

LUDViC

MANUAL OF DATA DIGITAL EDITION

Fernanda Pratas

Raïssa Gillier

December 2021

Index

0. Introduction.....	3
1. <teiHeader>.....	3
1.1. <fileDesc>.....	3
1.1.1. <titleStmt>.....	3
1.1.2. <editionStmt>.....	4
1.1.3. <publicationStmt>.....	4
1.1.4. <recordingStmt>.....	5
1.2. <profileDesc>.....	5
1.2.1. <particDesc>.....	5
1.2.2. <textClass>.....	6
2. <text>.....	8
2.1. Utterance: <u>.....	8
2.2. Vocalizations: <vocal>.....	9
2.3. Repetitions and Reformulations:	9
2.4. Incident: <incident>.....	9
2.5. Extralinguistic information: <kinesic>.....	9
2.6. Speech overlaps: <overlap>.....	10
2.7. Unclear sequences: <unclear>.....	10
2.8. Incomprehensible sequences: <gap>.....	10
2.9. Foreign sequences: <foreign>.....	10
3. Edition of the materials on TEITOK.....	11
3.1. Text tokenization.....	11
3.2. Tokens' edition.....	12
3.3. Transcription <i>versus</i> Edited Version.....	13
3.4. <dtok>.....	13
3.5. Linguistic Annotation.....	14
References:.....	14

0. Introduction

The criteria adopted in the digital edition of LUDViC materials are briefly presented in this manual, in what concerns the XML encoding (points 1 and 2) and the editing on the TEITOK platform (point 3).

Data were encoded in XML-TEI format, in accordance with the rules proposed by the Text Encoding Initiative (TEI) Consortium for editing digital texts.

Each XML file is composed of two main parts: the *header* (<teiHeader>), which gathers the metadata of each interview, and the *text* (<text>) that contains the orthographic transcription of the interview, including the encoding of some typical events of oral speech (repetitions, overlaps, hesitations, etc.).

The elements that make up each of these parts are described in the following sections.

1. <teiHeader>

The <header> is composed of two main parts: <fileDesc> groups the bibliographic information of each file, and <profileDesc> describes the information related to the participants of each interview.

1.1. <fileDesc>

This element is composed of <titleStmt>, <editionStmt>, <publicationStmt>, and <recordingStmt>.

1.1.1. <titleStmt>

It contains the element <title type="main">, which content corresponds to the project's name, and the element <title type="short">, that concerns to the title of each interview.

Example:

<title_tei type="main">LUDViC - Language Unity and Diversity: Variation in Caboverdean and beyond</title_tei>

<title_tei type="short">Clarisse on her old life and work in Fogo and how she went to live in Sal.</title_tei>

<funder> refers to the project's financing entity. Some materials come from the previous project *Events and Subevents in Caboverdean*¹, so in these cases both funding references are mentioned.

Example:

<funder>Fundação para a Ciência e a Tecnologia, IF/00066/2015</funder>

<funder>Fundação para a Ciência e a Tecnologia, IF/00066/2015; Fundação para a Ciência e a Tecnologia, PTDC/CLE-LIN/103334/2008</funder>

¹ The project *Events and Subevents in Caboverdean* (PTDC/CLE-LIN/103334/2008) was developed at the Centro de Linguística da Universidade Nova de Lisboa, from March 2010 to February 2013, under the coordination of the researcher Fernanda Pratas (<https://clunl.fcsh.unl.pt/en/investigacao/projetos-concluidos/eventos-subeventos-caboverdiano/>)

<principal> mentions the researcher responsible for the project, and its content is invariable.

Example:

```
<principal id="FP">Fernanda Pratas</principal>
```

<respStmt> identifies the persons responsible for the different tasks performed when editing the materials through the respective **<name>** and **<resp>** elements. The **@id** attribute, included in **<name>**, refers to the identification abbreviation of each person.

Example:

```
<respStmt>
  <name id="MG">Mireida Gomes</name>
  <resp>Transcription</resp>
</respStmt>
<respStmt>
  <name id="RG">Raïssa Gillier</name>
  <resp>XML-TEI encoding</resp>
  <resp>Orthographic uniformization</resp>
  <resp>POS tagging</resp>
</respStmt>
<respStmt>
  <name>Fernanda Pratas</name>
  <resp>Revision</resp>
</respStmt>
```

1.1.2. <editionStmt>

It refers to the type of edition and contains the project's website address.

Example:

```
<editionStmt>
  <edition>Electronic Edition<link target="teitok.clul.ul.pt/ludvic/"></edition>
</editionStmt>
```

1.1.3. <publicationStmt>

The elements **<publisher>** and **<pubPlace>** refer to the entity responsible for publishing the materials and its place of publication.

The element **<availability status="restricted">** indicates that the use of the materials is protected and subject to the conditions determined by the Creative Commons license, which is referred to in **<licence @target>**.

Example:

```
<publicationStmt>
  <publisher>CLUL - Centro de Linguística da Universidade de Lisboa</publisher>
  <pubPlace>Lisboa</pubPlace>
  <availability status="restricted">
    <licence target="https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode">Creative
  Commons</licence>
  </availability>
</publicationStmt>
```

1.1.4. <recordingStmt>

The element **<media>** contains the attribute **@mimeType** that indicates the file type and its extension via the **@mimeType** attribute; the **@url** attribute specifies the file name.

The element **<equipment>** refers to the equipment used to record the materials.

The element **<date>** mentions the year in which the interview was carried out.

The element **<settingDesc>** indicates the location and the district where the interview was carried out through the respective **<name type="place">** and **<name type="district">** elements.

Example:

```

<recordingStmt>
  <recording>
    <media mimeType="audio/mp3" url="Inês.mp3"/>
    <equipment>
      <p>TASCAM DR-100 recorder, with integrated microphone</p>
    </equipment>
    <date>2018</date>
    <settingDesc>
      <setting>
        <name type="place">São Vicente</name>
        <name type="district">São Vicente</name>
      </setting>
    </settingDesc>
  </recording>
</recordingStmt>

```

1.2. <profileDesc>

This element is composed by <particDesc> and <textClass>.

1.2.1. <particDesc>

The element **<person>** identifies the interview participants, whose role is specified through the attribute **@role="interviewer"** or **@role="participant"**. When this element concerns the participant, it also includes the attribute **@sex**.

The element **<personName>** contains the name of each of the participants. Participants' names are fictitious, in order to preserve their anonymity.

Example:

```

<person role="interviewer">
  <persName>Helderyse Évora</persName>
</person>
<person role="participant" sex="M">
  <persName id="Firmino">Firmino</persName>
</person>

```

When `<person>` refers to the participant, other elements relevant to the characterization of their sociolinguistic profile are included:

`<age>` specifies the participant's age, at the date of the interview, through the attribute `@value`.

`<birth>` refers to the participant's year of birth, through the attribute `@when`, and the name of the place where he/she was born.

`<nationality>` identifies the nationality of the participant.

`<education>` concerns the participant's level of education, which is referred to as one of four levels established: *null* (no education), *elementary* (primary education), *secondary* (secondary education) or *superior* (higher education).

`<occupation>` indicates the participant's profession.

`<residence>` identifies the location where the participant resides.

`<langKnowledge>` refers to the participant's knowledge of other languages. Given the specific situation of diglossia in Cabo Verde, participants with origin or residence in this country will always have this field filled in with *contact with Portuguese*, besides mentioning other languages.

`<langUsage>` contains the element `<language>` that refers to the participant's linguistic variety. To designate each linguistic variety, the name of each of the islands of the Cabo Verdean archipelago was adopted (Fogo, São Vicente, etc.). This element includes the attribute `@ident="kea"`, which identifies the Caboverdean language in the standardized nomenclature ISO 639-3. It also includes the attribute `@type` that specifies whether the participant is a first language or heritage language speaker: *L1 – First Language* or *HL – Heritage Language*.

`<event>` includes information related to the participant's possible stays on other islands or abroad.

Example:

```

<person role="participant" sex="M">
  <persName xml:id="Adriano">Adriano</persName>
  <age value="51"/>
  <birth when="1960"/>
  <nationality>Cabo Verdean</nationality>
  <education>elementary</education>
  <occupation>craftsman</occupation>
  <residence>Sal</residence>
  <langKnowledge>contact with Portuguese</langKnowledge>
  <langUsage>
    <language ident="kea" type="L1 (First Language)">Brava</language>
  </langUsage>
  <event>He lives in Sal for 12 years.</event>
</person>

```

1.2.2. `<textClass>`

It contains the element `<keywords>`, within which the keywords related to the main topics covered during the interview are mentioned.

Example:

```

<textClass>
  <keywords>Caboverdean language, emigration, family, New England</keywords>
</textClass>

```

Example of a complete <teiHeader>:

```

<teiHeader>
  <fileDesc>
    <titleStmt>
      <title_tei type="main">LUDVIC - Language Unity and Diversity: Variation in
Caboverdean and beyond</title_tei>
      <title_tei type="short">Pablo on his adaptation to the USA, on his studies and his
future professional plans, and on the differences between Caboverdean dialects</title_tei>
      <funder>Fundação para a Ciência e a Tecnologia, IF/00066/2015</funder>
      <principal id="FP">Fernanda Pratas</principal>
      <respStmt>
        <name id="MG">Mireida Gomes</name>
        <resp>Transcription</resp>
      </respStmt>
      <respStmt>
        <name id="RG">Raïssa Gillier</name>
        <resp>XML-TEI encoding</resp>
        <resp>Orthographic Uniformization</resp>
        <resp>POS tagging</resp>
      </respStmt>
      <respStmt>
        <name>Fernanda Pratas</name>
        <resp>Revision</resp>
      </respStmt>
    </titleStmt>
    <editionStmt>
      <edition>Electronic edition<link target="teitok.clul.ul.pt/ludvic"/></edition>
    </editionStmt>
    <publicationStmt>
      <publisher>CLUL - Centro de Linguística da Universidade de Lisboa</publisher>
      <pubPlace>Lisbon</pubPlace>
      <availability status="restricted">
        <licence target="https://creativecommons.org/licenses/by-nc-
sa/4.0/legalcode">Creative Commons</licence>
      </availability>
    </publicationStmt>
    <recordingStmt>
      <recording>
        <media mimeType="audio/mp3" url="Pablo.mp3"/>
        <equipment>
          <p>TASCAM DR-100 recorder, with integrated microphone.</p>
        </equipment>
        <date>2019</date>
        <settingDesc>
          <setting>
            <name type="place">Boston</name>
            <name type="district">Massachusetts</name>
          </setting>
        </settingDesc>
      </recording>
    </recordingStmt>
  </fileDesc>

```

```

<profileDesc>
  <particDesc>
    <listPerson>
      <person role="interviewer">
        <persName>Fernanda Pratas</persName>
      </person>
      <person role="participant" sex="M">
        <persName id="Pablo">Pablo</persName>
        <age>21</age>
        <birth>1997</birth>
        <nationality>Cabo Verdean</nationality>
        <education>secondary</education>
        <occupation>student</occupation>
        <residence>Boston</residence>
        <langKnowledge>contact with Portuguese and with English</langKnowledge>
        <langUsage>
          <language ident="kea" type="L1 (First Language)">Santo
Antão</language>
        </langUsage>
        <event>He lives in Boston for one year.</event>
      </person>
    </listPerson>
  </particDesc>
  <textClass>
    <keywords><term>emigration, language, food</term></keywords>
  </textClass>
</profileDesc>
</teiHeader>

```

2. <text>

This element contains the orthographic transcription of the interview. The tags used to encode some textual aspects are presented below, such as phenomena of oral speech (repetitions, dropouts, vocalizations, production overlaps, etc.), extralinguistic events (gestures, speech disturbances, etc.), incomprehensible sequences, among others.

2.1. Utterance: <u>

The speech of each participant is delimited by the element <u>, which attribute @who specifies its interlocutor. The inquirer is always referred to as INQ.

Example:

```
<u who="INQ">I bosê ta stod senpr li na Espargos o bosê ta bai pa Santa
Maria?</u>
```

```
<u who="Carla">Nãu, só na Espargos.</u>
```

```
<u who="INQ"><gap reason="inaudible"/> dret?</u>
```

```
<u who="Carla">Mas o menus.</u>
```


2.2. Vocalizations: <vocal>

The <vocal> element is used to encode filled pauses (eh, ah, etc.) and other vocalizations without lexical content (coughing, etc.). The same element is used to encode other events that may be discursively relevant, such as laughter, singing, etc..

Example:

<u who="Firmino">N naxsê na Santanton mas <vocal>eh</vocal> sink one N tiv k bá pa Sonsent.</u>

<u who="Liziane">Mi, N ka ta skesê ese koza li <vocal>[laughing]</vocal>.</u>

2.3. Repetitions and Reformulations:

The element is used to encode repeated or reformulated sequences, which attribute **@reason** specifies whether it refers to a reformulation or a repetition.

Example:

<u who="Adriano">Mas, pruntu, <del reason="reformulation">Djabraba mas N ka ta skesê nos Djabraba.</u>

<u who="José">Ka faltaba txeu pa noventa, mas <del reason="repetition">ka ka dá pa txiga lá.</u>

2.4. Incident: <incident>

The <incident> element encodes extralinguistic events that interfere and/or disturb communication.

Example:

<u who="INQ"><incident>[someone interrupts]</incident> <gap reason="inaudible"/> bosê podi fala.</u>

<u who="Gabriela">Agó jent skesê <incident>[someone else talks]</incident> agó jent skesê té d skrevê.</u>

2.5. Extralinguistic information: <kinesic>

The <kinesic> element is used to encode non-verbal events that are part of the communication, such as facial expressions or gestures.

Example:

<u who="Inês">Ahan. <kinesic>[moved]</kinesic></u>

<u who="Adriano">Enton loja ta fika asin <kinesic>[he makes a gesture]</kinesic>, mi, N ta fika na kel otu lugar sintadu.</u>

2.6. Speech overlaps: <overlap>

The <overlap> element is used to encode speech overlaps.

Example:

```
<u who="Lurdes">Sin, nha konxe-l, ki ten un varanda <overlap>grandi</overlap>.</u>
```

```
<u who="INQ"><overlap>Grandi</overlap>.</u>
```

2.7. Unclear sequences: <unclear>

The <unclear> element is used to encode sequences whose transcription is uncertain, either because it is an unknown phonetic sequence or because some event makes hearing difficult, such as the existence of noise.

Example:

```
<u who="Aldina">Si bu ten un <unclear>raka</unclear>, bu ta poi o mandioka o batata.</u>
```

```
<u who="Gabriela">Kel ves mi era branka, jent tá ta vrá preta pa el ta bá ta <unclear>franji</unclear>?</u>
```

2.8. Incomprehensible sequences: <gap>

The element <gap reason="inaudible"> is used to encode incomprehensible sequences that are, therefore, impossible to transcribe.

Example:

```
<u who="Edson"> (...) mas situason finanseira stá un bokadinhu asin <gap reason="inaudible"/> djá <gap reason="inaudible"/> es ka ta djuda (...)</u>
```

```
<u who="Clarisse"> (...) pamó kel fumu ben fiká riba nha kaza asin na São Filipe, na bila <gap reason="inaudible"/>.</u>
```

2.9. Foreign sequences: <foreign>

The <foreign> element encodes sequences that occur in languages other than Caboverdean, which attribute @lang specifies the language in question.

This element is not used to encode 'loan words'.

Example:

```
<u who="Letícia"><foreign lang="en">So</foreign>, kuand mia mãi i nha pai ben pa Merka, en ben na un 1967.</u>
```

```
<u who="Rafaela">Mas ten txeu é k ta deskonkordá k mi, <foreign lang="en">it's ok</foreign>, mas, mi, N ta sinti-l dent d min.</u>
```

Example of <text>:

```

<text id="Miguel">
  <u who="Miguel">Ami <del reason="repetition">é</del> é trabaui di orta,
  trabaui di limária, trabaui di fora.</u>
  <u who="INQ">Trabadja tudu?</u>
  <u who="Miguel"><vocal>[laughing]</vocal> Tudu N ta faze. Agó nau, fora
  <del reason="repetition">iá</del> iá N para, fora <del
  reason="repetition">nha</del> nha mininus ka krê.</u>
  <u who="INQ">Kantu anu ki nho djá tene?</u>
  <u who="Miguel">Oitenta anu.</u>
  <u who="INQ">Stá na ora di <overlap>diskansa</overlap>.</u>
  <u who="Miguel"><overlap>Diskansa</overlap>.
  <vocal>[laughing]</vocal></u>
  <u who="INQ"><vocal>Eh</vocal> kuzé ki nho ta kultiva <del
  reason="repetition">na</del> na orta?</u>
  <u who="Miguel">Kana, batata, mandioka, tudu kuza.</u>
  <u who="INQ">Nho ta fazê grogu?</u>
  <u who="Miguel">N ta faze. <vocal>[laughing]</vocal></u>
  <u who="INQ">Di terra?</u>
  <u who="Miguel">Di kana, sin. <vocal>[laughing]</vocal></u>
  <u who="INQ">Mas nhu nase <del reason="repetition">i</del> i kria li,
  senpri?</u>
  <u who="Miguel"><vocal>Nah</vocal> N nase na Santa Catarina,
  Somada.</u>
  <u who="INQ">Na Somada?</u>
  <u who="Miguel"><vocal>Aian</vocal>, ki N nasê. N ben kiria li, N kiria li, li
  asin na kes kaza, uns pardueru ki stá lá. Kel pardueru lá, lá ki N kiria.
  <vocal>[laughing]</vocal> Nha mai era senpri di li. Nha pai ki era di Santa Catarina.
  Lá di uns kau, lá na undi ki ta fladu Xã di Tanki.</u>
  <u who="INQ"><vocal>Ah</vocal> djá N obi fala.</u>
  <u who="Miguel">Xã di Tanki. <incident>[someone else talks]</incident> Xã di
  Tanki. Lá djá, N ben ta ben: N ta bai, N ta bá trabaia azágua, djá kel oki fin di azágua,
  nu ta ben <del reason="repetition">pa</del> pa sidadi ti tenpu txuba, nu ta bai; nu ta
  buska só karga, nu ta ben. <del reason="reformulation">Era moda</del> ka tinha
  moda di karru, era moda burru, na <del reason="repetition">burru</del> burru.</u>
  [...]
</text>

```

3. Edition of the materials on TEITOK

The second phase of editing the materials takes place directly on TEITOK platform, developed for the creation, visualization and editing of corpora with textual markup and linguistic annotation (Janssen 2014).

3.1. Text tokenization

Once encoded with TEI-XML tags, the XML file is imported into the TEITOK platform, as well as the corresponding audio file. Tokenization, performed automatically, consists of

delimiting any textual element (words or punctuation) with the `<tok></tok>` element, which obligatorily includes an `@id` attribute.

Example of sentence before and after tokenization:

`<u who="Daiana">Trabadju di skola, pamó mi é profesora di EBI.</u>`

`<u who="Daiana" id="u-1"><tok id="w-1">Trabadju</tok> <tok id="w-2">di</tok> <tok id="w-3">skola</tok><tok id="w-4">,</tok> <tok id="w-5">pamó</tok> <tok id="w-6">mi </tok> <tok id="w-7">é</tok> <tok id="w-8">profesora</tok> <tok id="w-9">di</tok> <tok id="w-10">EBI</tok><tok id="w-11">.</tok></u>`

3.2. Tokens' edition

The encoding system used in TEITOK consists of representing all types of information (metatextual and linguistic) through `<tok>` attributes. Thus, the XML-TEI tags used to encode discursive events correspond to a specific attribute in TEITOK. Table 1 establishes the correspondence between the TEI tags and the respective `<tok>` attributes.

TEI tags	TEITOK attributes
<code><vocal></code>	<code>@vcform</code>
<code></code>	<code>@delform</code>
<code><incident></code>	<code>@inform</code>
<code><kinesic></code>	<code>@kiform</code>
<code><overlap></code>	<code>@ovform</code>

Table 1 – Correspondence between TEI tags and TEITOK attributes

The encoding of the attributes above has a direct correspondence in the visualization of the text through a set of buttons that can be selected according to the research objectives of each user.

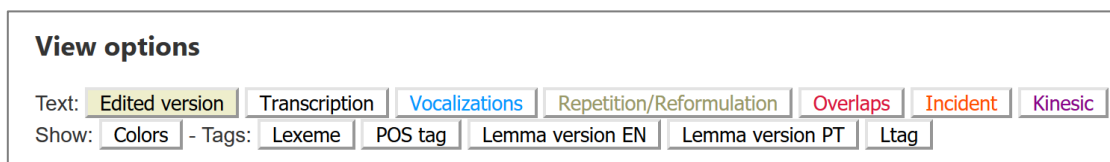


Figure 1 – Text view options in TEITOK platform

The remaining TEI tags, mentioned in points 2.7. to 2.9., don't have corresponding attributes since those encoded events are visualized directly on the platform. Unclear sequences (`<unclear>`) are displayed in a different color from the rest of the text, and

incomprehensible sequences (<gap>) are represented by [...]. The purpose of encoding sequences in languages other than Caboverdean (<foreign>) is to exclude these excerpts from the morphosyntactic annotation.

Examples:

El manda-m un pé asin BUUM, da-m un pé na boka d stom, dá un kola na pared.

Nu ta ten gana pa algen grandi bá nos kaza [...] pamodi nu ta saí.

3.3. Transcription versus Edited Version

The Edited Version consists of a 'clean' version of the Transcription, that is, it results from the elimination of the phenomena of oral speech present in the Transcription.

Example:

Transcription

Es ten uns koza k tip, es ten uns koza k es ta dzê k asves N ta feká ta spiá, N ta feká ta dzê "uk é k es dzê? Uk é kes kes palavra?".

Edited Version

Tip, es ten uns koza k es ta dzê k asves N ta feká ta spiá, N ta feká ta dzê "uk é k es dzê? Uk é kes palavra?".

To hide an element from the Edited Version, the **@form** attribute is filled in with two dashes. (--).

Token value (w-345): Es		
pform	Transcription (Inner XML)	Es
nform	Edited version	
form	Written	--
vcform	Vocalizations	
delform	Repetition/Reformulation	Es
ovform	Overlaps	
inform	Incident	
kiform	Kinesic	

Figure 2 – Procedure for hiding elements from the Edited Version

3.4. <dtok>

In the case of contracted forms (e.g. *né, del*) or enclitic verbal forms (e.g. *da-m; traze-l; akonpanha-s*), the <tok> element is 'splited' into two <dtok> so that each one of the simple lexical units that make up this complex sequence can receive the different layers of annotation. Each <dtok> contains the @form attribute filled in with the form of each of these units when they occur in their free form.

Example:

`<tok id="w-4210">oia-m`

`<dtok id="d-4210-1" form="oiá" lemma="odja" lemmaen="see" lemmapt="ver" pos="V" ltag="dj-j-i, fin-stress"/>`

`<dtok id="d-4210-2" form="mi" lemma="mi" lemmaen="1SG" lemmapt="1SG" pos="PPl.1SG" />`

`</tok>`

3.5. Linguistic Annotation

The linguistic annotation applies only to the Edited Version. The tagging levels include:

- lexeme
- POS (*part of speech*)
- English translation of the lexeme
- Portuguese translation of the lexeme
- Itag (phonetic-phonological phenomena)

For the first two levels, see the LUDViC Annotation Manual (lexematization and POS) at: http://teitok.clul.ul.pt/ludvic/pdf/Manual%20POS%20tagging_EN.pdf

For the Itag level, see: http://teitok.clul.ul.pt/ludvic/pdf/Itags_LUDViC_EN.pdf

As for the translation of lexemes into English and Portuguese, created specifically for this corpus (no other corpus in this platform has so far offered this functionality), it has three complementary purposes; these translations: (i) facilitate the future design of glosses to accompany any corpus data in linguistic works by other authors; (ii) make the corpus fully accessible to users who are not speakers of Caboverdean, not only by providing the immediate contextual meaning to each lexical item, but also because (and this was perhaps the most decisive fact when we decided to create these two levels) the searches can now be performed using lexical items in these two languages as a starting point; finally, they (iii) become indeed very useful also for Caboverdean-speaking users, who may be unsure about the spelling of the internal variants (these are used in the Word/Palavra field) or their lexemes.

Except for the linguistic tags (Itag), the annotation of the remaining levels is performed automatically using the NeoTag annotator (Janssen 2012). Manual revision is, however, essential to correct errors or to add missing information, since data specificity, which include different spelling variants, affects the performance of the annotator.

After running the annotator, the attributes corresponding to each of the linguistic annotation levels are generated, as shown in the examples below.

Examples:

```
<tok id="w-212" lemma="ningen" pos="PRON" lemmaen="nobody" lemmapt="ninguém">ningen</tok>
```

```
<tok id="w-771" lemma="kria" pos="V:PART" lemmaen="raise" lemmapt="criar" ltag="a-o, C-V fin">kriod</tok>
```

References:

Janssen, Maarten (2012). NeoTag: a POS tagger for grammatical neologism detection. In Proceedings of LREC 2012, Istanbul.

Janssen, Maarten (2014). TEITOK: Text-Faithful Annotated Corpora. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium, 2019. Available at: <https://tei-c.org/guidelines/P5/>