The DDI Corpus

The DDI corpus is a semantically annotated corpus of documents describing drug-drug interactions from the DrugBank database and MEDLINE abstracts on the subject of drug-drug interactions. This corpus is intended for training Information Extraction (IE) systems which are used to identify and classify pharmacological substances as well as extract drug-drug interactions from biomedical literature.

The corpus has been manually annotated with pharmacological substances (drugs) and interactions between them. Full details of the annotation scheme can be found in the annotation guidelines.

The use of the DDI corpus is subject to the terms of the DDI licence.

The organizers will provide human-annotated documents for the training and evaluation of participating systems. Additional training/test data sets will be provided together with those for the previous DDIExtraction2011task. The additional datasets include descriptions of DDIs from the DrugBank database as well as MedLine abstracts containing the keywords “drug drug interaction”.

For training, a data set based on the publicly available portion of the DDI corpus is provided in XML. It will be downloaded here: DDICorpusTraining.zip

The directory contains 2 subdirectories:

- DrugBank – contains 572 documents describing drug-drug interactions from the DrugBank database.
- MedLine – contains 142 abstracts on the subject of drug-drug interactions.

The following tables summarize the main features of the training dataset:

	DRUGBANK Training			**MedLine Training**		
	Total	**Avg. Doc**	**Avg. Sentences**	**Total**	**Avg. Doc**	**Avg. Sentences**
Documents	572			142		
Sentences	5675			1301		
Drugs	8197			1228		
Brand	1423			14		
Group	3206			193		
No human	103			401		
Total Drugs:	**12929**	**22.6**	**2.3**	**1836**	**12.9**	**1.4**
DDIs						
ddi	178			10		
advice	819			8		
effect	1548			152		
mechanism	1260			8162		
Total DDIs:	**3805**	**6.6**	**0.7**	**232**	**1.6**	**0.2**

The directory also contains the DTD of the XML annotation files (DDIunified.dtd)

The XML annotation format for the corpus follows the dtd shown below:

```
<!ELEMENTdocument          (sentence*) >
<!ELEMENTsentence  (entity*,ddi*) >
<!ELEMENTentity      EMPTY>
<!ELEMENTddi                 EMPTY>

<!ATTLISTdocument
        id                   ID          #REQUIRED>

<!ATTLISTsentence
        id                   ID          #REQUIRED
        text        CDATA    #IMPLIED>

<!ATTLISTentity
        id                   ID          #REQUIRED
        charOffset  CDATA    #IMPLIED
        type        CDATA    #IMPLIED
        text        CDATA    #IMPLIED>

<!ATTLISTddi
        id                   ID          #REQUIRED
        e1          CDATA    #IMPLIED
        e2          CDATA    #IMPLIED
        typeCDATA            #IMPLIED>
```

An example of an annotated document is shown:

```
- <document id="DDI-DrugBank.d505">
  - <sentence id="DDI-DrugBank.d505.s0" text="No formal drug/drug interaction studies with Plenaxis were performed.">
      <entity id="DDI-DrugBank.d505.s0.e0" charOffset="45-52" type="brand" text="Plenaxis"/>
    </sentence>
  - <sentence id="DDI-DrugBank.d505.s1" text="Cytochrome P-450 is not known to be involved in the metabolism of
    Plenaxis.">
      <entity id="DDI-DrugBank.d505.s1.e0" charOffset="66-73" type="brand" text="Plenaxis"/>
    </sentence>
  - <sentence id="DDI-DrugBank.d505.s2" text="Plenaxis is highly bound to plasma proteins (96 to 99%).">
      <entity id="DDI-DrugBank.d505.s2.e0" charOffset="0-7" type="brand" text="Plenaxis"/>
    </sentence>
  - <sentence id="DDI-DrugBank.d505.s3" text="Laboratory Tests Response to Plenaxis should be monitored by measuring
    serum total testosterone concentrations just prior to administration on Day 29 and every 8 weeks thereafter.">
      <entity id="DDI-DrugBank.d505.s3.e0" charOffset="29-36" type="brand" text="Plenaxis"/>
      <entity id="DDI-DrugBank.d505.s3.e1" charOffset="83-94" type="drug" text="testosterone"/>
    </sentence>
  - <sentence id="DDI-DrugBank.d505.s4" text="Serum transaminase levels should be obtained before starting treatment with
    Plenaxis and periodically during treatment.">
      <entity id="DDI-DrugBank.d505.s4.e0" charOffset="76-83" type="brand" text="Plenaxis"/>
    </sentence>
    <sentence id="DDI-DrugBank.d505.s5" text="Periodic measurement of serum PSA levels may also be considered."/>
  </document>
```

Figure 1 Example of an annotated document

<document> element

The root element is the <document> element with the following attributes:

- **id**: an unique id which is composed by the name of the corpus (DDI-DrugBank or DDI-MedLine) and an identifier beginning with "d" and followed by a number.

<sentence> element

Each sentence of the document is contained within a <sentence> element.

Each <sentence> element has the following attributes:

- **id** – A unique id which is composed by the name of the corpus (DDI-DrugBank or DDI-MedLine), the id of the document (d505), and an id beginning with "s" and followed by the index of the sentence (the index of the first sentence should be 0).
- **text**- contains the text of the sentence.

Within the <sentence> element, there are the following elements: <entity> and <ddi>. Elements of type <entity> correspond to all annotated pharmacological substances. Elements of type <ddi> correspond to all annotated drug-drug interactions.

<entity> element

Each <entity> element has the following attributes:

- **id** – A unique id which is composed by the name of the corpus (DDI-DrugBank or DDI-MedLine), the id of the document (d505), the id of the sentence, and an id beginning with "e" and followed by the index of the entity in the sentence (the first entity of the sentence should have the index 0).
- **charOffsets** - contains the start and end positions, separated by a dash, of the mention in the sentence. When the mention is a discontinuous name, it will contain the start and end positions of all parts of the mention separated by semicolon (see Figure 2).
- **text** – stores the text of the mention.

- **type** – stores the type of the pharmacological substance (drug, brand, group or no-human).

Discontinuos names

Sometimes the names of pharmacological substances can appear as **discontinuous names** in texts. Discontinuous names usually arise from some of the following examples:

- **Coordinators**. The sentence bellow contains a coordinate structure (aluminum and magnesium hydroxides) with two different pharmacological substances (aluminum hydroxide, magnesium hydroxide). The first entity (aluminum hydroxide (DDI-DrugBank.d42.s5.e4)) presents a discontinuous name. Note that its charOffset attribute contains the start and end positions of the two parts of the mention (separated by semicolon).

```
- <sentence id="DDI-DrugBank.d42.s5" text="If a patient requires TIKOSYN and anti-ulcer therapy, it is suggested that omeprazole, ranitidine, or antacids (aluminum and magnesium hydroxides) be used as alternatives to cimetidine, as these agents have no effect on the pharmacokinetic profile of TIKOSYN.">
    <entity id="DDI-DrugBank.d42.s5.e0" charOffset="22-28" type="brand" text="TIKOSYN"/>
    <entity id="DDI-DrugBank.d42.s5.e1" charOffset="34-43" type="group" text="anti-ulcer"/>
    <entity id="DDI-DrugBank.d42.s5.e2" charOffset="75-84" type="drug" text="omeprazole"/>
    <entity id="DDI-DrugBank.d42.s5.e3" charOffset="87-96" type="drug" text="ranitidine"/>
    <entity id="DDI-DrugBank.d42.s5.e4" charOffset="102-109" type="group" text="antacids"/>
    <entity id="DDI-DrugBank.d42.s5.e5" charOffset="112-119;135-143" type="drug" text="aluminum hydroxide"/>
    <entity id="DDI-DrugBank.d42.s5.e6" charOffset="125-143" type="drug" text="magnesium hydroxide"/>
    <entity id="DDI-DrugBank.d42.s5.e7" charOffset="174-183" type="drug" text="cimetidine"/>
    <entity id="DDI-DrugBank.d42.s5.e8" charOffset="251-257" type="brand" text="TIKOSYN"/>
    <ddi id="DDI-DrugBank.d42.s5.d0" e1="DDI-DrugBank.d42.s5.e0" e2="DDI-DrugBank.d42.s5.e7" type="advise"/>
  </sentence>
```

Figure2 Discontinuous names (“aluminum and magnesium hydroxide”). The charOffset attribute (see entity DrugDDI.d42.s5.e4) contains the start and end positions of the two part of the mention (aluminum hydroxide) separated by semicolon.

- **Abbreviations**. In some cases, an abbreviation, acronym or part of the name appears in parentheses. As noted in our guidelines, a mention containing abbreviation or acronym of its drug name is considered as a unique entity (see Figure 3, entity DDI-DrugBank.d230.s1.e1; monoamine oxidase (MAO) inhibitors).

```
- <sentence id="DDI-DrugBank.d230.s1" text="All vasopressors should be used cautiously in patients taking monoamine oxidase (MAO) inhibitors.">
    <entity id="DDI-DrugBank.d230.s1.e0" charOffset="4-15" type="group" text="vasopressors"/>
    <entity id="DDI-DrugBank.d230.s1.e1" charOffset="62-95" type="group" text="monoamine oxidase (MAO) inhibitors"/>
    <ddi id="DDI-DrugBank.d230.s1.d0" e1="DDI-DrugBank.d230.s1.e0" e2="DDI-DrugBank.d230.s1.e1" type="advise"/>
  </sentence>
```

Figure 3 Abbreviations and acronyms can appear in the middle of a mention.

<ddi> element

Each <ddi> element has the following attributes:

- **id** – A unique id which is composed by the name of the corpus (DDI-DrugBank or DDI-MedLine), the id of the document (d505), the id of the sentence, and an id beginning with "d" and followed by the index of the ddi in the sentence (the first ddi of the sentence should have the index 0).
- **e1**- stores the id of the first interacting entity.
- **e2**- stores the id of the second interacting entity
- **type** – stores the type of the drug-drug interaction (ddi, advice, effect, mechanism).

An example of an annotated sentence (containing DDIs with type "advise") within the XML file is shown below:

```
- <sentence id="DDI-DrugBank.d568.s3" text="Dosage adjustment of STRATTERA may be necessary when
  coadministered with CYP2D6 inhibitors, e.g., paroxetine, fluoxetine, and quinidine.">
    <entity id="DDI-DrugBank.d568.s3.e0" charOffset="21-29" type="brand" text="STRATTERA"/>
    <entity id="DDI-DrugBank.d568.s3.e1" charOffset="98-107" type="drug" text="paroxetine"/>
    <entity id="DDI-DrugBank.d568.s3.e2" charOffset="110-119" type="drug" text="fluoxetine"/>
    <entity id="DDI-DrugBank.d568.s3.e3" charOffset="126-134" type="drug" text="quinidine"/>
    <ddi id="DDI-DrugBank.d568.s3.d0" e1="DDI-DrugBank.d568.s3.e0" e2="DDI-DrugBank.d568.s3.e1"
    type="advise"/>
    <ddi id="DDI-DrugBank.d568.s3.d1" e1="DDI-DrugBank.d568.s3.e0" e2="DDI-DrugBank.d568.s3.e2"
    type="advise"/>
    <ddi id="DDI-DrugBank.d568.s3.d2" e1="DDI-DrugBank.d568.s3.e0" e2="DDI-DrugBank.d568.s3.e3"
    type="advise"/>
  </sentence>
```

Figure 4 Example of an annotated sentence with four entities and three ddis.

Note that the drug-drug interactions were only annotated at the sentence level and, thus, any interactions spanning over several sentences are not annotated.

DDI Corpus Licence

Please attribute the corpus by citing the following paper:

Isabel Segura-Bedmar, Paloma Martínez, César de Pablo-Sánchez, (2011). Using a Shallow Linguistic Kernel for Drug-Drug Interaction Extraction, Journal of Biomedical Informatics, 44(5), 789 – 804.