Detailed Task Description and System Response Guidelines

VERSION 1.1
(subject to minor changes)

The task

The Multilingual Named Entity Recognition task consists of three subtasks:

1. **Named Entity Mention Detection and Classification**: recognizing all unique named mentions of entities of five types:
    - persons (PER),
    - organizations (ORG),
    - locations (LOC),
    - events (EVT), and
    - products (PRO)
2. **Name Lemmatization:** computing for each detected named entity mention its corresponding base form/lemma,
3. **Entity Matching:** assigning to each detected named entity mention an identifier in such a way that detected mentions referring to the same real-world entity should be assigned the same identifier, which we will refer to as cross-lingual ID.

There is no need to return positional information of named entity mentions.

IMPORTANT: Systems may be tuned to solving all three subtasks or just a subset of the subtasks. Analogously, systems may cover all or just a subset of the languages.

System response

The following rules/conventions should be followed when designing the system/solution:

1. General rules

a. The system should not return more than one annotation for all occurrences of the same text form of a mention (e.g., inflected variant, acronym or abbreviation) of a named entity within the same document, unless the different occurrences thereof have different entity types (different readings) assigned to them or refer to different entities of the same type.

b. Since the evaluation will be case-insensitive, it is not relevant whether the system response includes case information. In particular, if the text includes lowercase, uppercase and/or mixed-letter named mention variants of the same entity, the system response should include only one annotation for all of these mentions. For instance, for "*ISIS*", "*isis*", and "*Isis*" (provided that they refer to the same named-entity type), only one annotation should be returned (e.g., "*Isis*").

c. Lemmatisation of a named entity mention refers to lemmatisation of the surface form extracted from the input text, e.g., the lemma for "*UE*" and "*Unii Europejskiej*" (eng. European Union) is "*UE*" and "*Unia Europejska*" respectively, whereas the cross-lingual ID assigned to the two aforementioned NE mentions should be the same. Analogously, lemma of a plural mention of an entity is expected to be nominative plural form (e.g., lemma of the word "*Japoncích*" in Czech should be "*Japonci*"), whereas lemma of a singular mention should be nominative singular (e.g., lemma of the word "*Japoncem*" in Czech should be "*Japonec*")

d. Recognition of the following entities is **is not part of the task:**

    - nominal or pronominal mentions of entities,
    - temporal and numerical expressions (e.g. currency expressions, measurements, quantities), identifiers such as email addresses, URLs, postal addresses, etc.

e. When assigning the type to a named entity (ORG, LOC, PER, EVT or PRO) in general it is assumed that the local document context and common knowledge is considered and exploited for resolving ambiguities, e.g., in "*Twitter announced revenues for 2018* ..." the name "*Twitter*" refers to a company (ORG), whereas in the phrase "*I posted it on Twitter*" the name "*Twitter*" refers to a product (PRO), **unless explicitly specified otherwise** in the remaining part of these guidelines.

f. In cases, in which the local document context and common knowledge does not provide sufficient information to disambiguate the named entity type (e.g., in the phrase "*Opel announced that …*" the mention of "*Opel*" could potentially refer either to an organisation (ORG) or a person (PER), namely, Adam Opel, the founder of the company, **the more probable**

interpretation should be considered, i.e., ORG, since Adam Opel died in 1837 and could not announce anything recently, unless the document is a historical one (unlikely). In case both NE type interpretation appear to be equally probable the following **NE type disambiguation rules** should be applied:

Possible interpretation	**Choose**
ORG + PER	PER
ORG + PRO	ORG

2. Person names (PER)

a. This category covers named references to individual people (e.g., "*Donald Trump*") or families (e.g. "*Kaczyńscy*"), and certain named references to groups of people.

b. Person names should not include titles, honorifics, and functions/positions. For example, in the text fragment "*CEO Dr. Jan Kowalski*", only "*Jan Kowalski*" should be recognized as a person name. However, initials and pseudonyms are considered named mentions of person names and should be recognized. Analogously, in the text fragment "*The Prime Minister of the United Kingdom Theresa May*", only "*Theresa May*" should be tagged as PER, whereas "*United Kingdom*" should be tagged as LOC.

c. Personal possessives derived from a named mention of a person should be classified as a person, and the base form of the corresponding person name should be extracted. For instance, for "*Piskorskijev mejl*" (Croatian - Piskorski's email) it is expected to recognize "*Piskorskijev*", classified with PER and extract the base form: "*Piskorski*".

d. Toponym-based (e.g. country name adjectives) references to **groups of people that are linked to geopolitical entities**[1] should also be recognized and tagged as PER, e.g., "*Ukrainians*". In this context, mentions of a single

[1]Geo-Political Entities are considered to be complex entities consisting of a population, a government or some administrative body, and a physical location. In the context of this shared task geo-political entities comprise countries, provinces, states, counties and cities, and suchlike entities. In the context of this task international bodies and organisations are not considered geopolitical entities.

member belonging to such groups, e.g., "*Ukrainian*" should be assigned the same cross-lingual ID as plural mentions, i.e., "*Ukrainians*". Furthermore, it should not matter whether "*Ukrainians*" refer to the entire nation, some "unspecified" part thereof or an organisation related to the geopolitical entity. In all these cases PER category should be used and the cross-lingual ID should be the same as assigned to other toponym-based references to the same geopolitical entity, e.g., "*Ukrainians*" and "*Ukraine*" should be assigned the same cross-lingual ID (different mention type, but the same cross-lingual ID).

e. Although continents ("*Europe*") and geographical regions ("*Eastern Europe*") are not geo-political entities, as regards named references to groups of people derived from such toponyms the same rule as specified in 2.d. applies by analogy, e.g., "*Europejczycy*" (Europeans) and "*Europa*" (Europe) should be classified as PER and LOC respectively, and both should be assigned the same cross-lingual ID.

f. Named mentions of other groups of people that do **have a formal organization unifying them** should be tagged as PER and associated with the same cross-lingual ID as the mentions of the corresponding organisation, e.g., in the phrase "*Sparťané vyhráli*..." (Spartans won...) the mention of "*Sparťané*" should be tagged as PER and have the same cross-lingual ID as the corresponding sport team, e.g., "*AC Sparta Praha*" (football club). Analogously phrases like "*Europoslanci*" (Members of the European Parliament) should be cross-linked with the mentions of the European Parliament (ORG).

g. Mentions to groups of people that **do NOT have a formal organization unifying them** should not be extracted, e.g., phrases like "*Muslims*", which refers to a religious group not linked to any particular organisation should not be recognized.

h. Fictive persons and characters (e.g., "*James Bond*") are considered as persons.

3. Locations (LOC)

a. This category includes all toponyms (e.g., cities, counties, provinces, regions, bodies of water, geological formations, etc.) and named mentions of facilities, i.e., functional and primarily man-made structures, such as: stadiums, parks, museums, theaters, hotels, hospitals, transportation hubs (e.g. airports, sea ports, train stations), churches, streets, railroads, highways, bridges, tunnels, parkings, and other similar urban and non-urban facilities. For instance, "*Łazienki Królewskie w Warszawie*" (a park in Warsaw, Poland) should be tagged as LOC, whereas general references to facilities

without a concrete location as "*parki w Warszawie*" (parks in Warsaw) should not be recognized, i.e, only "Warszawie" should be recognized as LOC in the latter case.

b. Even in case of named mentions of facilities that refer to an organization, the LOC tag should be used. For example, from the text phrase "*The Schipol airport has acquired new electronic gates*" the mention "*The Schipol airport*" should be extracted and classified as LOC.

c. By analogy to 3b., toponyms, in particular, country names (e.g., "*Polska*") that refer to geopolitical entities (physical location, population of the country, respective government, nation, or a sport team representing a country) should be extracted and classified as LOC disregarding the specific named mention role ("*Poland*" refers to a range of concepts) and assigned the same unique cross-lingual ID. In this context, the relevant toponyms and toponym-derived adjectives (see the rule in 2.d) referring to the same geo-political entity should be assigned the same cross-lingual ID. In all other contexts, i.e., when a country name (or other toponym) is used to refer to an organisation that has no specific link to the respective geopolitical entity (unless it is accidental), e.g. a band named "Russia", it should be tagged as ORG.

d. When recognising named mentions of facilities potential mentions of the location are considered to be part of the full mention, e.g., the entire phrase "*St. Stephen Church in Istanbul*" should be extracted and tagged as LOC.

4. Organizations (ORG)

a. This category covers all kind of organizations such as: political parties, public institutions, government units, non-governmental organizations, international organizations (e.g., European Union, NATO, united Nations, etc.) military organizations, companies, religious organizations, sport teams and organizations, education and research institutions, music groups, entertainment and media organizations, etc.

b. Organization designators and potential mentions of the seat of the organization are considered to be part of the organization name. For instance, from the text fragment "*Citi Handlowy w Poznaniu*" (a bank in Poznań), the full phrase "*Citi Handlowy w Poznaniu*" should be extracted.

5. Events (EVT)

a. This category covers named mentions of events, including: (a) occasions such as conferences, e.g. "*24. Konference Žárovného Zinkování*", concerts, festivales, holidays, e.g., "*Święta Bożego Narodzenia*" (eng. Christmas), (b) incidents such as wars, battles, and man-made disasters, e.g., "*Katastrofa Czernobylska*" - Chernobyl catastrophe, and (c) natural disasters and phenomena.

b. Future, speculative and fictive events, e.g., "*Czexit*" or "*Polexit*" are considered as event mentions too.

c. In case a named mention of the event does also refer to a location, then it should be tagged as LOC. For example, in the phrase "*He died in Waterloo, just before the end of the battle*", "*Waterloo*" should be tagged as LOC, not as EVT. However, the cross-lingual ID assigned to such a mention, i.e., "*Waterloo*" should be the same as in the case of other mentions to the battle, e.g., "*The Battle of Waterloo*".

d. When recognising named mentions of events the potential mentions of the location are considered to be part of the full mention, e.g., the entire phrase "*2004 Winter Olympics in Canada*" should be extracted and tagged as EVT.

6. Products(PRO)

a. This category covers product names, including for instance: electronics (e.g., "*Motorola Moto Z Play*", cars (e.g. "*Subaru Forester XT*"), vehicles (e.g., "*Fiat Panda*"), weapons (e.g., "*Kalashnikov AK-47*"), web-based services (e.g., "*Twitter*"), books (e.g., "*Harry Potter and the Sorcerer's Stone*"), software (e.g., "*MS Office*"), films (e.g., "*Gone with the Wind*"), TV programmes (e.g., "*Wiadomości TVP*"), , newspapers (e.g., "*The New York Times*") and other pieces of art, etc.

b. Names of legal documents (e.g., "*dyrektywy 2001/14/we Parlamentu Europejskiego i Rady*"), treaties, e.g., "*Traktat Lizboński*" (eng. Treaty of Lisbon), initiatives/programmes (e.g., "*Horizon 2020*") are also considered product names

7. Other aspects related to system response

a. In case of complex named entities, consisting of nested named entities, only the top-most (longest) entity should be recognized. For example, from the text fragment "*George Washington University*" one should not extract "*George Washington*", but the entire name, namely, "*George Washington University*".

b. In case of coordinated phrases like for instance "*European and British Parliament*" two names should be extracted (as ORG), i.e., "*European*" and "*British Parliament*". The respective lemmas would be "European" and "British Parliament".

8. Input texts

The input texts are the result of downloading HTML pages (mainly news articles or fragments thereof) and converting them into pure text using a hybrid HTML parser, which might have resulted in extracting texts that not only include the core body text of a Web page, but also some additional pieces of text (e.g., a list of labels from a menu, user comments, etc.) that might not necessarily constitute well-formed utterances in a given language. This phenomenon applies to a small fraction of texts in training/test collection. Such texts were included in the training/test document collections and will be included in the test data in order to maintain the flavour of "real-data". However, obvious HTML parser failure, e.g., extraction of javascript code or extraction of empty texts, were removed from the data sets. As regards the task at hand, it is important to emphasize that the entire text that is available (regardless of whether it contains more than the main text of the Web page) should be considered for matching names, including the title of the document.

Each of the input texts in the collections will be available in the following format:

- The first five lines of the document contain the following metadata,

  <ID>
  <LANGUAGE>
  <CREATION-DATE>
  <URL>
  <TITLE>

- The core text to be processed is available from the 6th line till the end of the file.

- Please note that both **<CREATION-DATE>** and **<TITLE>** information might be missing (since the HTML parsers might not had been able to extract it for various reasons). In such cases the corresponding lines are empty.

- All **input texts are encoded using UTF-8** encoding.

9. Output format

The system response should contain for each file in the input test corpus a corresponding file in the following format:

The first line should contain only the ID of the file in the test corpus.
Each subsequent line should be of the format:

Named-entity-mention <TAB> base-form <TAB> category <TAB> cross-lingual ID

The files with system response should be **encoded using UTF-8** encoding.

A file containing a system response should have the same name as the corresponding input file with an **additional extension ".out"**.

Two examples of input texts and corresponding system responses are given below, one for Czech and one for Polish (both texts are related to Brexit). The cross-lingual IDs present in both documents are marked in red in the system response examples.

An input in Czech language:

> *Japonci se ptají na czexit, říká Špicar ze Svazu průmyslu. 'Odešli bychom z Česka,' varovali ho*
>
> *Případné vystoupení České republiky z Evropské unie by bylo podle ekonomů, Hospodářské komory i Svazu průmyslu a dopravy ekonomickou sebevraždou. Odchod z EU by znamenal ztrátu stovek tisíc pracovních míst a česká ekonomika by se podle některých dostala na úroveň Běloruska. Praha 21:18 7. února 2018*
>
> *Evropská vlajka a státní vlajka České republiky (ilustrační foto) | Zdroj: Fotobanka Profimedia. Zpochybňování členství Česka v EU znervózňuje některé zahraniční investory. Například ve středu musel viceprezident Svazu průmyslu a dopravy Radek Špicar vysvětlovat japonským investorům aktuální politickou situaci a její dopady na ekonomiku, upozornil ve středu na twitteru. Japonce zneklidnily informace z médií, kde se mluví v souvislosti s novou českou vládou i o setrvání ČR v EU. „Odmítáme jakékoliv zpochybňování našeho členství v EU. Vysílá to negativní signál zahraničním investorům. Zpochybňování našeho členství v EU a z něj vyplývající nejistota by zahraniční investory mohlo vést k pozastavení plánovaných investic a v krajním případě i k odchodu ze země," řekl později Špicar.*
>
> *„Vedle ekonomických přínosů je pro nás EU také důležitou právní*

pojistkou a do jisté míry nás tak může chránit před námi samými, což se někdy také může hodit,“ dodal ekonom.

Ekonomická sebevražda. Podřizujeme se Bruselu, říká poslanec SPD Kobza. Neví ale, který orgán Evropské unie odhlasoval kvóty Číst článek Česko podle Hospodářské komory není Británie, proto odcházet z EU by pro zemi byla ekonomická sebevražda. Stejné srovnání použil i hlavní ekonom UniCreditu Pavel Sobíšek. „Zatímco brexit bude pro Británii znamenat podle vládní zprávy 'jen' zpomalení ekonomického růstu proti situaci setrvání v EU, czexit by byl pro Česko ekonomickou sebevraždou.“ Pouhé úvahy o czexitu jdou proti snaze omezit odliv dividend z Česka ve prospěch vyšších reinvestic," uvedl Sobíšek. Kdyby se czexit měl opravdu uskutečnit, vyvolalo by to podle něj masivní přeskupení kapacit zpracovatelského průmyslu mimo území Česka, znamenající ztrátu stovek tisíc pracovních míst. „Česko by se ocitlo v pozici Běloruska,“ dodal. Fatální dopady. Podle Hospodářské komory by potenciální czexit zhoršil přístup českým exportérům na trhy a zkomplikoval mezinárodní obchodní i politické vazby, na kterých je malá, otevřená ekonomika jako ČR zcela závislá. „Pro Česko by případný czexit byl zejména z pohledu českého exportu při podílu jednotného trhu téměř 85 procent naprosto fatální, proto nikdo rozumně uvažující by asi nechtěl českou ekonomiku závislou na exportu uvrhnout do doby temna,“ dodal mluvčí komory Miroslav Diro. ČTK

Expected system response:

```
cs-10
Japonci                 Japonci                 PER GPE-Japan
Czexit                  czexit                  EVT EVT-Czexit
Špicar                  Špicar                  PER PER-Radek-Spicar
Svazu průmyslu          Svaz průmyslu           ORG ORG-Svaz-Prumyslu
Svazu průmyslu a dopravy   Svaz průmyslu a dopravy
                                                ORG ORG-Svaz-Prumyslu
Česka                   Česko                   LOC GPE-Czech-Republic
České republiky         Česká republika         LOC GPE-Czech-Republic
Evropské unie           Evropská unie           ORG ORG-European-Union
Hospodářské komory      Hospodářská komora      ORG ORG-Hospodarska-Komora
EU                      EU                      ORG ORG-European-Union
Běloruska               Bělorusko               LOC GPE-Belarus
Praha                   Praha                   LOC GPE-Prague
Fotobanka Profimedia    Fotobanka Profimedia    ORG ORG-Profimedia-Photobank
Radek Špicar            Radek Špicar            PER PER-Radek-Spicar
Japonce                 Japonci                 PER GPE-Japanese
Twitteru                twitter                 PRO PRO-Twitter
ČR                      ČR                      LOC GPE-Czech-Republic
```

Bruselu	Brusel	LOC GPE-Brussels
SPD	SPD	ORG ORG-SPD-Party-CZ
Kobza	Kobza	PER PER-Kobza
Česko	Česko	LOC GPE-Czech-Republic
Británie	Británie	LOC **GPE-Great-Britain**
Británii	Británie	LOC **GPE-Great-Britain**
UniCreditu	UniCredit	ORG ORG-Unicredit
Pavel Sobíšek	Pavel Sobíšek	PER PER-Pavel-Sobisek
Brexit	brexit	EVT **EVT-Brexit**
Czexitu	czexit	EVT EVT-Czexit
Sobíšek	Sobíšek	PER PER-Pavel-Sobisek
Miroslav Diro	Miroslav Diro	PER PER-Miroslav-Diro
ČTK	ČTK	ORG ORG-CTK

An input in Polish language:

Farage: A może zrobić drugie referendum ws. Brexitu?

Farage, jeden z najbardziej gorących zwolenników wyjścia Wielkiej Brytanii z UE, były lider Partii Niepodległości Wielkiej Brytanii (UKIP), był jedną z twarzy kampanii na rzecz opuszczenia UE przez Londyn. Po zwycięstwie zwolenników Brexitu ustąpił ze stanowiska lidera UKIP uznając, że wypełnił swoją misję.

W ostatnim czasie, w obliczu trudnych negocjacji ws. warunków Brexitu, niektórzy brytyjscy politycy zaczęli sugerować, że na Wyspach powinno zostać zorganizowane drugie referendum ws. wyjścia Wielkiej Brytanii z UE. Były premier Tony Blair argumentuje, że byłoby to uzasadnione faktem, iż w kampanii przed poprzednim referendum zwolennicy Brexitu podawali m.in. że pieniądze, które przestaną trafiać do unijnego budżetu w formie składki, zostaną skierowane do NHS (brytyjskiego odpowiednika NFZ), co okazało się nieprawdą. Teraz sam Farage na Twitterze przyznał, że taką możliwość należałoby rozważyć, by "ostatecznie zamknąć sprawę". Od wpisu Farage'a odciął się obecny lider UKIP, Henry Bolton, który stwierdził, że drugie referendum byłoby "szkodliwe dla narodu". Z kolei eurodeputowany Gerard Batten zasugerował, że Farage'owi brakuje uwagi mediów, stąd jego kontrowersyjny wpis. Farage'a chwalą natomiast jego polityczni przeciwnicy. Chuka Ummuna, parlamentarzysta Partii PRacy stwierdził, że "Farage, być może po raz pierwszy w życiu, zgłosił słuszną uwagę".

Expected system response:

```
pl-10
Brexitu             Brexit          EVT EVT-Brexit
Chuka Ummuna        Chuka Ummun     PER PER-Chuka-Ummuna
Farage'a            Farage          PER PER-Nigel-Farage
Farage'owi          Farage          PER PER-Nigel-Farage
Farage              Farage          PER PER-Nigel-Farage
Gerard Batten       Gerard Batten   PER PER-Gerard Batten
Henry Bolton        Henry Bolton    PER PER-Henry-Bolton
Londyn              Londyn          LOC GPE-London
NFZ                 NFZ             ORG ORG-NFZ
NHS                 NHS             ORG ORG-NHS
Partii Niepodległości Wielkiej Brytanii
                    Partia Niepodległości Wielkiej Brytanii
                                    ORG ORG-UK-Independence-Party
Partii PRacy        Partia Pracy    ORG ORG-Labour-Party-UK
Tony Blair          Tony Blair      PER PER-Tony-Blair
Twitterze           Twitter         PRO PRO-Twitter
UE                  UE              ORG ORG-European-Union
UKIP                UKIP            ORG ORG-UK-Independence-Party
Wielkiej Brytanii   Wielka Brytania LOC GPE-Great-Britain
Wyspach             Wyspy           LOC GPE-Great-Britain
```

The cross-lingual identifiers may consist of an arbitrary sequence of alphanumeric characters. The form of the identifiers is not relevant; what will be relevant for the evaluation is that mentions of the same entity across documents — in any language — are assigned the same cross-lingual identifier.

Additional hints for the annotators and participants

This “dynamic” part of the guidelines includes any relevant hints and instructions for the annotators on how to proceed in special cases, as well as documentation of relevant questions posted by the participants and related clarifications.