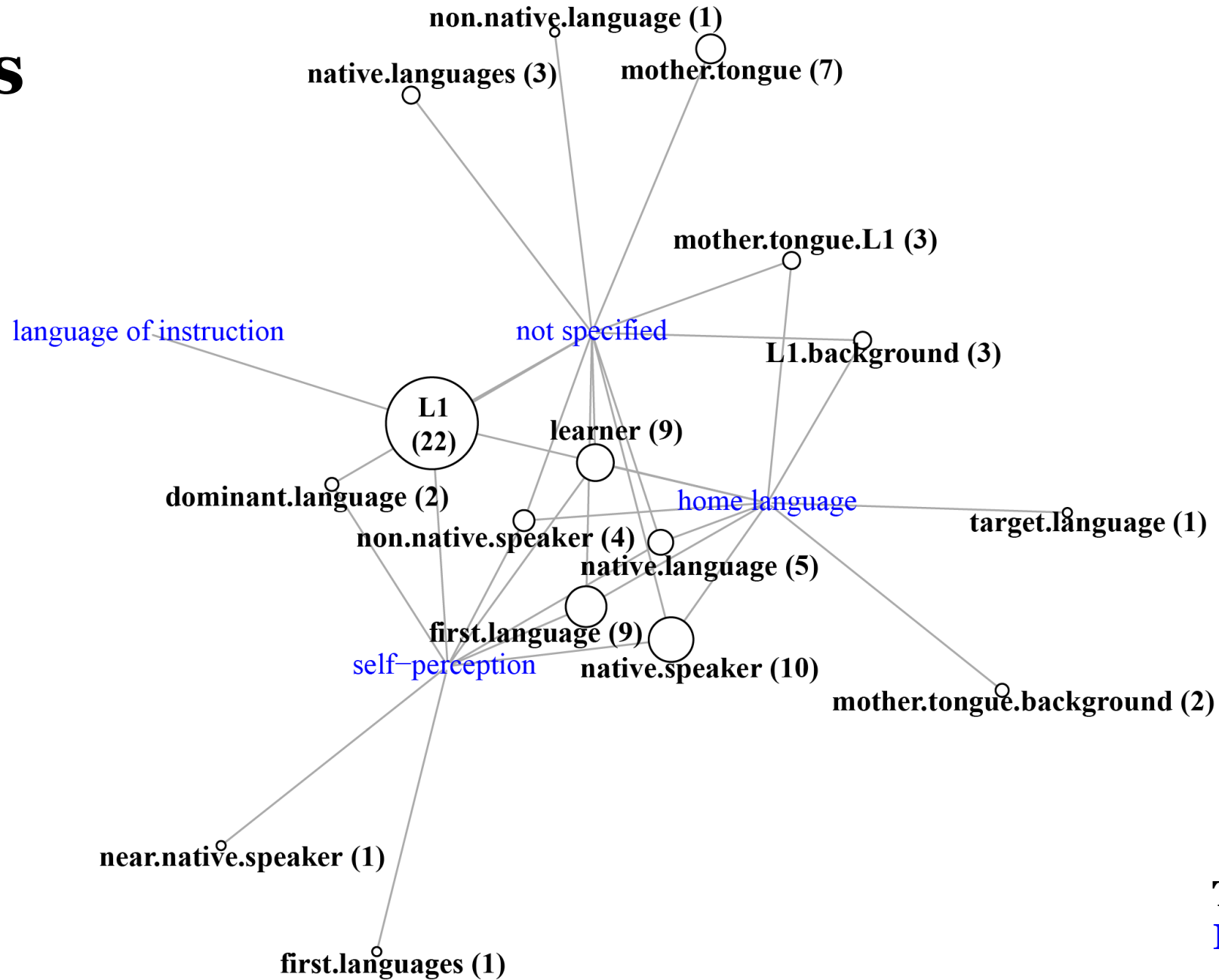
... How are corpora and their metadata **described and documented** in general?

Our review process



- original description** of the variable 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 categorized;
- instrument** of metadata elicitation (e.g., questionnaire).

Results

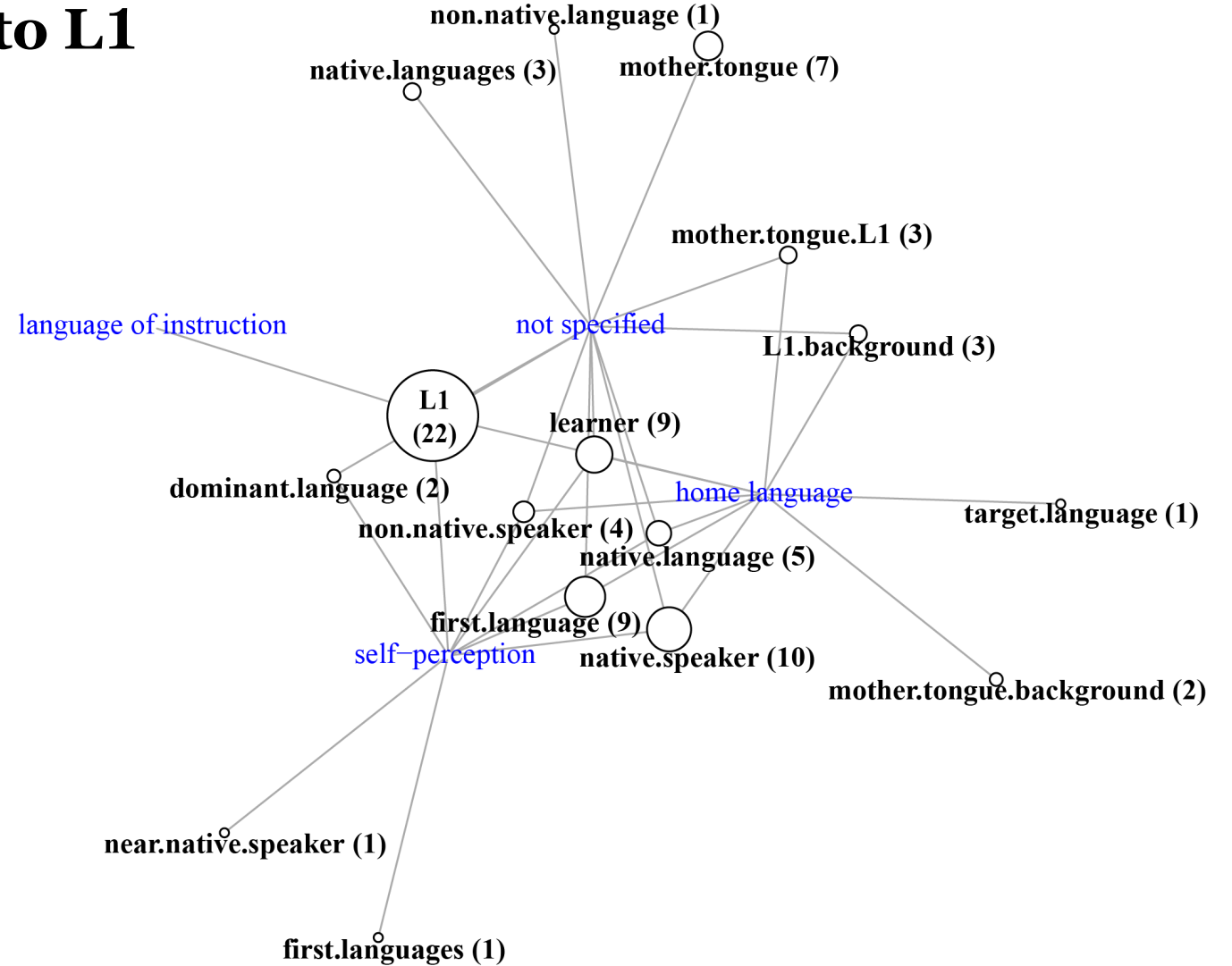


Terms
Definitions

Terminology used to refer to L1

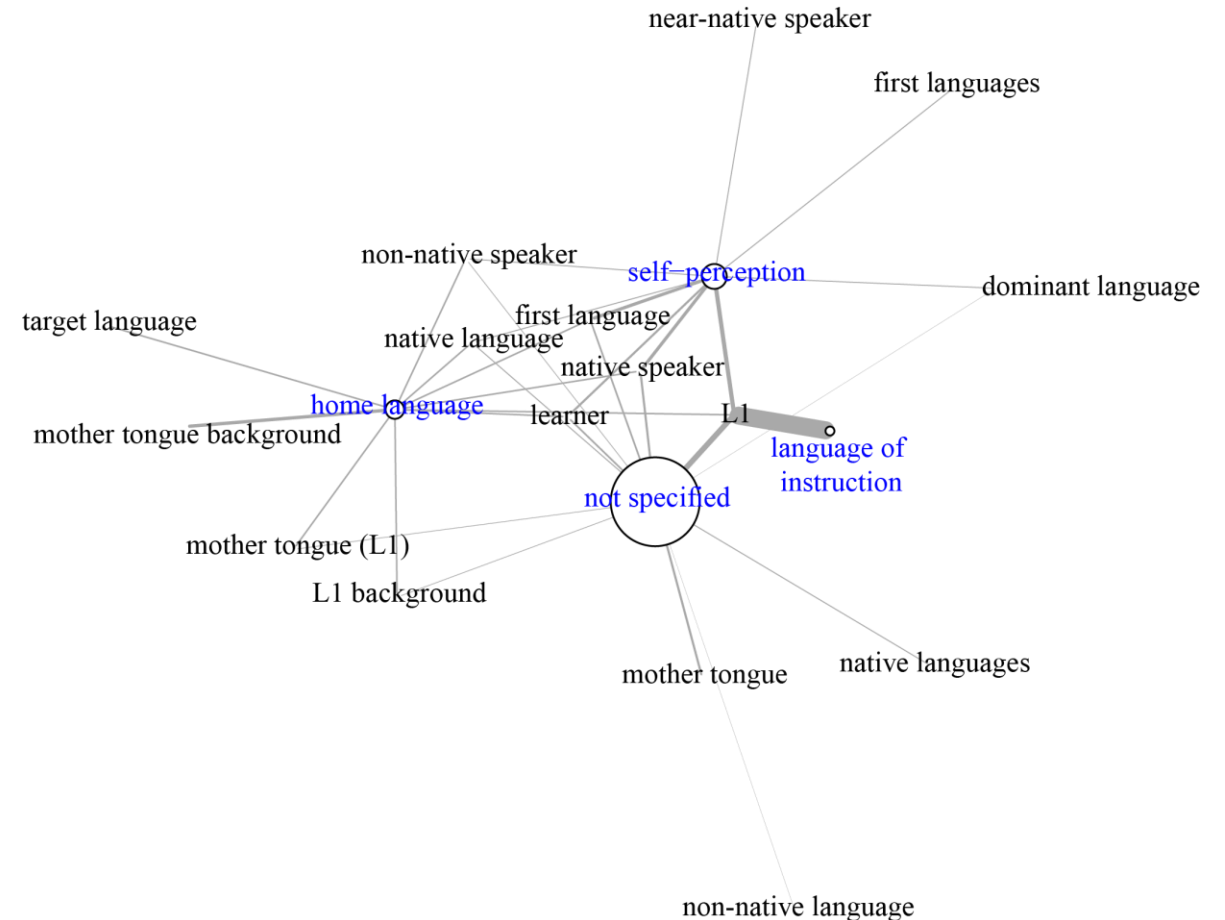
Variants and combinations of:

- L1
- native speaker/language
- mother tongue
- background
- first or dominant language
- learner (e.g., x learner of y)



Definitions given for L1 are based on...

- *Self-perception* (n=6): based on perceived language identity.
- *Home language* (n=4): language(s) spoken at home or in the family setting.
- *Language of instruction* (n=1): language first taught (by order and level) or pre-dominant in a certain school system.
- ??? -> *Undefined* (n=26): no explanation about the metadata is specified.



Observation 1:

Most corpora leave the definition of the *L1* metadata employed in their documentation implicit.

26 of 38 corpora (~70%) were not defined

APPENDIX A

List of Sociological and Task-Related Variables in Longdale (2010 version)

Sociological variables

Participants:

- DOB
- Sex
- Home country
- L1
- Home language
- L2s

Importantly, the Longdale research team strives to maintain comparability with the International Corpus of Learner English (ICLE), a data source that has frequently been used in learner corpus research (see Granger, Dagneux, & Meunier, 2002 and Granger et al., 2009). Thus, like the ICLE corpus, the Longdale database includes detailed learner profile information on learner and task variables (see Appendix A). However, unlike

Meunier & Litte (2013). Tracking Learners' Progress: Adopting a Dual 'Corpus Cum Experimental Data' Approach. *The Modern language journal* (Boulder, Colo.) 97.S1: 61–76.

Observation 2:

Terminology and Definitions of the *L1* metadata depend on researchers' perspectives.

We found many different terms and definitions.

All metadata describing the language background of the students (Table 6) focused on the official languages of the province – German (DE), Italian (IT) and Ladin (LAD) – and offered also a DE-IT bilingual option. Other languages were grouped into one single category (OTHER) to maintain the anonymity of this smaller group. Metadata regarding the **learner's L1 (author_L1)** is based on what authors perceive as their first language(s). To know more about the family languages, the students were also asked to indicate the first language(s) of their mother (**author_mother_L1**) and father (**author_father_L1**). Furthermore, the students were asked to indicate their

Glaznieks et al. (in print, 2024). The Kolipsi Corpus Family: Resources for Learner Corpus Research in Italian and German. *Italian Journal of Computational Linguistics*.

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. Norwegian is the language of schooling, English the first additional language introduced in school and French, German and Spanish second additional languages. However, the students may have learnt other languages elsewhere. 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 (see also Section 5). To avoid the possibility that individuals may be recognized, rare L1s have been collapsed into the category “other”.

Dirdal et al. (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.

Observation 3:

Terms and definitions (implicitly) act as criteria to build categories for analysis.

→ but not only...

NATIVE SPEAKER STATUS		
NATIVE SPEAKER	NS	Native speakers of North American English
NATIVE SPEAKER OTHER	NSO	Native speakers of non-American English
NEAR NATIVE SPEAKER	NRN	Non-native speakers who consider English as their current dominant language and who appear to have native-like fluency and grammatical proficiency.
NON-NATIVE SPEAKER	NNS	Non-native speaker of English other than near-native speakers
FIRST LANGUAGE		
Only shown when first language is other than North American English.		

from other participants. Demographic information (gender, age group, university position, and native language) was collected from each speaker on a form distributed at the end of each event. The speaker information is included in the header of each transcript

Simpson et al. (2002) *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Observation 4:

Very few corpora make the possibility of multiple L1s explicit.

...and representations used to record multiple L1s pose methodological challenges:

L1 = German+Italian

L1s = German

L1_1 = German, L1_2 = Italian

Observation 4:

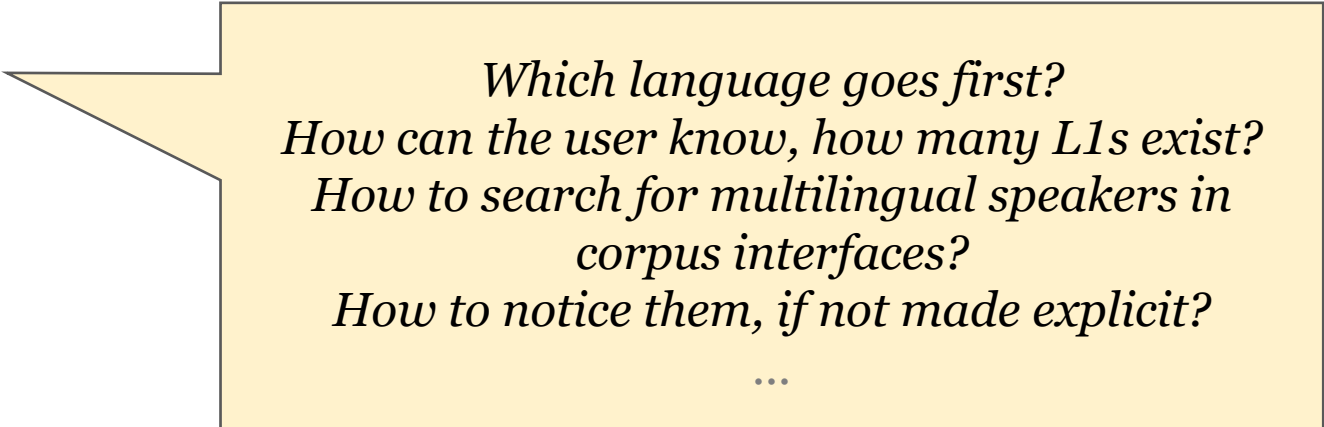
Very few corpora make the possibility of multiple L1s explicit.

...and representations used to record multiple L1s pose methodological challenges:

L1 = German+Italian

L1s = German

L1_1 = German, L1_2 = Italian



*Which language goes first?
How can the user know, how many L1s exist?
How to search for multilingual speakers in
corpus interfaces?
How to notice them, if not made explicit?*

...

not yet been attained. Subjects of six native or L1 languages are represented: Arabic, Mandarin Chinese, French, English, Portuguese and Russian. In its current form the corpus contains a total of over 570,000 words, including data from participants of all levels and L1s. The original data had to be carefully filtered since there were samples of students with a different L1 to those considered, as well as other potential participants whose data were deemed invalid for a variety of reasons (incomplete or unclear tasks, difficulty in certifying level of proficiency, no understanding of the tasks to be done, etc.).⁸ The current CAES version contains samples produced by 1,423 students of Spanish as a foreign language who wrote two or three texts in keeping with their level; this led to a total of 3,881 written tasks integrated in 1,423 samples.

⁸ This was particularly so in the case of the universities since the groups of students were most often multilingual, hence making the control of the L1 variable difficult.

Rojo et al. (2016) Learner Spanish on computer. The CAES 'Corpus de Aprendices de Español' project. In Alonso-Ramos M. (ed.) Spanish Learner Corpus Research: Current trends and future perspectives, 55-87.

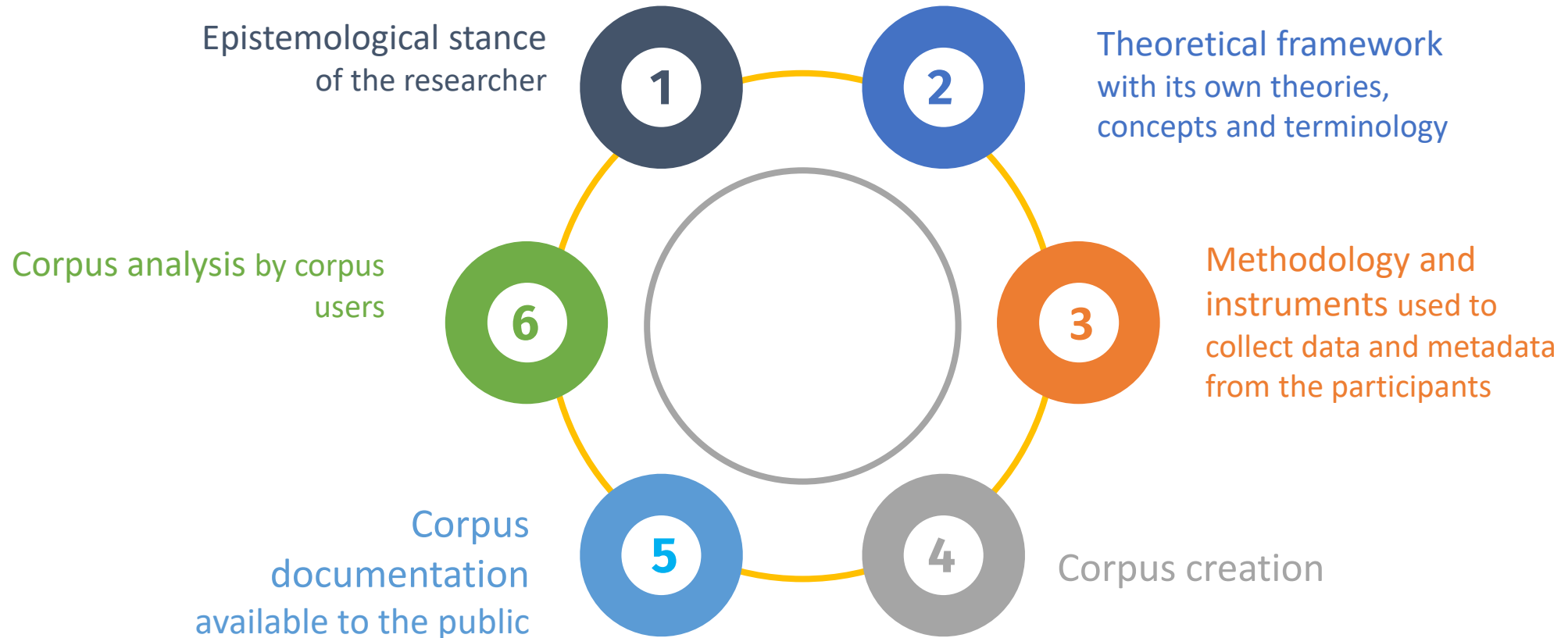
Observation 5:

Many corpora are lacking comprehensive and consistent documentation.

-
- No documentation available for 3 corpora
 - Documentation between corpus descriptions and corpus interfaces differed
 - Not always specified how metadata was elicited (questionnaire, combination of various sources, e.g. proficiency tests, ...)
 - Corpus descriptions not available in English (comparability?)

Summary

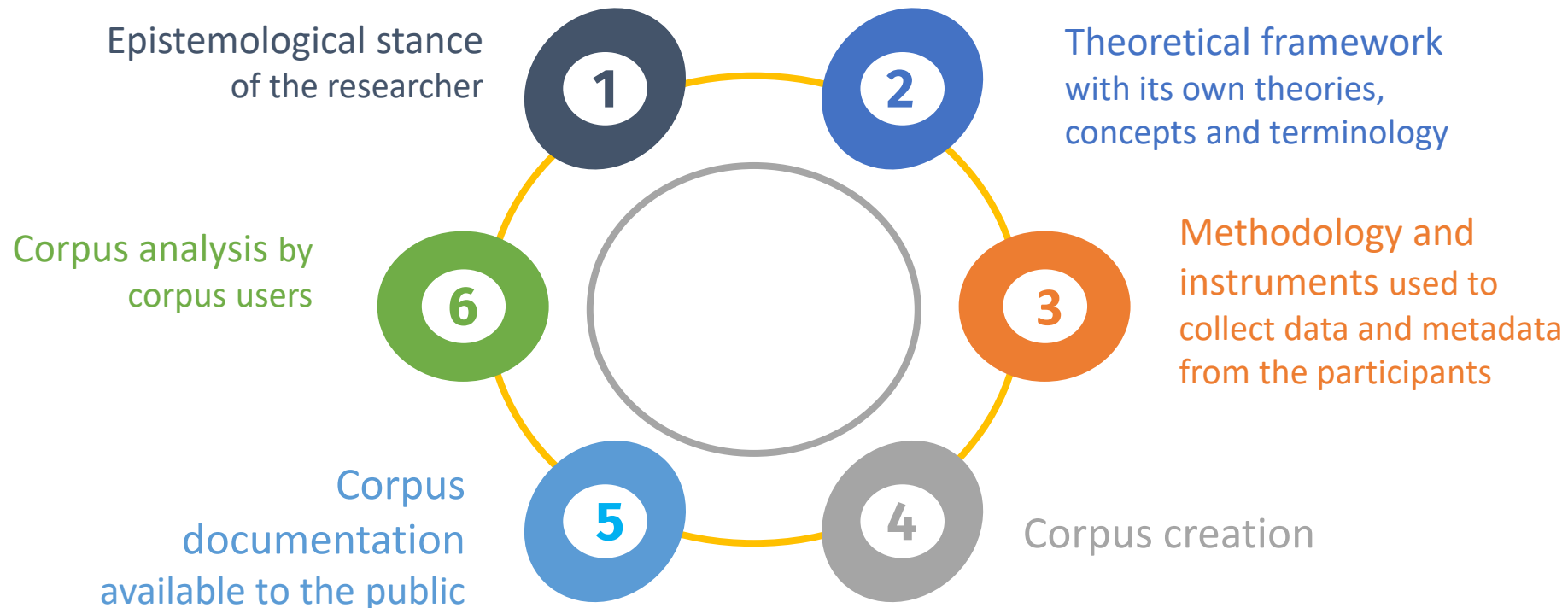
The L1 metadata and its definition and use has a strong impact on all steps of the research cycle in Learner Corpus Research...



How about empirically testing it?

Two empirical studies conducted on South Tyrolean learner corpus data (Leonide) (Glaznieks et al. 2022):

- 1) Simulation of a traditional LCR study to investigate the impact that different categorizations of the speakers' language background can have on study results;
- 2) Integration of a multilingual view on speakers' language background.



Discussion

What does this mean for...

- **(Re-)usability** of learner corpora
- **Replication** of learner corpus studies
- **Follow-up studies** with comparable methodologies
- **Cross-study comparisons** drawn from different corpora
- Focus on language practices of **multilingual speakers**
- **Standardization and interoperability** initiatives such as Core Metadata Schemas

???

Future Outlook

- More detailed learner profiles
- Share metadata elicitation methods (questions)
- Transparency on filtering decisions

Work towards transparency and interoperability in the field:
Community efforts for Core Metadata Schema

Thank you for your
attention!

Olga Lopopolo

olga.lopoplo@eurac.edu

Jennifer-Carmen Frey

jennifer.frey@eurac.edu

References

- Bauman, R., & Briggs C. (2003). *Voices of modernity: Language ideologies and the politics of inequality*. Cambridge: Cambridge University Press.
- 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.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.
- 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.
- Glaznieks, A. Frey, J., Stopfner, M., Zanasi, L. & Nicolas, L. (2022). Leonide. *International Journal of Learner Corpus Research* 8 (1): 97 – 120.
- Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1): 7-24.
- Gumperz, J. & Hymes, D. (eds.) (1972). *Directions in sociolinguistics*. Holt, Rinehart and Winston.
- Hackert, S. (2012). *The Emergence of the English Native Speaker: A Chapter in the Nineteenth-Century Linguistic Thought*. De Gruyter-Mouton: Boston, Berlin.
- May, S. (2012). *Language and minority rights: Ethnicity, nationalism and the politics of language* (2nd. ed.). New York: Routledge/Taylor & Francis.
- Meunier & Litte (2013). Tracking Learners' Progress: Adopting a Dual 'Corpus Cum Experimental Data' Approach. *The Modern language journal* 97 (S1): 61–76.
- Ortega, L. (2014). Ways forward for a bi/multilingual turn in SLA. In S. May (Ed.), *The multilingual turn: Implications for SLA, TESOL, and bilingual education*, 32–53. New York: Routledge/Taylor & Francis.
- Paquot, M., König, A., Stemle, E., & Frey, J.-C. (2023). *Core Metadata Schema for Learner Corpora* [dataset]. Open Data @UCLouvain.
- Pallaver, G. (2014) South Tyrol's changing political system: From dissociative on the road to associate conflict resolution. *The Journal of Nationalism and Ethnicity* 42 (3): 376–398.
- Pennycook, A., & Otsuji, E. (2015). *Metrolingualism: Language in the City* (1st ed.). Routledge.
- Rojo, G. & Palacios, I. M. (2016) Learner Spanish on computer. The CAES 'Corpus de Aprendices de Español' project. In Alonso-Ramos M. (ed.) *Spanish Learner Corpus Research: Current trends and future perspectives*, 55-87.
- Simpson, S. L., Briggs, J. O., & Swales, J. M. . (2002) *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.
- Wisthaler, V. (2015) South Tyrol: The importance of boundaries for immigrant integration. *Journal of Ethnic and Migration Studies* 42 (8): 1271–1289.