UNIVERSITÉ DE BELGRADE
FACULTÉ DE PHILOLOGIE
Département d'études romanes

avec le concours de

Ministère de la recherche, du développement technologique
et des innovations de la République de Serbie
& Agence universitaire de la Francophonie (AUF)

journées serbes des DICTIONNAIRES

COLLOQUE INTERNATIONAL FRANCOPHONE

(Méta)phraséographie et phraséomatique : exploration et innovation dans le projet DiCoP

CHEN Lian 陈恋
loselychen@gmail.com

LLL, Université d'Orléans

Membre associé aux laboratoires LT2D-Centre Jean Pruvost, Cergy Paris Université

Phraséologie et phraséoculturologie

Unité phraséologique (UP)

shúyǔ 熟语

Unité phraséologique : « […] **séquence polylexicale** constituée de deux ou plusieurs mots graphiques catégoriellement liés, contigus ou non. Les UP se caractérisent linguistiquement par : (i) **un certain degré de fixité syntaxique** (blocage des propriétés transformationnelles et ordre des constituants inaltérable) ; et/ou (ii) un certain degré **de figement sémantique** (non-compositionnalité au moins partielle) ; et/ou (iii) un certain degré **de figement lexical** (restriction paradigmatique) ; et / ou (iv) une contrainte sur l'emploi en situation de communication.» (BOLLY, 2011 : 28)

Ex : collocations, proverbes, expressions idiomatiques

Ex: 谚语*yànyǔ*,成语 *chéngyǔ*, 惯用语 *guànyòng yǔ*,, 歇后语*xiēhòuyǔ,* etc.

L'étude de ce phénomène culturel idiomatique spécifique en phraséologie s'appelle :
« **phraséoculturologie** »熟语文化学 *shú yǔwén huàxué* (L. Chen 2022).

(Méta)phraséographie et phraséomatique

Métaphraséographie (L. Chen 2022)

une (sous-)discipline dont l'objectif est l'étude des types de dictionnaires UP et des méthodes par lesquelles ils sont construits, ainsi qu'un objet de réflexion et de recherche sur la conception phraséographique.

→ théorique

Phraséographie (L. Chen 2022, 2023)

une branche de la phraséologie appliquée dont l'objet d'étude est le développement de collections, de glossaires et de dictionnaires d'UP

→ appliqué

Phraséomatique (L. Chen 2023)

À l'ère numérique et de l'intelligence artificielle (IA) contemporaine, le domaine de la phraséologie informatisée se trouve confronté à des défis importants, mettant en lumière l'émergence d'une discipline particulière : **la phraséomatique, ou la phraséologie informatique** (Chen 2023)

1. Unités phraséologiques et *shúyǔ* dans le DiCoP

Unités phraséologiques (UP) **VS** *Shúyǔ* 熟语

Défigements /Jeux de mots
tout feu tout **femme,** défigé de [être]
tout feu tout **flamme**

Collocations

	handicapé
	déficitaire
gravement	brûlé
	préjudiciable
	intoxiqué
	blessé

Proverbes
Le chat sauvage ne sait pas où il a pris une poule.

Expressions idiomatiques
avoir le coeur sur la main

Degré de figement

熟语活用 *Shúyǔ huóyòng*
默默无蚊 *mòmòwú**wén*** défigé de 默默无闻 *mòmòwú**wén***
faire quelque chose sans être entendu/rester inconnu/tomber dans l'oubli.
Le caractère « 蚊 wén » (moustique) a remplacé le « 闻 wén » (entendre).

惯用语Guànyòng yǔ (expression usuelle)
戴高帽 *dàigāomào* (porter, haut, chapeau) : porter un chapeau haut/ flatter quelqu'un/lécher les bottes, encenser

歇后语Xiēhòuyǔ (calembour à tiroir) :
和尚打伞 – 无发(法)无天 héshàngdǎsǎn – wúfà(fǎ)wú tiān : Un moine tient un parapluie – il n'a ni cheveu (discipline) ni ciel/ni dieu/il est sans respect ; sans foi ni loi

谚语Yànyǔ (proverbe)
虚心使人进步，骄傲使人落后xūxīn shǐ rén jìnbù, jiāo'ào shǐ rén luòhòu : L'orgueil provoque l'échec l'humilité engendre le progrès.

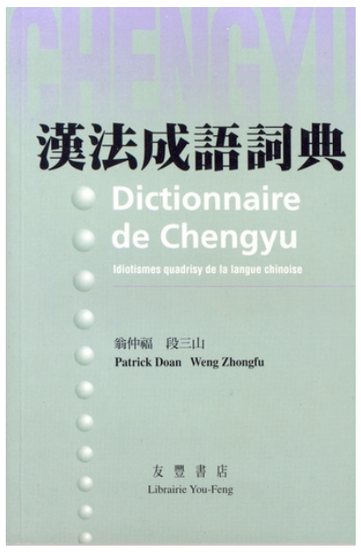
成语Chéngyǔ (expression toute faite)
龙腾虎跃*lóngténg-hǔyuè* (le dragon, sauter, le tigre, sauter): le dragon et le tigre bondissent/être plein de dynamisme, travailler ferme et viser haut.

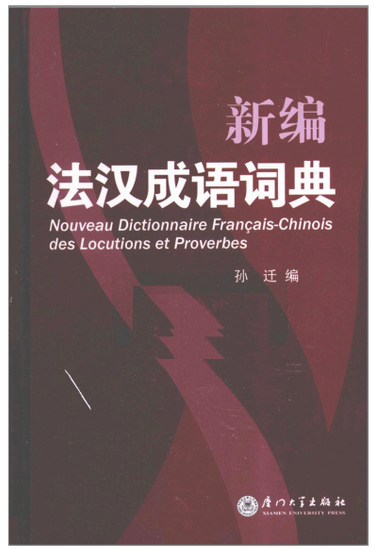
Les ressources numériques propres aux UP restent majoritairement monolingues

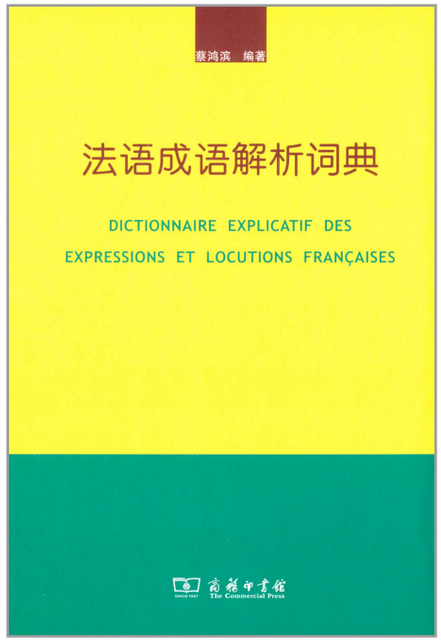
Expressio.fr

linternaute

Dictionnaires papiers bilingues chinois-français et français-chinois







Français- Portugais

CNRTL Centre National de Ressources Textuelles et Lexicales

Ortolang Outils et Ressources pour un Traitement Optimisé de la LANGue

Accueil | Portail lexical | Corpus | Lexiques | Dictionnaires | Métalexicographie | Outils | Contact

Dictionnaire d'expressions idiomatiques

Français - Portugais - Français

accueil | parcourir | **rechercher** | classement par concepts | télécharger | aide

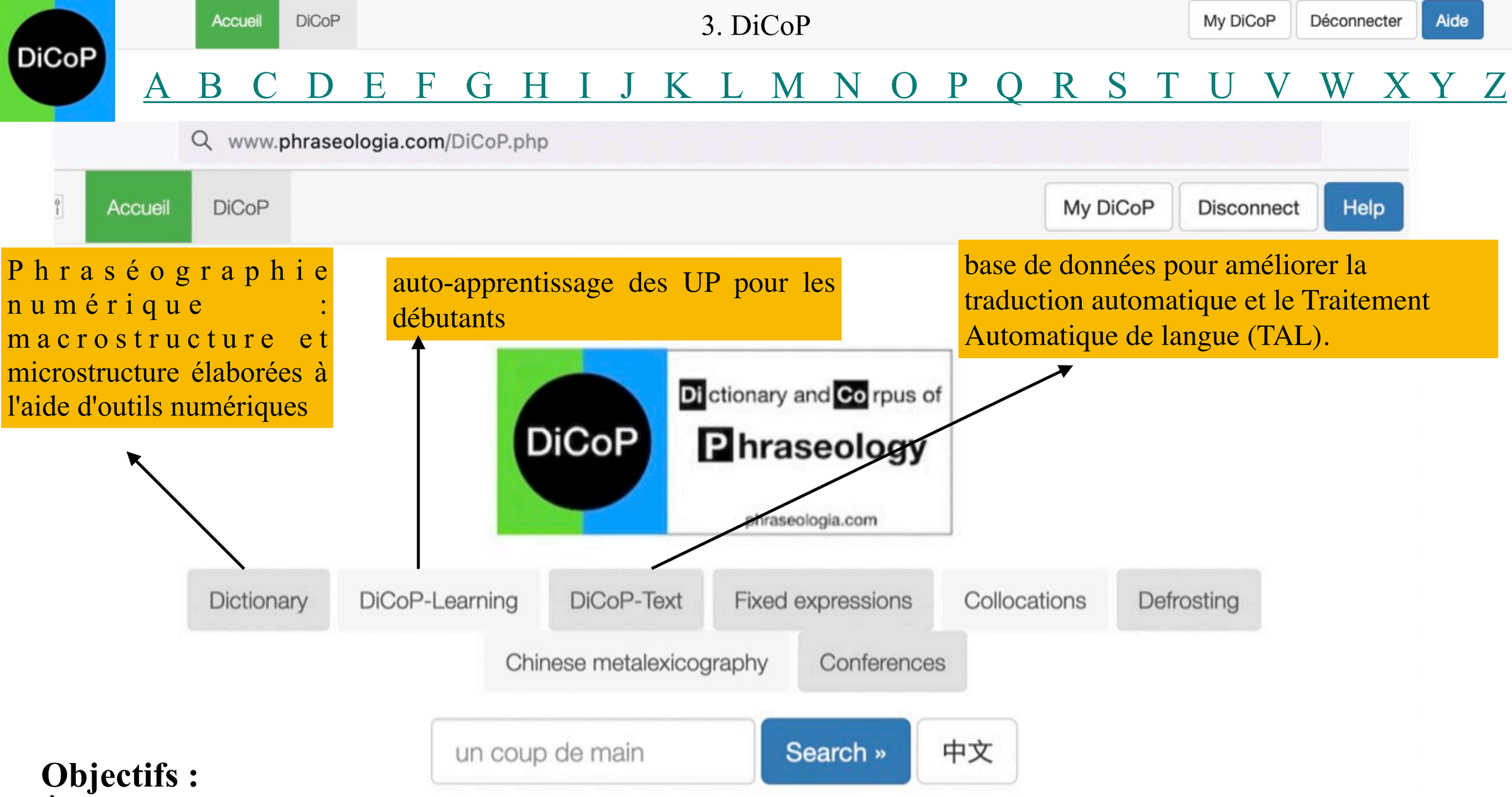
Recherche d´une expression ou d'un terme

en français : un coup de main rechercher en portugais : rechercher

Error connecting to mysql

https://www.cnrtl.fr/dictionnaires/expressions_idiomatiques/recherche.php

Français- Chinois ?
Chinois -Français ?



Objectifs :

À l'ère numérique et de l'intelligence artificielle (IA) contemporaine, le projet DiCoP (**Dictionnaire et Corpus de la Phraséologie (**disponible sur phraseologia.com, en cours de développement) vise à développer **un dictionnaire numérique de phraséologie multilingue** (dans un premier temps français-chinois/chinois-français, et à terme multilingue), **basé sur un corpus d'unités phraséologiques** et **des bases de données associées pour déterminer leur fréquence d'utilisation** (dans les journaux, les œuvres littéraires, les manuels scolaires, etc.) **dans la pratique et donc leur vitalité**, pour améliorer leur traduction automatique, et dans le but de faciliter l'accès aux unités phraséologiques.

3.1 Macrostructure du DiCoP

Papier -> numérique

Corpus numérisé monolingue/bilingue (6000 UP)

.doc

Xml-TEI-DTD -> csv

python

Extraction des données (2428 UP) du dictionnaire numérique existant

. CSV

Classement thématique

animaux (ordre alphébatique)

corps humain

chiffres ....

Classement alphabétique

A

B

C....

Macrostructure du DiCoP

Classement thématique

Ordre alphabétique

Exemple

绞尽脑汁

jiǎo jìn nǎo zhī

requête monolingue

mot clé

1)脑

Rechercher le *chéngyǔ* complet

绞尽脑汁

EI1	EI2	EI3
EI4	EI5	EI6
EI7	EI8	EI9

绞尽脑汁

Pinyin : jiǎo jìn nǎo zhī

Source : 出自老舍《四世同堂·愉生》

Signification : 形容苦思积虑，费尽脑筋，想尽办法；费尽心思去思考一件事情。

Exemple 1:

Exemple 2:

requête interlingue

Traduction en français

→ ◉ se creuser la cervelle

□ ◉ se casser la tête

◐ se pressurer le cerveau (Fam.)

□ ◐ se torturer les méninges (l'esprit, l'intellect).

□ mettre son esprit à la torture pour…

Phraséoculture :
Influencés par la médecine traditionnelle et la philosophie dans les deux pays, les chengyu chinois relatifs au cœur correspondent souvent aux EI françaises relatives à la tête.
Par exemple : les expressions « se creuser la tête », « se creuser la cervelle » ou « se casser la tête » équivalent aux expressions suivantes :
费尽心思 fèi jìn xīnsī (coûter, tout, cœur, pensée)
煞费苦心 shà fèi kǔ xīn (très, dépenser, avec acharnement, cœur)
挖空心思 wā kōng xīn sī (creuser, vide, cœur, pensée)
用心良苦 yòng xīn liáng kǔ (utiliser, cœur, très, dur)

Le dualisme « esprit-cœur » ou la dichotomie « raison-émotion » (Yu 2007b : 27), concepts opposés dans les cultures occidentales, n'existe pas dans la tradition chinoise.

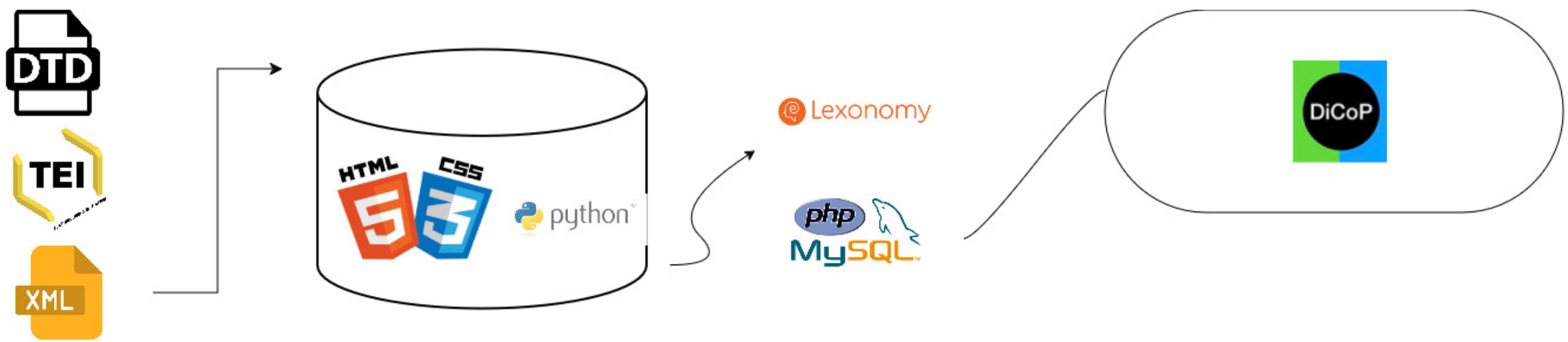
Traduction en anglais

Traduction en ...

La macrostructure est **la structure d'accès principale du dictionnaire : ordre alphabétique, traits des caractères en chinois, …**

Phraséoculture : dimension culturelle des UP (L. Chen 2022).

Le processus global



Extraction des EI existantes du laboratoire ATILF du format XML au CSV à l'aide de Python.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE TEI SYSTEM "dico_tei.dtd">
<TEI>
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>Dictionnaire électronique d'expressions idiomatiques</title>
                <author>Claudia Maria Xatara</author>
                <respStmt>
                    <name>Etienne Petitjean</name>
                </respStmt>
            </titleStmt>
            <publicationStmt>
                <distributor>
                    <name>ATILF UMR7118 – CNRS/UN2/UHP</name>
                </distributor>
            </publicationStmt>
            <sourceDesc>
                <p/>
            </sourceDesc>
        </fileDesc>
    </teiHeader>
    <text>
        <body>
            <entry>
                <form>
                    <orth>abandonner la partie</orth>
                </form>
                <gramGrp>
                    <pos>loc. verb.</pos>
                </gramGrp>
                <sense>
                    <def>renoncer à poursuivre une action</def>
                    <cit type="example">
                        <quote> Il n'a plus jamais remarché après son hospitalisation en février
                            2001. Il a lutté même si <oVar> les médecins ont abandonné la partie
                            </oVar> et jugé qu'il mourrait. </quote>
                        <bibl>
                            <ref target="www.meningitis.ca/fr/my_story/ernest.asp"/>
                            <date>28/07/05</date>
                        </bibl>
                    </cit>
                    <usg type="dom">RENONCEMENT</usg>
                    <xr type="syn">
                        <ref>quitter la partie</ref>
                    </xr>
```

1) Extraction des données au format XML conformément à la TEI-DTD.


```
<sans nom> *   EI-ATILF.py

1
2  from lxml import etree
3  infile = open("fp.xml", "r")
4  outfile = open("fp.csv", "w")
5  tree = etree.parse(infile)
6  outfile.write("entrees idiomatiques ATILF\tsynonymes\n")
7  for entry in tree.findall("//entry"):
8      form = entry.findall("form/orth")
9      for orth in form:
10         idiom = orth.text
11         outfile.write(idiom+"\t")
12     synonymes = ""
13     for syn in entry.findall("sense/xr"):
14        attriType = syn.get("type")
15        if attriType == "syn":
16            for ref in syn.findall("ref"):
17                synonymes += ref.text+" - "
18     outfile.write(synonymes.strip(" - "))
19     outfile.write('\n')
20
21 outfile.close()
22
```

2) Développement d'un programme avec python pour les traiter

3) enregistrement les données au format CSV. par ordre alphabétique.

J10

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
277	avoir les dents longuesavoir la dent - avoir les crocs - avoir l,Äôestomac dans les talons														
278	avoir les deux pieds sur terreavoir les pieds sur terre														
279	avoir les foies (blancs)avoir les jetons														
280	avoir les jetonsavoir les foies (blancs)														
281	avoir les jambes coupV©es														
282	avoir les mains dans les poches														
283	avoir les mains libres														
284	avoir les pieds sur terreavoir les deux pieds sur terre														
285	avoir les portugaises ensablV©es														
286	avoir l,Äôesprit d,Äôescalier														
287	avoir l,Äôesprit mal tournV©														
288	avoir l'estomac bien accrochV©avoir le coeur bien accrochV©														
289	avoir l,Äôestomac dans les talonsavoir la dent - avoir les crocs														
290	avoir le vent en poupealler comme sur des roulettes														
291	avoir le verbe facile														
292	avoir le verbe haut														
293	avoir le vin triste														
294	avoir [en] par-dessus la tV™teavoir [en] assez - avoir les boules - avoir [en] marre - avoir [en] plein le dos - avoir [en] plein les bottes - avoir [en] plein les couilles - avoir [en] ras la casquette														
295	avoir pignon sur rue														
296	avoir [en] plein le dosavoir [en] assez - avoir les boules - avoir [en] marre - avoir [en] par-dessus la tV™te - avoir [en] plein les bottes - avoir [en] plein les couilles - avoir [en] ras la casquette														
297	avoir plus d,Äôun tour dans son sac														
298	avoir plusieurs casquettes														
299	avoir plusieurs cordes √† son arc														
300	avoir quelque chose dans le ventre														
301	avoir [en] ras la casquetteavoir [en] assez - avoir les boules - avoir [en] marre - avoir [en] par-dessus la tV™te - avoir [en] plein le dos - avoir [en] plein les bottes - avoir [en] plein les couilles - avoir [en														
302	avoir ses tV™tes														
303	avoir son plein														
304	avoir sous la main														
305	avoir [en] sous les bigoudis														
306	avoir toute sa tV™te														
307	avoir tout faux														
308	avoir un chat dans la gorge														
309	avoir un cheveu sur la langue														
310	avoir un coeur d,Äôartichaut														
311	avoir un coup dans l,Äôaileavoir un coup dans le nez - avoir un verre dans le nez														

3.2 Microstructure informatisée de DiCoP

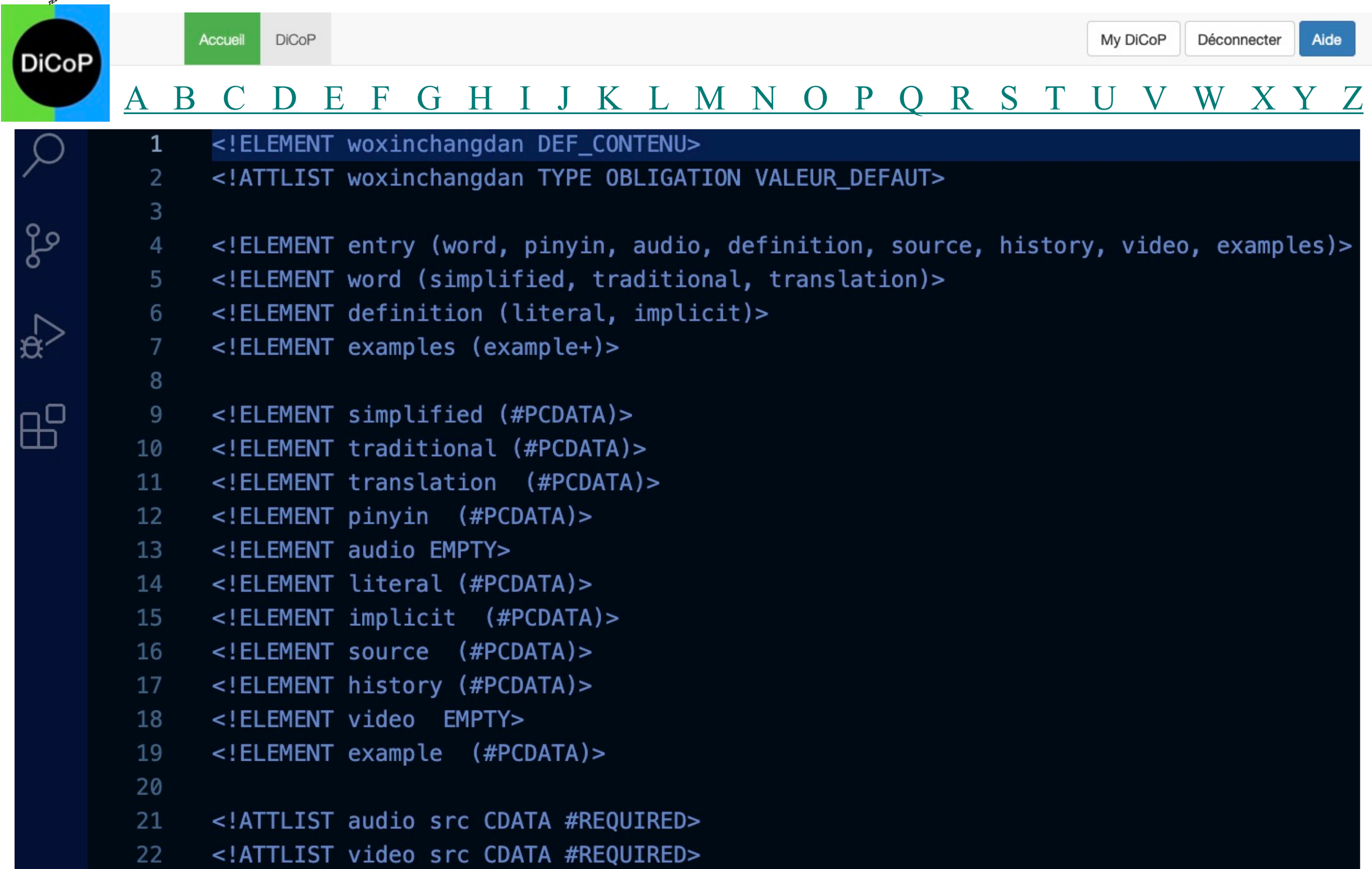
Microstructure : l'ensemble des informations ordonnées de chaque article, réalisant un programme d'informations constant pour tous les articles, et qui se lisent horizontalement à la suite de l'entrée


```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE product SYSTEM "product.dtd">

<entry>
  <word>
    <simplified> <b> 回头是岸 </b></simplified><br style="display:block;"/>
    <traditional><b> 回頭是岸 </b></traditional><br style="display:block;"/>
    <translation>retourner, tête, être, rivage</translation><br/>
  </word><br/>
  <pinyin font="italic">huí tóu shì àn</pinyin> <audio src="audio.mp3" /><br/>
  <definition><br/>
    <literal> si on tourne la tête, on retrouve le rivage</literal><br/>
    <implicit>il suffit de revenir sur son erreur pour recouvrer le droit chemin/□ revenir dans
    le droit chemin/il n'est jamais trop tard pour se corriger (pour s'amender, pour bien
    faire).
</implicit><br/>
  </definition><br/>
  <source> Premier acte de 《度翠柳》 (Du Cuiliu), dynastie des Yuan, anonyme </source><br/>
  <history>Le texte originel est «<0xa0>苦海无边，回头是岸<0xa0>Kǔ hǎi wú biān, huí tóu shì àn » :
  La mer de souffrances est immense, mais il suffit de reculer pour trouver terre ferme. Cette
  expression vient de la conception bouddhiste, selon laquelle une personne coupable qui
  accepte de se repentir, va pouvoir arriver de « l'autre côté » et libérer son âme en expiant
  ses péchés. Cette croyance est assez semblable à celle de la religion chrétienne qui insiste
  sur la nécessité du repentir pour « sauver son âme ».</history><br/>
  <video src="video.mp4" /> <br/>
  <example> 奉劝那些赌徒们，赌海无边，回头是岸，还是早些醒悟吧! Je voudrais dire à ces joueurs que la
  mer des jeux d'argent est sans limites, et que le retour en arrière est le rivage ; ils
  feraient donc mieux de se réveiller plus tôt !</example><br/>
  <example> 爱上一个这样的人多不幸，回头是岸 。C'est malheureux de tomber amoureux d'une telle
  personne, mais il n'est jamais trop tard pour se corriger et il est facile de se retourner.</
  example><br/>
</entry>
```

L'Extensible Markup Language, généralement appelé XML, « langage de balisage extensible » en français, est un métalangage informatique de balisage générique qui est un sous-ensemble du Standard Generalized Markup Language.

La norme Extensible Markup Language (XML) basée sur la Document Type Definition - Text Encoding Initiative (DTD-TEI) en XML pour les dictionnaires a été retenue, notamment dans l'optique d'assurer la pérennité de la ressource.



3.2 Microstructure informatisée de DiCoP

pf.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE TEI SYSTEM "dico_tei.dtd">
<TEI>
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>Dictionnaire électronique d'expressions idiomatiques</title>
                <author>Claudia Maria Xatara</author>
                <respStmt>
                    <name>Etienne Petitjean</name>
                </respStmt>
            </titleStmt>
            <publicationStmt>
                <distributor>
                    <name>ATILF UMR7118 – CNRS/UN2/UHP</name>
                </distributor>
            </publicationStmt>
            <sourceDesc>
                <p/>
            </sourceDesc>
        </fileDesc>
    </teiHeader>
```

Enfin, nous utilisons le format TEI-DTD **pour faciliter l'utilisation dans d'autres langages, car il s'agit d'une norme universelle qui s'applique à différents types de textes et à des langues différentes**.

dico_tei.dtd

```dtd
<?xml version="1.0" encoding="UTF-8"?>


<!ENTITY % _type "(reg| hint | gram | syn | ant | freq | dom | style)">

<!ELEMENT TEI (teiHeader, text)>
<!ELEMENT teiHeader (fileDesc)>
<!ELEMENT fileDesc (titleStmt, publicationStmt, sourceDesc)>
<!ELEMENT titleStmt (title, author, respStmt) >
<!ELEMENT respStmt (name) >
<!ELEMENT title (#PCDATA) >
<!ELEMENT author (#PCDATA) >
<!ELEMENT publicationStmt (distributor) >
<!ELEMENT distributor (name) >
<!ELEMENT name (#PCDATA) >
<!ELEMENT p (#PCDATA) >
<!ELEMENT sourceDesc (p) >
<!ELEMENT text (body) >
<!ELEMENT body (entry+) >
<!ELEMENT entry (form, gramGrp,sense+, usg*)>
<!ELEMENT form (orth) >
<!ELEMENT orth (#PCDATA) >
<!ELEMENT gramGrp (pos, gen*, number*) >
<!ELEMENT pos (#PCDATA)>
<!ELEMENT gen (#PCDATA) >
<!ELEMENT number (#PCDATA) >
<!ELEMENT sense (def, usg*,cit, usg, xr*, cit+) >
<!ATTLIST sense n CDATA #IMPLIED >
<!ELEMENT def (#PCDATA) >
<!ELEMENT cit ((quote, bibl) | (tr, ref*, usg*)) >
<!ATTLIST cit type (example|translation) #REQUIRED >
<!ELEMENT quote (#PCDATA |oVar)* >
<!ELEMENT oVar (#PCDATA) >
<!ELEMENT  bibl (ref, date) >
<!ELEMENT ref (#PCDATA) >
<!ATTLIST ref target CDATA #IMPLIED >
<!ELEMENT date (#PCDATA) >
<!ELEMENT usg (#PCDATA)>
<!ATTLIST usg type %_type; #REQUIRED>
<!ELEMENT xr (ref, usg*,ref?) >
<!ATTLIST xr type (ant | syn) #REQUIRED >
<!ELEMENT tr (#PCDATA) >

```

3.2 Microstructure informatisée du DiCoP

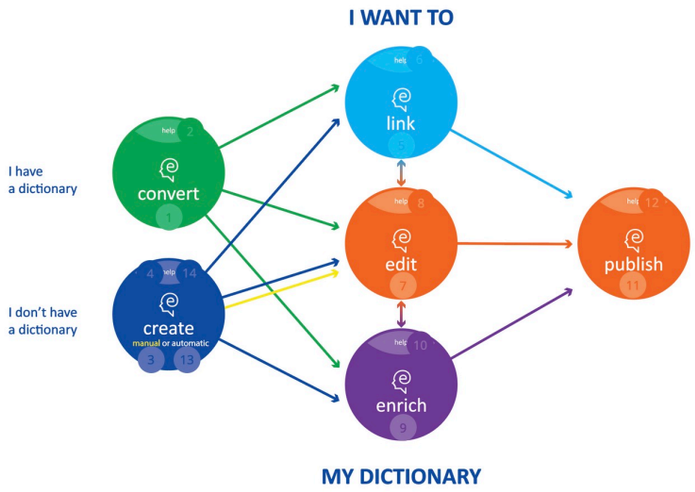
Lexonomy

https://lexonomy.elex.is/

Un système de rédaction de dictionnaires en ligne pour rédiger et publier des dictionnaires.

elexis

https://elex.is/



un maximum de 500 entrées

php MySQL

PHP : Hypertext Preprocessor, plus connu sous son sigle PHP, **est un langage de programmation libre, principalement utilisé pour produire des pages Web dynamiques via un serveur web.**

MySQL : un système de gestion de bases de données relationnelles.

python

langage de programmation

grue (n.f.) Faire le pied de grue « attendre debout » [exp.] v. 鹄立，久等

1. 该表达式来源于17世纪的固定表达faire la jambe de grue和 faire de la grue (和16世纪的faire la grue)。在Faire de la grue中，grue这个名词有动词"等待"的含义 (如在诗人 Maurice Scève作品中), 但是Bonaventure des Périers认为，在16世纪的语境中faire de la grue中的grue一词也含有鸟类:鹤 的比喻义。

2. 在 16 世纪，鹤的比喻用途通常是贬义的(être grue 愚蠢，suivre la multitude comme les grues [Calvin], s'en aller comme des grues [ibid.] 像鹤一样跟随众人，没有自己的想法。以及 Le Roux 在 1752 年给出的变体:être planté comme une grue，与 être planté comme un sot同义 : 一动 不动地站着，等待很久，久等。最后，"妓女"的意思来自faire le pied de grue，人行道上妓女等 待客人的形象比喻。

   Et d'indiquer des yeux les dizaines de Noires et de mûlatresses qui font, tout près, le pied de grue en se pavanant lascivement. (Le Point, 18-03-1995)

   并使眼色表明附近几十个黑人妇女和黑白混血女人正在作出淫荡姿态等客上门呢。

Actuellement, le site Web (PHP+MySQL) n'a pas encore été rendu public.

DiCoP | Accueil | DiCoP | My DiCoP | Déconnecter | Aide

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

localhost:8888/AFPC2/sangumaolu3.php

Photo

UP + Phraséoculture

三顾茅庐 三顧茅廬

trois, se rendre, chaumière

sāngùmáolú

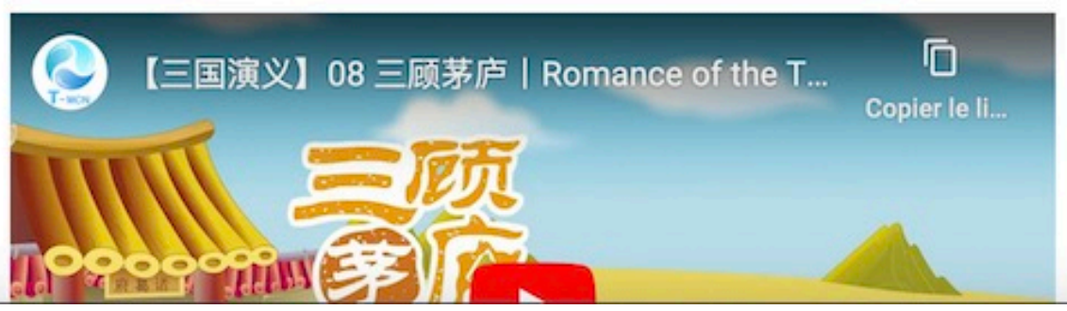
Aller visite qqn à trois reprises dans sa chaumière

rendre plusieurs visites à un homme de talent pour le prier de sortir d'une vie recluse ou de la retraite/inviter qqn plusieurs fois de suite à prendre une fonction importante.

「出师表 Chūshībiǎo」 (mémoriaux) écrits par Liang Zhuge (227-228)

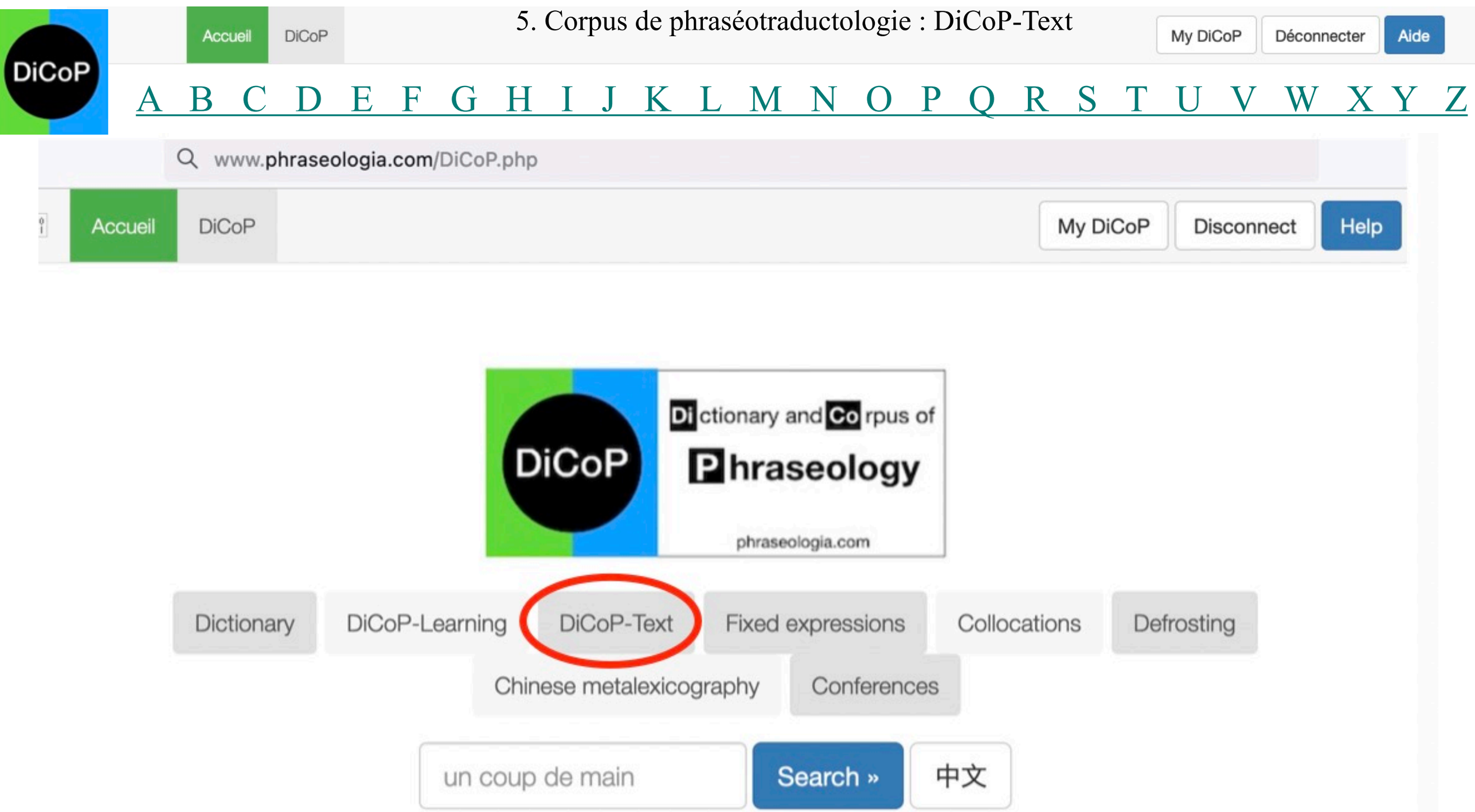
LIU Bei (161 - 223), puissant seigneur chinois de la fin de la dynastie Han et du début de la période des Trois Royaumes, veut que le célèbre stratège ZHU Geliang devienne son conseiller militaire. Lors de ses deux premières visites, ZHU Geliang est intentionnellement absent. La troisième fois, LIU Bei réussit à le rencontrer et à le persuader d'accepter ce poste.

> *这一年冬天有人三顾茅庐，感恩知己，一夕倾谈遂相许以驱驰了。（唐弢《桥》）*



Vidéo

5. Corpus de phraséotraductologie : DiCoP-Text



DiCoP-Text : La phraséologie pose particulièrement des difficultés en matière de traduction, car influencée tant par la linguistique que la culture ou la stylistique, et reflète fortement les choix et la technique de traduction du traducteur. **DiCoP-Text** sera **une base de données** (corpus monolingue, parallèle, contrastive, multilingue) **permettant de connaître la fréquence d'emploi des UP**, afin d'en vérifier la vitalité dans la pratique. Il doit permettre de scruter sans peine l'utilisation des UP dans les traductions, pour des études lexicométriques et contribuer aux recherche scientifiques en matière de la **phraséotraduction automatique** et du TAL (traitement automatique de la langue).

La collecte des textes et le traitement numérique sont actuellement en cours. Le choix du corpus doit refléter la diversité linguistique de la langue et donc être suffisamment large pour garantir une représentation adéquate des UP et améliorer la précision et la fiabilité du DiCoP.

- **représenter une variété de genres (littérature, poésie, discours, roman, science fiction, etc.)** pour trouver un équilibre entre le langage formel et informel (textes officiels, académiques, ainsi que dialogues, conversations quotidiennes, etc.).
- ressources modernes (20e siècle et suivant)
- références déjà accessibles sous forme numérique (ou déjà ocerisées).

Journaux : le « défigement» (活用 huóyòng en chinois), phénomène que l'on rencontre par exemple à travers les jeux de mots contenus **dans les titres, les slogans et les publicités .**

人民日报 PEOPLE'S DAILY

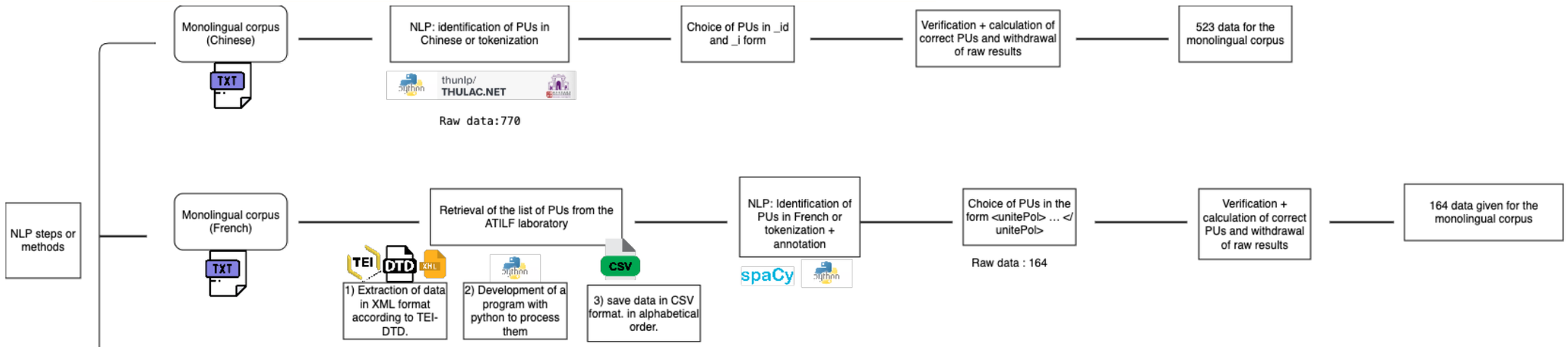
新华社 XINHUA NEWS AGENCY

微博 weibo.com

Le Canard enchaîné — Journal satirique paraissant le mercredi

Corpus en cours de développement + TAL

Méthode de TAL suivie dans le projet DiCoP et la base de données DiCoP-Text



python

spaCy

TXM

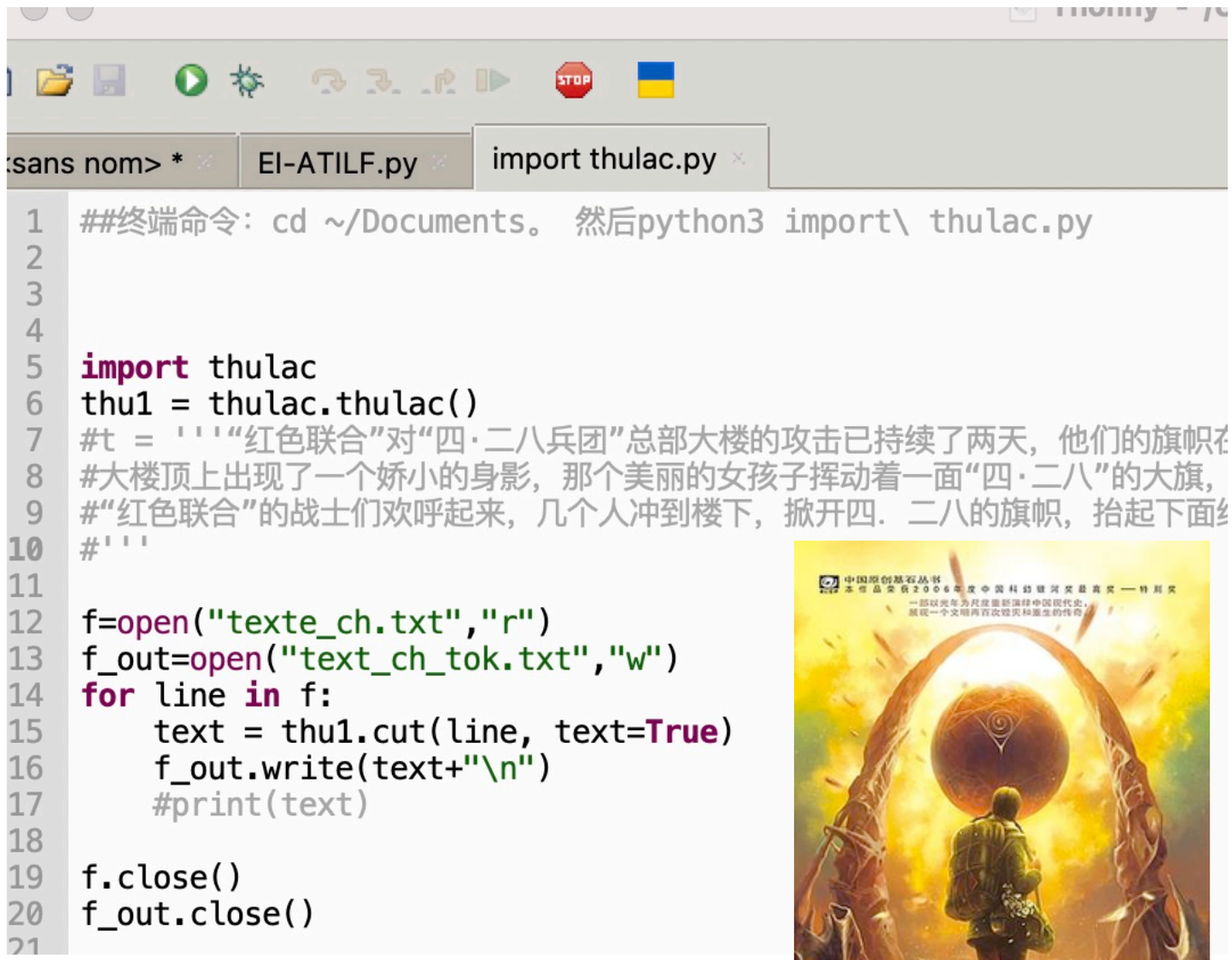
thunlp/ THULAC.NET

LEXICO3

汉语助研 Chinese Research Helper

AntConc

TAL : Identifier les UP en chinois ou token



python™

thunlp/
THULAC.NET

	UP chinoises identifiées par Thulac			
Total	771			
Taux de réussite	65.43% (504/771)			
Taux d'erreur	34.57% (267/771)			
	Sous forme **_id** (567 UP)		Sous forme **_i** (204 UP)	
	Correct	Erreur	Correct	Erreur
Nombre d’EI	321	246	183	21
Pourcentage	56.61 %	43.39%	89.71%	10.29%

TAL : Identifier les UP en français ou token

ce ne fut qu'après de longues minutes qu'elle retacha ses bras encore tend
lentement vers la tribune, s'assit à côté du cadavre et serra la main déjà
les yeux vides se perdaient dans le lointain.
Lorsque quelqu'un vint pour emporter le corps, Ye Wenjie sortit de la poche
qu'elle glissa dans la main de son père : sa pipe.
Quand Wenjie quitta silencieusement les lieux, il n'y avait déjà plus perso
terrain de sport.
Elle prit le chemin qui la ramenait chez elle.
Lorsqu'elle arriva devant le dortoir des enseignants, elle entendit au prem
provenir de leur maison : ce rire émanait de la bouche de la femme qu'elle
mère.
Wenjie <unitePol>tourna les talons</unitePol> sans bruit, laissant ses pied
semblait.
Elle s'aperçut que ceux-ci l'avaient transportée jusque devant la porte de
principal lorsqu'elle était en quatrième année de licence, et aussi son ami
À compter des deux années de maîtrise d'astrophysique, puis ensuite, lorsqu
avec le lancement de la révolution et jusqu'à aujourd'hui, le Pr Ruan avait
personne dont elle s'était sentie le plus proche.
Ruan Wen était jadis partie étudier à Cambridge.

Tourner les talons

identifié 164 EI

par la balise « <unitePol> </unitePol> ».

```python
import spacy
from spacy.matcher import PhraseMatcher

nlp = spacy.load("fr_core_news_sm")
matcher = PhraseMatcher(nlp.vocab, attr="LEMMA")
terms = ["écorcher le renard", "retourner à ses moutons", "tourner les truies au foin"]
# Only run nlp.make_doc to speed things up
patterns = [nlp(text) for text in terms]
matcher.add("TerminologyList", patterns)

f = open("rabelais.txt","r")
for ligne in f:
    doc = nlp(ligne)
    matches = matcher(doc)
    for match_id, start, end in matches:
        span = doc[start:end]
        print(span.text)
f.close()
```

python™

spaCy

Corpus parallèles : Alignement des textes à l'aide de BERTAlign

bfsujason/**bertalign**

Multilingual sentence alignment using sentence embeddings

```
[7]  src = """中国，1967年。
     “红色联合”对“四·二八兵团”总部大楼的攻击已持续了两天，他们的旗帜在大楼周围躁动地飘扬着，仿佛渴望干柴的火种。“红色联合”的
     大楼顶上出现了一个娇小的身影，那个美丽的女孩子挥动着一面“四·二八”的大旗，她的出现立刻招来了一阵杂乱的枪声，射击的武器五花
     “红色联合”的战士们欢呼起来，几个人冲到楼下，掀开四．二八的旗帜，抬起下面纤小的遗体，作为一个战利品炫耀地举了一段，然后将

     tgt = """
     Chine, 1967.


     L'assaut de l'Union rouge contre le quartier général de la brigade du 28 Avril durait depuis déjà de
     Tout autour de l'édifice, les drapeaux de la brigade claquaient au vent, telles des torches attendan
     L'officier en chef de l'Union rouge était terriblement inquiet, non qu'il ait peur des miliciens en
     Ce qu'il craignait, c'étaient les quelques dizaines de fours de métal en fusion entreposés à l'intér
     Ceux-ci, reliés par des détonateurs électriques, étaient pleins à ras bord d'explosifs.
     Il ne pouvait les voir, mais il pouvait sentir leur présence magnétique.
     Un doigt sur un bouton et tous seraient réduits en cendres.
     Les jeunes gardes rouges de la brigade du 28 Avril étaient capables d'une telle folie.
     Par comparaison avec les vétérans qui avaient mûri à mesure qu'ils affrontaient les tempêtes, cette
     Plus fous que la folie elle-même.
     Une silhouette gracile émergea au sommet de la tour, celle d'une belle jeune fille brandissant un gr
     Son irruption fut accueillie par une volée de coups de feu.
     C'était un cocktail hétérogène d'armes : des vieilles carabines à l'américaine, en passant par des Z
     On trouvait même des armes blanches comme des coutelas et des lances, qui formaient avec les armes à
     Les miliciens de la brigade du 28 Avril connaissaient la chanson par cœur.
```

```
aligner.print_sents()
```

terminées

中国，1967年。
Chine, 1967.

“红色联合”对“四·二八兵团”总部大楼的攻击已持续了两天，他们的旗帜在大楼周围躁动地飘扬着，仿佛渴望干柴的火种。
L’assaut de l’Union rouge contre le quartier général de la brigade du 28 Avril durait depuis déjà deux jour

“红色联合”的指挥官心急如焚，他并不惧怕大楼的守卫者，那二百多名“四·二八”战士，与诞生于l966年初、经历过大检阅和大串联的“红色联合
L’officier en chef de l’Union rouge était terriblement inquiet, non qu’il ait peur des miliciens en faction

他怕的是大楼中那十几个大铁炉子，里面塞满了烈性炸药，用电雷管串联起来，他看不到它们，但能感觉到它们磁石般的存在，开关一合，玉石俱
Ce qu’il craignait, c’étaient les quelques dizaines de fours de métal en fusion entreposés à l’intérieur du

Les jeunes gardes rouges de la brigade du 28 Avril étaient capables d’une telle folie.

比起已经在风雨中成熟了许多的第一代红卫兵，新生的造反派们像火炭上的狼群，除了疯狂还是疯狂。
Par comparaison avec les vétérans qui avaient mûri à mesure qu’ils affrontaient les tempêtes, cette générat

大楼顶上出现了一个娇小的身影，那个美丽的女孩子挥动着一面“四·二八”的大旗，她的出现立刻招来了一阵杂乱的枪声，射击的武器五花八门，有
Une silhouette gracile émergea au sommet de la tour, celle d’une belle jeune fille brandissant un grand éte

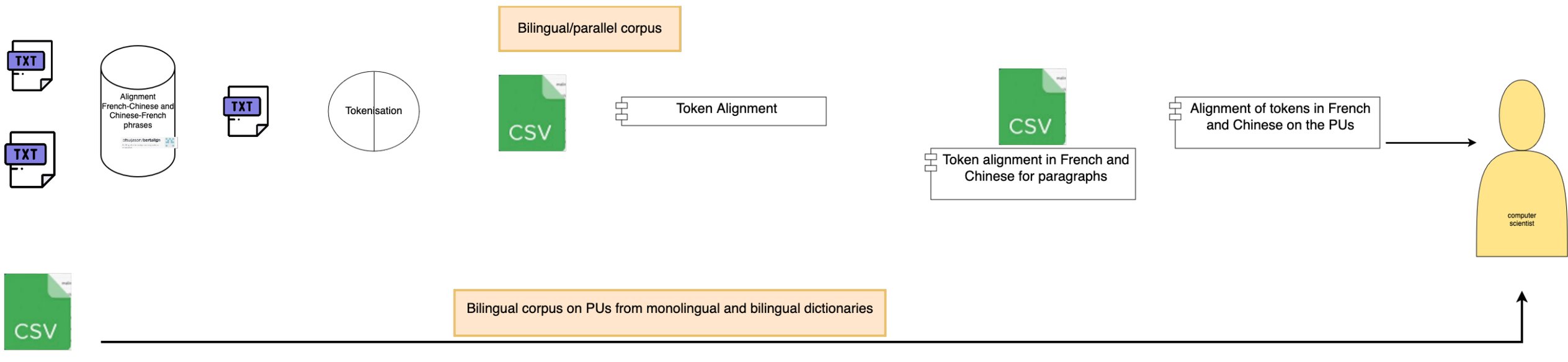
——连同那些梭标和大刀等冷兵器，构成了一部浓缩的近现代史……“四·二八”的人在前面多次玩过这个游戏，在楼顶上站出来的人，除了挥舞旗帜外
On trouvait même des armes blanches comme des coutelas et des lances, qui formaient avec les armes à feu un

La fille qui venait d’apparaître sur le toit croyait sans doute qu’elle connaîtrait aujourd’hui le même des

这次出来的女孩儿显然也相信自己还有那样的幸运她挥舞着战旗，挥动着自己燃烧的青春，敌人将在这火焰中化为灰烬，理想世界明天就会在她那

Tableau : Phrases comportant les 549 UP, et leur traduction

mot_chinois	sent_chinois	sent_fr	mot_fr
心急如焚	“红色联合”的指挥官心急如焚.	L’officier en chef de l’Union rouge était terriblement inquiet.	était terriblement inquiet
玉石俱焚	开关一合，玉石俱焚.	Un doigt sur un bouton et tous seraient réduits en cendres.	tous seraient réduits en cendres
五花八门	射击的武器五花八门.	C’était un cocktail hétérogène d’armes .	un cocktail hétérogène
愈演愈烈	射击的武器五花八门：有陈旧的美式卡宾枪、捷克式	C’était un cocktail hétérogène d’armes : des vieilles carabines à l’américaine, en passant par des	qui avait déclenché une explosion des conflits
全身而退	每次他们都能在弹雨中全身而退，为自己挣到了崇高	Chaque fois, ceux qui parvenaient à sortir indemnes de la pluie de projectiles étaient couverts de	ceux qui parvenaient à sortir indemnes de la pluie de projecti
化为灰烬	敌人将在这火焰中化为灰烬，理想世界明天就会在她	L’ennemi serait bientôt consumé par les flammes et demain, un monde idéal jaillirait de son sang	consumé par les flammes
错综复杂	任何一处都有错综复杂的对立派别在格斗。	En ces temps où coexistaient de multiples factions, des conflits complexes éclataient partout entre différentes obédiences.	complexes
牛鬼蛇神	与其他的牛鬼蛇神相比，反动学术权威有他们的特点	Les membres de l’Autorité académique réactionnaire possédaient des particularités qui les distinguaient des autres adversaires malfaisants de la révolution.	des autres adversaires malfaisants
肃然起敬	都自己结束了他们那曾经让人肃然起敬的生命。	Hai Mo et tant d’autres décidèrent de mettre eux-mêmes un terme à des vies autrefois très estimées.	très estimées
旷日持久	旷日持久的批判将鲜明的政治图像如水银般注入了他们的意识	après une séance de critique ayant duré jusqu’au soir, la glorieuse idéologie politique de la révolution se déversait comme du mercure dans leur conscience, ébranlant les fondations d’un édifice intellectuel jadis façonné par le savoir et la raison.	e ayant duré jusqu’au soir
痛哭流涕	并为此痛哭流涕，	et ils pleuraient de douleur.	et ils pleuraient de douleur
牛鬼蛇神	他们的忏悔往往比那此非知识分子的牛鬼蛇神要深刻得多	Les confessions des intellectuels étaient toujours beaucoup plus profondes et sincères que celles des autres ennemis de la révolution.	des autres ennemis
牛鬼蛇神	只有处于第一阶段的牛鬼蛇神才能对他们那早已过度兴奋的神经产生有效的刺激	Seuls les dangereux ennemis qui restaient bloqués à la première étape étaient capables de susciter une ferveur dans leurs esprits depuis longtemps fanatisés.	les dangereux ennemis
死路一条	顽固下去是死路一条！	, vous devez bien vous rendre compte que la force unie contre laquelle vous résistez est si puissante que si vous continuez à vous obstiner, elle vous conduira à la mort !	elle vous conduira à la mort
有奶便是	，他有奶便是娘	Et, comme les mauvais esprits se rencontrent	Et, comme les mauvais esprits se rencontrent
有备而来	并且显然有备而来	Elle avait manifestement préparé sa tirade	manifestement préparé sa tirade
彻头彻尾	更是彻头彻尾的反动唯心主义……”	Elle prétend que l'univers a des limites, ce qui n'est autre chose que de l'idéalisme réactionnaire…	autre chose que
滔滔不绝	听着妻子滔滔不绝的演讲	En entendant son épouse déverser son éloquent discours,	déverser son éloquent discours
两手空空	你两手空空地上来，	Tu es venue les mains vides	les mains vides
迫不及待	绍琳迫不及待地要继续下去了	Shao Lin n'attendait que cette occasion	n'attendait que
摇摇欲坠	以维持自己那摇摇欲坠的精神免于彻底垮掉	elle devait continuer à parler sans relâche pour préserver son esprit d'un effritement total.	d'un effritement total
不知所措	他这突然的反击令批判者们一时不知所措。	Cette soudaine contre-attaque plongea un moment ses accusateurs dans l'embarras.	dans l'embarras
实践出真知	这等于说正确的哲学是从天上掉下来的，这反对实践出真知	C'est donc remettre en cause l'idée que les connaissances véritables proviennent de la pratique,	les connaissances véritables proviennent de la pratique,
无言以对	绍琳和两名大学红卫兵无言以对	Ni Shao Lin ni ses deux étudiants gardes rouges n'étaient en mesure de justifier convenablement ce paradoxe.	n'étaient en mesure de justifier convenablement ce paradoxe
无坚不摧	但来自附中的四位小将自有她们“无坚不摧”的革命方式	Mais ce n'était pas le cas des quatre jeunes filles, persuadées que la révolution ne devait souffrir aucune attaque.	ne devait souffrir aucune attaque



Conclusion et perspectives