Preserving Endangered Linguistic Heritage: Developing a Corpus for the Study of Contact-induced Changes in Corfioto

Giorgio Maria Di Nunzio[1], **Georgios Vardakis**[2]
[1]Department of Information Engineering, University of Padua
Via Gradenigo 6a, 35131, Padova, Italy
[2]Department of Foreign Languages, Translation and Interpreting, Ionian University
Tsirigoti Sq., 49132, Corfu, Greece
giorgiomaria.dinunzio@unipd.it, georgios.vardakis@ionio.gr

Abstract

This paper presents current results of a work-in-progress project on the aims, goals, and methods for compiling a state-of-the-art morphosyntactically annotated corpus of Corfioto, the endangered Balkan Venetan variety of the Corfiot Jews. It gives an outline of the workflow for building, archiving, managing and annotating the first mixed-language corpus of original oral and written data of the Corfiot Jews, based on the Universal Dependencies (UD) framework and introduces the design and the implementation of an application for the Interactive MorPhosyntactic Annotation of Corfioto (IMPACT). The creation and the annotation of the corpus serves three goals: i) attain a quantitative analysis of variation in available data for the analysis of contact-induced syntactic change in clausal complementation in Corfioto; ii) enable the creation of a gold standard and the training of a model for the linguistic annotation of all data in the Universal Dependencies framework; and iii) contribute to the ever-growing research in the development of language resources and tools for endangered and low-resource contact varieties via the collaboration of computational, theoretical and fieldwork linguists.

Keywords: Endangered Languages, Universal Dependencies, Morphosyntactic Annotation, Corfioto

1. Introduction

The intersection of language technology and digital humanities increasingly enables the preservation and analysis of endangered linguistic varieties through the creation of open, interoperable and computationally annotated resources (Ajani et al., 2024). To this end, the present work focuses on Corfioto, the critically endangered Italo-Romance variety historically spoken by the Corfiot Jewish community. Despite recent significant progress made on its general documentation (Mücke, 2019, 2025; Vardakis, 2019, 2021, 2023, 2025), Corfioto remains severely underdocumented within digital linguistic resources. Hence, the complex sociolinguistic context of its emergence is yet to be systematically studied using computational and corpus-based methods (Vardakis and Di Nunzio, 2022).

The development of a morphosyntactically annotated corpus of Corfioto therefore addresses both linguistic and documentary research objectives. From a linguistic point of view, it enables a quantitative exploration of contact-induced variation and syntactic change, providing empirical foundations for studying typological shifts in clausal complementation and other phenomena of contact-induced syntactic change (D'Alessandro, 2021; Andriani et al., 2022). From a cultural and digital humanities standpoint, this initiative seeks to reclaim, document, and disseminate the textual and oral heritage of a community whose language has been historically and sociopolitically marginalized. In this sense, the project aligns with ongoing initiatives in digital heritage preservation, wherein computational infrastructures act as mediators between past linguistic realities and present analytical and ethical needs.

Our research is also motivated by the growing awareness of structural inequalities in NLP performance cross-linguistically around the world (Blasi et al., 2022), which have particularly affected endangered, contact, and low-resource varieties. Thus, by providing a reproducible workflow and an interoperable annotated corpus, we aspire to support equitable resource development.

To operationalize these cultural and linguistic objectives, we present a methodology that combines principles of digital documentation with computational reproducibility. The corpus development follows the Universal Dependencies (UD) framework (Nivre et al., 2016, 2020) which ensures consistency, transparency, and interoperability across linguistic datasets. This choice facilitates not only cross-linguistic comparability, but also compliance with the FAIR principles for scientific data management and stewardship (Wilkinson et al., 2016), which are increasingly seen as the foundational element for sustainable resource design in digital humanities and language technologies.

The development of the Interactive MorPhosyntactic Annotation of Corfioto (IMPACT) interface integrates fieldwork data, computational modeling, and linguistic analysis. This combination is designed to support both expert and community-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 985–996
11–16 May 2026.

based annotation, thereby reinforcing inclusivity and accessibility in corpus creation. To promote reproducibility and collaborative annotation, the interface incorporates best practices from comparable initiatives targeting dialectal and diachronic variation (Çelikkol et al., 2024).

Finally, the project shares broader efforts made within language sciences and the digital humanities to promote data sharing and re-use through standardized representations and documentation (Forkel et al., 2018). The corpus and interface described here aim to serve as both a linguistic resource and a model of interdisciplinary collaboration, bridging computational linguistics, fieldwork, and cultural preservation.

This paper outlines the corpus design and annotation workflow, describes the implementation of IMPACT as a collaborative interface, and discusses its implications for the development of language resources for endangered and contact varieties. More specifically, we address two coupled objectives:

- First, we support a corpus-based analysis of contact-induced change in clausal complementation in Corfioto.
- Second, we aim to produce a UD-compliant gold standard that can seed automatic annotation and model training for the broader corpus.

The current annotated release is intentionally construction-driven: the 117 gold sentences were selected to maximize coverage of *ke*-introduced complements, which introduces a sampling bias with respect to general morphosyntactic distributions. We make this bias explicit and treat the present dataset as a seed resource; subsequent releases will expand the gold standard with more diverse structures, speakers, and text types to support both linguistic analysis and robust parser development.

The paper is organized as follows: Section 2 provides the theoretical and methodological background for the study, outlining the role of corpus-based approaches in research on contact-induced linguistic change. Section 3 introduces Corfioto, the endangered Italo-Romance variety of the Corfiot Jews, setting the historical and sociolinguistic context of language contact among Romance, Greek, and Hebrew within the language communities of the Corfiot Jews, addressing the 'hybrid' structure of its clausal complementation system, which serves as the focus of the present corpus study. Section 4 presents the central linguistic phenomenon investigated in the project, namely the retreat of the Romance infinitive and its replacement by finite clausal complements introduced by the particle *ke*. In Section 5, we present the methodology adopted for the creation and annotation of the Corfioto corpus, describing the data collection, preprocessing, and annotation workflow based on the Universal Dependencies framework. In Section 6, we present the Interactive MorPhosyntactic Annotation of Corfioto (IMPACT) system, describing its architecture, functionalities, and implementation, together with the structure of the dataset and annotation rules that guide the creation of the first gold-standard corpus for Corfioto and related varieties.

2. Contact Induced Linguistic Change

Contact-induced linguistic change (Thomason and Kaufman, 1988; Seifart, 2019) is a central aspect of human languages ranging from phonological, morphological, syntactic, semantic and pragmatic change to the emergence of hybrid grammars (Aboh, 2015). Contrary to traditional approaches to language contact, focusing on the different range of mechanisms and outcomes of contact including pidgins, creoles, mixed languages and new varieties, universalist approaches to the study of language contact treat recombination of linguistic features in settings where speakers are exposed to different varieties, registers and repertoires in different ecologies (Mufwene, 2001) as the norm (Aboh, 2015, 2019, 2020). The quantitative and qualitative power of corpus linguistic methods has been crucial in our understanding of central issues in language contact, including code-switching in different contact settings in top-down and bottom-down approaches (Adamou, 2019; Bullock et al., 2020). While nonstandard, minority and endangered languages (Himmelmann, 1998, 2006; Woodbury, 2011; Nevins, 2022; Adamou, 2024) have maintained a central position in contact linguistics since its foundation as a field (Thomason and Kaufman, 1988; Aikhenvald, 2020; Adamou, 2021), at the time of the most accelerating decline of linguistic diversity in the history of the world's languages (Kik et al., 2021; Adamou, 2024), research in developing and deploying computational tools and methods for resource creation and mobilization in language documentation (Berez-Kroeker et al., 2023) via the interdisciplinary collaboration of NLP researchers, computational, formal, and fieldwork linguists (Anastasopoulos et al., 2020; Galliot et al., 2022; Moeller et al., 2026) is an ever-growing domain of inquiry.

Despite the growing literature on the creating workflows for managing, building and preparing minority, endangered and under-resourced language corpora (DARIAH working groups[1]) and the annotation of UD and SUD treebanks for widely spoken corpora (Kahane et al., 2021; Dobrovolc, 2022),

[1] https://www.dariah.eu/activities/working-groups-list/

documentation of languages typically referred to as "contact languages" remains limited (Buzato, 2023). In this direction, the creation of the first morphosyntactically annotated corpus of written and oral data of Corfioto will offer linguists access to a previously undocumented variety within a FAIR data approach, and will enable statistically significant results on the contact-induced effect of clausal complementation based on the annotation of the whole corpus, via automatic annotation, on the basis of a gold standard. In fact, ensuring interoperability and long-term accessibility of linguistic resources is increasingly framed within the FAIR data principles, which promote the creation of resources that are findable, accessible, interoperable and reusable. Recent work has explored how these principles can be applied to terminology and linguistic resources within large research infrastructures (Di Nunzio et al., 2024; Di Nunzio and Vezzani, 2025).

3. Corfioto, a critically endangered variety

(Italián) Corfióto /korf'joto/, Corfiot Italian, Talián and Italiká (Mücke, 2019; Vardakis, 2025) are glossonyms describing the Romance variety historically spoken by the Corfiot Jews in Corfu Town, Greece. In his most recent assessment of the vitality (Lee and Van Way, 2018) of Corfioto in Corfu and in Israel, Vardakis characterizes Corfioto a critically endangered language (Vardakis, 2025). Today, Corfioto is still spoken by the generation of the direct descendants of the Corfiot Jews who were born and raised in Corfu or who were born to Corfiot Jew parents during the first half of the 20th century and today live in Corfu, in other areas of Greece, and diaspora, mainly Israel, as well as in Italy, Switzerland and South America.

Contemporary Corfioto is characterized by a large, though not neat, divergence between its lexical, morphological and morphosyntactic elements, most of which are inherited from Venetan, on the one hand, and some of its syntactic features, which characterize the diachrony of Greek and extreme southern Italian dialects/varieties (ESIDs). Although linguistic descriptions of certain features based on oral data have been presented in studies of the last 20 years (Nachtmann, 2001; Mücke, 2019, 2025; Vardakis, 2021, 2023, 2025), remarks on features of oral varieties of the Corfiot Jews which indicate divergence from the diachrony of the Romance languages spoken by the Corfiot Jews and their ancestors originating in the Spanish and the Italian peninsula date back to the early 20th century (Belleli, 1905; Gottheil and Belleli, 1904). The co-presence of different languages, including the local (Modern Greek) Heptanesian (Krimpas, 2021) variety of Corfu and different Italo-Romance varieties spoken by different components of the Jewish community, is largely supported by philological, historical, sociolinguistic and linguistic evidence. Based on the comparison of certain morphological and morphosyntactic properties, Vardakis classifies Corfioto as the indigenous Venetan Balkan variety of Corfu, whose emergence points to the contact of Romance, Greek and Hebrew during a period starting with the Venetian domination of Corfu (1386–1797) and reaching the early 20th century (Vardakis, 2025).

4. The Linguistic Phenomenon: Clausal Complementation

The recent description and analysis of the morphosyntactic and syntactic properties of Corfioto (Vardakis, 2025) have contributed majorly to the analysis of the most well-attested linguistic phenomenon in all previous descriptive studies of Corfioto: the retreat and loss of the Romance infinitive in clausal complementation and its replacement by morphologically finite constructions introduced by the particle *ke* (Gottheil and Belleli, 1904; Belleli, 1905; Cortelazzo, 1948; Levi, 1961; Nachtmann, 2001; Mücke, 2019). Unlike Venetan (1), from which Corfioto has inherited most of its lexical and morphological elements, clausal complementation in contemporary Corfioto lacks infinitives. In contemporary Corfioto, all different types of clausal complements are introduced by a single homophonous subordinator *ke*, irrespective of the semantic properties of the matrix verb or other morphosyntactic properties of the *ke*-finite complement, while the complement predicate is encoded via an inflected verb form indexing person and number morphologically, either in subject co-reference or disjoint reference with the matrix, as shown in (2).

(1) te= voj-o parl-ar
2SG want-1SG talk-INF
'I want to talk to you'

(2) vój-o ke= te= párl-o
want-1SG PRT= 2SG= talk-1SG
'I want to talk to you'

Although our corpus shows a robust dominant presence of finite complementation constructions across informants, limited cases of infinitive complement clauses are attested in spontaneous speech data, as illustrated in (3) and (4), and need therefore to be compared with similar data in other corpora, suggesting infinitive reduction but no clear loss (Mücke, 2019; Vardakis, 2023).

(3) no bizǫ́n-a parl-ár italián
NEG need-3SG speak-INF Italian
'we must not speak Italian'

(4) le tsinkue so ke
DET.PL.F five know-1SG PRT
skriv-er
write-INF
'I can write in these five (languages)'

Interestingly, oral data show an alternation between infinitival and finite complements of the same matrix verb, even in intrasentential contexts, as shown in (5).

(5) bizǫ́ŋ-a liɣ-ár,
need.3SG tie-INF
'It is necessary to tie, for them to tie him.'
ex

ke lo lígh-ano
PRT 3SG.M tie-3PL

'It is necessary to tie, for them to tie him.'

The creation of a mixed-data corpus will enable a quantitative approach to variation between finite and infinitival complementation, revealing the environments which are more prone to the retreat or the retention of the phenomenon.

5. Methodology

This section presents a structured methodology on the progress made in building and annotating the Corfioto corpus, as well as directions for the future development of the project. The methodology as conceived in its current form means to:

- integrate the different tasks required for the annotation of morphosyntactic features in any language variety;
- facilitate the reproducibility of the experiments; and
- include an in-depth linguistic analysis to furthering our understanding of contact-induced syntactic change in Corfioto.

The creation of the corpus relies on collaboration between computational, theoretical and fieldwork linguists in order to combine linguistic analysis with computational infrastructure for resource creation. Similar interdisciplinary approaches have been successfully applied in previous projects aiming at documenting minority language varieties through curated digital corpora (Agosti et al., 2012, 2016, 2010, 2011). At the time of writing, the gold-standard annotation comprises 117 sentences drawn from two transcribed oral interviews, focusing on clausal complementation contexts. This first release is intended as a seed dataset for guideline refinement and for bootstrapping model-based pre-annotation; the next steps are: i) to enlarge the gold portion with additional interviews and spontaneous speech segments; ii) to broaden coverage beyond complementation to improve representativeness; and iii) to iteratively retrain and evaluate automatic pre-annotation as the gold standard grows.

We conceptualize the workflow in four macro-phases (building, managing, annotating, archiving). These phases are operationalized through the seven concrete tasks listed below, which specify the data flow from collection to gold-standard creation and model-based pre-annotation. Reproducibility is ensured through explicit annotation guidelines, stable sentence identifiers linked to source recordings, and UD-compatible exports of each release of the gold data. The methodology of the management and analysis of the linguistic data collected with the interviews is the following:

1. Data collection and compilation: gather a substantial dataset of texts, representing various text types and registers. Ensure that the dataset covers a wide range of syntactic structures and linguistic features.
2. Preprocessing: Clean and preprocess the dataset, including tokenization, sentence segmentation, and lowercasing. Ensure that the text is in a format suitable for parsing.
3. Election of a Universal Dependency (UD) parser: Choose a UD parser that supports Corfioto or a closely related language or family, e.g., Italo-Romance.
4. Annotation guidelines: Develop annotation guidelines specific to Corfioto's morphosyntactic features. Define dependency relations, part-of-speech tags, and other linguistic categories relevant to the language.
5. Create a gold standard: Annotate a representative portion of the Corfioto dataset manually to create a gold standard. This annotated data will serve as the basis for training and evaluation.
6. Train, validate, and evaluate model: Train the selected UD parser using the annotated Corfioto data. Fine-tune the model to ensure it captures the language's unique morphosyntactic features.
7. Documentation and sharing: Document the integration process, guidelines and the trained model. Make the parser accessible to other researchers interested in analyzing Corfiot or similar languages. Engage with the linguistics and NLP research community, sharing insights, best practices, and resources.

The ongoing work has been founded on steps 1-5 and partially step 7 for sharing the annotated

dataset produced during the gold standard annotation.

The methodological framework lays the foundation for the development of the Interactive MorPhosyntactic Annotation of Corfioto (IMPACT) system, designed to operationalize the corpus workflow through a dedicated environment for data visualization, collaborative annotation, and quality control.

In the following sections, we present the architecture and functionalities of the IMPACT system, describe the composition and structure of the collected dataset, and discuss the first outcomes of the ongoing annotation process, which illustrate both the challenges and the potential of applying computational methods to an endangered, contact-induced variety.

6. IMPACT Annotation System

The Interactive MorPhosyntactic Annotation of Corfioto (IMPACT) system is an integrated Web environment which has been designed and implemented to allow for on-the-fly adjustments, enabling users to fine-tune annotation results or modify display preferences. Its conception and use enables experts to explore dependency trees, part-of-speech tagging, and syntactic relationships in an intuitive and interactive manner. This Web application, that can also be used in a stand-alone mode, will serve as a valuable tool for linguistic analysis and facilitate research on language contact and change, particularly in the context of Corfioto or similar languages. The source code of the Web application is constantly updated and made available to the research community. The development of this platform has been part of a wider initiative for the design and implementation of interactive platforms for linguistic and philological research has recently attracted growing attention in the digital humanities community, where dedicated tools are designed to integrate corpus management, annotation and visualization within a single environment (Milazzo and Di Nunzio, 2024, 2025).

IMPACT is an application designed and implemented using the R shiny (Chang et al., 2025) framework for Web Interactivity, the datatable (DT) package (Xie et al., 2025) to review the annotation, and the textplot package (Wijffels, 2022) for text visualization and analysis. IMPACT is an integrated Web environment which has been designed and implemented to allow for on-the-fly adjustments, enabling experts to fine-tune annotation results or modify display preferences.

Figure 1 shows a screenshot of the actual implementation of the IMPACT system. The user can:

1. select the sentence to analyze, then
2. transliterate and translate the original sentence;
3. parse or update the text and the plot;
4. interact (edit and modify) with the morphosyntactic annotations produced by the dependency parser, and
5. analyze the corresponding dependency parser tree.

The steps for the implementation of the annotation process focus on integrating a Universal Dependency (UD) parser for the tagging of morphosyntactic features of Corfioto, or similar varieties, to keep this resource up to date with the current state-of-the-art in linguistic annotation, make annotation guidelines available as language change occurs, and make this resource available for training a syntactic parser with new data.

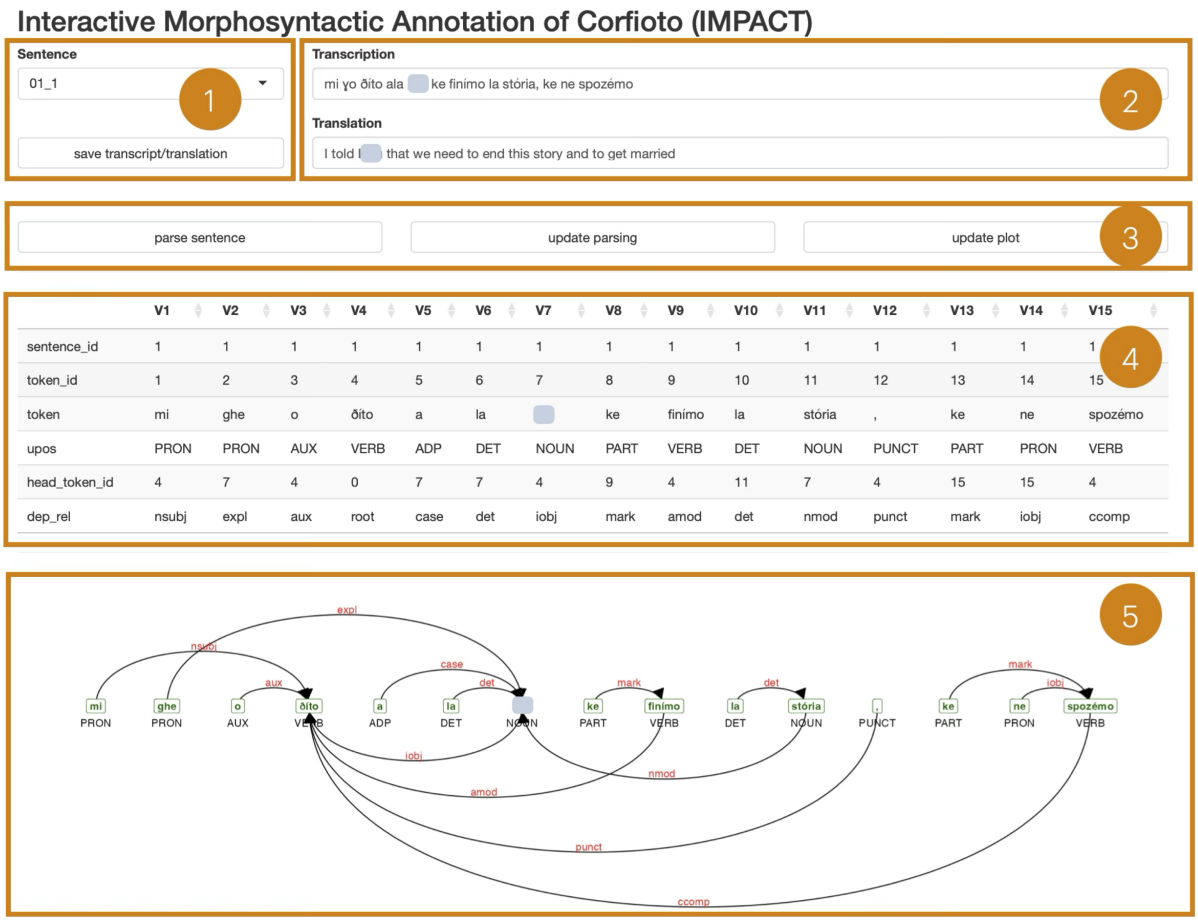
Although the application has been created to serve as a valuable tool for linguistic analysis of Corfioto, it aspires to become a valuable tool for automatic morphosyntactic annotation, especially of genetically affiliated languages. The work in progress made here aims at contributing to the creation of a first gold standard of the variety, as has already been done for other minority and under-resourced languages see (Anastasopoulos et al., 2020; Karahóga et al., 2022; Millour et al., 2024).

6.1. Dataset

The output of the workflow aims to create the first corpus of mixed oral and written data of Corfioto which is based on a) the transcription of more than 30 hours of i) original oral linguistic data (produced via free, semi-structured methods or elicitation) collected in fieldwork with the collaboration of at least 10 consultants in Greece, Israel and Italy from 2019 to date and ii) audio and video recordings dated to the previous century found in institutional and personal archives and offered to us by the informants and; b) written glossaries compiled by speakers or groups of speakers, which contain words and short phrases in Latin, Greek and Hebrew script together with translation equivalents in Italian, French, Greek and Hebrew.

6.2. Annotation Rules

The annotation of the golden standard today is limited to a total of 117 sentences coming from transcribed oral data from two distinct interviews. In most cases, the sentences are segmented at the prosodic level whereas syntactic and semantic criteria are taken into consideration for the segmentation of sentences that are too long. All sentences include clausal complements of the types described



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
sentence_id	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
token_id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
token	mi	ghe	o	ðíto	a	la		ke	finímo	la	stória	,	ke	ne	spozémo
upos	PRON	PRON	AUX	VERB	ADP	DET	NOUN	PART	VERB	DET	NOUN	PUNCT	PART	PRON	VERB
head_token_id	4	7	4	0	7	7	4	9	4	11	7	4	15	15	4
dep_rel	nsubj	expl	aux	root	case	det	iobj	mark	amod	det	nmod	punct	mark	iobj	ccomp

Figure 1: An example of the use of the IMPACT system for the sentence (translated in English) "I told [X] that we need to end the story and to get married". The small gray squares are used to anonymize the names of the interviewees.

in this thesis, which are marked by the subordinator/modal marker *ke*.

Prior to annotation, we performed some basic preprocessing of the data, which includes lowercase for all tokens in the corpus but for the initial letter of proper names, which is capitalized. A first automatic annotation is performed via the use of an existing UD treebank used from genetically affiliated language, preferably a Romance language. At this stage,automatic preannotation has been performed using the UD_Italian-VIT corpus.[2]

After the automatic tokenization/annotation, we performed manual checking and corrections according to the following rules:

- Since no prior spelling conventions of the variety exist, tokenization of the dataset is based on 'syntactic words', which are defined at a language-specific level.
- Tokenization includes separation of clitic particles and argumental clitics from the verb host.[3]
- Morphological contractions are subject to further segmentation; for instance, the portmanteau form *al* (PREP.DEF.M) is segmented as *a* (PREP) and *el* (DEF.M).
- Syllabic accent is kept on words with more than two syllables.
- Proper nouns are represented only using the initial letter to protect anonymity of the participants.
- Punctuation is kept only whenever relevant at the prosodic and/or syntactic and semantic level e.g., interrogative, exclamative or subordinate clauses.

[2] https://github.com/UniversalDependencies/UD_Italian-VIT

[3] the clause 'that you give it to him' (ke= ti= ghe= lo= dágha', (PREP= SCL.2SG= IO.3SG= OCL.3SG.M= GIVE.SUBJ.2/3SG), is manually separated into five words.

6.3. Interoperability and I/O Specifications

IMPACT is designed around UD-style annotation and supports an annotation loop that combines automatic pre-annotation with manual correction. Input data are managed at sentence level and include the original transcription, optional transliteration and translation fields, and metadata linking each sentence to its recording and speaker context. The system supports token-level editing to enforce project-specific segmentation decisions (e.g., separation of clitics from verbal hosts where required). The primary export is a UD-compatible representation that can be used to archive the gold standard and to train or evaluate UD parsers; the same export can also be used for downstream quantitative analyses of morphosyntactic variation. This ensures that annotation produced within IMPACT remains portable across UD toolchains and comparable to other UD resources.

6.4. Comparison and Justification

Several annotation environments currently support morphosyntactic analysis and dependency parsing within the Universal Dependencies (UD) framework. Widely used platforms include WebAnno (Eckart de Castilho et al., 2014), INCEpTION (Klie et al., 2018), ArboratorGrew (Guibon et al., 2020), and UD Annotatrix (Tyers et al., 2017), which provide general-purpose interfaces for part-of-speech tagging, dependency editing, and visualization. Other tools, such as ELAN (Wittenburg et al., 2025) and FLEx (SIL International, 2026), are commonly adopted in documentary and descriptive linguistics for transcription, interlinear glossing, and lexical management. While these tools have proven effective for resource-rich languages or for phonological and lexical annotation, their direct applicability to endangered or contact varieties remains limited.

Most existing solutions rely on pre-trained models and standardized orthographies, which are typically unavailable for languages such as Corfioto. Their data models are not optimized for hybrid or mixed-language corpora, nor do they easily support project-specific tokenization decisions that are critical in corpora with non-standardized orthographies and frequent cliticization, where segmentation choices must be applied consistently across annotators and releases. Furthermore, the separation between data collection, annotation, and visualization tools hinders interoperability and reproducibility, particularly in interdisciplinary research that combines computational and fieldwork-based methods.

The IMPACT system was designed to address these limitations by providing a single integrated environment for corpus management, morphosyntactic annotation, and dependency visualization. Its Web-based interface supports real-time editing and adjustment of tokenization, tagging, and tree structures, enabling users to refine the parser output interactively. Implementation in the R Shiny framework ensures transparency, reproducibility, and compatibility with quantitative analysis tools, while the modular architecture allows adaptation to other low-resource or contact languages. Similar approaches have been adopted in the development of specialized digital linguistic resources, where computational infrastructures support the management and exploration of complex linguistic datasets (Vezzani and Di Nunzio, 2019, 2020a,b; Vezzani et al., 2018).

Developing a new system, rather than relying on existing off-the-shelf platforms, was necessary to ensure flexibility and compatibility with the linguistic properties of Corfioto. The language presents non-standard orthography, complex cliticization, and high intraspeaker variation, which require customized annotation workflows and tokenization procedures. Existing systems do not adequately support these characteristics within a UD-compliant framework. IMPACT therefore offers a tailored solution that accommodates linguistic variability and enables collaborative refinement of annotations and hypotheses directly within the platform.

7. Conclusion

The creation of a morphosyntactically annotated corpus of Corfioto serves multiple goals. Besides the urgent need to preserve the rich linguistic heritage of the Corfiot Jews, it will enable a quantitative approach to the analysis of contact-induced phenomena, including infinitival loss and the emergence of Balkan-type complementation. Finally, it will contribute to current goals on the wider representation of endangered and low-resource languages within the fields of computational linguistics, highlighting the joint opportunities that the collaboration between linguists, fieldworkers and computer scientists can bring.

In this study, we presented a methodology for the creation of a morphosyntactically annotated corpus of Corfioto via IMPACT, designed to facilitate the analysis of variation and change in clausal complementation in Corfioto. The steps for the implementation of the annotation process focus on integrating a Universal Dependency (UD) parser for the tagging of morphosyntactic features of Corfioto, or similar varieties, to keep this resource up-to-date with the current state-of-the-art in linguistic annotation, make annotation guidelines available as language change occurs, and make this resource available

for training syntactic parser with new data.[4]

Future work will focus on expanding the annotated dataset and refining the annotation guidelines for Corfioto, with particular attention to clausal complementation and variation in finite versus infinitival constructions. The next phase will also involve training a dedicated UD parser on the emerging gold-standard data and integrating semiautomatic validation modules to enhance consistency. These developments will consolidate IMPACT as both a research tool and a scalable model for the annotation of endangered and contact-induced language varieties.

8. Bibliographical References

Enoch Aboh. 2020. Lessons From Neuro-(a)-Typical Brains: Universal Multilingualism, Code-Mixing, Recombination, and Executive Functions. *Frontiers in Psychology*, 11(488).

Enoch O. Aboh. 2019. Our creolized tongues. In Edit Doron, Malka Rappaport Hovav, Yael Reshef, and Moshe Taube, editors, *Language contact, continuity and change in the genesis of Modern Hebrew*. Benjamins, Amsterdam.

Enoch Oladé Aboh. 2015. *The Emergence of Hybrid Grammars: Language Contact and Change*. Cambridge Approaches to Language Contact. Cambridge University Press, Cambridge.

Evangelia Adamou. 2019. 52. Corpus linguistic methods. In Jeroen Darquennes, Joseph Salmons, and Wim Vandenbussche, editors, *Language Contact. An international handbook. Volume 1*, pages 638–653. De Gruyter Mouton.

Evangelia Adamou. 2021. *The adaptive bilingual mind*. Cambridge University Press, Cambridge.

Evangelia Adamou. 2024. *Endangered Languages*. MIT Press, Cambridge MA.

Maristella Agosti, Birgit Alber, Giorgio Maria Di Nunzio, Marco Dussin, Diego Pescarini, Stefan Rabanus, and Alessandra Tomaselli. 2011. A Digital Library of Grammatical Resources for European Dialects. In *Digital Libraries and Archives - 7th Italian Research Conference, IRCDL 2011, Pisa, Italy, January 20-21, 2011. Revised Papers*, pages 61–74.

[4]The data and the source code of the IMPACT system will be available online at the following link https://github.com/gmdn/LREC2026/tree/main/IMPACT

Maristella Agosti, Birgit Alber, Giorgio Maria Di Nunzio, Marco Dussin, Stefan Rabanus, and Alessandra Tomaselli. 2012. A Curated Database for Linguistic Research: The Test Case of Cimbrian Varieties. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2230–2236, Istanbul, Turkey. European Language Resources Association (ELRA).

Maristella Agosti, Paola Benincà, Giorgio Maria Di Nunzio, Riccardo Miotto, and Diego Pescarini. 2010. A Digital Library Effort to Support the Building of Grammatical Resources for Italian Dialects. In *Digital Libraries - 6th Italian Research Conference, IRCDL 2010, Padua, Italy, January 28-29, 2010. Revised Selected Papers*, pages 89–100.

Maristella Agosti, Emanuele Di Buccio, Giorgio Maria Di Nunzio, Cecilia Poletto, and Esther Rinke. 2016. Designing A Long Lasting Linguistic Project: The Case Study of ASIt. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4479–4483, Portorož, Slovenia. European Language Resources Association (ELRA).

Alexandra Aikhenvald. 2020. Language contact and endangered languages. In Anthony Grant, editor, *The Oxford Handbook of Language Contact*. Oxford University Press.

Yusuf Ayodeji Ajani, Bolaji David Oladokun, Shuaib Agboola Olarongbe, Margaret Nkechi Amaechi, Nafisa Rabiu, and Musediq Tunji Bashorun. 2024. Revitalizing Indigenous Knowledge Systems via Digital Media Technologies for Sustainability of Indigenous Languages. *Preservation, Digital Technology & Culture*, 53(1):35–44.

Antonios Anastasopoulos, Christopher Cox, Hilaria Cruz, and Graham Neubig. 2020. Endangered Languages meet Modern NLP. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 39–45, Barcelona, Spain (online).

Luigi Andriani, Roberta D'Alessandro, Alberto Frasson, Brechje van Osch, Luana Sorgini, and Silvia Terenghi. 2022. Adding the microdimension to the study of language change in contact. Three case studies. *Glossa: a journal of general linguistics*, 7(1).

Lazarus Belleli. 1905. *Greek and Italian Dialects as Spoken by the Jews in Some Places of the Balkan Peninsula*. London.

Andrea Berez-Kroeker, Shirley Gabber, and Aliya Slayton. 2023. Recent Advances in Technologies for Resource Creation and Mobilization in

Language Documentation. *Annual Review of Linguistics*, 9:195–214.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Barbara Bullock, Almeida Jacqueline Toribio, Jacqueline Serigos, and Gualberto Guzmán. 2020. Processing Multilingual Data. In Yaron Matras and Evangelia Adamou, editors, *The Routledge Handbook of Language Contact*, pages 7–27. Taylor & Francis, Oxon & New York.

Dalmo Buzato. 2023. Universal Dependencies and Language Contact Annotation: Experience from Warao refugees signs in Brazil. In *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 509–519, Belo Horizonte, Brazil. Association for Computational Linguistics.

Melis Çelikkol, Lydia Körber, and Wei Zhao. 2024. Exploring Diachronic and Diatopic Changes in Dialect Continua: Tasks, Datasets and Challenges. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 12–22, Bangkok, Thailand. Association for Computational Linguistics.

Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2025. *shiny: Web Application Framework for R*. R package version 1.11.1.

Manlio Cortelazzo. 1948. Caratteristiche dell'italiano parlato a Corfù. *Lingua Nostra*, 9:29–34.

Roberta D'Alessandro. 2021. Syntactic Change in Contact: Romance. *Annual Review of Linguistics*, (7):309–28.

Giorgio Maria Di Nunzio, Eszter Papp, Federica Vezzani, and Ellie Kemp. 2024. Fair terminology meets clear global. In *Linking Theory and Practice of Digital Libraries: 28th International Conference on Theory and Practice of Digital Libraries, TPDL 2024, Ljubljana, Slovenia, September 24–27, 2024, Proceedings, Part II*, pages 173–182, Berlin, Heidelberg. Springer-Verlag.

Giorgio Maria Di Nunzio and Federica Vezzani. 2025. FAIRterm 2.0: Towards FAIR Terminologies Resources for EOSC. In *IEEE International Conference on Cyber Humanities (IEEE-CH)*, IEEE Explore, Florence, Italy. IEEE.

Kaja Dobrovolc. 2022. Spoken Language Treebanks in Universal Dependencies: an Overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Richard Eckart de Castilho, Seid Muhie Yimam, Iryna Gurevych, and Chris Biemann. 2014. Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the CLARIN Annual Conference (CAC 2014)*. Latest release version 3.6.11 (Dec 2021) available at https://github.com/webanno/webanno/releases.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, and others. 2018. Cross-linguistic data formats, advancing data sharing and reuse in comparative linguistics. *Scientific Data*, 5:180205.

Benjamin Galliot, Guillaume Wisniewski, Séverine Guillaum, Guillaume Jacques, and Alexis Michaud. 2022. Facilitating NLP specialists' access to language archive materials: an update. In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, pages 109–118. Marseille, France.

Richard Gottheil and Lazarus Belleli. 1904. Judaeo-greek and judaeo-italian. In Isidore Singer, editor, *The Jewish Encyclopedia*, volume 7, pages 310–313. Funk & Wagnalls, New York. Online version available at `https://www.jewishencyclopedia.com/articles/8950-judaeo-greek-and-judaeo-italian` (accessed 2026-03-05).

Gaël Guibon, Khensa Daoudi, Kim Gerdes, Bruno Guillaume, and Kirian Guiller. 2020. When collaborative treebank curation meets graph-rewriting: Arborator-grew. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. The tool integrates the Grew API and supports collaborative dependency treebank annotation.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–95.

Nikolaus P. Himmelmann. 2006. Language documentation: What it is and what it is good for? In J Gippert, Nikolaus P. Himmelmann, and

U Mosel, editors, *Essentials of Language Documentation*, pages 1–30. De Gruyter Mouton, Berlin.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.

Ritván Jusúf Karahóga, Panagiotis Krimpas, Vivian Stamou, Vasileios Arampatzakis, Dimitrios Karamatskos, Vasileios Sevetlidis, Nikolaos Constantinides, Nikolaos Kokkas, George Pavlidis, and Stella Markantonatou. 2022. Morphologically annotated corpora of pomak. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 179–186.

Alfred Kik, Martin Adamec, Alexandra Y Aikhenvald, Jarmila Bajzekova, Nigel Baro, Claire Bowern, Robert Colwell K., Pavel Drozd, Pavel Duda, Sentiko Ibalim, Leonardo R. Jorge, Jane Mogina, Ben Ruli, Katerina Sam, Hannah Sarvasy, Simon Saulei, George D. Weiblen, Jan Zravy, and Vojceth Novotny. 2021. Language and ethnobiological skills decline precipitously in Papua New Guinea, the world's most linguistically diverse nation. *PNAS*, 118(22).

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Panagiotis Krimpas. 2021. [The Modern Greek Heptanesian variety: A holistic approach]. In *Proceedings of the ICGL14*, pages 646–660, Patras. University of Patras.

Nala H. Lee and John R. Van Way. 2018. The language Endangerement Index. In Lyle Campbell and Anna Belew, editors, *Cataloguing the World's Endangered Languages*. Routledge.

Leo Levi. 1961. Tradizioni liturgiche, musicali e dialettali a Corfù. *La Rassegna Mensile di Israel*, 27(1):20–31.

Marta Milazzo and Giorgio Maria Di Nunzio. 2024. The Onomastic Repertoire of the Roman d'Alexandre (ORNARE). Designing an Integrated Digital Onomastic Tool for Medieval French Romance. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15982–15987, Torino, Italia. ELRA and ICCL.

Marta Milazzo and Giorgio Maria Di Nunzio. 2025. ORNARE: Toward a Digital Methodology for Onomastic Data in Medieval French Romance. *Umanistica Digitale*, (21):141–158.

Alice Millour, Lorenza Brasile, Alberto Ghia, and Laurent Kevers. 2024. Aggetivu, Aggitivu o Aghettivu? POS Tagging Corsican Dialects. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 600–608. ELRA and ICCL, Torino, Italia.

Sarah Moeller, Godfred Agyapong, Jarrod Cruz, Alexis Palmer, and Mans Hulden. 2026. Computational Methods for Language Documentation and Description. *Annual Review of Linguistics*, (12):147–70.

Johannes Mücke. 2019. Infinitive reduction in Corfiot Italian: a case of areal convergence? In *Proceedings of the 5th Patras International Conference of Graduate students in Linguistics (PICGL5)*, pages 214–238, Patras. University of Patras.

Johannes Gregor Mücke. 2025. *Das italoromanische Idiom auf Korfu. Sprachverdrängung und Sprachkontakt in Südosteuropa.* PhD Thesis, Universität Graz, Graz.

Salikoko S. Mufwene. 2001. *The ecology of language evolution*. Cambridge University Press, Cambridge.

Jenny Nachtmann. 2001. Italienisch als Minderheitensprache: Fallbeispiel Korfu. Unpublished Master's Thesis, University of Freiburg, Freiburg.

Andrew Nevins. 2022. *When Minoritized Languages Change Linguistc Theory*. Cambidge University Press.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Pyysalo Sampo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Frank Seifart. 2019. Contact-induced change. In Jeroen Darquennes, Joseph Salmons, and Wim Vandenbussche, editors, *Language Contac: An International Handbook Volume 1*, pages 13–23. De Gruyter Mouton, Berlin, Boston.

SIL International. 2026. Fieldworks Language Explorer FLEx (version 9.3.7). https://software.sil.org/fieldworks/. Accessed 12 March 2026.

Sarah Grey Thomason and Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. University of California Press, Berkeley, Los Angeles, London.

Francis M. Tyers, Mariya Sheyanova, and Jonathan North Washington. 2017. Ud annotatrix: An annotation tool for universal dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT 2017)*.

Georgios Vardakis. 2019. Linguistic variation in Jewish-Romance and Greek dialects: A structural analysis of the Judeo-Italian dialect of Corfu. Unpublished Master's Thesis, Sorbonne Université, Paris.

Georgios Vardakis. 2021. A formal analysis of complementation in Corfioto. Unpublished Master's Thesis, University of Padua, Padua.

Georgios Vardakis. 2023. Documenting Corfioto: Evidence for Contact-Induced Grammaticalization in the Romance Variety of the Jewish Community of Corfu. In Nikolaos Lavidas, Alexander Bergs, Elly van Gelderen, and Ioanna Sitaridou, editors, *Internal and External Causes of Language Change: The Naxos Papers*, pages 247–285. Springer International Publishing, Cham.

Georgios Vardakis. 2025. *The hybrid verb complementation system in Corfioto, a Romance contact variety of the Jewish community of Corfu*. Unpublished PhD Thesis, University of Padua, Padua.

Georgios Vardakis and Giorgio Maria Di Nunzio. 2022. A Methodology for the Management of Contact Languages Data. The Case Study of the Jews of Corfu. In *Linking Theory and Practice of Digital Libraries*, pages 538–542, Cham. Springer International Publishing.

Federica Vezzani and Giorgio Maria Di Nunzio. 2019. Computational Terminology in eHealth. In *Digital Libraries: Supporting Open Science*, Communications in Computer and Information Science, pages 72–85, Cham. Springer International Publishing.

Federica Vezzani and Giorgio Maria Di Nunzio. 2020a. Methodology for the standardization of terminological resources: Design of TriMED database to support multi-register medical communication. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 26(2):265–297.

Federica Vezzani and Giorgio Maria Di Nunzio. 2020b. On the Formal Standardization of Terminology Resources: The Case Study of TriMED. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4903–4910, Marseille, France. European Language Resources Association.

Federica Vezzani, Giorgio Maria Di Nunzio, and Geneviève Henrot. 2018. TriMED: A Multilingual Terminological Database. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jan Wijffels. 2022. *textplot: Text Plots*. R package version 0.2.2.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2025. Elan (eudico linguistic annotator) – multimedia

annotation tool, version 7.0. Latest version downloadable as of 2025; manual version 6.3 updated Jan 2022.

Anthony C. Woodbury. 2011. Language Documentation. In PK Austin and J Sallbank, editors, *The Cambridge Handbook of Endangered Languages*. Cambridge University Press, Cambridge.

Yihui Xie, Joe Cheng, Xianying Tan, and Garrick Aden-Buie. 2025. *DT: A Wrapper of the JavaScript Library 'DataTables'*. R package version 0.34.0.