A Universal Dependencies corpus for Ligurian

Stefano Lusito
Universität Innsbruck
Institut für Romanistik
Innsbruck, Austria
stefano.lusito@uibk.ac.at

Jean Maillard
University of Cambridge
Dept. of Computer Science and Technology
Cambridge, United Kingdom
jean@maillard.it

Abstract

Ligurian is a minority Romance language spoken in the homonymous region of Northern Italy and the Principality of Monaco, amongst others. In this paper we present the first Universal Dependencies treebank for Ligurian, consisting of 316 sentences and 6 928 tokens, extracted from a wide variety of sources to reflect variation in syntax and register.

Along with the corpus, we contribute a short analysis of the varieties and spelling systems of Ligurian, as well as a set of recommendations and annotation guidelines for certain constructions with non-trivial analyses. We hope that these will serve as a foundation for further research, to encourage the development of NLP technologies for a language that has so far been under-served.

1 Introduction

Ligurian is a minority Romance language originating from the Northern part of the Italian peninsula, considered to be "definitely endangered" by UNESCO.[1] In spite of its relatively extensive usage throughout the centuries, no methodical corpus whatsoever exists for Ligurian, and no advanced NLP technologies have been developed for it.

Universal Dependencies (UD) (Nivre et al., 2016) is a cross-lingual framework for consistent annotations of parts-of-speech, morphological features, and syntactic dependencies. The project aims to facilitate the development of parsing technologies, enabling the use of techniques such as cross-lingual transfer.

In this paper we present the first ever digital corpus of Ligurian,[2] consisting of 316 sentences annotated according to the UD framework. We also contribute an analysis of the current state of the language, including its varieties and spelling system, and provide recommendations to serve as a foundation for future research.

The creation of a UD treebank for Ligurian enables the development of parsers and taggers for it, unlocking NLP technologies as well as software which is fundamental for linguistic research, such as advanced search tools for corpus linguistics (Guillaume, 2019).

The complete lack of technological support for Ligurian is – we believe – partly to blame for its endangered status. With this project we hope to encourage further research in the language and the development of NLP tools for it, in the hope of playing a small role in helping reverse a course which could otherwise lead to its complete disappearance.

2 The Ligurian language

2.1 Definition of *Ligurian*

Ligurian denotes the ensemble of Romance varieties traditionally spoken within the homonymous region of Liguria in Northern Italy. In its local forms, it is also the historical language of the Principality of Monaco as well as of the Tabarkin communities of Southern Sardinia, amongst others (Toso, 2003a; Toso, 2001; Toso, 2003b).

[1] http://www.unesco.org/languages-atlas/

[2] Available at https://github.com/UniversalDependencies/UD_Ligurian-GLT/.

Despite being traditionally associated to the so-called Gallo-Italic Romance dialects (Ascoli, 1876), the Ligurian varieties distinguish themselves from the other members of that group (Piedmontese, Lombard, Emilian and Romagnol) for their characters of conservatism and, at the same time, innovation in their evolution from vulgar Latin (Toso, 1995, p. 30). In this respect, it was the pioneering work of Diez (1836, p. 86) to first identify Ligurian as the transition area between North and Central Italian dialects.

The more recent division of Ligurian Romance varieties (Toso, 2002) distinguishes between a central, linguistically dynamic area (known as "Genoese" in the literature – *zeneise* in Ligurian) and marginal zones which have not received many of the innovative traits spread out from the central zone to a considerable portion of the region.[3] The Genoese dialects – whose extension includes the whole coastline from Noli to Framura and a sizeable portion of the corresponding inland region – cover more than a third of the administrative region's surface (encompassing many of the main urban areas) and represent by far the most widespread Ligurian variety as for number of speakers.

2.2 Role of Genoese and its literary production

In fact, Genoese is nowadays the only Ligurian dialect with a written corpus – mainly literary – which continuously stretches from the 13th century to the present day (Toso, 2009). Being traditionally considered the most prestigious Ligurian variety, it has also historically served as the *koinè* language for speakers of other Ligurian dialects – the usage of Italian having reached general oral diffusion only in the second half of the last century[4] – and functions still today as the Ligurian reference dialect when no particular diatopic information is required or specified (Toso, 1997).

2.3 Genoese spelling system

The long history of Genoese as a written tongue has led to the development of a spelling system which has evolved over the years together with the language itself (Toso, 2009, p. 27-32).

The influence of a relatively modest, but high quality literature among those who usually write in Genoese for public purposes is still such that all the main features of its traditional spelling system are generally accepted (e.g. ⟨o⟩ for [u], ⟨u⟩ for [y], ⟨æ⟩ for [ɛ(ː)] or ⟨x⟩ for [ʒ]). Nevertheless, the freedom allowed by the lack of both state recognition and prescriptive institutions still results in several disputed aspects, such as the writing of pre-tonic double consonants, always pronounced as singleton (e.g. *accattâ* vs. *acatâ* [akaˈtaː] 'to buy') and vowel-length markers, especially when a long vowel comes before the main stress of a word (e.g. *mäveggia* vs. *mâveggia* [maːˈveʤːa] 'wonder'). While, on the one hand, this situation leaves the field open to stimulating debates among the speakers, on the other it can sometimes generate confusion for the general public and even jeopardise meritorious projects. An illustrative case is the Ligurian edition of Wikipedia,[5] where the lack of uniform spelling guidelines (along with the use of a multitude of different local dialects) leads to a disorganised appearance (Lusito, 2021).

Driven by the aim to find a possible solution to these issues, a slight reform of Genoese spelling was recently proposed by a diverse group of journalists, writers, and academics (Acquarone, 2015). It has been adopted by the main Ligurian newspaper for its Ligurian-language columns (*Il Secolo XIX*), the book series *E restan forme* (poetry) and *Biblioteca zeneise* (prose)[6], the magazine *O Stafî* as well as the research project *GEPHRAS* currently running at the University of Innsbruck.[7]

Since the texts collected in this corpus come in large part from some of the aforementioned sources, this is also the spelling system adopted in this work.

[3] An in-depth outline of the evolutionary differences and features of the Ligurian dialects is to be found in Toso (1995, p. 30-42).

[4] Estimates for the percentage of people with an adequate knowledge of Italian at the time of the political unification of the country (1861-1870) range between 2.5% (De Mauro, 1991) and 9.5% (Castellani, 1982).

[5] `https://lij.wikipedia.org/`

[6] Respective pubishers Zona (`http://www.editricezona.it/`) and De Ferrari (`https://www.deferrarieditore.it/`).

[7] `https://romanistik-gephras.uibk.ac.at/`

Genre	Documents	Paragraphs	Sentences	Tokens
Fiction	4	19	76	2 216
News	2	7	59	1 472
Bible	1	13	46	1 241
Grammar examples	2	—	77	851
Wikipedia	2	8	18	754
Spoken	1	20	40	394
Total	12	67	316	6 928

Table 1: Composition of the Universal Dependencies corpus for Ligurian. *Documents* refer to chapters for the Fiction and Bible genres, and articles for the News and Wikipedia genres.

2.4 General Ligurian syntactic features

As already mentioned, the phonology, morphology, and syntax of Ligurian show features in the middle between those of North Italian dialects, on the one hand, and Tuscan and the South Italian ones, on the other.

Following Forner (1997, p.250-252), among some of the main features in contrast with standard Italian we find:

1. the presence of subject clitics,
2. compound demonstrative pronouns (although one-word pronouns also exist: *veuggio sto chì* besides the less frequent *veuggio questo* 'I want this one')[8],
3. bicomposed verbs, especially to express direction (*dâ quarcösa inderê* 'to give something back', Italian 'restituire qualcosa', *piccâ drento à quarcösa* / *quarchedun* 'to crash against something / somebody', Italian *scontrare qualcosa* / *qualcuno*), and
4. periphrastic structures to create progressive forms, with different possibile solutions (for 'I am working' one could use a construction with verb, adverb and infinitive, like *son derê à travaggiâ* or *son apreuvo à travaggiâ*, or a cleft sentence like *son chì che travaggio*; Italian has *stare* followed by a verb in gerund form instead: *sto lavorando*).[9]

3 Corpus development

3.1 Collection

The texts included within the corpus (see Table 1) cover several genres and have been extracted from the most varied sources, in order to reflect variation in syntax and register. All texts were already written according to the aforementioned spelling system, which was maintained with minimal interventions to increase uniformity among them. The largest category, fiction, consists of four excerpts from three texts by contemporary authors or translators (Lusito, 2020; Toso, 2018; Iacopone, 2017). We also include one news and one magazine article (Canessa, 2016; Toso, 2020); an excerpt from a translation of the Gospel of Mark (Toso, 2019); two articles from the Ligurian edition of Wikipedia; a number of example sentences from a Genoese grammar book (Toso, 1997) selected to demonstrate a variety of characteristic syntactic constructions; and the transcript of a short comedy sketch, expressly conceived to be broadcast on the radio, but which accurately reflects oral usage. Finally, we include translations of the 20 example sentences making up the Cairo CICLing Corpus,[10] a multilingual parallel treebank of short sentences.

The relatively low fraction of texts coming from Wikipedia – a common source of content for textual corpora – is due to its inconsistent use of orthography and dialects, as mentioned in Section 2.3, as well as to quality issues with some of its contents, which appear to be written by novice language learners (Lusito,

[8]The non-marked Italian respective form is *voglio questo*. The construction comprising the adverb – *voglio questo qui* – is possible; in that case, if necessary, Ligurian would use a cleft sentence to mark the focus: *l'é sto chì che veuggio*.

[9]All the Ligurian examples are in Genoese.

[10]`https://github.com/UniversalDependencies/cairo`

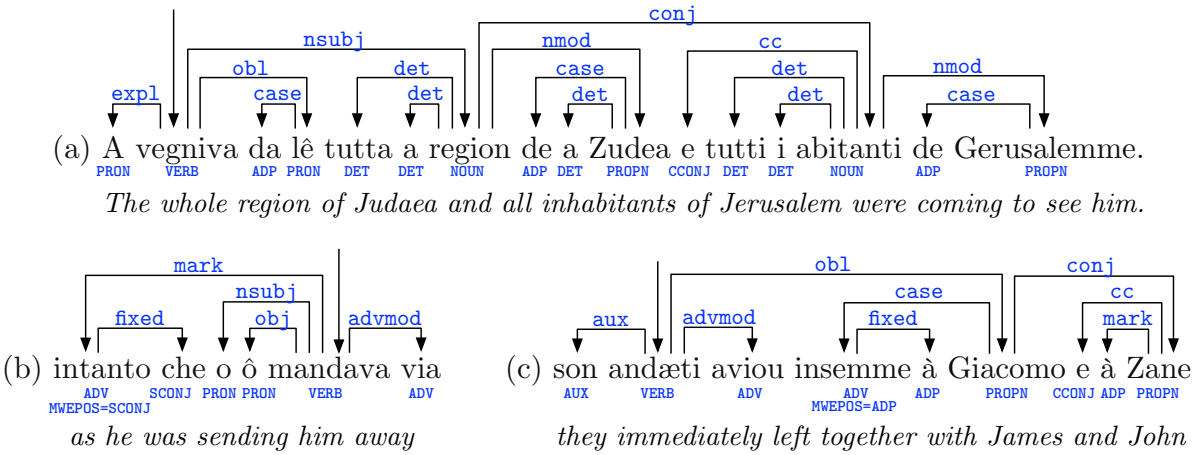

Figure 1: Some examples of annotated Ligurian sentences drawn from the corpus.

2021).

3.2 Annotation

The annotation was performed entirely manually by two trained linguists – both native Ligurians and intimately familiar with the language – using the CoNLL-U Editor tool (Heinecke, 2019). A two-step process was used: a first pass of annotations was followed by discussions, after which both annotators went back separately over their initial annotations. Inter-annotator agreement was measured on a sample of 60 sentences from the fiction domain, which was found by the annotators to be by far the most complex and difficult part of the corpus to label. Agreement, calculated with Cohen's kappa, was 0.97 for POS tags and 0.84 for labelled dependencies in the first round. After the second round of annotations, agreement increased to 0.99 for POS tags and 0.97 for dependencies. One of the most frequent sources of disagreement involved the clitics *ghe* and *ne*, which are traditionally treated as adverbs but can sometimes be seen as demonstrative pronouns (Toso, 1997).

We discuss here the key aspects of the guidelines developed for the UD annotation of Ligurian, focussing on the analyses which might not be immediately self-evident. Some of these are exemplified in Fig. 1.

Tokenisation Tokenisation is performed by whitespace and punctuation, analogously to other Romance languages. Multi-word tokens are used for clitics (*andemmosene* → *andemmo se ne*, 'let us go away from here'; *pensâghe* → *pensâ ghe* 'to think about it';) as well as for adpositions fused with articles (*in sciô* → *in sce o* 'on the'; *do* → *de o* 'of the'; *a-a* → *à a* 'at the').

Articles They are marked by `PronType=Art`, and can be definite (`Definite=Def`, *o, a, l', e, i*) or indefinite (`Definite=Ind`, *un, unna, do, da, di, de*). They have grammatical gender and number.

Adjectives These can have grammatical gender, number, and degree. Comparative and superlatives which differ from their positive form (*megio, pezo*) are marked `Degree=Cmp`. Absolute superlatives (*braviscima* 'very good') are marked `Degree=Abs`. All other cases are denoted by the absence of the `Degree` feature.

Numerals Ordinal numerals are tagged `ADJ` with `NumType=Ord`, and have gender and number. Cardinals are tagged `NUM` with `NumType=Card`. Some cardinals – *un/uña* 'one', *doî/doe* 'two', *trei/træ* 'three' and their composites (e.g. *vintidoî* 'twenty-two' masc., *vintidoe* 'twenty-two' fem.) – also have grammatical gender.

Auxiliaries We mark as `AUX` the copular verbs *ëse* ('to be') and *stâ* (functionally equivalent to the Spanish 'estar') when functioning as copula; the passive auxiliaries *ëse* and *an(d)â* 'to go'; the tense auxiliaries *ëse* and *avei* 'to have'; and the passive auxiliary *vegnî* and *an(d)â*. We also mark as `AUX` the

modals *poei* ‘to be able to’, *dovei* ‘to have to’, *voei* ‘to want’, and *savei* ‘to know’, following the treatment of analogous verbs in Universal Dependencies treebanks of other Romance languages.

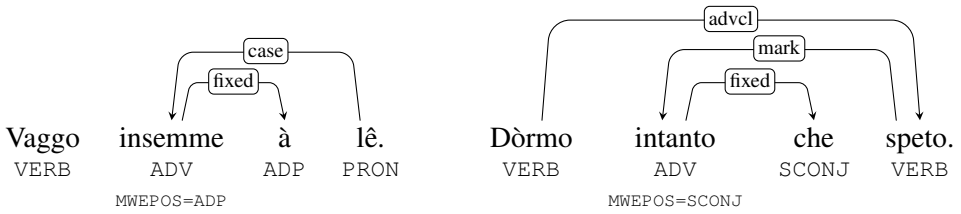

Figure 2: Grammaticalised multi-word expressions. *I go along with her* and *I sleep while I wait*.

Multi-word expressions Expressions which have undergone grammaticalisation are joined with the `fixed` relation. When the part-of-speech annotation of the head token does not match that of the expression as a whole, we use the additional annotation `MWEPOS` (in the `MISC` column of the CoNLL-U format) to indicate the part-of-speech of the expression as a whole. Examples of multi-word expressions include conjunctions (*intanto che* ‘while’, *de za che* ‘since’), adverbs (*de longo* ‘always’, *in derê* ‘behind’), and prepositions (*in sce* ‘on’, *in cangio de* ‘instead of’).

Clitic doubling In Ligurian, subject pronominal doubling is normally mandatory for the third person singular (*o Gioan o mangia* ‘Gioan eats’) and in some dialects for the third person plural (*i mæ amixi i mangian* ‘my friends eat’). In sentences where both the clitic and the lexical subject appear, the former is marked `expl`.

Ghe* and *ne The clitic *ghe*, when not acting as personal pronoun, is traditionally seen as an adverb, but can in many cases be interpreted as a demonstrative pronoun: *cöse ghe pòsso fâ?* (‘what can I do about that?’). The particle *ne* represents an analogous case. Due to the subtlety of these distinctions, it was decided that these clitics, when not acting as personal pronouns, would be tagged `ADV`.

Euphonic *l’* Whenever clitic doubling occurs, if the verb starts with a vowel it is usually preceded by the particle *l’* (*a lalla a l’ammia o mâ* ‘the aunt looks at the sea’). As it merely plays a euphonic role, we tag it `PART` and attach it to the verb with the relation `dep`.

Language-specific relations We use `expl:pv` for clitics attached to pronominal verbs (*assunnâse* ‘to dream’, *fâghela* ‘to achieve something’), `expl:impers` for the impersonal usage of the pronoun *se* (*in scî cotidien se parla de sti fæti* ‘these facts are being discussed in newspapers’), and `expl:pass` for all uses of *se* as passive marker (*d’autunno se mangia e rostie* ‘roast chestnuts are eaten in autumn’). Similarly to other treebanks, heads of relative clauses are attached to the nominals they modify via `acl:relcl` (Nivre et al., 2016).

4 Corpus statistics

The annotated corpus contains 316 sentences, 6 928 tokens (syntactic words), 1 563 unique surface forms, and 1 192 unique lemmas. Part-of-speech tag and dependency relation statistics for the annotated treebank are shown in Table 2.

In order to get an indication of the quality and consistency of the treebank’s annotations, we test the performance of a standard dependency parser trained on the corpus (Straka and Straková, 2017) using 10-fold cross-validation. The parser, which was trained using the default hyperparameters, achieves 100.0% F1 for tokenisation, 92.00% F1 for lemmatisation, 86.62% F1 for POS tagging, 83.45% F1 for feature prediction, 69.96% UAS and 60.74% LAS. While these scores are not as high as those commonly seen for high-resource languages, they compare favourably to the performance observed for other corpora of similar or even larger sizes (Straka and Straková, 2017; Jónsdóttir and Ingason, 2020, *inter alia*), confirming the consistency of the annotations. An exciting direction for future research would be to explore the possibility of boosting parsing performance via cross-lingual transfer on Italian or Spanish UD data.

Label	Count
det	849
punct	792
case	748
nsubj	416
advmod	412
obl	411
root	316
obj	287
nmod	267
mark	245
cc	243
conj	234
aux	216
amod	150
expl	149
dep	134
expl:pv	125
iobj	105
fixed	103
acl:relcl	102
cop	100
advcl	95
xcomp	95
parataxis	83
ccomp	53
flat	43
acl	35
discourse	23
expl:impers	23
appos	22
nummod	19
dislocated	16
csubj	8
vocative	6
orphan	3

(a) Dependency labels

Tag	Count	Example lemmas
PRON	928	*o che se ghe me*
ADP	904	*de à da in pe*
NOUN	896	*giorno paise parte gio çittæ*
DET	850	*o un quello tutto mæ*
PUNCT	792	*, . ! : ?*
VERB	762	*fâ ëse avei anâ dî*
ADV	459	*no ciù ben tanto ghe*
AUX	318	*ëse avei poei stâ voei*
CCONJ	240	*e ma ò ni comme*
ADJ	219	*bello antigo mæximo santo cao*
PROPN	189	*Zena Gexù Segnô Zane Arbâ*
SCONJ	148	*che se comme perché quande*
PART	136	*l'*
NUM	42	*doî eutto 1929 quaranta quattro*
INTJ	23	*scì ben eh ah no*
X	22	*Tintin aventures de del les*

(b) Part-of-speech tags

Table 2: Corpus annotation statistics

5 Conclusions

We have presented the first corpus of Ligurian annotated according to the Universal Dependencies framework, as well as a set of instructions for the annotation of the less trivial constructions. Additionally, to motivate our choice of linguistic variant and spelling system, we contributed an analysis of the dialects and orthographic standards of Ligurian, setting some guidelines which we hope will prove themselves useful for future contributions of corpora in this language. While the size of this corpus is small compared to the datasets of high-resource Romance languages such as French or Italian, it will now be possible to use this data to bootstrap any future Ligurian annotation efforts.

References

Andrea Acquarone. 2015. Scrivere la lingua. In Andrea Acquarone, editor, *Parlo Ciæo. La lingua della Liguria*, pages 87–94. De Ferrari / Il Secolo XIX, Genoa, Italy.

Graziadio Isaia Ascoli. 1876. Del posto che spetta al ligure nel sistema dei dialetti italiani. *Archivio glottologico italiano*, 2:111–160.

Fabio Canessa. 2016. L'inno naçionâ da Liguria. In Andrea Acquarone, editor, *Riso Ræo, l'antologia de Parlo Ciæo*. De Ferrari.

Arrigo Castellani. 1982. Quanti erano gl'italofoni nel 1861? *Studi linguistici italiani*, 8:3–26.

Tullio De Mauro. 1991. *Storia linguistica dell'Italia unita*. Laterza, Roma-Bari, Italy.

Friedrich Christian Diez. 1836. *Grammatik der romanischen Sprachen*, volume 1.

Werner Forner. 1997. Liguria. In Martin Maiden and Mair Parry, editors, *The dialects of Italy*, pages 198–225. Routledge, London, United Kingdom, and New York, United States of America.

Bruno Guillaume. 2019. Graph Matching for Corpora Exploration. In *JLC 2019 - 10èmes Journées Internationales de la Linguistique de corpus*, Grenoble, France, November.

Johannes Heinecke. 2019. ConlluEditor: a fully graphical editor for Universal dependencies treebank files. In *Universal Dependencies Workshop 2019*, Paris.

Bartolomeo Iacopone. 2017. Derê à scappâ. In *Vuxe de Ligüria*, volume 9. Comune di Pontedassio, Pontedassio, Italy.

Hildur Jónsdóttir and Anton Karl Ingason. 2020. Creating a parallel Icelandic dependency treebank from raw text to Universal Dependencies. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2924–2931, Marseille, France, May. European Language Resources Association.

Stefano Lusito. 2020. Lazarillo de Tormes. In *Cabirda. Lengue e lettiatue romanse*, volume 5. Zona, Genoa, Italy. Translation of an anonymous Spanish novella from 1554.

Stefano Lusito. 2021. Tipologie testuali e modalità di circolazione della prosa contemporanea in genovese. In Giuliano Bernini, Federica Guerini, and Gabriele Iannaccaro, editors, *La presenza dei dialetti italo-romanzi nel paesaggio linguistico: ricerche e riflessioni*, pages 155–174. Bergamo University Press / Sestante Edizioni.

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

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Fiorenzo Toso. 1995. *Storia linguistica della Liguria*. Le Mani, Recco, Italy.

Fiorenzo Toso. 1997. *Grammatica del genovese*. Le Mani, Recco, Italy.

Fiorenzo Toso. 2001. *Isole tabarchine. Gente, vicende e luoghi di unavventura genovese nel Mediterraneo*. Le Mani, Recco, Italy.

Fiorenzo Toso. 2002. Liguria. In Manlio Cortelazzo, Carla Marcato, and Nicola De Blasi, editors, *I dialetti italiani: storia, struttura, uso*, pages 245–252. UTET, Turin, Italy.

Fiorenzo Toso. 2003a. *Da Monaco a Gibilterra. Storia, lingua e cultura di villaggi e città-stato genovesi verso Occidente*. Le Mani, Recco, Italy.

Fiorenzo Toso. 2003b. *I Tabarchini della Sardegna. Aspetti linguistici ed etnografici di una comunità ligure doltremare*. Le Mani, Recco, Italy.

Fiorenzo Toso. 2009. *La letteratura ligure in genovese nei dialetti locali. Profilo storico e antologia*, volume 1. Le Mani, Recco, Italy.

Fiorenzo Toso. 2018. *A bocca do lô*. De Ferrari, Genoa, Italy. Translation of 1892 novel *La bocca del lupo* by G. Invrea.

Fiorenzo Toso. 2019. O santo evangëio segondo Marco. Forthcoming.

Fiorenzo Toso. 2020. I freschi de gexe do ponente, sensa destin. *O Stafî*, 2(2):4.