EVALITA 2020
KIParla Part of Speech tagging (KIPOS)
TASK GUIDELINES

Eugenio Goria – *Dipartimento di Studi Umanistici, Università degli Studi di Torino*
Massimo Cerruti – *Dipartimento di Studi Umanistici, Università degli Studi di Torino*
Silvia Ballarè – *Dipartimento di Filologia Classica e Italianistica "Alma Mater Studiorum", Università di Bologna*
Caterina Mauri – *Dipartimento di Lingue, Letterature e Culture Moderne, Università di Bologna*
Cristina Bosco - *Dipartimento di Informatica, Università degli Studi di Torino*

1. INTRODUCTION

The following are the guidelines for the **KIPOS** task of the EVALITA 2020 evaluation campaign.

Participants to the evaluation task are required to use the data provided by the organization to set up their systems.

The organisation will provide three data sets, all drawn from the KIParla corpus: the first one, referred henceforth to as **Development Set (DS),** contains data automatically annotated using a specific tagset (see a following section for the tagset description) and manually revised and must be used to train participants' systems; the second one, referred henceforth to as **Silver Set (SS)**, contains data only automatically annotated using the same tagset applied in the DS; the third one, referred henceforth to as **Test Set (TS)** contains instead the test data in blind format, tokenized but without annotation, for the evaluation and will be given to participants in the date scheduled for the evaluation.

Participants are allowed to use other resources both for training and to enhance final performances, as long as their results apply the tagsets and are compliant with the format described in this document.

Each participant team is also required to send a **paper** (in electronic format) which contains a brief description of the system, especially considering techniques and resources used and a bibliographic reference. Some error analysis may also do the report more interesting for the community.

2. DATA RELEASE

All data (DS, SS and TS) will be made available for download on the **GitHub** repository of KIPOS2020 (https://github.com/boscoc/kipos2020) according to the timetable published in the KIPOS2020 website (http://www.di.unito.it/~tutreeb/kipos-evalita2020/).

All data are covered by a **Creative Commons license** (Attribution-NonCommercial-ShareAlike 4.0 International) that can be found in this same GitHub.

Following the indications published in the GitHub of KIPOS2020, the participants must fill in a **form** for accepting the licence and to be registered as authorized data users. Only after filling in the form, they will receive by email the password necessary to unzip the data downloaded from the GitHub repository.

Participants are not allowed to re-distribute the KIPOS2020 data to other non registered users.

During the evaluation campaign, and before the date scheduled for the evaluation, **all participants are encouraged to communicate to the organizers** (using the Google group kipos-evalita2020@googlegroups.com - https://groups.google.com/forum/#!forum/kipos-evalita2020) **any error found in the DS's data**. This will allow the organizers to update and redistribute it to the participants in an enhanced form.

3. DATA and TASK DESCRIPTION

All the data will be provided by the organizers as plain text files in UTF-8 format.
Data where organized in two portions, one including data for **formal speech** (DS-formal, SS-formal and TS-formal) and the other for **informal speech** (DS-informal, SS-informal and TS-informal). The task is indeed organized in three tracks as follows:

- **Main task - general**: training on all given data (both DS-formal and DS-informal) and testing on all test set data (both TS-formal and TS-informal)
- **Subtask A - crossFormal**: training on data from DS-formal only and testing separately on data from formal register (TS-formal) and from informal register (TS-informal)
- **Subtask B - crossInformal**: training on data from DS-informal only and testing separately on data from formal register (TS-formal) and from informal register (TS-informal).

The SS data can be also used if considered as helpul by participants, considering that they are not manually revised and are therefore prone to errors generated by the tools used for their annotation.
The DS's and SS's data provided by the organizers are tokenized (see section 3.1) and annotated according to the tagset described below (see section 3.2).
The TS's data will be provided, according to the scheduling, without annotation but after the application of a tokenization strategy similar to that applied on the DS's and SS's data described below (see section 4.1).

The strategies applied for tokenization and tagging in the KIPOS@Evalita2020 datasets are the result of an adaptation of those applied in the Universal Dependencies (UD) treebanks for Italian (http://universaldependencies.org/it/pos/index.html) and in the task about Part of Speech tagging of social media (POSTWITA) held in the Evalita2016 campaign (see the report of the organizers of this task). This makes the KIPOS2020 gold and silver standard (i.e. DS and SS datasets) compliant with the UD tokenization and allows the conversion towards this format suitable with a very small effort.

3.1. Tokenization

In KIPOS@Evalita2020, we decided to follow the tokenization strategy applied in the Universal Dependencies (UD) treebanks for Italian.
This means in particular that expressions including more than one token each, like the articulated prepositions (e.g. *dalla, nell', al...*) and clitic cluster attached to the end of a verb form (e.g. *suonargliela, regalaglielo, dandolo...*) in the DS and SS datasets are split in all the necessary tokens, as in the following example (see especially tokens in boldface: 2-3 and 4-5 which are respectively split in 2 and 3 and 4 and 5):

1	dovresti	AUX
2-3	**parlarmi**	**VERB_PRON**
2	**parlarVERB**	
3	**mi**	**PRON**
4-5	**della**	**ADP_A**
4	**di**	**ADP**
5	**la**	**DET**
6	tua	DET
7	casa	NOUN

Apostrophe is tokenised together with the preceding character.

3.2 Tagset

In KIPOS@Evalita2020, we decided to broadly follow the tagset applied in the POSTWITA@Evalita2016, which is in turn an adaptation to the informal text from social media of the tagset proposed in the Universal Dependencies (UD) project for Italian treebanks. The notes released for POSTWITA@Evalita2016 inspired the present ones (http://corpora.ficlit.unibo.it/PoSTWITA/index.php?slab=guidelines).
The full tagset applied in KIPOS is described in the following table with some example.

TAG	**Part of Speech**	**Examples**
ADJ	• Adjectives • Interrogative Adjectives • Also used in question answering	• ***che*** *gelato vuoi?* ***quanti*** *anni hai?* • ci vediamo domani? **– Esatto**!
ADP	• Primary and secondary Prepositions • Postpositions	• *di, a, da, in, con, su, per…* • *senza, tranne, …* • *dieci anni* ***fa***
ADP_A	• Articulated Prepositions	*dalla, nella, sulla, dell'…*
ADV	• Adverbs • Interrogative Adverbs	• *lo scrivo* ***qui, …*** • *mi chiedo* ***dove*** *sia andato,* • *non so* ***come*** *mi chiamo,* • *mi dici* ***quando*** *partiamo?,* • *chiedo* ***perché*** *abbiano fatto così*
AUX	• Auxiliary Verbs • Modal Verbs • Periphrastic structures	• *essere, avere* • *potere, volere, dovere* • ***sta*** *mangiando,* ***viene*** *visto, …*
CCONJ	• Coordinating Conjunctions • Discourse markers with connective function	• *e, ma, o* • *però, anzi, quindi, dunque, …*
DET	• Demonstrative Adjectives • Quantifiers • Articles • Possessive Adjectives	• *ho letto* ***dei*** *libri* • ***questo*** *corso si tiene il lunedì,* • ***quella*** *ragazza è* ***sua*** *sorella, …* • ***alcuni*** *studenti,* ***vari*** *studenti,*

	- Numerals when precede the noun	***qualche** studente* - *il, la, un, una, …* - ***mio** padre, **tuo** fratello, …* - ***tre** ragazze, …*
.DIA	Italian Dialectal Words	***Seinë*** → INTJ.DIA
INTJ	Interjections	*sì, no, okay, ecco, ...*
.LIN	Foreign Words	***House*** → NOUN.LIN
NEG	Negation (associated with a phrase)	***non***
NOUN	Common Nouns	*cane, tavolo, ...*
NUM	Numerals occuring without a Noun (non Adjective)	- es. quanti sono? ***tre***
PARA	Paraverbal communication	*eh, mh, oh, bla bla, …*
PRON	- Strong pronouns and clitics - Interrogative Pronouns - Relative Pronouns	- *io, me, tu, te, …* - ***glielo** dico* - *chi?, cosa?, quale?, che?* - *il **quale**, dove, cui*
PROPN	Proper Nouns	*Mario Rossi, Michele, Bologna* ...
SCONJ	Subordinating Conjunctions	- *dove, quando, perché* in contesti non interrogativi (es. *so dove sei stato*) - *ho detto **che**…* - ***se** vuoi…* - ***che*** polivalente (anche in costruzioni relative: ad es. *la ragazza **che** vedi*)
VERB	Verbs (including *essere* and *avere* in non auxiliary role)	- ***aveva** vent'anni* - **era** molto stanco - **era** una persona sola
VERB_PRON	Verb + Clitic pronoun cluster	***mangiarlo**, **donarglielo**…*
X	Other	***fior***- (parole interrotte)

The annotation of **named entities** are handled considering each one as a unique token assigning to it the PROPN tag, like in the following example:

1	son	AUX
2	stata	VERB
3	una	DET
4	volta	NOUN
5	ad	ADP
6	**ascoli**	**PROPN**
7	**piceno**	**PROPN**
8	quando	SCONJ
9	ero	VERB
10	più	ADV
11	piccola	ADJ

All **named entities** referring to persons have been substitued, for achieving the anonimity of data, while the names of cities are provided in lowercase.

Non-Italian **foreign words** are annotated, when possible, following the same PoS tagging criteria adopted in UD guidelines for referring language. Example: "good-bye" INTJ

Some challenging issues follow.

1) **Interjection** or adverb?
   The tag INTJ has been used only when the form cannot be properly classified as ADV or otherwise.
   For instance, *ciao* INTJ; *okay* INTJ; *bene!* ADV; *esatto* ADJ; *certo* ADJ

2) The verb ***essere***: main or auxiliary?
   The verb *essere* has been considered as auxiliary and tagged with AUX when it is associated with past participle or occurs in passive structures, as main verb tagged with VERB otherwise.
   For instance, *è* AUX *arrivato* VERB; *è* VERB *sul* ADP_A *tavolo* NOUN

3) **Phraseological constructions**
   They must be split in more tokens to be tagged separately.
   For instance, *per* ADP *favore* NOUN; *va* VERB *bene* ADV

4) **Determiners**
   All the elements that can be displaced in the position where usually the article occurs can be tagged as determiners using the tag DET.
   For instance, *vari* DET *studenti* NOUN; *alcuni* DET *studenti* NOUN; *diversi* DET *studenti* NOUN; *tre* DET *studenti* NOUN

5) **Auxiliary**
   The auxiliary verbs *essere* and *avere*, modal verbs (*potere, volere, dovere*) and the auxiliary verbs of periphrastic constructions (e.g. *stare* + gerundio) must be annotated with the tag AUX, while causative structures must be annotated otherwise.
   For instance, *siamo* AUX *andati* VERB; *abbiamo* AUX *mangiato* VERB; *voglio* AUX *vedere* VERB; *dovete* AUX *ascoltare* VERB; *possiamo* AUX *andare* VERB; *lasciamo* VERB *perdere* VERB.

3.3 The data format

Data for KIPOS are extracted from the KIParla corpus (http://kiparla.it/), a resource for the study of spoken Italian, composed of transcribed conversations[1].
In particular, each file included in the KIPOS datasets is the transcription of the turns of a single conversation where two or more speakers were involved. Each turn in each conversation is encoded with three main identifiers, respectively indicating the **conversation** (alphanumeric), the **speaker** (alphanumeric) and the **position of the turn** (numeric) within the context of the conversation. This means that all turns of a specific conversation begin with the same identifier, and all the turns pronounced by the same

[1] *All speakers were informed of the aims of the project, agreed to the recording and signed a consent form. In the recordings performed after May 2018, the consent complies with the European Union's General Data Protection Regulation (G.D.P.R.), which allows us to freely share the (anonymized) data as far as their use is noncommercial.*

speaker will include the same speaker identifier. The full **text** of the turn follows the identifiers in a separate line which is in turn followed by the tokens of the turn associated with proper part of speech. The following example shows the first three turns of a KIPOS file where the encoding of conversation, speaker, turn, full text and tokenized and annotated text are provided.

```
# conversation = BOD2018
# speaker = 1_MP_BO118
# turn = 1
# text = dovresti parlarmi della tua casa
1       dovresti        AUX
2-3     parlarmi        VERB_PRON
2       parlar  VERB
3       mi      PRON
4-5     della   ADP_A
4       di      ADP
5       la      DET
6       tua     DET
7       casa    NOUN

# conversation = BOD2018
# speaker = 2_MP_BO118
# turn = 2
# text = attuale
1       attualeADJ

# conversation = BOD2018
# speaker = 3_AM_BO140
# turn = 3
# text = mh sì
1       mh      PARA
2       sì      INTJ
```

For what concerns the lines where tokens are annotated, a CoNLL-like format one has been applied using only the first three of the ten CoNLL columns. The following pattern has been indeed applied:

index tab word tab tag

where *index* is the position of the token within the linear order of the turn, *word* is the word form occurring in the turn, *tag* is the part of speech tag associated with *word*, *tab* are tab keys used for separating *index* from *word* and *word* from *tag*. Each line finishes with a line break and no space character is allowed on a line.

3.4 The data file names

All data provided for KIPOS2020 are organized in files each corresponding to a single conversation. For what concerns the names of these files, the following pattern has been applied:

- the first two letters correspond to the city where the recording was collected: TO (Torino) and BO (Bologna)
- folowing two characters corresponding to the type of activity: A1 (office hours), A3 (random conversation), C1 (exams), D1 (lessons) and D2 (interviews); for the aim of KIPOS2020 C1, A1 and D1 are considered FORMAL, while A3 and D2 are considered INFORMAL.

4. EVALUATION

4.1 Format of the Test Set

The TS, that will be provided according to the published scheduling, will contain only the tokenized words but not the tags, that have to be added by the participant systems to be submitted for the evaluation.
The tokenization applied in the TS for the purpose of the evaluation is slightly different from that used in the development data in order to have in the TS a format more adequate to the standard scripts and metrics. While the format applied in the DS and SS for each multiple token word includes both a multiple token line and the lines where not split in the different tokens, the format applied in the TS only includes a single token for each line, but all multiple token lines were removed and substituted by that necessary for hosting the tokens not split. An example of this format follows:

```
# conversation = BOD2018
# speaker = 1_MP_BO118
# turn = 1
# text = dovresti parlarmi della tua casa
1 dovresti
2 parlarmi
3 di
4 la
5 tua
6 casa
```

The participants are requested to return the TS file using exactly the same tokenisation format, containing exactly the same tokens of the TS provided by the organizers and paying attention to the UTF-8 encoding (e.g. to newline character format). The comparison with the TS gold standard reference file will be performed line-by-line, thus a misalignment will produce wrong results.
The correct tokenized and tagged data of the TS (called *gold standard TS*) will be exploited for the evaluation and will be provided to the participants after the evaluation, together with their score.

4.2 Evaluation metrics

The evaluation is performed in a "black box" approach: only the systems' output is evaluated. A **single run** will be accepted and evaluated for each track for each participant.
The evaluation metric will be based on a token-by-token comparison and only one tag is allowed for each token.

The considered metric is the *Tagging accuracy*: it is defined as the number of correct PoS tag assignment divided by the total number of tokens occurring in the TS.