Inception:
How to provide correct input

Linguistic Resources for Natural Language Processing
LM Language Technologies and Digital Humanities
2024-25

Cristina Bosco

Missing tokens

For some sentences the automatic analysis (generated by UDPipe or other tools) is not correct since it does not include all the tokens it should include.

Missing tokens

For example, the analysis provided by UDPipe for the following Italian sentences is incorrect since at least one of the clitics encompassed in the word in bold (that begins with the verb) is missing:

1) Ho comprato un libro interessante e Giorgio ha detto di **mandarglielo** presto.
(*I bought a book interesting and Giorgio has tell to* ***send-him-it*** *soon*)

2) **Avvisatelo**!
(***Warn-him***!)

Missing tokens: Ex. I

Ho comprato un libro interessante e Giorgio ha detto di **mandarglielo** presto.
(*I bought a book interesting and Giorgio has tell to* ***send-him-it*** *soon*)

The morphological analysis of the word **mandarglielo i**ncludes 3 tokens:

mandar VERB
glie PRON
lo PRON

Missing tokens: Ex. I

Ho comprato un libro interessante e Giorgio ha detto di **mandarglielo** presto.
(*I bought a book interesting and Giorgio has tell to* ***send-him-it*** *soon*)

According to the UD format the expected annotation is as follows (also including an extra line for the entire word):

```
11 - 13  mandarglielo     < extra line
11  mandar VERB
12  glie PRON
13  lo PRON
```

Missing tokens: Ex. I

Ho comprato un libro interessante e Giorgio ha detto di **mandarglielo** presto.
(*I bought a book interesting and Giorgio has tell to* ***send-him-it*** *soon*)

The analysis provided by UDPipe **only includes 2 tokens.**
The first clitic is not correctly recognised and separated.

```
11-12   mandarglielo   _   _   _   _   _   _   _   TokenRange=56:68
11  mandarglie  mandargliere  VERB  V  VerbForm=Inf  9  xcomp  _  _
12  lo  lo  PRON  PC  Clitic=Yes|Gender=Masc|Number=Sing|Person=3|PronType=Prs  11
13  presto  presto  ADV  B  _  11  advmod  _  SpaceAfter=No|TokenRange=69:
14  .  .  PUNCT  FS  _  2  punct  _  SpacesAfter=\r\n\s\r\n|TokenRange=75:76
```

Missing tokens: Ex. I

Ho comprato un libro interessante e Giorgio ha detto di **mandarglielo** presto.
(*I bought a book interesting and Giorgio has tell to* ***send-him-it*** *soon*)

The analysis provided by UDPipe **only includes 2 tokens.**
The first clitic is not correctly recognised and separated.

```
11-12   mandarglielo   _   _   _   _   _   _   _   TokenRange=56:68
11  mandarglie  mandargliere  VERB  V  VerbForm=Inf  9  xcomp  _  _
12  lo  lo  PRON  PC  Clitic=Yes|Gender=Masc|Number=Sing|Person=3|PronType=Prs  11
13  presto  presto  ADV  B  _  11  advmod  _  SpaceAfter=No|TokenRange=69:
14  .  .  PUNCT  FS  _  2  punct  _  SpacesAfter=\r\n\s\r\n|TokenRange=75:76
```

Missing tokens: Ex. I

Ho comprato un libro interessante e Giorgio ha detto di **mandarglielo** presto.
(*I bought a book interesting and Giorgio has tell to* ***send-him-it*** *soon*)

The analysis provided by UDPipe **only includes 2 tokens.**
The first clitic is not correctly recognised and separated.
This also induces an error in lemmatisation.

```
11-12   mandarglielo                                                    _   _   _   _   _   TokenRange=56:68
11  mandarglie  mandargliere  VERB  V  VerbForm=Inf  9  xcomp  _  _
12  lo  lo  PRON  PC  Clitic=Yes|Gender=Masc|Number=Sing|Person=3|PronType=Prs  11
13  presto  presto  ADV  B  _  11  advmod  _  SpaceAfter=No|TokenRange=69:
14  .  .  PUNCT  FS  _  2  punct  _  SpacesAfter=\r\n\s\r\n|TokenRange=75:76
```

Missing tokens: Ex. I

Ho comprato un libro interessante e Giorgio ha detto di **mandarglielo** presto.
(*I bought a book interesting and Giorgio has tell to* ***send-him-it*** *soon*)

But have you tried different models?

They can provide different analysis.

Missing tokens: Ex. I

Using ITALIAN-ISDT only 2 tokens are generated and the token for the clitic GLIE is missing

```
11-12   mandarglielo   _   _   _   _   _   _   _   TokenRange=56:68
11  mandarglie  mandargliere  VERB  V  VerbForm=Inf  9  xcomp  _  _
12  lo  lo  PRON  PC  Clitic=Yes|Gender=Masc|Number=Sing|Person=3|PronType=Prs  11  obj  _
13  presto  presto  ADV  B  _  11  advmod  _  SpaceAfter=No|TokenRange=69:75
14  .  .  PUNCT  FS  _  2  punct  _  SpacesAfter=\r\n\s\r\n|TokenRange=75:76
```

while using POSTWITA **3 tokens** are generated and the analysis is correct

```
11-13   mandarglielo   _   _   _   _   _   _   _   TokenRange=56:68
11  mandar  mandare  VERB  V  VerbForm=Inf  9  xcomp  _  _
12  glie  gli  PRON  PC  Clitic=Yes|Gender=Masc|Number=Sing|Person=3|PronType=Prs  11  iobj  _  _
13  lo  lo  PRON  PC  Clitic=Yes|Gender=Masc|Number=Sing|Person=3|PronType=Prs  11  obj  _  _
14  presto  presto  ADV  B  _  11  advmod  _  SpaceAfter=No|TokenRange=69:75
15  .  .  PUNCT  FS  _  2  punct  _  SpacesAfter=\r\n\s\r\n|TokenRange=75:76
```

Missing tokens: Ex. I

Using ITALIAN-ISDT only 2 tokens are generated and the token for the clitic GLIE is missing

```
11-12   mandarglielo   _   _   _   _   _   _   _   TokenRange=56:68
11  mandarglie  mandargliere  VERB  V  VerbForm=Inf  9  xcomp  _  _
12  lo  lo  PRON  PC  Clitic=Yes|Gender=Masc|Number=Sing|Person=3|PronType=Prs  11  obj  _
13  presto  presto  ADV  B  _  11  advmod  _  SpaceAfter=No|TokenRange=69:75
14  .  .  PUNCT  FS  _  2  punct  _  SpacesAfter=\r\n\s\r\n|TokenRange=75:76
```

while using POSTWITA **3 tokens** are generated and the analysis is correct

```
11-13   mandarglielo   _   _   _   _   _   _   _   TokenRange=56:68
11  mandar  mandare  VERB  V  VerbForm=Inf  9  xcomp  _  _
12  glie  gli  PRON  PC  Clitic=Yes|Gender=Masc|Number=Sing|Person=3|PronType=Prs  11  iobj  _  _
13  lo  lo  PRON  PC  Clitic=Yes|Gender=Masc|Number=Sing|Person=3|PronType=Prs  11  obj  _  _
14  presto  presto  ADV  B  _  11  advmod  _  SpaceAfter=No|TokenRange=69:75
15  .  .  PUNCT  FS  _  2  punct  _  SpacesAfter=\r\n\s\r\n|TokenRange=75:76
```

Missing tokens: Ex. 2

Avvisatelo!
(***Warn-him***!)

1	Avvisatelo	avvisatelo	NOUN	S	Gender=Masc|Number=Sing			
2	!	!	PUNCT	FS	_	1	punct	_	SpaceAfter=No|TokenRa

The analysis provided by UDPipe **only includes 1 token** rather than 2 because the **clitic is not correctly recognised** and separated.

Missing tokens: Ex. 2

Avvisatelo!
(***Warn-him***!)

1	Avvisatelo	avvisatelo	NOUN	S	Gender=Masc|Number=Sing	0	root	
2	!	!	PUNCT	FS	_	1	punct	_	SpaceAfter=No|TokenRange=10:11

The analysis provided by UDPipe **only includes 1 token** rather than 2 because the **clitic is not correctly recognised** and separated.

Missing tokens: Ex. 2

Avvisatelo!
(***Warn-him***!)

1	Avvisatelo	avvisatelo	NOUN	S	Gender=Masc|Number=Sing	0	root		
2	!	!	PUNCT	FS	_	1	punct	_	SpaceAfter=No|TokenRange=10:11

The analysis provided by UDPipe **only includes 1 token** rather than 2 because the **clitic is not correctly recognised** and separated.

This also induces an error in morphology

Missing tokens: Ex. 2

Avvisatelo!
(***Warn-him!***)

1	Avvisatelo	avvisatelo	NOUN	S	Gender=Masc|Number=Sing	0	root	
2	!	!	PUNCT	FS	_	1	punct	_	SpaceAfter=No|TokenRange=10:11

The analysis provided by UDPipe **only includes 1 token** rather than 2 because the **clitic is not correctly recognised** and separated.

This also induces an error in morphology

This also induces an error in lemmatisation.

Missing tokens: Ex. 2

Avvisatelo!
(***Warn-him***!)

1	Avvisatelo	avvisatelo	NOUN	S	Gender=Masc|Number=Sing	0	root	
2	!	!	PUNCT	FS	_	1	punct	_ SpaceAfter=No|TokenRange=10:11

What to do when all the models provide incorrect analysis?

We can apply the analysis to some **similar cases** looking for one that is **correctly analysed**. Then we can copy the annotation generated for that one, and modify what is needed.

Missing tokens: Ex. 2

Avvisatelo!
(***Warn-him***!)

1	Avvisatelo	avvisatelo	NOUN	S	Gender=Masc|Number=Sing	0	root		
2	!	!	PUNCT	FS	_	1	punct	_	SpaceAfter=No|TokenRange=10:11

For example, the analysis for "Colpiscilo!" Is correct:

1-2	Colpiscilo	_	_	_	_	_	_	_	SpaceAfter=No|TokenRange=15:25
1	Colpisci	colpire	VERB	V	Mood=Imp|Number=Sing|Person=2|Tense=Pres|VerbForm=				
2	lo	lo	PRON	PC	Clitic=Yes|Gender=Masc|Number=Sing|Person=3|PronType=Prs	1			

Missing tokens: Ex. 2

```
1-2   Colpiscilo   _  _  _  _  _  _  _  SpaceAfter=No|TokenRange=15:25
1  Colpisci colpire  VERBV  Mood=Imp|Number=Sing|Person=2|Tense=
2  lo  lo  PRON   PCClitic=Yes|Gender=Masc|Number=Sing|Person=3|P
3  !   !   PUNCT  FS_  1  punct   _  SpaceAfter=No|TokenRange=25:26
```

Missing tokens: Ex. 2

```
1-2   Colpiscilo   _  _  _  _  _  _  _  SpaceAfter=No|TokenRange=15:25
1  Colpisci colpire  VERBV  Mood=Imp|Number=Sing|Person=2|Tense=
2  lo  lo  PRON   PCClitic=Yes|Gender=Masc|Number=Sing|Person=3|P
3  !   !   PUNCT  FS_  1  punct   _  SpaceAfter=No|TokenRange=25:26

1-2   Avvisatelo _  _  _  _  _  _  _  SpaceAfter=No|TokenRange=15:25
1  Colpisci colpire  VERBV  Mood=Imp|Number=Sing|Person=2|Tense=
2  lo  lo  PRON   PCClitic=Yes|Gender=Masc|Number=Sing|Person=3|P
3  !   !   PUNCT  FS_  1  punct   _  SpaceAfter=No|TokenRange=25:26
```

Missing tokens: Ex. 2

```
1-2   Colpiscilo   _  _  _  _  _  _  _  SpaceAfter=No|TokenRange=15:25
1  Colpisci colpire  VERBV  Mood=Imp|Number=Sing|Person=2|Tense=
2  lo  lo  PRON   PCClitic=Yes|Gender=Masc|Number=Sing|Person=3|P
3  !   !   PUNCT  FS_  1  punct   _  SpaceAfter=No|TokenRange=25:26

1-2   Avvisatelo _  _  _  _  _  _  _  SpaceAfter=No|TokenRange=15:25
1  Colpisci colpire  VERBV  Mood=Imp|Number=Sing|Person=2|Tense=
2  lo  lo  PRON   PCClitic=Yes|Gender=Masc|Number=Sing|Person=3|P
3  !   !   PUNCT  FS_  1  punct   _  SpaceAfter=No|TokenRange=25:26

1-2   Avvisatelo _  _  _  _  _  _  _  SpaceAfter=No|TokenRange=15:25
1  Avvisate   avvisare   VERBV  Mood=Imp|Number=Sing|Person=2|T
2  lo  lo  PRON   PCClitic=Yes|Gender=Masc|Number=Sing|Person=3|P
3  !   !   PUNCT  FS_  1  punct   _  SpaceAfter=No|TokenRange=25:26
```

Missing tokens: Ex. 2

Be careful!

As seen in the examples, **an error can have impact on more than** one kind of information / **column**.

Be careful to correct all columns and rows that may be affected by the error once you have detected its presence.

For example, when you change the content of the column word, check whether the lemma is correct or change it according to the correction done for the word.

Missing tokens: Ex. 2

Be careful!

The CoNLL-U format is based on **10 columns**.
Columns are separated by **tab characters**.

Inception only reads files in correct CoNLL-U format, with lines composed of 10 columns and separated by tabs!

If a tab or column is missing on a line, Inception cannot read that line, or open the file and import it to create a project.
When you try to import it, Inception generates a **warning** message in which you can find precise information about the error.

Missing tokens: Ex. 2

Be careful!

It is easy to do errors in adding / modifying / removing columns also because the characters for tabs are usually invisible.

```
1-2   Colpiscilo   _  _  _  _  _  _  _  SpaceAfter=No|TokenRange=15:25
1  Colpisci colpire  VERB  V  Mood=Imp|Number=Sing|Person=2|Tense=Pres
2  lo  lo  PRON   PC Clitic=Yes|Gender=Masc|Number=Sing|Person=3|PronT
3  !   !   PUNCT  FS _  1  punct   _  SpaceAfter=No|TokenRange=25:26
```

But you can work with a text editor that can make them visible:

```
1-2Δ Colpiscilo Δ _Δ_Δ_Δ_Δ_Δ_Δ_Δ_ΔSpaceAfter=No|TokenRange=15:25¬
1ΔColpisciΔcolpireΔ VERBVΔMood=Imp|Number=Sing|Person=2|Tense=Pr
2ΔloΔloΔPRONΔ  PCClitic=Yes|Gender=Masc|Number=Sing|Person=3|Pron
3Δ!Δ !Δ PUNCTΔFSΔΔ1Δpunct  _ΔSpaceAfter=No|TokenRange=25:26
```