LSDT: a Dependency Treebank of Lombard Sinti

Marco Forlano
University of Bergamo/Pavia
marco.forlano@unibg.it

Luca Brigada Villa
University of Bergamo/Pavia
luca.brigadavilla@unibg.it

Abstract

Lombard Sinti is a variety of Romani spoken in Italy, in an area that can be identified with the Italian region of Lombardy. In this paper we present LSDT, a dependency treebank of Lombard Sinti consisting of 100 sentences manually annotated both morphologically and sintactically.

1 Introduction

Since its publication, Universal Dependencies (de Marneffe et al., 2021) provides annotated resources for 130 languages. The data comes in form of treebanks, that is, collections of sentences with morphological and syntactic annotation. For such annotation, UD has developed a scheme which has the goal to be cross-linguistically consistent in order to facilitate research from a typological perspective. Despite the amount of languages in UD, minority languages are still underrepresented in this project. This paper presents LSDT, a treebank of Lombard Sinti, a variety of Romani.

Romani is an Indo-Aryan language spoken across Europe by socio-ethnic communities whose members mostly call themselves *Roma* or *Sinte/i*. Nowadays, it is widely acknowledged that such groups have a common Indian origin (see Matras, 2002: 14-18 for an overview on the theories about the origin of Romani populations). According to scholars, their migration from India began by the first millennium AD, and after they arrived in Greece in the Middle Ages, they started to move all over Europe in different groups, slowly getting to those areas where they still largely reside (see Beníšek, 2020: 25-26). As a non-territorialized language, Romani shows a very high rate of geographic variation (see Matras et al., 1997; Matras, 2002: 5-13; Elšík and Beníšek, 2020 for an overview on Romani dialects and their classification criteria). Furthermore, all Romani varieties are affected in each country by asymmetric contact with their various co-territorial majority languages (Matras and Adamou, 2020), which ultimately leads to different degrees of language shift (see e.g. Adamou, 2010, 2016).

The paper is structured as follows: in Section 2 we introduce Lombard Sinti providing the typologycal classification of this language and its historical background; in Section 3 we present the data and how we converted the sentences to CoNLL-U format; in Section 4 we explain in detail the process of annotation of the treebank; finally, Section 5 concludes the paper and presents the work we plan to do in the future.

2 Lombard Sinti

Lombard Sinti (endonym: *sinto/sintu*[1]) is a variety of Romani spoken in Northern Italy. Its core area corresponds to the Italian region of Lombardy, though Lombard Sinti speakers are also found in the adjacent areas of Eastern Piedmont and Emilia. Lombard Sinti is commonly assigned by scholars to the Northern branch of Romani dialects (see Matras, 2002: 9). In this section, we discuss some typological features of Lombard Sinti and we offer a historical background of the variety.

2.1 Typological classification

Romani is a language with rich inflectional morphology; as for the nominal inflection, "lexical nouns inflect for the paradigmatic categories of nominal case and number and have an inherent gender, masculine vs feminine" (Elšík, 2020: 163). There are eight cases, commonly referred to as nominative, vocative, oblique, dative, locative, comitative/instrumental, and genitive. From a syntactic point of view, Romani shows several typological features that have been shaped through contact

[1]The alternation between the vowels -o e -u in Lombard Sinti should be probably understood as a free variation, although a preference for one vowel over the other in terms of norm and frequency seems to be partially linked to sub-group membership within Lombard Sinti.

with non-Indic languages, especially with Bizantine Greek (Adamou and Matras, 2020: 235). The most prominent is the nominative-accusative and neutral alignment, as opposed to the ergative one that is found in modern Indo-Aryan languages and that can be even reconstructed for Proto-Romani (Beníšek, 2020). Other innovations include the emergence of definite and indefinite articles and the development of an unmarked VO order, alongside some cross-linguistic tendencies associated with it, e.g., the presence of prepositions (Adamou and Matras, 2020). Furthermore, in all Romani varieties, subordinate clauses follow independent clauses. The outlined typological features are generally found in Lombard Sinti too. However, in Lombard Sinti, a complete loss of cases in the nominal inflection has occurred. Moreover, the variety shows a neutral morphosyntactic alignment, whereby A, S, and P arguments are expressed identically (i.e. with a zero-mark, see e.g. Malchukov et al., 2010), while in most Romani varieties, animated P arguments are expressed differently from A and S through a dedicated oblique suffix. From the above overview, it appears that Lombard Sinti and Italian are similar in many respects as far as typology is concerned (see also Sorrenti, 2014). Nonetheless, some major differences can be found. Just to mention a few, while pronominal clitics in Lombard Sinti are consistently placed after the verb, in Italian their positioning is constrained (and appear after the verb only when the mood is non-finite). Moreover, in Lombard Sinti, the rich inherited case system has been retained for personal pronouns, which limits the use of prepositions in such contexts. Finally, if compared to Italian, Lombard Sinti disposes of a smaller range of non-finite verbal moods, which tend to be expressed analytically.

2.2 Available sources on Lombard Sinti

The available sources on Lombard Sinti are rather scarce. The glossary by Partesani (1973) and the sketchy grammatical description by Soravia (1977) represent the first attempts in this direction. Moreover, Lombard Sinti lemmas appear in Soravia and Fochi (1995) *Dizionario Sinottico delle Parlate Zingare in Italia*[2], the only existing vocabulary on Romani varieties spoken in Italy. More recently, some research by Andrea Scala has focused on specific structural aspects of Lombard Sinti, such as the numeral system (Scala, 2017) or some morphosyntactic constructions, such as the negative one (Scala, 2020). The most complete work to date is perhaps Sorrenti (2014), which proposes a full grammatical description of Lombard Sinti, pointing out similarities and differences with Italian.

2.3 Historical background

As is often the case with Romani varieties, the history of Lombard Sinti and its speakers can be partially reconstructed on linguistic evidence. The high rate of borrowings from Germanic languages and, to a lesser extent, from Slavic languages, suggests that after the Greek diaspora Lombard Sinti speakers resided for some time in German-speaking areas via the Western Balkans. Their arrival in Italy is estimated during the Modern Era (Piasere, 2004).

Official data on Lombard Sinti speakers is not available. Indicatively, in Northern Italy, there are an estimated 30.000 Sinti[3], divided into different groups speaking different varieties, including the Lombard ones. In the panorama of the Sinti varieties spoken in Italy, Lombard Sinti appears to be one of the best preserved. Indeed, it is still learned as L1 by children and maintains strong identity functions. According to the Expanded Graded Intergenerational Disruption Scale (Lewis and Simons, 2010), a measure that is used to assess the vitality or endangerment rate of languages, Lombard Sinti might thus be classified as Vigorous. Currently, Lombard Sinti speakers exhibit a linguistic repertoire where Italian and Sinti are used together in informal domains, while in formal ones only Italian is used (Scala, 2012).

Lombard Sinti, as Romani varieties in general, is a non-standardized language, as its use is primarily oral. Therefore, neither normative grammar nor graphic standards are available in this variety, although some more or less spontaneous written productions exist (see the translation of the Gospel of Mark by don Mario Riboldi, or the use of Sinti on new media, cf. Scala, 2015). Among scholars, to transcribe Lombard Sinti data it is common to adopt a semi-conventionalized academic standard that promotes, for instance, the use of diacritics from Slavic alphabets to represent affricate sounds (Matras, 1999). Such a choice informs, among others, the *Dizionario Sinottico delle Parlate Zingare in Italia* (Soravia and Fochi, 1995) as well as

[2] Synoptic Dictionary of Gypsy Speeches in Italy.

[3] http://romafacts.uni-graz.at/

the Lombard Sinti sample in the Romani Morpho-Syntax Database (Matras et al., 2009).

Finally, it is worth mentioning that Romani varieties in Italy do not appear among the languages that enjoy the specific protection tools provided by national law 482/1999 on linguistic minorities (see e.g. Fiorentini, 2022). This inhibits the use of Roman Sinti in public domains, from schools to local and national media.

3 Data

The sentences included in the treebank[4] were collected from the Romani Morpho-Syntax Database (Matras et al., 2009) a resource that collects dictionaries, sentences and phrases from many Romani dialects.

We extracted the available data from the Lombard Sinti portion of the database[5], selecting 100 complete sentences from the sample. The goal of the selection was to include in our sample sentences that differ from each other for length, gender of the nominal elements, tenses of the verbs in order to obtain a diverse set.

num. tokens	num. sentences
4	1
5	4
6	9
7	14
8	11
9	13
10	11
11	11
12	6
13	5
14	5
15	2
16	2
17	3
18	1
19	1
20	1

Table 1: Distribution of the lengths of the sentences in the treebank in terms of number of tokens

After the selection of the sentences, we performed a shallow tokenization considering the spaces in the sentences which resulted in a CoNLL-U file with one token per line without annotation. Each sentence in the treebank was provided with a `sent_id` which contains a four-digit number that refers to the position of the sentence in the treebank and the number of the sample in the Lombard Sinti portion of the RMS Database (e.g. `sent_id = 0027@RMS-443` is the 27th sentence in the treebank and corresponds to sentence 443 in the RMS). Along with the `sent_id`, we included the text and the English translation in the metadata of each sentence.

4 Annotation

To annotate the treebank, we used UD Annotatrix (Tyers et al., 2017), a tool that allows to upload a file formatted in CoNLL-U and annotate it. First, we corrected the tokenization obtained considering the spaces between words, then we provided lemmatization and morphological features for each token. Finally, for each sentence, we annotated the syntactic dependencies.

In this Section we explain more in detail each phase of the annotation process.

4.1 Tokenization and lemmatization

The tokenization obtained considering the spaces between words was corrected by manually separating pronominal clitics from their host verbs (e.g. *selma* → *sel ma* '(he/she) visits us'; *dukadoma* → *dukadom ma* '(I) hurt myself') and articles from prepositions in case of articulated prepositions (e.g. *pur drom* → *par u drom* 'on the street'; *ki vierta* → *ka i vierta* 'to the pub') in order to adhere to Universal Dependencies' guidelines on tokenization[6].

number of tokens	990
number of unique lemmas	293

Table 2: Number of tokens and unique lemmas in the LSDT treebank

To carry out the lemmatization, we relied on the conventions that are used in Romani linguistics. Therefore, we considered the masculine singular forms as lemmas for nouns, adjectives and demonstrative pronouns and the third person of the present indicative for verbs. As for stressed personal pronouns, since in Lombard Sinti they are inflected by case and gender (only in the third person), we chose the nominative masculine as base form.

[4] https://github.com/unipv-larl/LSDT

[5] https://romani.humanities.manchester.ac.uk//rms/browse/phrases/phraselist

[6] https://universaldependencies.org/u/overview/tokenization.html

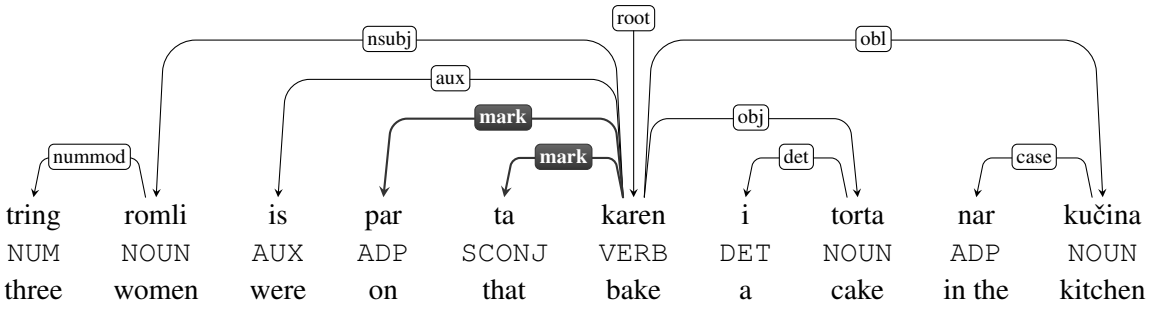

Figure 1: Dependecy tree of the sentence 'tring romli is par ta karen i torta nar kučina' ('three women were baking a cake in the kitchen').

4.2 POS tagging and morphological annotation

Each POS was then assigned the relevant morphological features (see Table 4 in the Appendix A for the complete list). Gender and number were annotated for nouns and adjectives, while verb form, mood, tense, and person were annotated for verbs. Among verbs, only past participles, which in Lombard Sinti have an adjectival meaning (Scala, 2011: 256), were also annotated for number and gender. In line with the literature on Romani linguistics, the feature `Mood=Subj` (subjunctive) was used to indicate the "subordinative mood" (Scala, 2011: 258; Sorrenti, 2014: 138), i.e., the mood of verbs in implicit completive clauses. As for personal pronouns, gender was annotated only for third-person singular, since a gender-based distinction is made in Lombardi Sinti only between the forms *joi* 'she' and *jo* 'he'. `Case` was annotated only for stressed personal pronouns, as unstressed clitics show one single morphological manifestation for the non-nominative case (Sorrenti, 2014: 135).

4.3 Syntactic annotation

Lombard Sinti shows some language-specific constructions. To express the progressive meaning, it uses a periphrasis where the verb 'to be' is followed by the preposition *par* 'on', by the complementizer *ta* (which is however optional) and the dependent subjunctive (e.g. *jom par ta sua*, 'I'm going to sleep', lit. 'I'm on that sleep'). In such constructions, we considered the dependent subjunctive as the root and annotated the relation between it and the two function words *par* and *ta* as `mark`, and the relation with the verb 'to be' as `aux` (as shown in Figure 1).

The label `fixed` was used to annotate, among others, the relation between the third person copula *i* 'is' and the oblique pronouns (Figure 2) that

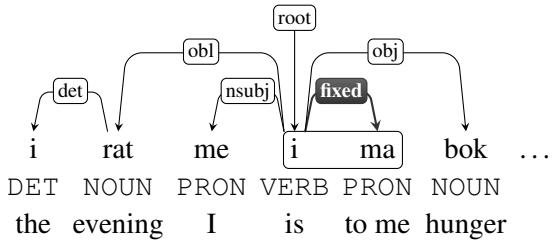

Figure 2: Dependecy tree of part of the sentence 'i rat me ima bok, ma se xava vava tulo' ('in the evening I get very hungry, but if I eat I'll get very fat').

in Lombard Sinti are used together to form the paradigm of the verb *to have* (e.g. *i-ma* 'I have', lit. 'is-to-me').

Finally, in Lombard Sinti, third-person unstressed pronouns are used as clitics to (re)activate a reference to the subject of the sentence, that may be overtly expressed or not (e.g. *u ker i-lu nevo eta baro*, 'the house is-it new and big', see Figure 3 in Appendix B). In such constructions, we used the label `expl` to annotate the relation between the clitic and the verb.

5 Conclusion and future work

In this paper, we presented LSDT, a dependency treebank of Lombard Sinti which constitutes the first resource of this kind for this language. We showed the process we followed to annotate the sentences both morphologically and syntactically, dealing with the issues that this language presents.

In the future, we plan to enlarge the treebank with more sentences including different sources (speech, written documents). A language resource like this, if sufficiently expanded, might be also useful to train a parser to automatically annotate other texts in Lombard Sinti and to build models that can

be used to transfer information and annotate texts in other Romani varieties.

Acknowledgments

We would like to thank the anonymous reviewers who provided great insights and suggestions. The paper is the result of close collaboration between the two authors. For academic purposes, Marco Forlano is responsible of the annotation of the treebank, Sections 2 and 4 and Luca Brigada Villa is responsible of Sections 1, 3 and 5.

References

Evangelia Adamou. 2010. Bilingual speech and language ecology in greek thrace: Romani and pomak in contact with turkish. *Language in society*, 39(2):147–171.

Evangelia Adamou. 2016. *A Corpus-Driven Approach to Language Contact*. De Gruyter Mouton, Berlin, Boston.

Evangelia Adamou and Yaron Matras. 2020. Romani syntactic typology. In Yaron Matras and Anton Tenser, editors, *The Palgrave Handbook of Romani Language and Linguistics*, pages 187–227. Palgrave Macmillan, London.

Michael Beníšek. 2020. The historical origins of romani. In Yaron Matras and Anton Tenser, editors, *The Palgrave Handbook of Romani Language and Linguistics*, pages 13–47. Palgrave Macmillan, London.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Viktor Elšík. 2020. Romanimorphology. In Yaron Matras and Anton Tenser, editors, *The Palgrave Handbook of Romani Language and Linguistics*, pages 155–186. Palgrave Macmillan, London.

Viktor Elšík and Michael Beníšek. 2020. Romani dialectology. In Yaron Matras and Anton Tenser, editors, *The Palgrave Handbook of Romani Language and Linguistics*, pages 389–417. Palgrave Macmillan, London.

Ilaria Fiorentini. 2022. *Socialinguistica delle minoranze in Italia: Un'introduzione*. Carocci, Rome.

Paul Lewis and Gary Simons. 2010. Assessing endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique*, 55(2):103–120.

Andrej Malchukov, Martin Haspelmath, and Bernard Comrie. 2010. Ditransitive constructions: a typological overview. In Andrej Malchukov, Martin Haspelmath, and Bernard Comrie, editors, *Studies in ditransitive constructions: A comparative handbook*, pages 1–64. De Gruyter Mouton, Berlin, New York.

Yaron Matras. 1999. Writing romani: The pragmatics of codification in a stateless language. *Applied linguistics*, 20(4):481–502.

Yaron Matras. 2002. *Romani: A Linguistic Introduction*. Cambridge University Press, Cambridge.

Yaron Matras and Evangelia Adamou. 2020. Romani and contact linguistics. In Yaron Matras and Anton Tenser, editors, *The Palgrave Handbook of Romani Language and Linguistics*, pages 329–352. Palgrave Macmillan, London.

Yaron Matras, Peter Bakker, and Hristo Kyuchukov. 1997. *The typology and dialectology of Romani*, volume 156 of *Amsterdam studies in the theory and history of linguistic science : Series 4, Current issues in linguistic theory*. John Benjamins Publishing Company, Amsterdam.

Yaron Matras, Christopher White, and Viktor Elšík. 2009. The Romani Morpho-Syntax (RMS) Database. In Martin Everaert, Simon Musgrave, and Alexis Dimitriadis, editors, *The Use of Databases in Cross-Linguistic Studies*, pages 329–362. De Gruyter Mouton, Berlin, New York.

Sergio Partesani. 1973. Glossario del dialetto zingaro lombardo. *Lacio Drom*, 8(4):2–29.

Leonardo Piasere. 2004. *I Rom d'Europa*. Laterza, Rome, Bari.

Andrea Scala. 2011. Così vicini, così lontani: i parlanti romanì, l'italiano e la scuola. In Rosella Bozzone Costa, Luisa Antonietta Fumagalli, and Ada Valentini, editors, *Apprendere l'italiano da lingue lontane: prospettiva linguistica, pragmatica ed educativa*, pages 249–265. Guerra Edizioni, Perugia.

Andrea Scala. 2012. Purché la lingua non sia una sola. Trasformazione dei repertori e conservazione del plurilinguismo presso i sinti italiani dall'Unità ad oggi. In *Coesistenze linguistiche nell'Italia pre-e postunitaria, Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI)*, pages 437–448, Rome. Bulzoni.

Andrea Scala. 2015. Se proprio dobbiamo scrivere, almeno facciamolo come gli altri italiani: i Sinti dell'Italia settentrionale e la grafizzazione della loro lingua. In Silvia Dal Negro, Federica Guerini, and Gabriele Iannaccaro, editors, *Elaborazione ortografica delle varietà non-standard. Esperienze spontanee in Italia e all'estero*. Bergamo University Press, Bergamo.

Andrea Scala. 2017. I numerali da 1 a 10 in sinto lombardo. In Giuseppe Sergio and Massimo Prada, editors, *Italiani di Milano. Studi in onore di Silvia Morgana*, pages 789–797. Ledizioni, Milano.

Andrea Scala. 2020. A lombard sinti ethno-text on mourning and marriage. *Eivista annuale dell'associazione Ethnorema*, 16:59–71.

Giulio Soravia. 1977. *Dialetti degli zingari italiani*. Pacini, Pisa.

Giulio Soravia and Camillo Fochi. 1995. *Vocabolario sinottico delle lingue zingare parlate in Italia*. Centro Studi Zingari, Rome.

Elena Sorrenti. 2014. Italiano e sinto lombardo a confronto: somiglianze, divergenze e prospettive didattiche. *Italiano LinguaDue*, 6(1):117–147.

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*, pages 10–17, Prague, Czech Republic.

A Tags used in the treebank

part-of-speech	count
ADJ	32
ADP	90
ADV	72
AUX	20
CCONJ	17
DET	150
NOUN	172
NUM	9
PRON	101
PUNCT	109
SCONJ	49
VERB	168

Table 3: Parts-of-speech and their frequencies in the treebank.

deprel	count
ADJ	Gender, Number
ADP	Gender, Number
AUX	Mood, Person, Tense, VerbForm
DET	Definite, Gender, Number, Poss PronType
NOUN	Gender, Number
NUM	NumType
PRON	Case, Clitic, Gender, Number, Person Poss, PronType, Reflex
PUNCT	Gender, Number, Person, PronType
VERB	Gender, Mood, Number, Person, Tense, VerbForm

Table 4: List of features annotated for each part-of-speech tag.

deprel	count
acl	14
advcl	35
advmod	57
amod	13
aux	13
case	70
cc	17
ccomp	9
conj	16
cop	14
csubj	1
det	147
expl	23
fixed	19
iobj	34
mark	81
nmod	9
nsubj	66
nummod	9
obj	60
obl	55
punct	108
root	100
xcomp	19

Table 5: Dependency relations and their frequencies in the treebank.

B Additional tree

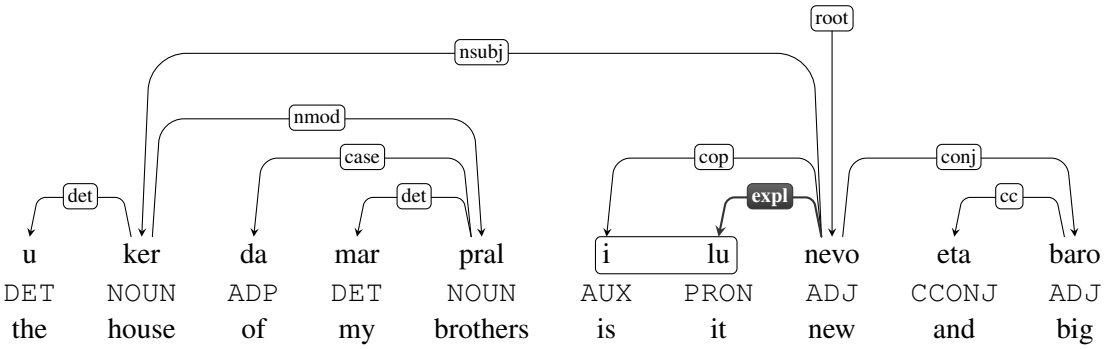

Figure 3: Dependecy tree of the sentence ‘u ker da mar pral ilu nevo eta baro’ (‘my brother’s house is new and big’).