Multilingual Resources

MICKUS, Timothee

Contents of the lecture

Encoding

Encoding

General caveat

- Multilingual NLP applications must deal with variations in writing:
  - Multiple alphabets or writing systems (latin, cyrillic, arabic, chinese...)
  - Multiple variants of similar writing systems: diacritics, etc.
- Computers deal with these variations by using different **encodings**: latin1, UTF-8, CP1251 ...
- Each encoding is a specific mapping of characters to binary representations, and vice-versa: low level text representation is done by manipulating the text in binary or byte format.
- Encodings generally cover a specific set of characters : latin1 covers only basic latin characters, CP1251 contains both latin and cyrillic letters, etc. The unicode standard defines both UTF-8 and UTF-16 encodings and tries to represent any possible character.
- The good practice is to keep track of the encoding of files, and, as much as possible, use **unicode** encoding (used by default in python 3)

Useful python library for detecting encoding: `chardet`

Word-level resources

Word-level Resources

Wiktionary

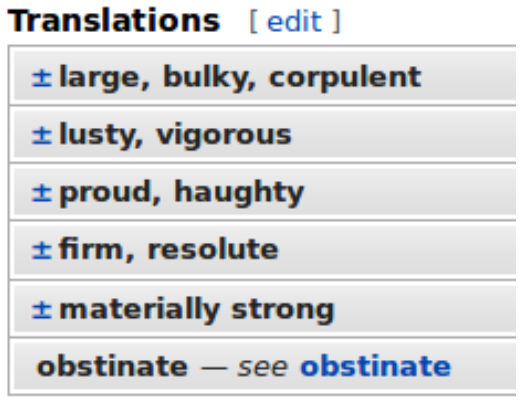
- Wiktionary is a collaboratively edited multilingual web-based project.
- The aim is to produce dictionaries for all the world's languages, currently it covers 171 languages
- Wiktionary data is frequently used in NLP, both in multilingual and in monolingual contexts

  cf. for instance GLAWI: http://redac.univ-tlse2.fr/lexiques/glawi.html which is a freely distributed resource for French, mapping morphological annotations from GLÀFF to definitions from the French wiktionary.
- As a consequence of the collaborative nature of the project, Wiktionary is generally deemed to have broad coverage, but unsystematic definitions.

Word-level Resources

Wiktionary

- Many dictionaries include relevant multilingual information that can be exploited in linguistic applications and experiments.
- In our case, wiktionary entries often have a "**Translations**" subsection



- ... which can be retrieved by parsing the XML dump

```
====Translations====
{{trans-top|large, bulky, corpulent}}
* Finnish: {{t|fi|pönäkkä}}, {{t+|fi|tanakka}}
* Greek: {{t+|el|}}
* Irish: {{t|ga|alpartha}}
```

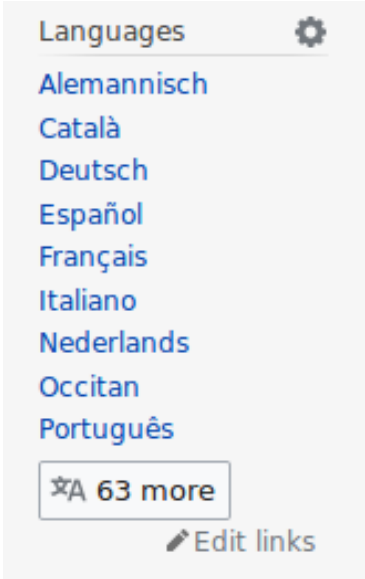
(dumps are available here: https://dumps.wikimedia.org/)

- but it's actually hard work.

Word-level resources

Wiki

- More generally, Wiki-based resources such as Wikipedia, Wiktionary, Wikimedia, etc. often display “interlanguage links” :



- Likewise, they can be retrieved by parsing the XML dump and it can quickly become a time-consuming task.
- More info : https://en.wikipedia.org/wiki/Help:Interlanguage_links

Word-level resources

Wordnet

The nltk implementation of wordnet boasts multilingual support

- The list of all codes for supported languages can be found using `wn.langs()`.
- Synsets can be queried with the `lang` keyword :

```
>>> wn.synsets('cane', lang='ita')
[Synset('dog.n.01'), Synset('cramp.n.02'),
Synset('hammer.n.01'), Synset('bad_person.n.01'),
Synset('incompetent.n.01')]
```

- It's possible to retrieve lemmas in a given language with the functions `synset.lemma_names()` and `synset.lemmas()` :

```
>>> dog = wn.synset('dog.n.01')
>>> dog.lemmas('ita')
[Lemma('dog.n.01.cane'), Lemma('dog.n.01.Canis_familiaris')]
>>> dog.lemma_names('ita')
['cane', 'Canis_familiaris']
```

- Lemmas have a `lemma.lang()` function that maps to their language code:

```
>>> lemma  = dog.lemmas('ita')[0]
>>> lemma.lang()
'ita'
```

- The function `wn.all_lemma_names()` can be restricted to a specific language using the `lang` keyword.

Word-level resources

Babelnet

Babelnet (https://babelnet.org/) is a network of concept mostly based on the integration of wikipedia, wiktionnary and wordnet, with an official Java API. It makes full use of the multilingual structures of Wordnet and Wiki-resources.

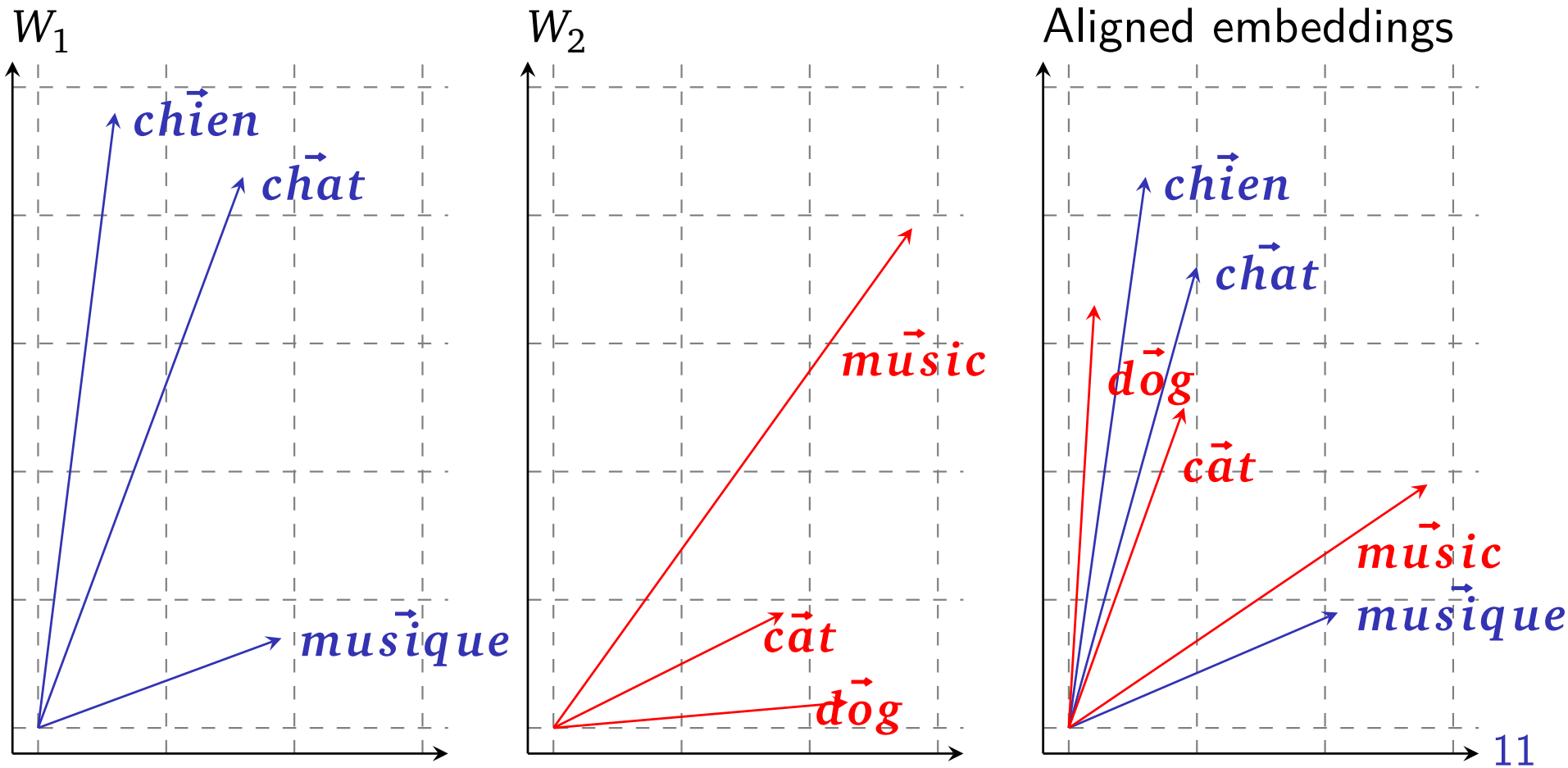
- Babelnet is a collection "Babel synsets", mapping of wordnet synsets to wikipedia pages.
- The mapping is initialized by first aligning pages and synsets which are monosemous (wikipedia pages with no disambiguation page associated, and synsets with only one lemmas).
- Redirections are mapped to the synset they redirect to.
- The rest of the mapping is computed by selecting the most probable sense in wordnet based on the content of the wikipedia page.

Word-level resources

Cross-lingual embeddings requirements

In some NLP applications, different sets of word embeddings from multiple languages are used jointly.

- To do this we need to project them in a **shared semantic space**, ie. we "**align**" them
- we want to make sure that items with a similar meanings are near one another: even more so when it comes to translation pairs



Word-level resources

Cross-lingual embeddings requirements

Aligning word embeddings for two different languages L_1 and L_2 require :

- a set of embeddings for each of the two languages L_1 and L_2,
- a set of word pairs (w_1, w_2) such that w_1 is a word of L_1 and w_2 is a translation for w_1 in L_2.

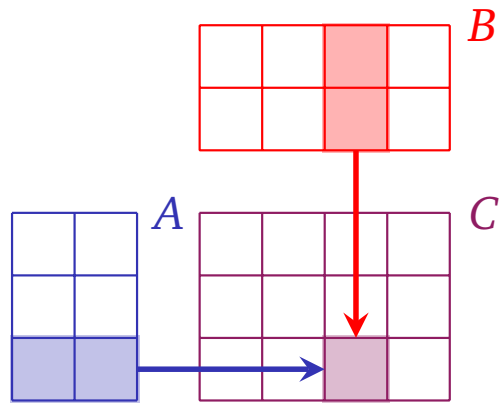
If these are available, various algorithms can be used to transfer the word embeddings in a common space, we'll focus on SVD (Smith et al., 2017).

Word-level resources

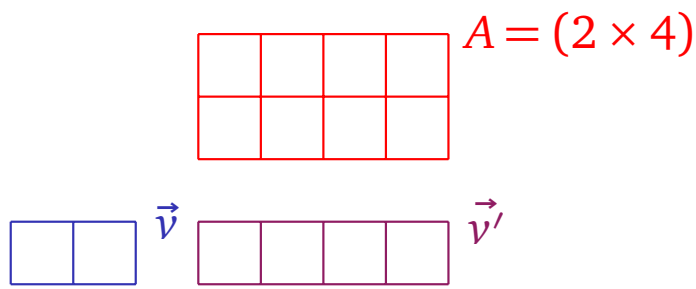
Matrix multiplication as a vector function

- The multiplication $C = A\,B$ of a matrix A of shape $(M \times N)$ and a matrix B of shape $(N \times P)$ is of shape $(M \times P)$. The cell $\langle i, j \rangle$ in C will have as value the dot product between the i^{th} **row** vector in A and the j^{th} **column** vector in B:

$$C_{ij} = \sum_{k=1}^{N} A_{ik} \times B_{kj}$$



- Therefore the multiplication of a vector $\vec{v}$ of shape $(1 \times d)$ and a matrix A of shape $(d \times d')$ is a vector $\vec{v'} = \vec{v}A$, of shape $(1 \times d')$

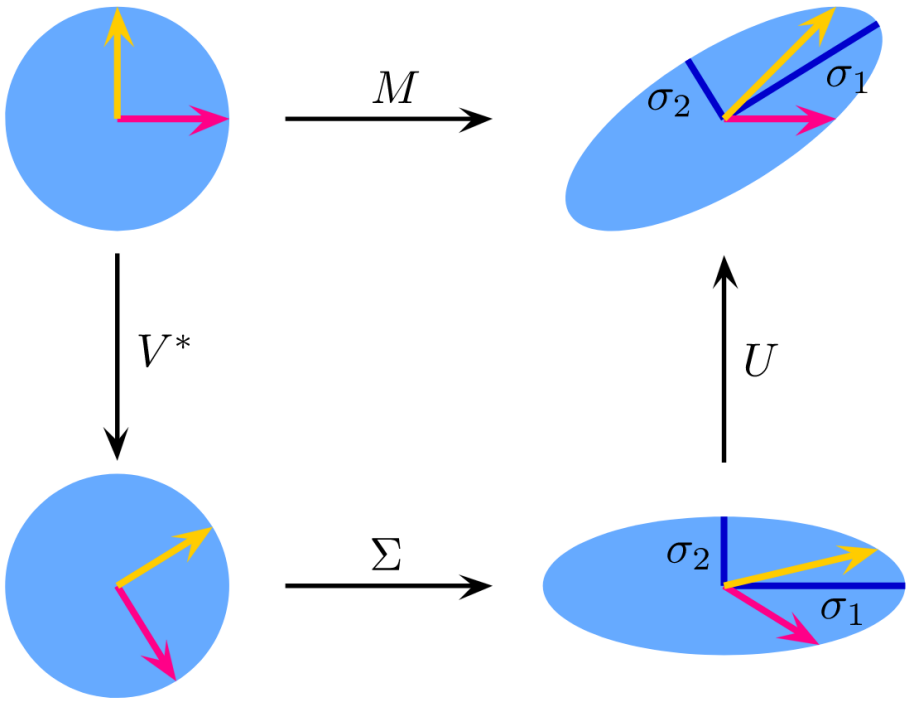


- Thus a matrix of shape $(d \times d')$ can be seen as a linear transformation, ie. a function mapping vectors from a space of dimension d to another space of dimension d'.

Word-level resources

Cross-lingual embeddings

- To align two word embedding spaces, we need to compute for each of them a linear transformation that projects the vectors into a shared space
- SVD decomposes a transformation into a rotation, followed by a scaling, and then a second rotation



$$M = U \cdot \Sigma \cdot V^*$$

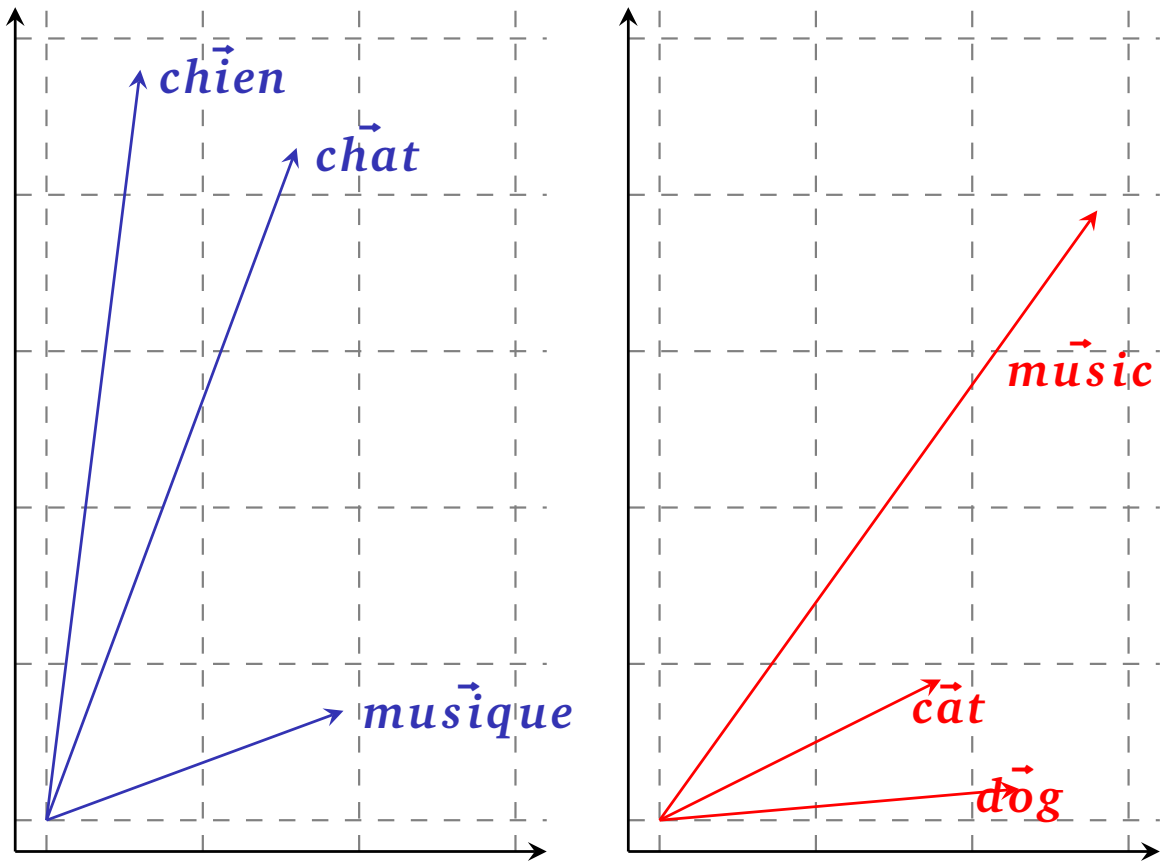
(image from wikipedia)

- it requires that the two sets of embeddings we align can be described as matrices where the i^{th} row in the one corresponds to the translation of the word for the i^{th} row in the other

Word-level resources

Understanding SVD alignment

We first need to assess what sort of function is needed to map one space to the other : we do that using the **matrix product**.



In this example, high values on the y axis in W_1 map to low values on the y' axis in W_2. This will be captured in the dimension wise-product:

$$P = W_1^T W_2$$

This product defines the conjunction of the distributional descriptions of word vectors: the cell $\langle y, y' \rangle$ corresponding to the importance of y in W_1 to compute y' in W_2 will be given a low coefficient. Rows in P will correspond to dimensions in W_1, and columns to dimensions in W_2.

Word-level resources

Understanding SVD alignment

We can then use SVD to see how we would need to rotate the two spaces so that they match.

- In linear algebra, SVD is a factorization of a Matrix M in three terms, U, Σ and V, such as $M = U\ \Sigma\ V^T$
- Σ is a diagonal matrix, and U and V are unitary matrix, ie. $U\ U^T = U^T\ U = I$ and $V\ V^T = V^T\ V = I$
- When M is a square matrix (of shape $K \times K$), U and V^T can be seen as rotations and Σ as a scaling factor.
- U is is a set of eigenvectors for the row vectors of M, and V likewise for the column vectors of M
- in the case of our two semantic spaces W_1 and W_2, if we define M as the conjunction of the effects in W_1 and W_2, ie. $M = W_1^T W_2$, we can therefore see U as the rotation mapping W_1^D to its natural description in a shared semantic space, and an approximation of the necessary rotation for W_1; likewise, we can see V as a natural description of W_2.

Word-level resources

Understanding SVD alignment

The function we saw previsouly was proposed by Smith et al. (2017). In detail, they use the following procedure to align two embedding matrices W_1 and W_2, using a bilingual lexicon $D = \langle w_1^i, w_2^i \rangle$:

I First compute W_1^D and W_2^D, the subsets of the matrices W_1 and W_2 containing only vectors of words present in D.

II Compute the matrix product $P = W_1^{DT}\ W_2^D$, which can be seen as pairing up W_1^D and W_2^D based on the vectors components.

III Then retrieve the rotations by computing the SVD: $P = U\Sigma V^T$

IV Apply the first rotation to W_1, and the second to W_2 : $W_1' = W_1 U$ and $W_2' = W_2 V$.

This is akin to rotating both word embedding spaces so that they are projected in the same space: we use the transformation U on the embedding space W_1, the superset of W_1^D as both relates to the rows of the dimension-wise product $P = W_1^{DT}\ W_2^D$; likewise we use V on W_2. This allows us to mesh together the semantic spaces.

Sentence-level multilingual resources

Sentence-level resources

Parallel corpora

Machine translation has been a long standing goal of NLP (The term was first coined by Warren Weaver in 1949)

- Machine translation (generally) requires parallel data: linguistic elements from a given source language must be mapped to another target language
- This alignment can be made at any linguistic level
- Today, most statistical & neural MT systems rely on parallel corpora of sentences in natural language

Sentence-level resources

Existing parallel corpora

This entails that many sentence-level parallel corpora can be found

- see for instance the corpora available at WMT: http://www.statmt.org/wmt19/translation-task.html
- ... or the opensubtitles datasets : http://opus.nlpl.eu/OpenSubtitles.php
- you can even find English-Inuktitut parliamentary parallel data : http://www.inuktitutcomputing.ca/NunavutHansard/info.php

Sentence-level resources

Parallel corpora

What does a parallel corpus look like? Europarl En ⟷ De

Source	**Target**
europarl-v7.de-en.en	europarl-v7.de-en.de
1 Resumption of the session	1 Wiederaufnahme der Sitzungsperiode
2 I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period.	2 Ich erkläre die am Freitag, dem 17. Dezember unterbrochene Sitzungsperiode des Europäischen Parlaments für wiederaufgenommen, wünsche Ihnen nochmals alles Gute zum Jahreswechsel und hoffe, da Sie schöne Ferien hatten.
3 Although, as you will have seen, the dreaded ...	3 Wie Sie feststellen konnten, ist der gefürchtete ...

Sentence-level resources

Universal annotations

Another type of multilingual resources are those concerned with universal annotation schemes.

- Although Chinese and Japanese have classifier whereas French doesn't, one can try to make an inventory of all the possible PoS-tags, and use it consistently across languages
- This idea has been seriously considered : cf. for instance the Universal PoS tagset of Petrov, Das, and McDonald (2012)
- Likewise, efforts have been made to consistently annotate morphosyntactical features across languages (for instance Interset by Zeman (2008) has been used to map features across languages, by deriving an "interlingua" representation)
- Lastly, much research has been made to present a cross-lingual dependency annotation scheme, called "**Universal Dependencies**" (UD, cf. http://universaldependencies.org).

Sentence-level resources

Universal Dependencies

- the UD project is an open collaboration, mainly coordinated by Joakim Nivre
- the UD project proposes dependency corpora in 79 languages across many linguistic phyla, from Akkadian to Yoruba and from Old French to Swedish Sign language.
- the annotation scheme is based on an integration of the Stanford Dependency annotations, the universal PoS tagset and the Interset interlingua for morphological features.
- the main idea of the UD project is to be both accessible to the non-specialist (learner, human annotator or NLP engineer) and linguistically accurate for each language (despite being universal).
- the corpora are each split in three (train, dev and test), and are available both as raw `.txt` format and as `.conllu` format

Sentence-level resources

`.conllu` format

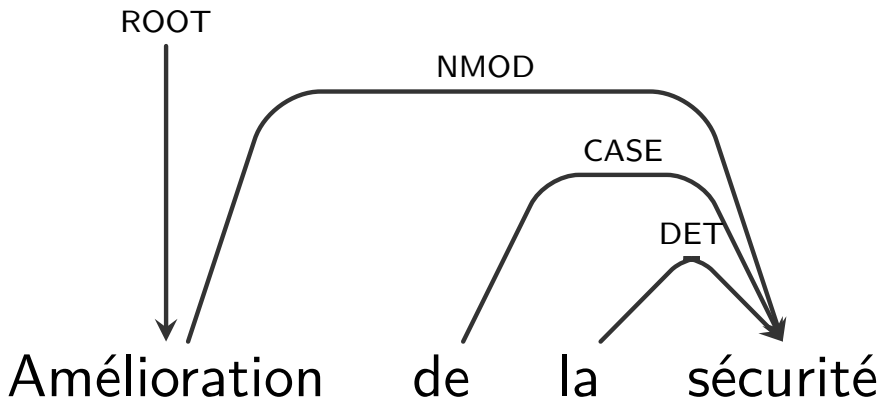
The `.conllu` format is a widely adopted format for dependency tree banks

- Each sentence is represented as the list of its tokens, eventually preceded by meta-information (eg. sentence ID or plain text) signalled by a # character at the start.
- Each token contains fields or "columns", separated by tabs, listed in a specific order :
  1. **ID**: its index in the sentence
  2. **FORM**: its word form
  3. **LEMMA**: its lemma, if available
  4. **UPOS**: its universal PoS tag
  5. **XPOS**: its language-specific PoS tag
  6. **FEATS**: its morphosyntactic features
  7. **HEAD**: the index of its head, or 0 if it is the root
  8. **DEPREL**: the dependency relation that it holds with respect to its head
  9. **DEPS**: an enhanced graph annotation
  10. **MISC**: any remaining miscellaneous annotation

  Any unspecified or missing information is represented using the _ character. ID cannot be missing. In UD tree banks, the UPOS, HEAD and DEPREL columns must not be unspecified or missing.
- blank lines separate sentences

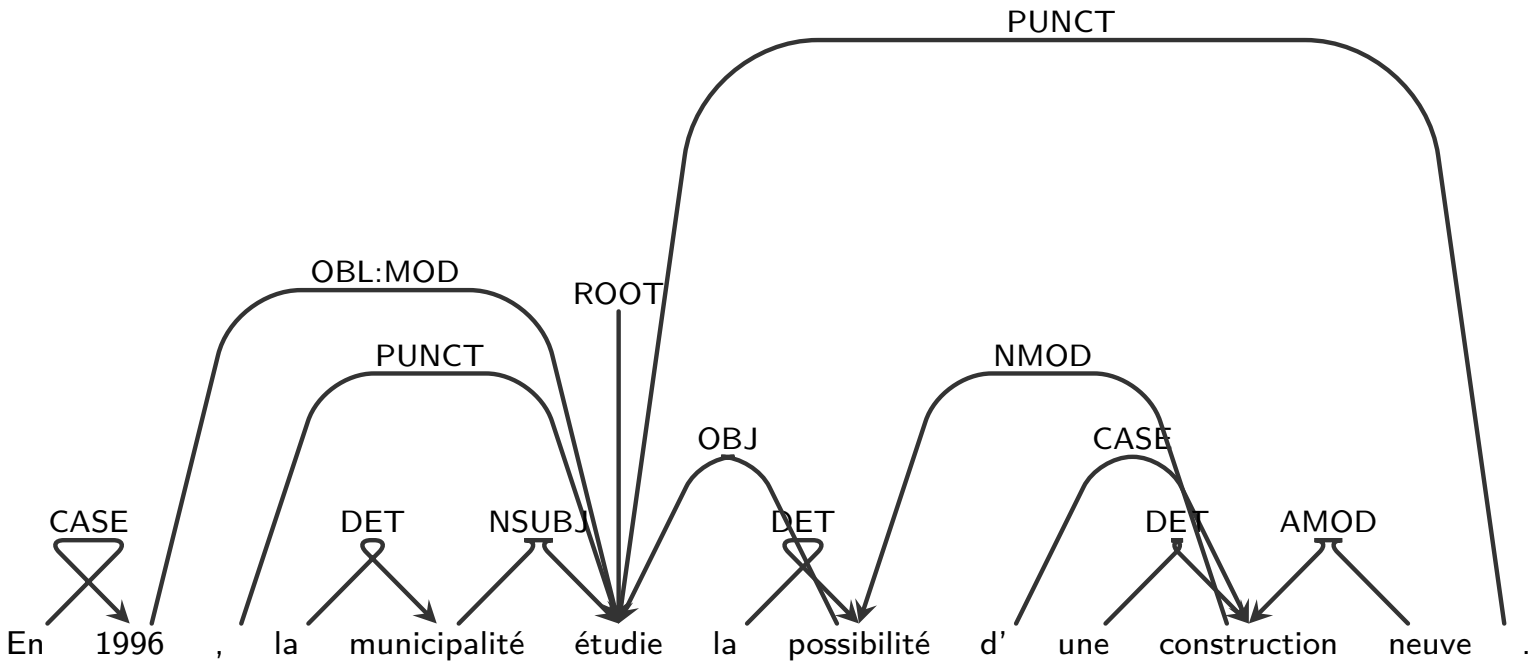
Sentence-level resources

UD example I




```
# sent_id = annodis.er_00007
# text = Amélioration de la sécurité
1       Amélioration    amélioration    NOUN    _       Gender=Fem|Number=Sing  0       root    _       _
2       de      de      ADP     _       _       4       case    _       _
3       la      le      DET     _       Definite=Def|Gender=Fem|Number=Sing|Pron    Type=Art        4       det     _       _
4       sécurité        sécurité        NOUN    _       Gender=Fem|Number=Sing  1       nmod    _       _
```

Sentence-level resources




```
# sent_id = annodis.er_00029
# text = En 1996, la municipalité étudie la possibilité d'une construction neuve.
1       En      en      ADP     _       _       2       case    _       _
2       1996    1996    NUM     _       NumType=Card    6       obl:mod _       SpaceAfter=No
3       ,       ,       PUNCT   _       _       6       punct   _       _
4       la      le      DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art        5       det     _       _
5       municipalité    municipalité    NOUN    _       Gender=Fem|Number=Sing  6       nsubj   _       _
6       étudie  étudier VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0       root    _       _
7       la      le      DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art        8       det     _       _
8       possibilité     possibilité     NOUN    _       Gender=Fem|Number=Sing  6       obj     _       _
9       d'      de      ADP     _       _       11      case    _       SpaceAfter=No
10      une     un      DET     _       Definite=Ind|Gender=Fem|Number=Sing|PronType=Art        11      det     _       _
11      construction    construction    NOUN    _       Gender=Fem|Number=Sing  8       nmod    _       _
12      neuve   neuf    ADJ     _       Gender=Fem|Number=Sing  11      amod    _       SpaceAfter=No
13      .       .       PUNCT   _       _       6       punct   _       _
```