

CLARIN Annual Conference Proceedings

2022

Edited by

Tomaž Erjavec, Maria Eskevich

10 – 12 October 2022
Prague, Czechia

Please cite as:
CLARIN Annual Conference Proceedings, 2022. ISSN 2773-2177 (online).
Eds. T. Erjavec and M. Eskevich.
Prague, Czechia, 2022.

Programme Committee

Chair:

- Tomaž Erjavec, Jožef Stefan Institute (SI)

Members:

- Starkaður Barkarson, Árni Magnússon Institute for Icelandic Studies (IS)
- Lars Borin, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Eva Hajičová, Charles University Prague (CZ)
- Marinos Ioannides, Cyprus University of Technology (CY)
- Neeme Kahusk, University of Tartu (EE)
- Krister Lindén, University of Helsinki (FI)
- Monica Monachini, Institute of Computational Linguistics “A. Zampolli” (IT)
- Karlheinz Mörth, Austrian Academy of Sciences (AT)
- Costanza Navarretta, University of Copenhagen (DK)
- Jan Odijk, Utrecht University (NL)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- Kiril Simov, IICT, Bulgarian Academy of Sciences (BG)
- Inguna Skadiņa, University of Latvia (LV)
- Koenraad De Smedt, University of Bergen (NO)
- Marko Tadić, University of Zagreb (HR)
- Jurgita Vaičėnienė, Vytautas Magnus University (LT)
- Vincent Vandeghinste, Instituut voor de Nederlandse Taal (Dutch Language Institute), the Netherlands & KU Leuven (BE)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Andreas Witt, University of Mannheim (DE)
- Friedel Wolff, South African Centre for Digital Language Resources, North-West University (ZA)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Lars Borin, SE
- António Branco, PT
- Tomaž Erjavec, SI
- Eva Hajičová, CZ
- Martin Hennelly, ZA
- Erhard Hinrichs, DE
- Marinos Ioannides, CY
- Nicolas Larrousse, FR
- Krister Lindén, FI
- Monica Monachini, IT
- Karlheinz Mörth, AT
- Costanza Navarretta, DK
- Jan Odijk, NL
- Stelios Piperidis, GR
- Eiríkur Rögnvaldsson, IS
- Kiril Simov, BG
- Inguna Skadiņa, LV
- Koenraad De Smedt, NO
- Marko Tadić, HR
- Jurgita Vaičėnienė, LT
- Tamás Váradi, HU
- Kadri Vider, EE
- Martin Wynne, UK

Subreviewers:

- Olivier Baude, FR
- Federico Boschetti, IT
- Angelo Mario Del Grosso, IT
- Dimitrios Galanis, GR
- Maria Gavriilidou, GR
- Zijian Győző Yang, HU
- Kinga Jelencsik-Mátyus, HU
- Bence Nyéki, HU
- Christophe Parisse, FR
- Valeria Quochi, IT
- Efstathia Soroli, FR
- Thorsten Trippel, DE

CLARIN 2022 submissions, review process and acceptance

- Call for abstracts: 16 December 2021, 21 February 2022
- Submission deadline: 29 April 2022
- In total 21 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 07 and 10 June 2022
- Notifications to authors: 30 June 2022
- 16 accepted submissions

More details on the paper selection procedure and the conference can be found at <https://www.clarin.eu/event/2022/clarin-annual-conference-2022>.

Table of Contents

Language resources and CLARIN centres

<i>ACTER 1.5: Annotated Corpora for Term Extraction Research</i> Ayla Rigouts Terryn, Veronique Hoste and Els Lefever	1
--------------------------------------------------------------------------------------------------------------------------------	---

<i>Linguistic autobiographies. Towards the creation of a multilingual resource family</i> Silvia Calamai, Rosalba Nodari, Claudia Soria and Alessandro Carlucci	5
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---

<i>CLARIN-LV: Many Steps till Operation</i> Inguna Skadiņa, Ilze Auziņa, Roberts Dargis, Eduards Lasmanis and Arnis Voitkāns	9
---------------------------------------------------------------------------------------------------------------------------------------	---

Tools and workflows. Part 1

<i>BabyLemmatizer: A Lemmatizer and POS-tagger for Akkadian</i> Aleksi Sahala, Tero Alstola, Jonathan Valk and Krister Lindén	14
----------------------------------------------------------------------------------------------------------------------------------------	----

<i>WebLicht-Batch – A Web-Based Interface for Batch Processing Large Input with the WebLicht Workflow Engine</i> Claus Zinn and Ben Campbell	19
-------------------------------------------------------------------------------------------------------------------------------------------------------	----

<i>The CLaDA-BG Dictionary Creation System: Specifics and Perspectives</i> Zhivko Angelov, Kiril Simov, Petya Osenova and Zara Kancheva	24
--------------------------------------------------------------------------------------------------------------------------------------------------	----

Tools and workflows. Part 2

<i>A Lightweight NLP Workflow Engine for CLARIN-BE</i> Adriaan Lemmens and Vincent Vandeghinste	29
----------------------------------------------------------------------------------------------------------	----

<i>Natural language processing for literary studies: Graph Literary Exploration Machine (GoLEM)</i> Agnieszka Karlińska, Wiktor Walentynowicz, Maciej Maryl and Jan Wiczorek	35
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----

<i>Supporting Ancient Historical Linguistics and Cultural Studies with EpiLexO</i> Valeria Quochi, Andrea Bellandi, Michele Mallia, Alessandro Tommasi and Cesare Zavattari ...	39
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----

Legal Questions

<i>EU Data Governance Act: New Opportunities and New Challenges for CLARIN</i> Pawel Kamocki and Krister Lindén	44
--------------------------------------------------------------------------------------------------------------------------	----

Curation of Language Resources

<i>CLARIN Depositing Guidelines: State of Affairs and Proposals for Improvement</i> Jakob Lenardič and Darja Fišer	48
-----------------------------------------------------------------------------------------------------------------------------	----

<i>The Resource Publishing Pipeline of the Language Bank of Finland</i> Ute Dieckmann, Mieta Lennes, Jussi Piitulainen, Jyrki Niemi, Erik Axelson, Tommi Jauhiainen and Kristen Lindén	53
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----

<i>TEI and Git in ParlaMint: collaborative development of language resources</i> Tomaž Erjavec and Matyáš Kopp	57
-------------------------------------------------------------------------------------------------------------------------	----

Research cases

<i>Analysing changes in official use of the design concept using SweCLARIN resources</i> Lars Ahrenberg, Daniel Holmer, Stefan Holmlid and Arne Jönsson	61
------------------------------------------------------------------------------------------------------------------------------------------------------------------	----

<i>A Snapshot of Climate Change Arguments: Searching for Recurring Themes in Tweets on Climate Change</i> Maria Skeppstedt and Robin Schaefer	65
------------------------------------------------------------------------------------------------------------------------------------------------------------	----

<i>Linguistic Framing of Political Terror: Distant and Close Readings of the Discourse on Terrorism in the Swedish Parliament 1993–2018</i> Magnus P. Ängsal, Daniel Brodén, Mats Fridlund, Leif-Jöran Olsson and Patrik Öhberg	69
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----

Linguistic Autobiographies. Towards the Creation of a Multilingual Resource Family

Silvia Calamai
Università di Siena, Italy
silvia.calamai@unisi.it

Rosalba Nodari
Università di Siena, Italy
rosalba.nodari@unisi.it

Claudia Soria
CNR-ILC, Italy
claudia.soria@ilc.cnr.it

Alessandro Carlucci
University of Bergen, Norway
alessandro.car-
lucci@uib.no

Abstract

This paper describes a project aimed at adding a new type of corpus to the CLARIN resource family tree, called ‘linguistic autobiographies’. In a linguistic autobiography the writer explicitly reflects on the relationship between him/herself and the language. This genre is fruitfully used in different educational settings, and research has shown that it helps to uncover the social, affective, and psychological dimensions of language learning. The potential of a multilingual collection is discussed starting from Italian data.

1 CLARIN Resource families and linguistic diversity

The CLARIN Resource Family (Fišer et al. 2018) is a user-friendly overview per data type of available language resources in the CLARIN infrastructure aimed at the needs of researchers from digital humanities, social sciences and human language technologies. Resource families are provided according to modality (spoken, multimodal, computer-mediated), genre (historical, academic, literary, newspaper, ...), multilingualism, and intended use (reference, L2 learners). These groupings of corpora, lexical resources and tools are meant to facilitate comparative research: for each resource family, a brief description is provided followed by a list of the resources belonging to the family, together with the most important metadata (name, size, annotation, license, language, description and availability). Thus, resource families provide a curated view of the available CLARIN resources. Over the years, this has proven to be a highly visible initiative appreciated by a broad spectrum of CLARIN users (Leonardič and Fišer 2020) and it therefore deserves to be maintained and enlarged. In this paper, we introduce a new textual genre (linguistic autobiographies) and we argue that a resource family devoted to this genre could be useful for both first and second language research and pedagogy, as well as for a better understanding of the linguistic landscape in European schools and universities, as it is shown in the following section.

2 The underutilised potential of linguistic autobiographies in shaping the European multilingual landscape

2.1 What are linguistic autobiographies

A linguistic autobiography is a non-fictional genre where the writer explicitly reflects on the relationship between him/herself and the language. In this self-reflective writing practice, language becomes the

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>

overarching organising principle for retracing salient moments in the writer's life. The idea behind this genre is that the acquisition and the interaction of different languages can be seen as the acquisition of selfhood (Ramsdell 2004). Linguistic autobiographies can be considered as both research and a pedagogical tool. They are used by professors and teachers in secondary schools and university classrooms to help students in developing their metalinguistic and metapragmatic abilities; within superdiverse multilingual classrooms, linguistic autobiographies allow students to narrate their multilingual selves and they help make their languages more visible. Linguistic autobiographies can also help linguists in gaining access to language ideologies and attitudes towards language varieties: in particular, these narratives can help understand how ideologies about languages can have an impact on the linguistic behaviour of speakers.

2.2 Linguistic autobiographies as a powerful teaching tool

Linguistic autobiographies are highly versatile in that they can easily be collected without requiring specific skills or academic knowledge. Several templates are available to facilitate the production of linguistic autobiographies, which offer some suggestions to think about languages and self-reflect on key points of the writers' own life (cf. Canobbio 2006; D'Agostino 2007; Luppi, Thüne 2022). For example, one possible template requires mentioning:

- 1) Family members and personal data (place of birth, eventual relocations, etc.);
 - 2) Family linguistic background: L1 of the grandparents, L1 of the parents;
 - 3) Family linguistic situation: parents and grandparents' linguistic preferences (they speak which language to whom and when); which languages are used for ordinary communication between family members (with children, with the rest of the family); family choices in linguistic education (which language is taught to children); which language is spoken at home; which other varieties are used at home, and for what need (communicative, expressive, affective needs, identity stances, etc.);
 - 4) Family and school attitudes and behaviours: repression of non-standard varieties; are any specific varieties preferred to other varieties? Are there any disfavoured languages? Are there any forbidden languages used in secret between friends or family members?
 - 5) Meeting with linguistic diversity (holiday trips to different regions, community of practices, peer groups, school environment, extended family etc.). Formal and informal language learning (foreign language schools, friends from abroad etc.); are different languages used with different groups of people? Are non-standard varieties used for performing specific identities? Etc.
 - 6) Personal evaluation of language learning in and out of school; ability to perceive different linguistic varieties, social evaluation of accents, stigma and stereotypes towards accents, varieties, languages, etc.
- The template (and the terminology used) can be adapted depending on the age, educational background and other characteristics of those involved. In any case, linguistic autobiographies are deeply personal texts that outline the writers' own thoughts and feelings, giving them the possibility for spontaneous expression.

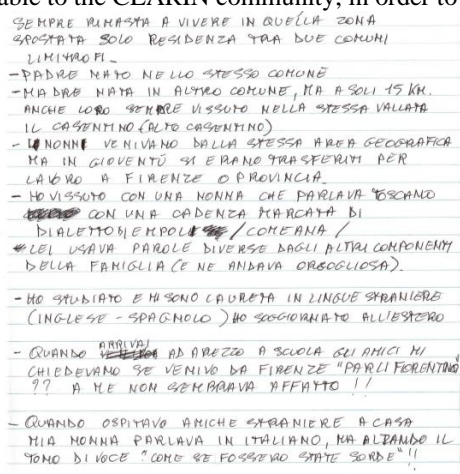
2.3 The societal potential of linguistic autobiographies

Linguistic autobiographies are fruitfully used in different educational settings, and research has shown that this tool helps to uncover the social, affective, and psychological dimensions of language learning (Franceschini, Mieczkowski 2004; Gropaldi 2010; Salvadori, Blondeau, Polimeni 2020). This kind of narrative permits students to develop an awareness of cultural and linguistic diversity, and to learn about the social value of languages. In superdiverse settings, linguistic autobiographies help students in understanding mechanisms of stereotyping and linguistic discrimination and are considered an empowering tool. Teachers can also gain access to the students' linguistic learning process, and they can discover students' learning practices, and reasons for studying languages, as well as their needs and expectations. For policy makers and stakeholders, linguistic autobiographies can help in understanding students' motivation for language learning, thus developing specific school curricula for addressing students' communicative needs. For those who are interested in the sociology of languages, linguistic autobiographies can also provide unique information about informal language-learning opportunities and the different values that speakers attach to different types of multilingualism.

3 Contributing linguistic autobiographies to CLARIN

The VLO repository already offers a selection of linguistic autobiographies collected in two (non-exclusively) Italian-speaking settings, namely Language Biographies from South Tyrol and from Basel. However, these two corpora consist of audio interviews, together with their transcriptions. The aim of this proposal is, instead, to create a new multilingual corpus that comprises several written linguistic autobiographies of both L1 and L2 speakers, collected in different languages and different national settings. This new corpus will offer comparable corpora of the same genre with the appropriate meta-data profile. For example, by searching the corpus according to speakers' L1, it will be possible to compare the biographies of speakers with the same L1 across different countries, educational settings, etc.

Firstly, a selection of linguistic autobiographies collected in Italy and Norway in different educational settings will be deposited in the CLARIN repository. Currently, the Italian corpus comprises almost 200 linguistic autobiographies collected during university linguistic courses, ca. 50 autobiographies of secondary school students and ca. 40 autobiographies of secondary school teachers (see a specimen in Fig. 1). The data will be digitised, anonymised, and the appropriate metadata description will be chosen. The metadata profile used for the Italian corpus will then be applied to the different collections. The anonymisation and metadatation will give the possibility to make the corpus downloadable and accessible under public licences. Together with linguistic autobiographies, a multilingual template will be made available to the CLARIN community, in order to facilitate the collection of linguistic autobiographies.



SEMPRE RIMASTA A VIVERE IN QUELLA ZONA
SPOSTATA SOLO RESIDENZA TRA DUE COMUNI
LIMITATO FL -
- PADRE NATO NELLO STESSO COMUNE
- MIA BABE NATO IN ALTRO COMUNE, MA A SOLI 15 KM.
ANCHE LOBO ~~PER~~ VISSUTO NELLA STESSA VALLATA
IL CASENTINO (ALTO CASENTINO)
- ~~MA~~ NONNE VENIVANO DALLA STESSA AREA GEOGRAFICA
MA IN GIOVENTÙ SI ERANO TRASFERITI PER
LAVORO A FIRENZE O PROVINCIA
- HO VISSUTO CON UNA NONNA CHE PARLAVA TOSCANO
~~PARLAVA~~ CON UNA CADENZA PARLATA DI
BIALLETTO DI EMPOLI / ~~COMERANA~~ /
LEI USAVA PAROLE DIVERSE DAGLI ALTRI COMPONENTI
DELLA FAMIGLIA (E NE ANDAVA ORGOGLIOSA).

- HO STUDIATO E MI SONO LAUREATA IN LINGUE STRANIERE
(INGLESE - SPAGNOLO) HO SOCIANIZATO ALL'ESTERO

- QUANDO ~~PARLAVA~~ AD AREZZO A SCUOLA GLI AMICI MI
CHIEDEVANO SE VENIVO DA FIRENZE "PARLI FIORENTINO"
?? A ME NON SEMBRAVA AFFATTO !!

- QUANDO OSPITAVO ANICHE STRANIERE A CASA
MIA NONNA PARLAVA IN ITALIANO, MA ALZANDO IL
TOMO DI VOCE "COME SE FOSSEVO STATE SORDE" !!

Figure 1. An example of linguistic autobiography from the University of Siena corpus

The creation of a new linguistic corpus in the CLARIN resource family initiative does carry the need for further consideration regarding metadata description. According to Leonardič and Fišer (2020), this is a crucial issue in building resource families. The curated resources tend to have different depths and breadths of metadata description, which in turn has consequences on their final usability. It is of utmost importance, therefore, that linguistic autobiographies are described in such a way that their collection under a resource family is straightforward and allows for their maximum comparability. We believe that a metadata description that is adequate for building resource families must satisfy two interacting main requirements: a) exhaustiveness and b) comparability.

Exhaustive metadata description is obviously important not only for the sake of accuracy, but also in order to maximise the impact on the traceability of linguistic autobiographies and the possibility for them to be discovered and included in future collections. In order to ensure the highest possible degree of comparability with other resources, either already present or to be added in the future, metadata description should also take into account the metadata sets used for describing resources that are partially overlapping for content and/or genre with linguistic autobiographies. It is likely that linguistic autobiographies have features in common with already existing or future resources (such as resources belonging to oral histories).

From a first analysis of the VLO we have identified oral interviews, general autobiographies and personal narratives as the most similar genres already represented in CLARIN collections. Linguistic autobiographies, on the other hand, show some peculiar features such as a) the written modality vs.

mainly oral one and b) their strong emphasis on the linguistic component: language is the key around which the narrative is built and articulated, and the recollection of one's life follows a linguistic path, while in general oral narratives focus on the main events in the lives of speakers.

4 Towards a CLARIN Resource Family for Linguistic Autobiographies

We are confident that in some of the countries involved in the CLARIN network, linguistic autobiographies are already used in school and university settings. With this project, thus, we would like to help uncover this eventually already existing material as well as to encourage the production of new one. A multilingual collection of such written material will indeed offer an invaluable picture from several perspectives. Firstly, it can be used as teaching material, from school classes of any grade to university courses, in order to raise awareness of heritage languages, accentism and glottophobia. Secondly, it can help teachers to better understand the most used and known languages in the classrooms. In addition, it represents a useful tool to verify among pupils, students, and teachers, the pervasiveness of the concept of linguistic error and deviation in describing linguistic repertoires.

Comparable corpora of linguistic autobiographies will also provide valuable quantitative and qualitative data to researchers interested in a variety of topics – such as language attitudes, language and migration, multilingualism and language contact. Finally, this new resource can help policymakers in designing linguistic policies more consistent with the different linguistic landscapes which are truly present in different European schools and universities.

References

- Canobbio, S. 2006. Dialecto dei giovani e politiche linguistiche delle famiglie, appunti dal Piemonte. In: Marcato, G.: *Giovani, lingue e dialetti, Atti del Convegno, Sappada - Plodn, 29 June - 3 July 2005*. Unipress, Padova: 239-244.
- D'Agostino, M. 2007. *Sociolinguistica dell'Italia contemporanea*. Il Mulino, Bologna.
- Fišer, D., Lenardič, J., and T. Erjavec. 2018. CLARIN's Key Resource Families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*: 1320–1325.
- Franceschini, R., and Miecznikowski, J. (eds) 2004. *Leben mit mehreren Sprachen. Vivre avec plusieurs langues. Sprachbiographien. Biographies langagières*. Lang, Bern.
- Groppaldi, A. 2010. L'autobiografia linguistica: strumento per una moderna didattica dell'italiano L2-LS. *Italiano LinguaDue*, 2(1): 89-103.
- Lenardič, J. and D. Fišer. 2020. Extending the CLARIN Resource and Tool Families. In: Navarretta, C. & Eskevich, M.: *Proceedings of CLARIN Annual Conference 2020, 05-07 October 2020, Virtual Edition*: 1–5.
- Luppi, R., and Thüne, E.-M. (eds) 2022. *Biografie linguistiche. Esempi di linguistica applicata*, Centro di Studi Linguistico-Culturali (CeSLiC), Bologna.
- Ramsdell, L. 2004. Language and Identity Politics: The Linguistic Autobiographies of Latinos in the United States. *Journal of Modern Literature* 28(1): 166-176.
- Salvadori, E.; Blondeau, N.; Polimeni, G. (eds.) 2020. La formazione e le competenze di un insegnante riflessivo. *Italiano LinguaDue*, 12(2): 352-389.