# SICILIAN3BANK

Translation and Annotation in Universal Dependencies Guidelines – v. 1.0

These are the annotation guidelines followed for the creation of the parallel treebank presented in the paper by Caterina Maria Cappello[1], Sabrina D'Alì[1], Mario Guglielmetti[1], Elisa Di Nuovo[2], Cristina Bosco[1], *under review* to the CLiC-it 2025 conference.

[1] *Università degli Studi di Torino*

[2] *European Commission, Joint Research Centre (JRC)*

# 1. INTRODUCTION

The guidelines presented here refer to the criteria adopted for the translation and annotation in Universal Dependencies (UD) of the texts contained in the Sicilian3bank. Due to the lack of standard references for Sicilian, as it happens for non-official languages, it was necessary to take some decisions, related especially to the resolution of ambiguous cases.

The annotation has been carried out starting from the output obtained using the parser UDPIPE, and the version 2.15 of the Italian ISDT model. A total of three annotators have been involved (the first three authors of the paper), one main annotator per text included in the resource. Each annotation was reviewed by another native-speaker annotator, and disagreement discussed in meetings with the three annotators, and occasionally with the other authors of the paper. To discourage subjective choices, annotators had to cross-check their decisions using various grammars (Baiamonte, 2024; Fortuna, 2002; Gerbino & Barone, 2011; Giacalone, 2009; Gorini, 2017; Lumia, 2010; Messina, 2007; Russo, 2003) and dictionaries (Biundi, 1851; Mortillaro, 1876; Rocca, 1839; Traina, 1868). In addition, we consulted some online sources, such as Wikizziunariu, Glosbe, Napizia-Chiù dâ Palora and Salviamo il siciliano.

Section 2 provides the translation guidelines. Section 3 the annotation guidelines for the application of the UD format.

# 2. TRANSLATION GUIDELINES

In the annotation of our treebank, we have chosen to include an additional comment line dedicated to translation. This translation is not intended to be a literal rendering of the original text, but rather a fluent and grammatically well-formed version in Italian. The aim is to provide a discursive translation that preserves the readability and syntactic naturalness of the target language, Italian specifically. The translation line is aligned 1:1 to the Sicilian sentences. At this stage of the project, we have

opted to retain proper names in Sicilian without translating them into Italian. This decision was made in light of the semantic complications that may arise from translating proper names and, given that producing an optimal translation is not the primary goal of this treebank, we have chosen to defer this issue for future consideration. Consequently, the current guidelines stipulate that proper names should remain untranslated. Exceptions are toponyms that have a direct equivalent in Italian and, for this reason, have been translated. Technically, the translation is inserted as a comment line immediately following the # text = line in the UD-format CoNLL-U file. This line begins with # translation = followed by the translated version of the sentence. The resulting translated sentences have been automatically annotated using UDPIPE and manual correction of this parallel treebank is left for future work. We report below some examples of the comment lines of the CoNLL-U file storing the Sicilian text and its corresponding translation into Italian. We decided to translate into Italian and not English, because despite it would open more the fruitability of the resource, it would reinforce the use of English as mainstream language, while we want to stress the importance of having a plurality of languages. Future work might involve the translation of these texts also to English.

```
# sent_id = 35
# text = Nuḍḍu di nuiautri sapìa soccu fari.
# translation = Nessuno di noi sapeva cosa fare.
```

```
# sent_id = 202
# text = Giuvannuzza si susìu, ma u dinocchiu l'abbannunau.
# translation = Giuvannuzza si alzò, ma il ginocchio l'abbandonò.
```

```
# sent_id = 118
# text = A Missina cc'era un picciriddru, figghiu di 'na lavannara.
# translation = A Messina c'era un bambino, figlio di una lavandaia.
```

# 3. APPLYING UD TO SICILIAN TEXTS

## 3.1 TOKENISATION

### Articulated Prepositions Annotation

The following table provides the articulated prepositions in the contracted form (signaled by the circumflex accent) that may appear in the treebank, along with their corresponding decomposition. The escape mark indicates that the same articulated preposition can be used for more than one feature (e.g. dî - di+lu - could be used with both masculine and feminine heads).

| Articulated prepositions | Composition | Lemmas | Feats |
|---|---|---|---|
| dû | di+lu | di+lu | Definite=Def\|Gender=Masc\|Number=Sing\|PronType=Art |
| dâ | di+la | di+lu | Definite=Def\|Gender=Fem\|Number=Sing\|PronType=Art |
| dî | di+li | di+lu | Definite=Def\|Gender=Masc/Fem\|Number=Plur\|PronType=Art |
| ô | a+lu | a+lu | Definite=Def\|Gender=Masc\|Number=Sing\|PronType=Art |
| â | a+la | a+lu | Definite=Def\|Gender=Fem\|Number=Sing\|PronType=Art |
| ê | a+li | a+lu | Definite=Def\|Gender=Masc/Fem\|Number=Plur\|PronType=Art |
| nô/nnô/ntô | ni+lu/nta+lu | ni+lu/nta+lu | Definite=Def\|Gender=Masc\|Number=Sing\|PronType=Art |
| nâ/nnâ/ntâ | ni+la/nta+la | ni+lu/nta+lu | Definite=Def\|Gender=Fem\|Number=Sing\|PronType=Art |
| nê/nnê/ntê | ni+li/nta+li | ni+lu/nta+lu | Definite=Def\|Gender=Masc/Fem\|Number=Plur\|PronType=Art |
| kû/cû | cu+lu | cu+lu | Definite=Def\|Gender=Masc\|Number=Sing\|PronType=Art |
| kâ/câ | cu+la | cu+lu | Definite=Def\|Gender=Fem\|Number=Sing\|PronType=Art |
| kî/chî | cu+li | cu+lu | Definite=Def\|Gender=Masc/Fem\|Number=Plur\|PronType=Art |
| pû | pi+lu | pi+lu | Definite=Def\|Gender=Masc\|Number=Sing\|PronType=Art |
| pâ | pi+la | pi+lu | Definite=Def\|Gender=Fem\|Number=Sing\|PronType=Art |
| pî | pi+li | pi+lu | Definite=Def\|Gender=Masc/Fem\|Number=Plur\|PronType=Art |

In Sicilian articulated prepositions can be formed also with indeterminative articles, e.g. *ôn*, consisting of a simple preposition and an indeterminative article (see the decomposition applied in CoNLL-U below).

```
7-8   ôn   _   _   _   _   _   _   _   _
7     a    a   ADP E   _   6   fixed _   _
8     nu   unu DET RD  Definite=Def|Number=Sing|PronType=Art   9
      det
```

## Contraction of clitic pronouns

The following table lists the pairs of pronouns that can be found in the contracted form (signaled by circumflex accent) that may occur in the text, along with their corresponding tokenisation. These contractions cannot be attached to the verb **dicimû*, but they can only occur separately, e.g. *mû dici* (*me lo dici*).

| Contraction of clitic pronouns | Compositio n | Feats |
|---|---|---|
| mû | mi + lu | Number=Sing\|Person=1\|PronType=Prs\|Clitic=Yes Gender=Masc\|Number=Sing\|Person=3\|PronType=Prs\|Clitic=Ye s |
| mâ | mi + la | Number=Sing\|Person=1\|PronType=Prs\|Clitic=Yes Gender=Fem\|Number=Sing\|Person=3\|PronType=Prs\|Clitic=Yes |
| mî | mi + li | Number=Sing\|Person=1\|PronType=Prs\|Clitic=Yes Gender=Fem/Masc\|Number=Plur\|Person=3\|PronType=Prs\|Cliti c=Yes |
| tû | ti + lu | Number=Sing\|Person=2\|PronType=Prs\|Clitic=Yes Gender=Masc\|Number=Sing\|Person=3\|PronType=Prs\|Clitic=Ye s |
| tâ | ti + | Number=Sing\|Person=2\|PronType=Prs\|Clitic=Yes |

| | | |
|---|---|---|
| | la | Gender=Fem\|Number=Sing\|Person=3\|PronType=Prs\|Clitic=Yes |
| tî | ti + li | Number=Sing\|Person=2\|PronType=Prs\|Clitic=Yes Gender=Fem/Masc\|Number=Plur\|Person=3\|PronType=Prs\|Clitic=Yes |
| sû | si + lu | Number=Sing/Plur\|Person=3\|PronType=Prs\|Clitic=Yes Gender=Masc\|Number=Sing\|Person=3\|PronType=Prs\|Clitic=Yes |
| sâ | si + la | Number=Sing/Plur\|Person=3\|PronType=Prs\|Clitic=Yes Gender=Fem\|Number=Sing\|Person=3\|PronType=Prs\|Clitic=Yes |
| sî | si + li | Number=Sing/Plur\|Person=3\|PronType=Prs\|Clitic=Yes Gender=Fem/Masc\|Number=Plur\|Person=3\|PronType=Prs\|Clitic=Yes |
| nû | ni + lu | Number=Plur\|Person=1\|PronType=Prs\|Clitic=Yes Gender=Masc\|Number=Sing\|Person=3\|PronType=Prs\|Clitic=Yes |
| nâ | ni + la | Number=Plur\|Person=1\|PronType=Prs\|Clitic=Yes Gender=Fem\|Number=Sing\|Person=3\|PronType=Prs\|Clitic=Yes |
| nî | ni + li | Number=Plur\|Person=1\|PronType=Prs\|Clitic=Yes Gender=Fem/Masc\|Number=Plur\|Person=3\|PronType=Prs\|Clitic=Yes |
| vû | vi + lu | Number=Plur\|Person=2\|PronType=Prs\|Clitic=Yes Gender=Masc\|Number=Sing\|Person=3\|PronType=Prs\|Clitic=Yes |
| vâ | vi + la | Number=Plur\|Person=2\|PronType=Prs\|Clitic=Yes Gender=Fem\|Number=Sing\|Person=3\|PronType=Prs\|Clitic=Yes |
| vî | vi + li | Number=Plur\|Person=2\|PronType=Prs\|Clitic=Yes Gender=Fem/Masc\|Number=Plur\|Person=3\|PronType=Prs\|Clitic=Yes |

## Verbs containing clitic pronouns

As foreseen in the UD guidelines, verbs containing one or more clitic pronouns are tokenised by systematically separating the clitics from the conjugated verb.

```
26-28  dimmillu  _  _  _  _  _  _  _
26  di  diciri  VERB  V
Mood=Imp|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin  16
parataxis  _  _
27  mi  mi  PRON  PC
Gender=Masc|Number=Sing|Person=1|PronType=Prs  26  iobj  _
Clitic=Yes
28  lu  lu  PRON  PC
Gender=Masc|Number=Sing|Person=3|PronType=Prs  26  obj  _
Clitic=Yes
```

```
10-12   Jamuninni        _       _       _       _       _       _       _       _
10      Jamu    jiri    VERB    V
Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin   6
parataxis       _       _
11      ni      ni      PRON    PC
Gender=Masc|Number=Sing|Person=3|PronType=Prs   10      iobj    _
Clitic=Yes
12      nni     ni      PRON    PC
Gender=Masc|Number=Sing|Person=3|PronType=Prs   10      obj     _
Clitic=Yes
```

One peculiarity in this respect about Sicilian is that in southern Sicilian there is the tendency to put the clitic pronoun *tu* (‘you’) in the indicative and subjunctive tenses, with the exception of the present and future tenses. Here too, therefore, we follow the tokenisation scheme outlined above, i.e. the division of verb and clitic pronoun into two tokens.

```
4-5     pinzàvatu       _       _       _       _       _       _       _
SpaceAfter=No
4       pinzàva         pinzariVERB     V
Mood=Ind|Number=Sing|Person=2|Tense=Imp|VerbForm=Fin    18
ccomp:reported  _       _
5       tu      tu      PRON    PE      Number=Sing|Person=2|PronType=Prs   4
nsubj   _       _
```

## Compound forms

During the annotation, we decided to break down some compound forms, such as *napocu* (shown in the example below), into two tokens. Decomposition was applied when the constituent elements presented a clear grammatical and functional autonomy, despite being written as a single word in the orthographic form. We did not treat this as “error”, adding in the MISC column the “CorrectSpaceAfter=Yes”, as it might be the sign of lexicalisation, like what happened to *ddassutta* (shown in the example below), which, despite it is straightforwardly recognisable as the composition of *dda* (*là*, ‘there’) and *sutta* (sotto, ‘down’), it is already in a more advanced phase of lexicalisation, with sources reporting it as lemma.

```
2-3     'napocu         _       _       _       _       _       _       _       _
2       'na     unu     DET     E
Definite=Ind|Gender=Masc|Number=Sing|PronType=Art       3       det     _
3       pocu    pocu    PRON    E       Gender=Masc|Number=Sing      10      obl     _
```

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | ddassutta | ddassutta | ADV | _ | _ | 5 | advmod | _ | _ |

## 3.2 LEMMATISATION

### Verbs

Verb lemmas must be reported in the infinitive form, as per standard conventions. In the current version of the treebank we have removed graphic accents from the lemma, e.g.

- *jucàvanu → jucari,*
- *passijàvanu → passijari,*
- *stricàvanu → stricari.*

### Determiners

All determiners have been lemmatised according to the standard, reducing them to their full canonical form. As a result, any abbreviations frequently occurring in the corpus have been normalised to their complete base form, in the masculine singular, as prescribed by the standard. Where necessary, the tables below indicate in bold the lemma used for those determiners. Below, we present tables listing the most common determiners found in the treebank, the lemmas are in bold. Given the high variation in Sicilian and the absence of a standardised form, these tables are not currently exhaustive. They may therefore be updated in the future to include additional forms.

| **Definite Articles** | **Indefinite Articles** |
|---|---|
| **Lu** / ’u / L’ (Masc/Sing) | **Unu** / ’nu (Masc/Sing) |
| La / ’a / L’ (Fem/Sing) | Un / ’n (Masc/Sing) |
| Li / ’i / L’ (Masc-Fem/Plur) | Una / ’na / ’nna (Fem/Sing) |

| **Possessives** | |
|---|---|
| Me’/ Mo’ / **Miu** | Masc/Sing |
| Me’ / Ma’ / Mia | Fem/Sing |
| Me’ | Masc-Fem/Sing-Plur |
| Miei | Masc/Plur |

| | |
|---|---|
| **Tuou** / To’ | Masc/Sing |
| Ta’ | Fem/Sing |
| To’ / Toi | Masc-Fem/Sing-Plur |
| So’ / **Suou** | Masc-Fem/Sing-Plur |
| Sa’ | Fem/Sing |
| Soi | Masc-Fem/Plur |
| **Nostru** | Masc/Sing |
| Nostra | Fem/Sing |
| Nostri | Masc-Fem/Plur |
| **Vostru** | Masc/Sing |
| Vostra | Fem/Sing |
| Vostri | Masc-Fem/Plur |
| **Suou** / So’ | Masc/Sing |
| Suou / So’ | Fem/Sing |
| So’ | Masc-Fem/Plur |

| **Demonstratives** | |
|---|---|
| **Chistu** / ’stu / Chissu / ’ssu | Masc/Sing |
| Chista / ’sta / Chissa / ‘ssa | Fem/Sing |
| Chisti / ’sti / Chissi / ‘ssi | Masc-Fem/Plur |
| **Chiddu** / ‘ddu | Masc/Sing |
| Chidda / ‘dda | Fem/Sing |
| Chiddi / ‘ddi | Masc-Fem/Plur |

| **Indefinites** | |
|---|---|
| Quàntu | Nenti |
| Tantu | Picca |
| Pocu | Quàrchi |
| Troppu | Certu |
| Nuddu | Quarchidunu |

| **Interrogatives** | |
|---|---|
| quali | chi |
| **cui** / cu’ | cuantu |

## Variants Annotation

In Sicilian, variants of the same lemma may be characterised by a doubled initial consonant, e.g. *ci* and *cci*, *ni* and *nni*, as shown in the examples below.

- Si scantava ca **ni** putìa succèdiri quarchi cosa, ca **cci** putìa tràsiri a mafia nta st'ammazzatina, e ca capaci ca **ci** putìa finiri videmma a iḍḍu accuḍḍì.
  - ni ni PRON
  - cci ci PRON
  - ci ci PRON

- [...] ma a parti chistu a nuḍḍu **ci** succidìu nenti.
  - ci ci PRON

- [...] u Zuccu **ni** sta 'spittannu!
  - ni ni PRON

- "Sparàgnati i palori, chi **nni** putìa sapiri iu?" dissi iḍḍu.
  - nni ni PRON

This is likely due to attempts to transfer phonetic aspects (long consonant) in the written form. Other variants due to different attempts to transcribe a phonetic aspect

(pronunciation of voiced retroflex stop), which we all lemmatised using *dd*, are *dd*, *ḍḍ, ddh, ddr*:

- *Iddu, iḍḍu, iddhu, iddru → iddu*

Furthermore, there may be instances where one or more sounds are omitted in a word, resulting in *apheresis* or *elision*. In such cases, we follow the same approach as before by normalising these forms to their extended variants, as shown in the following examples:

- *nciuria→ inciuria* (NOUN),
- *nfini → infini* (ADV),
- *assà → assai* (ADV),
- *diri → diciri* (VERB).

## 3.3 MORPHOLOGY

Following the guidelines adopted in UD, we did not annotate morphological features of invariable tokens. Thus, if a token is invariable by number, we have only noted the gender and *vice versa*. Instead, it was decided to maintain gender and number in the case of determinants, even if they might have several referents. When a determiner undergoes elision, and the final vowel is omitted, we have nevertheless annotated its morphological features, following what was done in VALICO (despite this is not done in the ISDT treebank).

- *granni → grande* (‘big’) – invariant for both gender and number

  «*Nardu*, quantu ti facisti **granni**!» (with a masculine singular referent, *Nardu*)
  granni granni ADJ A _ 25 amod _

  Pi fàrivi accapiri megghiu, cc'era Pippineḍḍu, ntisu "u Liotru", picchì avìa du’ *aricchi* accussì **granni** ca parìanu du’ paracqua; (with a masculine plural referent, *aricchi*)
  granni granni ADJ A _ 20 amod _

- *principi → principe* (‘prince’) – invariant for number

  Haju 'ntisu diri di certuni ca 'nna vota a Missina cci ji' un signuri 'rossu, unu di sti **principi**, ca pi passàrisi un crapicciu, macàri fannu mòriri un puvireddru[...] (with a masculine plural referent)
  principi principi NOUNS Gender=Masc 17 nmod _

  Ddu **principi** cci nni jiccò 'n'àutra[...] (with a masculine singular referent)
  principi principi NOUNS Gender=Masc 5 nsubj _

- *li* → *le/gli* (‘the.feminine/masculine.plural’) – used for both feminine and masculine referents (as shown in the examples)

  A viritati sulu una era: èramu tutti troppu nichi, sulu na cricca di picciotti, nun sapìamu comu funziunijàvanu **li** *cosi*, comu funziunijava lu munnu.
  li lu DET RD
  Definite=Def|Gender=Fem|Number=Plur|PronType=Art 23 det _

  Chistu avia summuzzatu nni tutti **li** *gurfi* di lu munnu, e ddoppu avilli firriatu tutti, vinni a Siculiana.
  li lu DET RD
  Definite=Def|Gender=Masc|Number=Plur|PronType=Art 7 det _

  Giuvannuzza ci fici **l'***occhi* duci.
  l' lu DET RD
  Definite=Def|Gender=Masc|Number=Plur|PronType=Art 5 det _

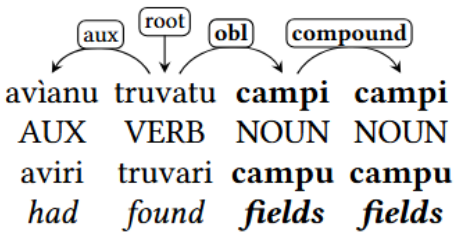
  Cc'era 'na corda cu 'na campana fora di **l'***acqua* [...]
  l' lu DET RD
  Definite=Def|Gender=Fem|Number=Sing|PronType=Art 11 det _

## 3.4 SYNTAX

### Reduplication

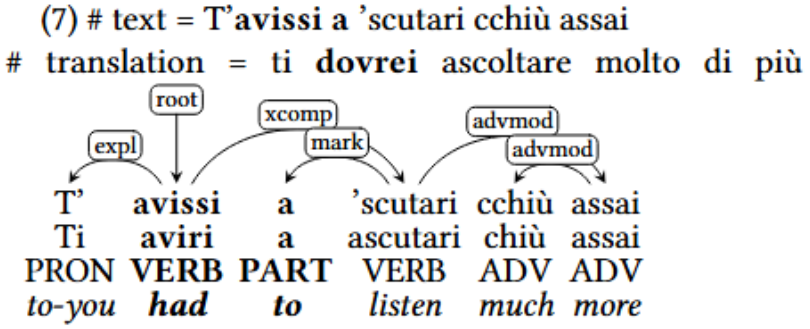
Reduplication is a phenomenon typical of Sicilian, as well as other languages, which consists in the repetition of a word, resulting in a shift or extension of meaning within the sentence. We use the relation *compound* and the relation *obl*, in line with UD guidelines. In addition we added *LOC=adv* in the last column of the CoNLL-U file to indicate that there is an adverbial locution, as done in VALICO.

(6) # text = avìanu truvato **campi campi**
# translation = avevano trovato **tra i campi**



### The periphrases ‘*aviri a* + infinitive’ and ‘*veniri a* + *diciri*’

The annotation in UD of such a resource allows for drawing a parallel with other languages, like the English *have to* construction, which is similarly used to express obligation and certainty. In the English UD treebanks ‘to’ is consistently annotated as a particle when used in this way. We therefore decided to treat the element ‘*a’*, which is usually tagged as a preposition in our corpus, as a particle in this specific construction.




Another periphrastic construction found in the treebank texts is *veniri a* + *diciri* (literal translation into Italian *venire a dire*), which can have the meaning of the Italian verb significare (‘to mean’). In such cases, we annotated *a* as particle to highlight the grammaticalisation of such structure.

# Bibliography

Baiamonte, S. (2024), *Documento per l’ortografia del siciliano. Documentu pi l’ortugrafìa dû sicilianu. Seconda edizione*, Cademia Siciliana.
Biundi, G. (1851), *Vocabolario manuale completo siciliano-italiano : seguito da un'appendice e da un elenco di nomi proprj siciliani: coll'aggiunta di un dizionario geografico in cui sono particolarmente descritti i nomi di città, fiumi, villaggi ed altri luoghi rimarchevoli della Sicilia: e corredato di una breve grammatica per gl'Italiani*, Palermo, Stamperia Carini.
Fortuna, A. (2002), *La Grammatica Siciliana: Principali regole grammaticali, fonetiche e grafiche*, Caltanissetta, Terzo Millennio Editore.
Gerbino, G. & Barone, N. (2011), *Cenni di ortografia siciliana*, Trapani, Jò A.L.A.S.D..
Giacalone, F. (2009), *Prammatica Siciliana. Storia della nostra lingua, proverbi, curiosità, modi di dire, consigli pratici per una corretta scrittura*, Trapani, Edizioni Colorgrafica.
Gorini, M. (2017), *Ortografia siculo-calabra*, https://michelegorini.blogspot.com/2017/08/ ortografia-siculo-calabra.html.
Lumia, V. (2010), *La Nostra Grammatica Siciliana*, Trapani, Jò A.L.A.S.D..
Messina, A. (2007), *Grammatica sistematica della lingua siciliana. Dall’ortoepia all’ortografia. Dall’analisi grammaticale all’analisi logica e del periodo. Con antologia esemplificativa dei poeti. Seconda edizione riveduta e ampliata con 30 chine sui mestieri d’una volta eseguite da Francesco Nania e Poesie*.

Mortillaro, V. (1876), *Nuovo dizionario siciliano-italiano*, volume unico, terza edizione, Palermo, Stabilimento tipografico Lao.
Rocca, R. (1839), *Dizionario siciliano-italiano compilato su quello del Pasqualino con aggiunte e correzioni*, volume unico, Catania, Pietro Giunti Editore.
Russo, N. (2003), *Corso di grammatica siciliana*.
Traina, A. (1868), *Nuovo vocabolario siciliano-italiano*, volume 1, Lauriel.

# Sitography

Glosbe, *Dizionario*, https://it.glosbe.com/
Lingua siciliana, *Come scrivere in siciliano*, https://linguasiciliana.com/come-scrivere-in-siciliano/
Napizia, *Chiù dâ Palora*, https://www.napizia.com/cgi-bin/cchiu-da-palora.pl
Salviamo il siciliano, *Dizionario*, http://www.salviamoilsiciliano.com/come-si-dice/dizionario/
Wikizziunariu, *Lu dizziunariu lìbbiru*, https://scn.wiktionary.org/wiki/P%C3%A0ggina_principali