Per-fide

Workshop:

como pesquisar em corpora

Per-fide

O que é um *corpus*?

Um *corpus* (*corpora* no plural) é uma **coleção de textos** em **formato eletrónico**:

- compilada segundo **objetivos específicos**,
- considerada representativa de uma **língua ou parte dela**;
- destinada à **pesquisa** (Sinclair, 2004).

Corpora orais

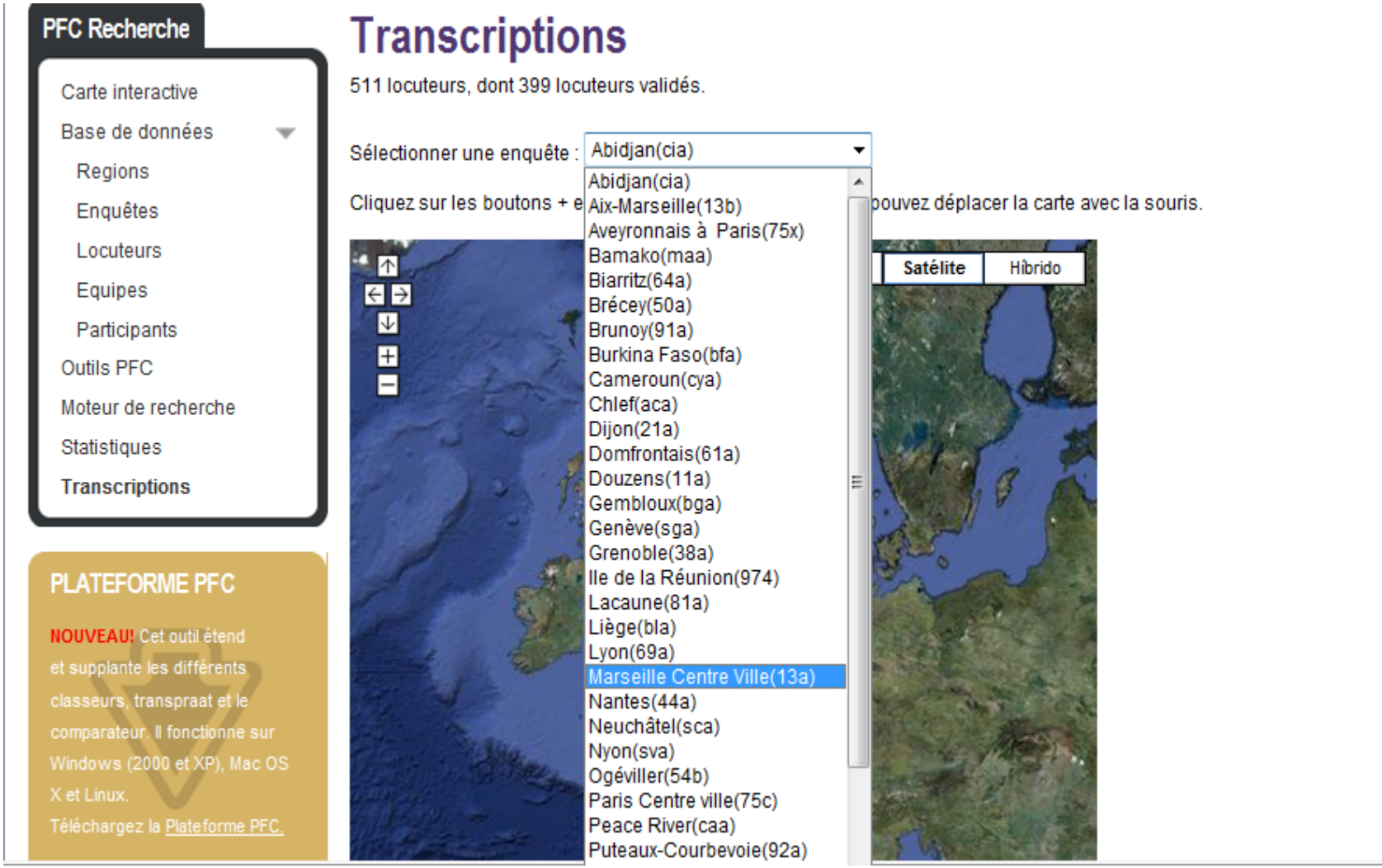
- Apresentam-se tipicamente num **formato eletrónico**. Estes textos não são para serem lidos de forma sequencial, como um livro, mas para serem **interrogados**.

- Alguns corpora estão disponíveis num **formato áudio** (com transcrição) e/ou **vídeo**.

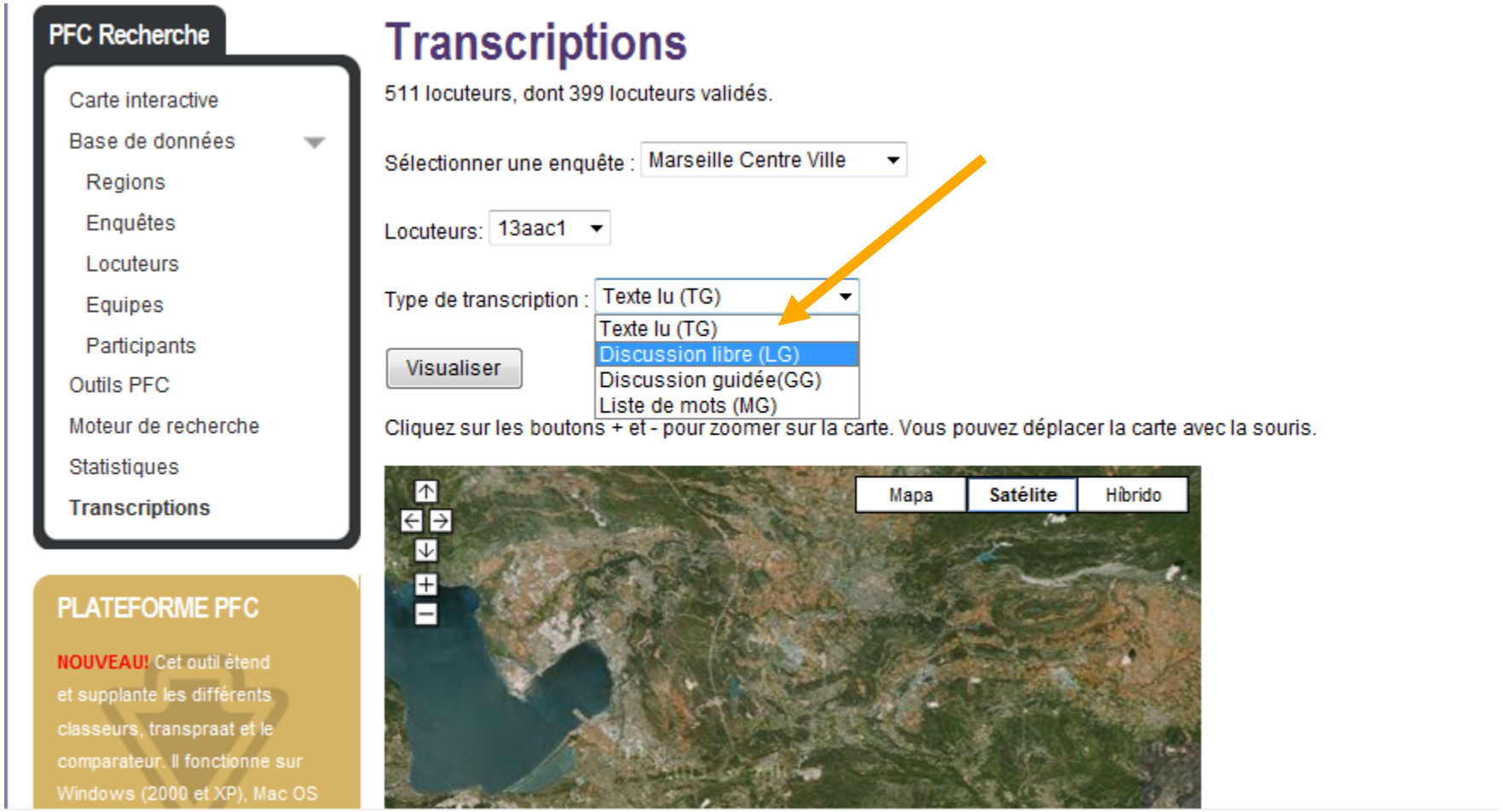
Centres de Ressource pour la Description de l'Oral (CRDO – http://crdo.risc.cnrs.fr/exist/crdo/crn.htm)
Corpus de Français Parlé Parisien (CFPP2000 – http://ed268.univ-paris3.fr/syled/ressources/Corpus-Parole-Paris-PIII/Corpus.html),
Phonologie du Français Contemporain (PFC – http://www.projet-pfc.net/),
Corpus de Langue Parlée en Interaction (CLAPI – http://clapi.univ-lyon2.fr/)
Projet nancéien TCOF (l' ATILF) (http://www.cnrtl.fr/corpus/tcof/).

Phonologie du Français Contemporain (PFC – http://www.projet-pfc.net/)

Phonologie du Français Contemporain



Phonologie du Français Contemporain



Phonologie du Français Contemporain

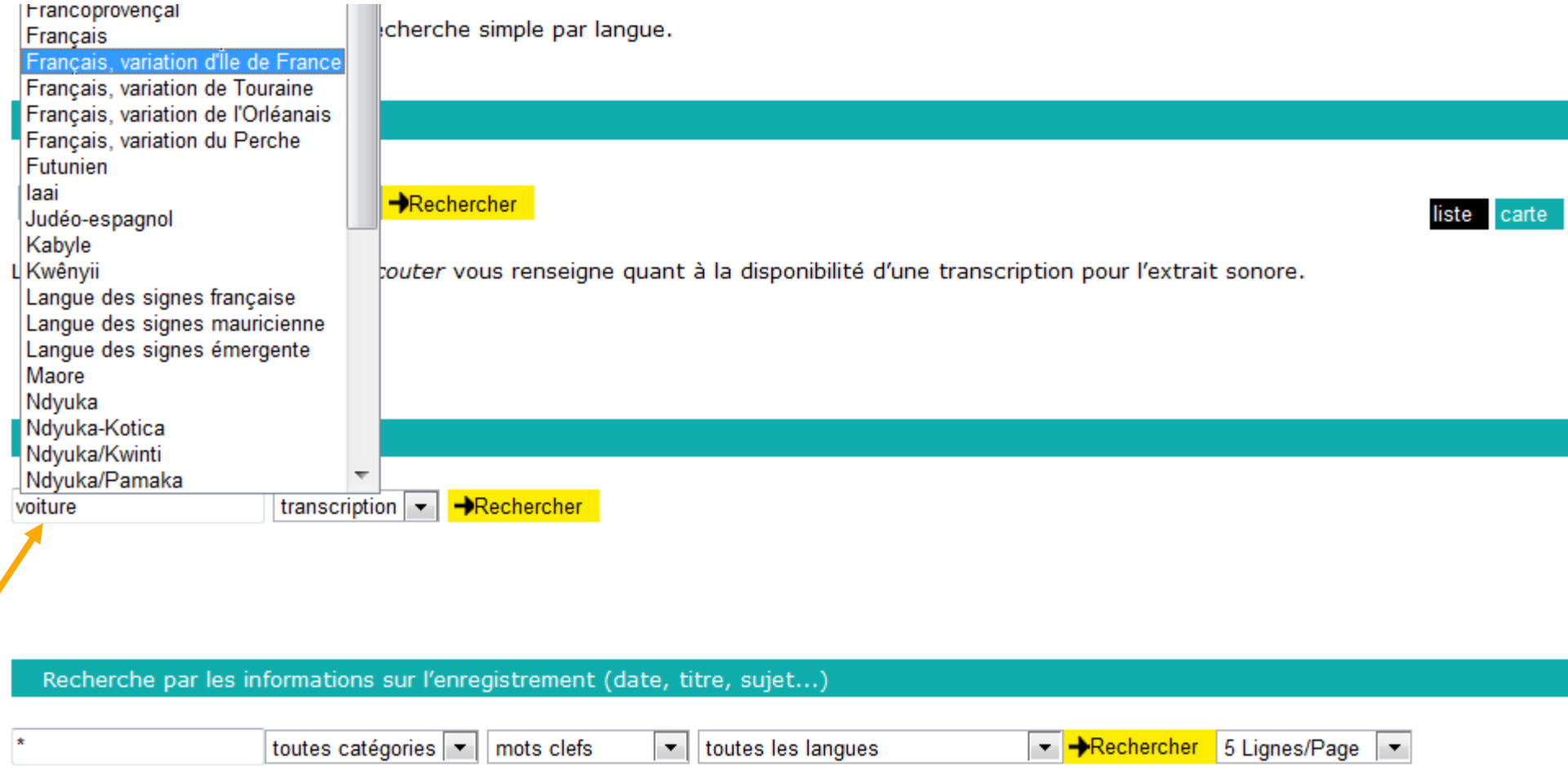
Le lecteur de texte synchronisé provient du logiciel libre Cantare. Cantare est un logiciel produit par le Centre collégial de développement de matériel didactique avec le soutien financier du ministère de l'Éducation, des Loisirs et du Sport du Québec.

Transcription à copier :

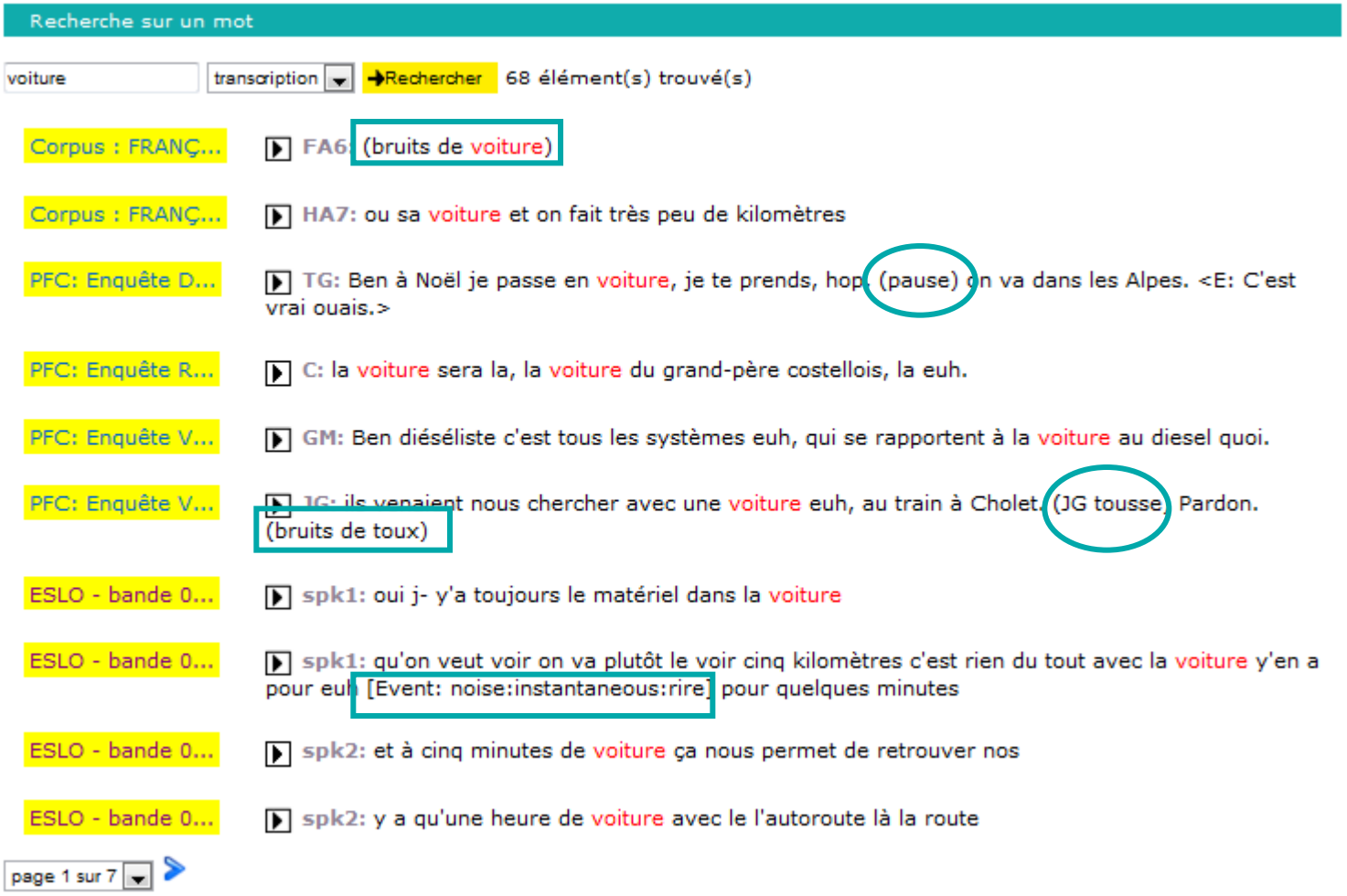
E1: Bon. <E2: Vous pouvez par/> Vous êtes venue ce matin parce que v/ vous allez où après-midi en balade ?
AC1: Euh cet après-midi euh, Pierre euh, il lui reste quelques esches,
AC1: et il veut aller vers la presqu'île de Cassis.
E1: Ah! bon. <AC1: Pour pêcher.> Il avait pêché l'autre jour ? <AC1: Non.> quand il m'a dit 'je, je vais,'
AC1: Non, il pêche pas <E1: il fait beau. Ah. (XX).> Enfin, il m'a, il m'a rien ramené en tout cas. <E2: (XX)>
AC1: Voilà, alors euh la soupe euh chaque fois qu'il pêche, je suis allée avec lui euh,
AC1: euh un peu là. Alors il pêche euh quatre poissons que je mets au congélateur,
AC1: et la soupe, c'est la soupe annuelle (rires).
E1: Oui, ben euh, oh A/ Anne fait pareil quand Denis pe/, va pêcher. <AC1: Ah oui, oui.>, oui, oh ben, il en pêche pas beaucoup.
E1: ou il les donne ou alors i/ il les met au congélateur. <AC1: Il aime pêcher euh, Denis ?> Denis mais il y va avec le bateau lui, tu comprends, voilà. <AC1: Ah oui, oui.>
E1: Maintenant il sors plus ton frère, en bateau ? <AC1: Non, non> Non hein, il sort plus.

http://corpusdelaparole.in2p3.fr/

CORPUS DE LA PAROLE: INTERFACE DE PESQUISA



CORPUS DE LA PAROLE: ACESSO ÀS GRAVAÇÕES E TRANSCRIÇÕES



CORPUS DE LA PAROLE: METADADOS

Informations	+
Corpus : FRANÇAIS DES ANNÉES 80, Enregistrement : FA 6	
Editeur(s)	Groupe ICOR/ Plateforme CLAPI
Langue	Français (fra)
Enregistré en	1984-06-01
Participant(s)	Groupe de recherche du CREDIF (researcher) Mochet, Marie-anne (depositor) Wittig, Gilberte (recorder) CLAPI - Equipe Médiathèque (data_inputter) Locuteur 1 (speaker) Locuteur 2 (speaker) Lesguillons, Mélanie (transcriber)
Description(s)	Interactions construites Interaction en face à face sur un thème imposé Support non anonymisé, de bonne qualité Accès dans CLAPI aux 75 descripteurs documentant le corpus En orthographe adaptee, Toilettée, Minutée La convention de transcription est accessible dans CLAPI
Lieu	France, Redon
Durée	0:09:27

TIPLOGIA DOS CORPORA

CORPUS MONOLINGUES	CORPUS BI- (ou MULTILINGUES)	
	COMPARÁVEIS	PARALELOS
		MONO OU BIDIRECCIONAIS
Uma obra ou várias na **mesma língua**	**originais** em duas ou mais línguas dentro de um mesmo domínio, tipologia textual semelhante, datas próximas etc.	**originais** e **respetivas traduções** em uma ou mais línguas
CETEMPúblico: **190 milhões de palavras extraídas do diário PÚBLICO**	**Scientext** Textos científicos/ académicos originalmente escritos em francês e em inglês. **Projeto COMET: CorTec** Textos técnicos e/ou científicos originalmente escritos em português	**Compara** **Opus** **Per-Fide**

Per-fide

- **Estado da língua:**

sincrónico
ou
diacrónico (Corpus do Português)

- **Objetivo:**

Corpus de referência

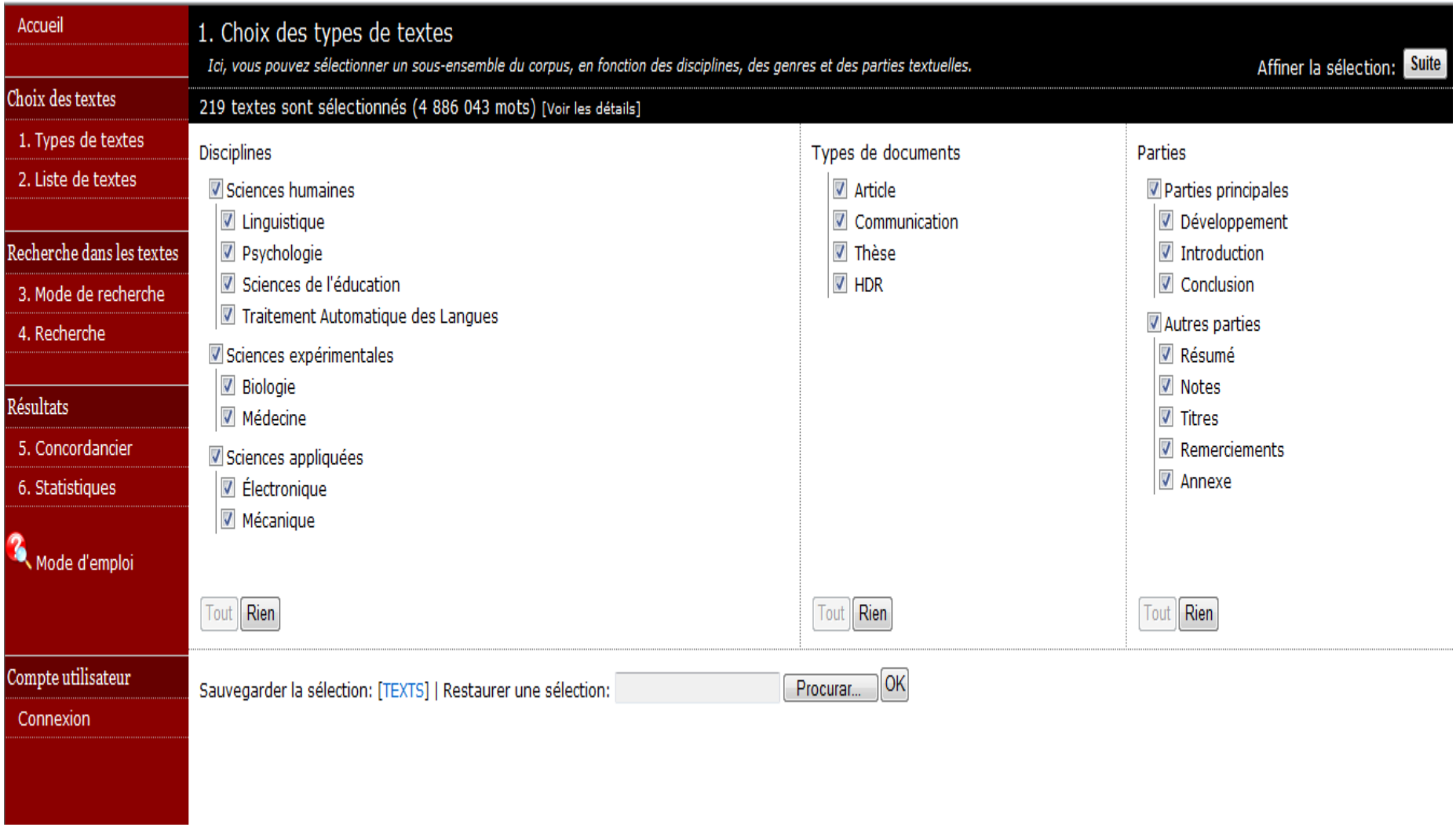
Corpus de Referência do Português Contemporâneo CRPC;
Corpus de Referência del Español Actual CREA)

Para fins específicos

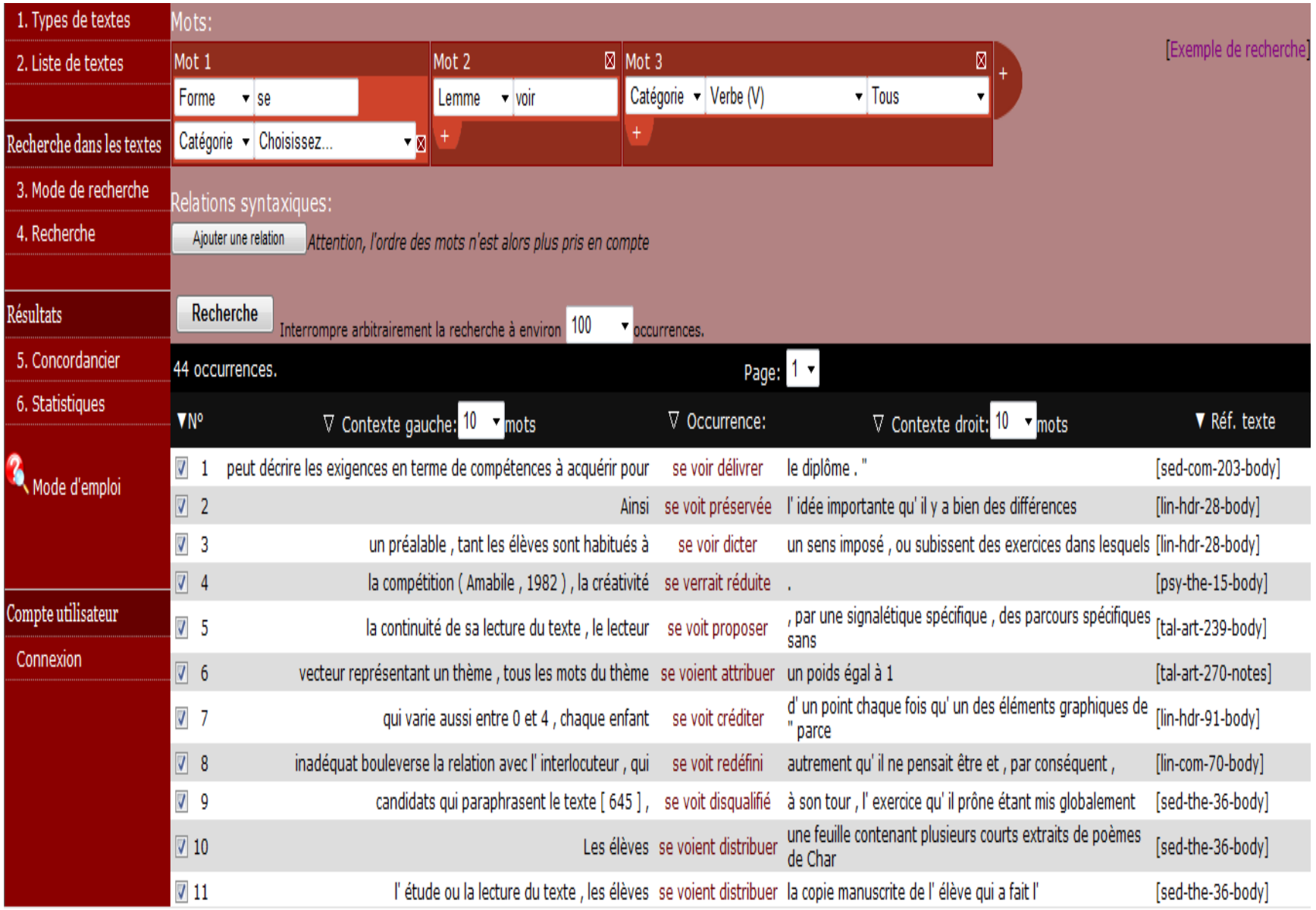
Scientext

http://scientext.msh-alpes.fr/scientext-site/spip.php?article9

SCIENTEXT FR-EN

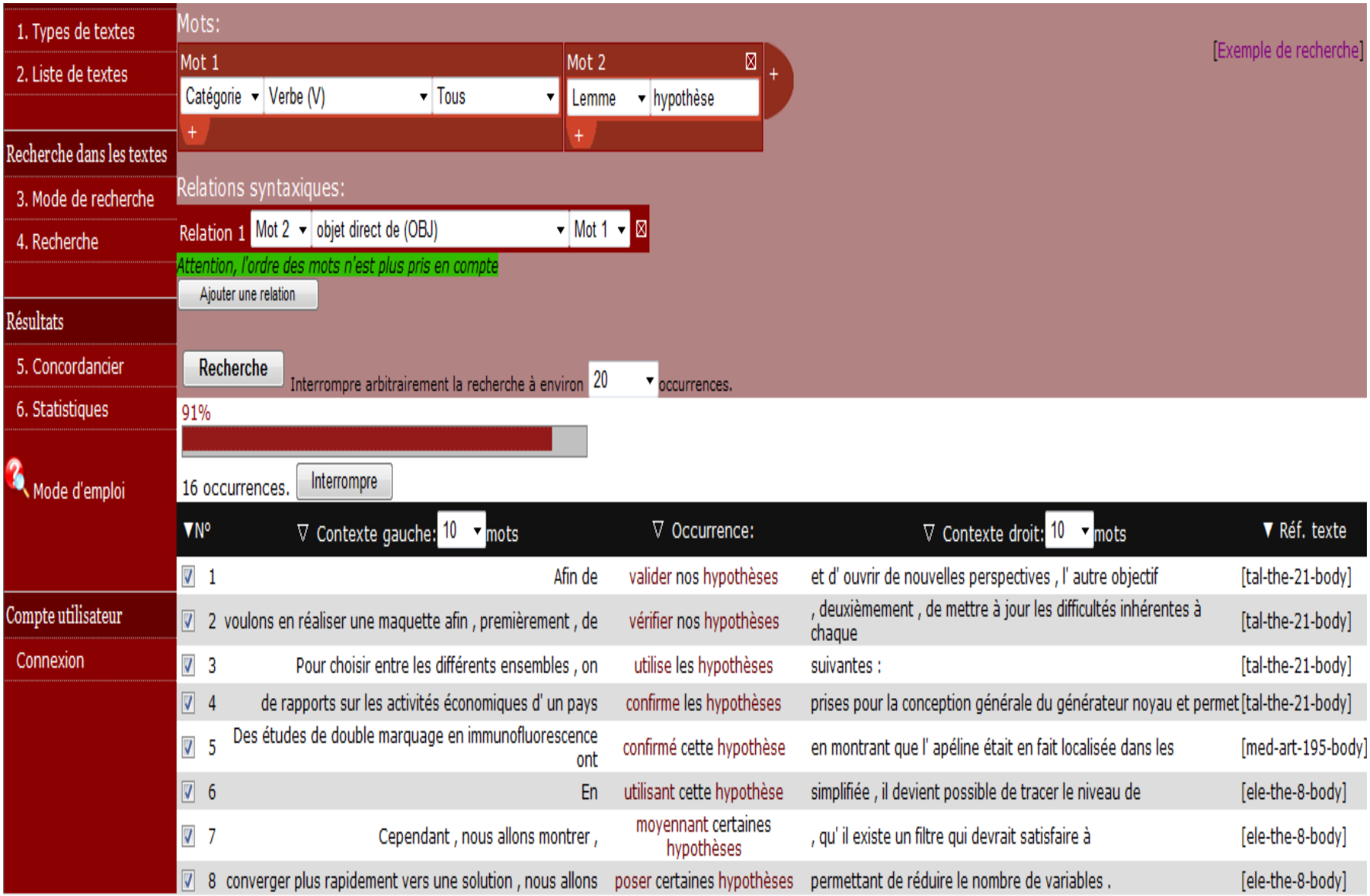


SCIENTEXT: CONCORDÂNCIA DE *SE VOIR V*



N°	Contexte gauche: 10 mots	Occurrence:	Contexte droit: 10 mots	Réf. texte
1	peut décrire les exigences en terme de compétences à acquérir pour	se voir délivrer	le diplôme . "	[sed-com-203-body]
2	Ainsi	se voit préservée	l' idée importante qu' il y a bien des différences	[lin-hdr-28-body]
3	un préalable , tant les élèves sont habitués à	se voir dicter	un sens imposé , ou subissent des exercices dans lesquels	[lin-hdr-28-body]
4	la compétition (Amabile , 1982) , la créativité	se verrait réduite	.	[psy-the-15-body]
5	la continuité de sa lecture du texte , le lecteur	se voit proposer	, par une signalétique spécifique , des parcours spécifiques sans	[tal-art-239-body]
6	vecteur représentant un thème , tous les mots du thème	se voient attribuer	un poids égal à 1	[tal-art-270-notes]
7	qui varie aussi entre 0 et 4 , chaque enfant	se voit créditer	d' un point chaque fois qu' un des éléments graphiques de " parce	[lin-hdr-91-body]
8	inadéquat bouleverse la relation avec l' interlocuteur , qui	se voit redéfini	autrement qu' il ne pensait être et , par conséquent ,	[lin-com-70-body]
9	candidats qui paraphrasent le texte [645] ,	se voit disqualifié	à son tour , l' exercice qu' il prône étant mis globalement	[sed-the-36-body]
10	Les élèves	se voient distribuer	une feuille contenant plusieurs courts extraits de poèmes de Char	[sed-the-36-body]
11	l' étude ou la lecture du texte , les élèves	se voient distribuer	la copie manuscrite de l' élève qui a fait l'	[sed-the-36-body]

SCIENTEXT: CONCORDÂNCIA DE *V* + *HYPOTHÈSE*



▼N°	▽ Contexte gauche: 10 mots	▽ Occurrence:	▽ Contexte droit: 10 mots	▼ Réf. texte
1	Afin de	valider nos hypothèses	et d' ouvrir de nouvelles perspectives , l' autre objectif	[tal-the-21-body]
2	voulons en réaliser une maquette afin , premièrement , de	vérifier nos hypothèses	, deuxièmement , de mettre à jour les difficultés inhérentes à chaque	[tal-the-21-body]
3	Pour choisir entre les différents ensembles , on	utilise les hypothèses	suivantes :	[tal-the-21-body]
4	de rapports sur les activités économiques d' un pays	confirme les hypothèses	prises pour la conception générale du générateur noyau et permet	[tal-the-21-body]
5	Des études de double marquage en immunofluorescence ont	confirmé cette hypothèse	en montrant que l' apéline était en fait localisée dans les	[med-art-195-body]
6	En	utilisant cette hypothèse	simplifiée , il devient possible de tracer le niveau de	[ele-the-8-body]
7	Cependant , nous allons montrer ,	moyennant certaines hypothèses	, qu' il existe un filtre qui devrait satisfaire à	[ele-the-8-body]
8	converger plus rapidement vers une solution , nous allons	poser certaines hypothèses	permettant de réduire le nombre de variables .	[ele-the-8-body]

Per-fide

- **Objetivo:**

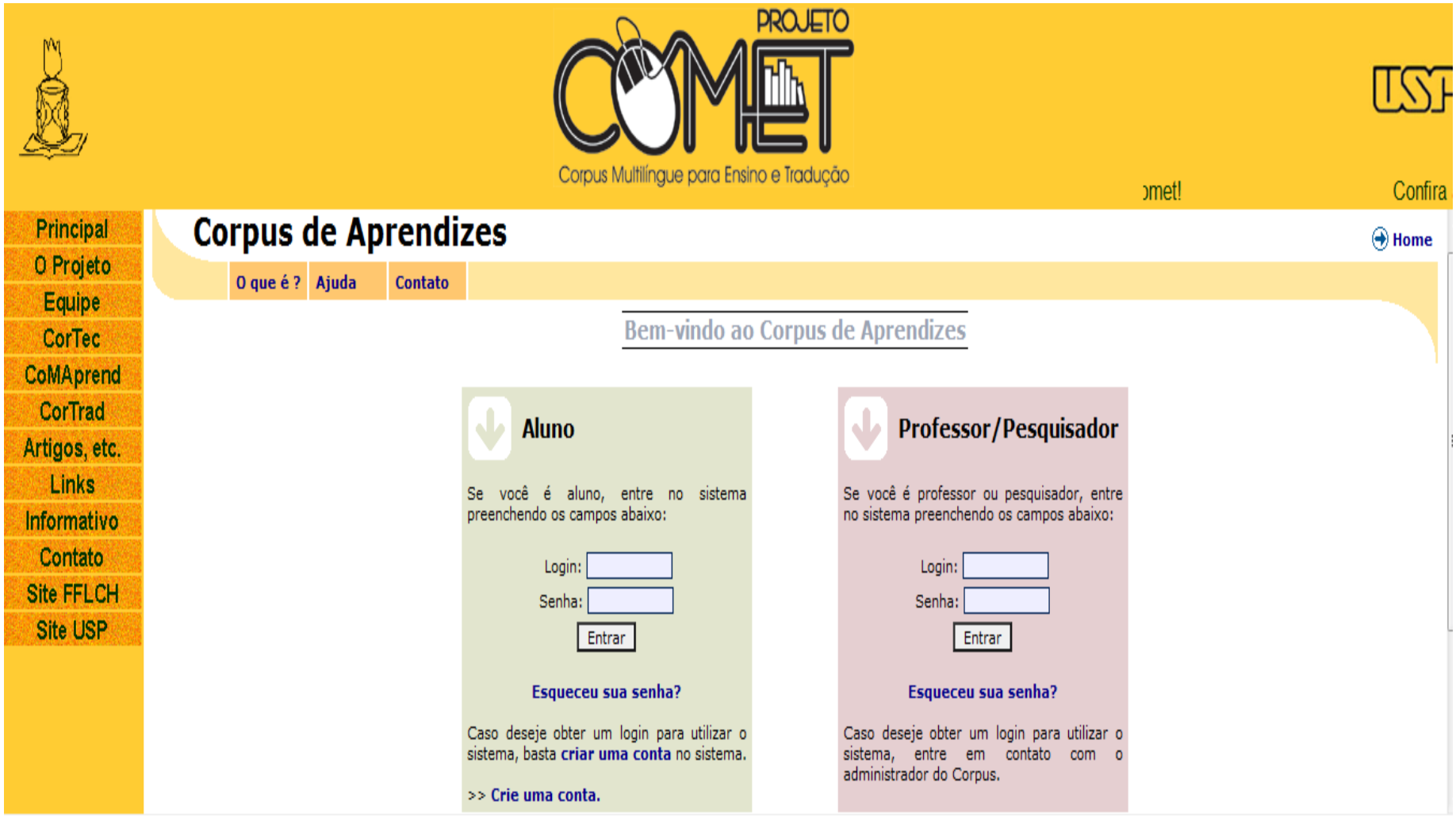
CORPORA & TADUÇÃO:

- ENSINO, TEORIA E PRÁTICA DE TRADUÇÃO:

Recolha de dados de PLE;

Corpus Multilingue de Aprendizes (CoMAprend) (de, es, fr, it, it))

Corpus Multilingue de Aprendizes (CoMAprend)



Per-fide

- **Objetivo:**

Corpora de aprendizagem:

- de línguas:

Projeto COMET: CorTrad;

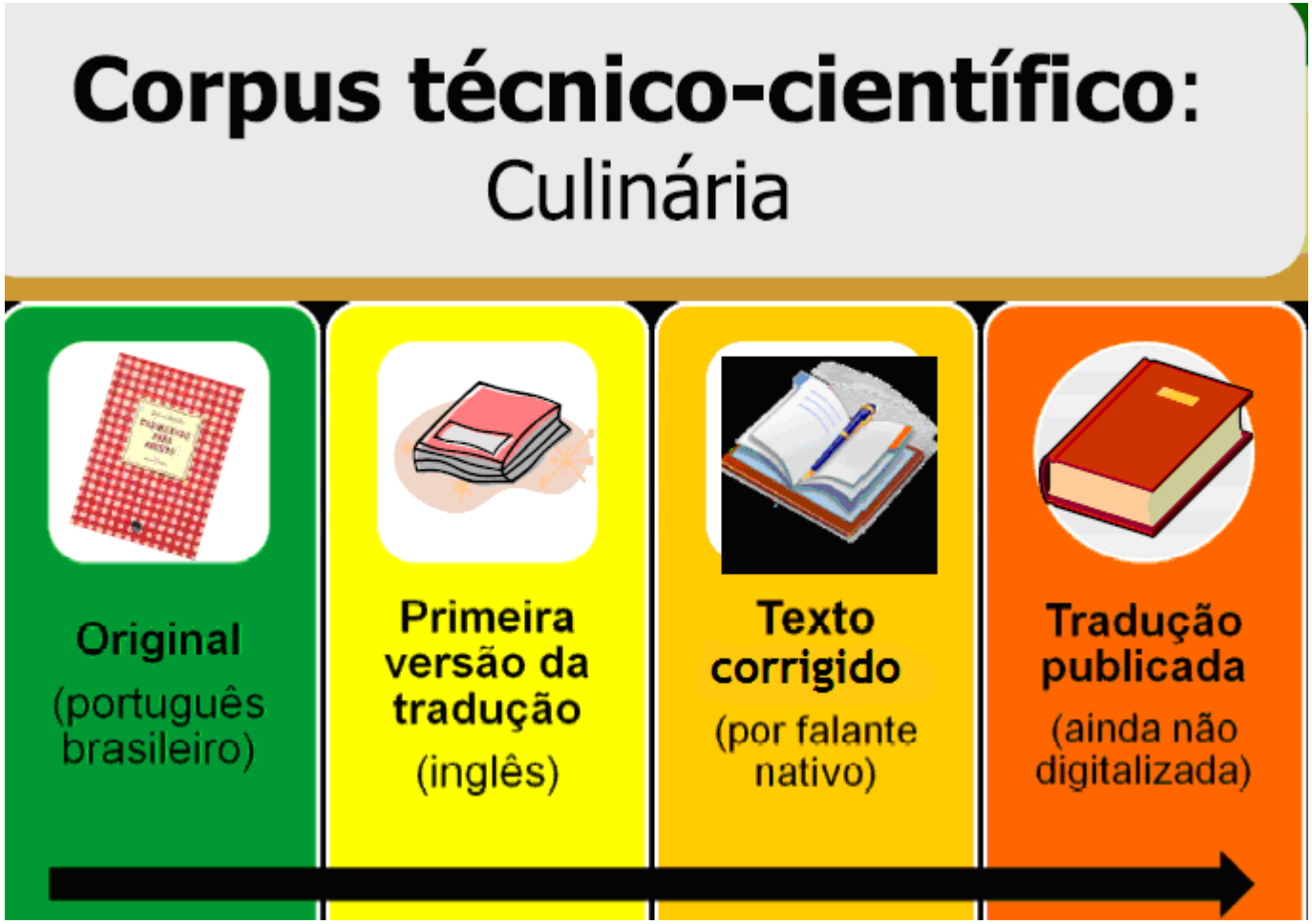
Projeto Mellange

PROJETO COMET
CORPUS MULTILINGUE DE ENSINO E TRADUÇÃO
(CORPUS DE TRADUÇÃO: CorTrad)

PROJETO COMET

Corpus de tradução (multiversão português-inglês): CorTrad

(http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html)



Casos em que ***natural*** em português não é traduzido por ***natural*** em inglês

natural vs !natural

CorTrad técnico-científico culinária

Expressão de busca: **"natural" %c :CORTRAD_CULI_TRAD1 ! "natural" %c :CORTRAD_CULI_TRAD2 ! "natural" %c**
Resultado escolhido: **concordância em contexto**
Corpus pesquisado: **originais** (versão 2.2)

Voltar

Nova pesquisa

15 ocorrências.

Original	Primeira tradução	Tradução revisada
um jantar, com o frescor da noite?	to one's mind but why not dinner, enjoying the freshness of the night?	one's mind but why not dinner, enjoying the freshness of the night?
1/2 xícara de iogurte **natural**	cup whole milk plain yogurt	cup plain whole milk yogurt
3/4 de xícara de iogurte **natural**	cup whole milk plain yogurt	cup plain whole milk yogurt
3/4 de xícara de iogurte **natural**	cup whole milk plain yogurt	cup plain whole milk yogurt
1 1/2 xícara de iogurte **natural**	1 cups whole milk plain yogurt	1 cups plain whole milk yogurt
2 1/3 de xícaras de iogurte **natural**	2 1/3 cups whole milk plain yogurt	2? cups plain whole milk yogurt
1 xícara de iogurte **natural**	1 cup whole milk plain yogurt	1 cup plain whole milk yogurt
600 ml de iogurte **natural** (3 copinhos)	600 ml (21 fl oz) whole milk plain yogurt	600 ml (21 fl oz) plain whole milk yogurt
100 ml de iogurte **natural**	100 ml (3.5 fl oz) whole milk plain yogurt	100 ml (3 fl oz) plain whole milk yogurt
600 ml de iogurte **natural** (3 copinhos)	600 ml (21 fl oz) whole milk plain yogurt	600 ml (21 fl oz) plain whole milk yogurt
200 ml de iogurte **natural** (1 copinho	200 ml (7 fl oz) whole milk plain yogurt	200 ml (7 fl oz) plain whole milk yogurt
200 ml de iogurte **natural** (1 copinho	200 ml (7 fl oz) whole milk plain yogurt	200 ml (7 fl oz) plain whole milk yogurt
1 litro de suco de laranja **natural**	1 L (1.1 qt freshly squeezed orange juice	1 L (1 qt freshly squeezed orange juice
1 litro de suco de laranja **natural**	1 L (1.1 qt freshly squeezed orange juice	1 L (1 qt freshly squeezed orange juice
800 ml de iogurte **natural** (4 copinhos)	800 ml (28 oz) whole milk plain yogurt	800 ml (28 oz) plain whole milk yogurt

MeLLANGE Learner Translator Corpus (LTC)

http://mellange.eila.jussieu.fr/index.fr.shtml

Education and Culture

Leonardo da Vinci

- About MeLLANGE
- MeLLANGE corpus resources
  - LTC
  - eCoLoRe TMX corpus
  - CTS large corpora
  - MeLLANGE query interface
- Additional corpus resources
- MeLLANGE useful links

About MeLLANGE

The MeLLANGE (Multilingual eLearning in Language Engineering) project has brought together academic and industry partners from France, Austria, the Czech Republic, Germany, Italy, Spain, Switzerland, and the UK to create innovative learning materials for trainee and professional translators.

Apart from designing and implementing interactive courses in Corpus Linguistics for Translators and Translation Memory, the project consortium has also contributed trainee and professional translations of four texts belonging to different domains. This collection of translations is the MeLLANGE Learner Translator Corpus (LTC). The LTC includes work done by trainees which was subsequently annotated for errors within the MeLLANGE project and according to a customised error typology. The corpus can be accessed via the MeLLANGE query interface. More detailed information about this corpus (including guidelines on using the query interface), as well as other corpora that have been repurposed during the project can be found via the MeLLANGE corpus resources page.

The consortium has also assembled a list of relevant corpora for the field of translation studies and translator training. This list, together with descriptions of the resources and links to them is available on the Additional corpus resources page.

Finally, there are a number of resources - such as source texts, corpora in raw, as well as processed formats, the MeLLANGE error typology, etc.- which are available via the MeLLANGE useful links page.

Did you know that ...

Students translating into German made the biggest number of mistakes: on average 32 mistakes per translation. Or does that mean that German tutors are the most demanding?

Tipologia de erros

http://corpus.leeds.ac.uk/mellange/images/mellange_error_typology_en.jpg

Error-Annotation Scheme

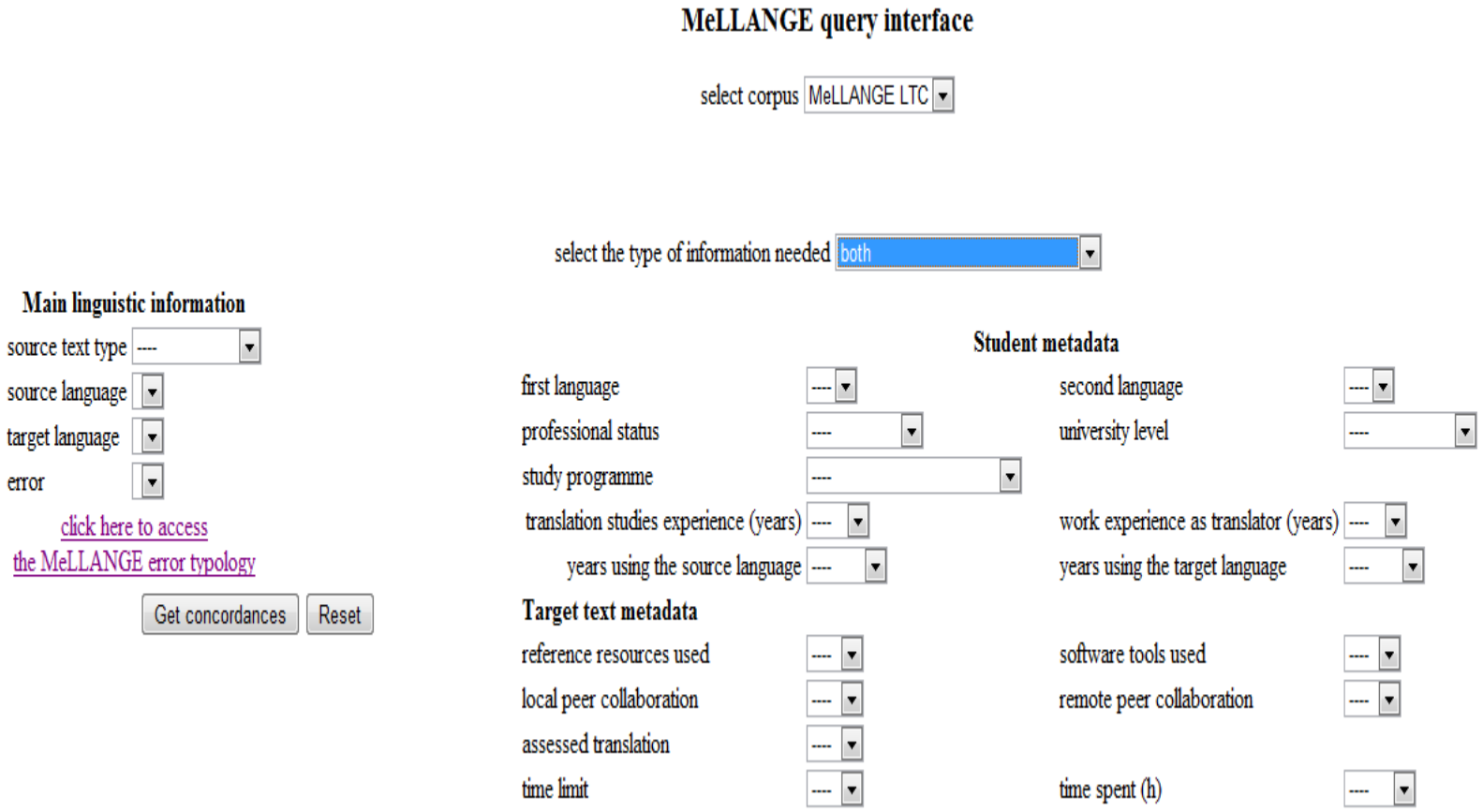
An Error-Annotation Scheme was developed by the consortium for the annotation of the **LTC**. Annotators may chose from two categories of errors: language and content transfer errors. They may also define customized categories if they feel the need to. In the near future, annotators will also be able to associate explanatory notes and comments to the errors they mark up.

Here are a few practical examples of error annotations:

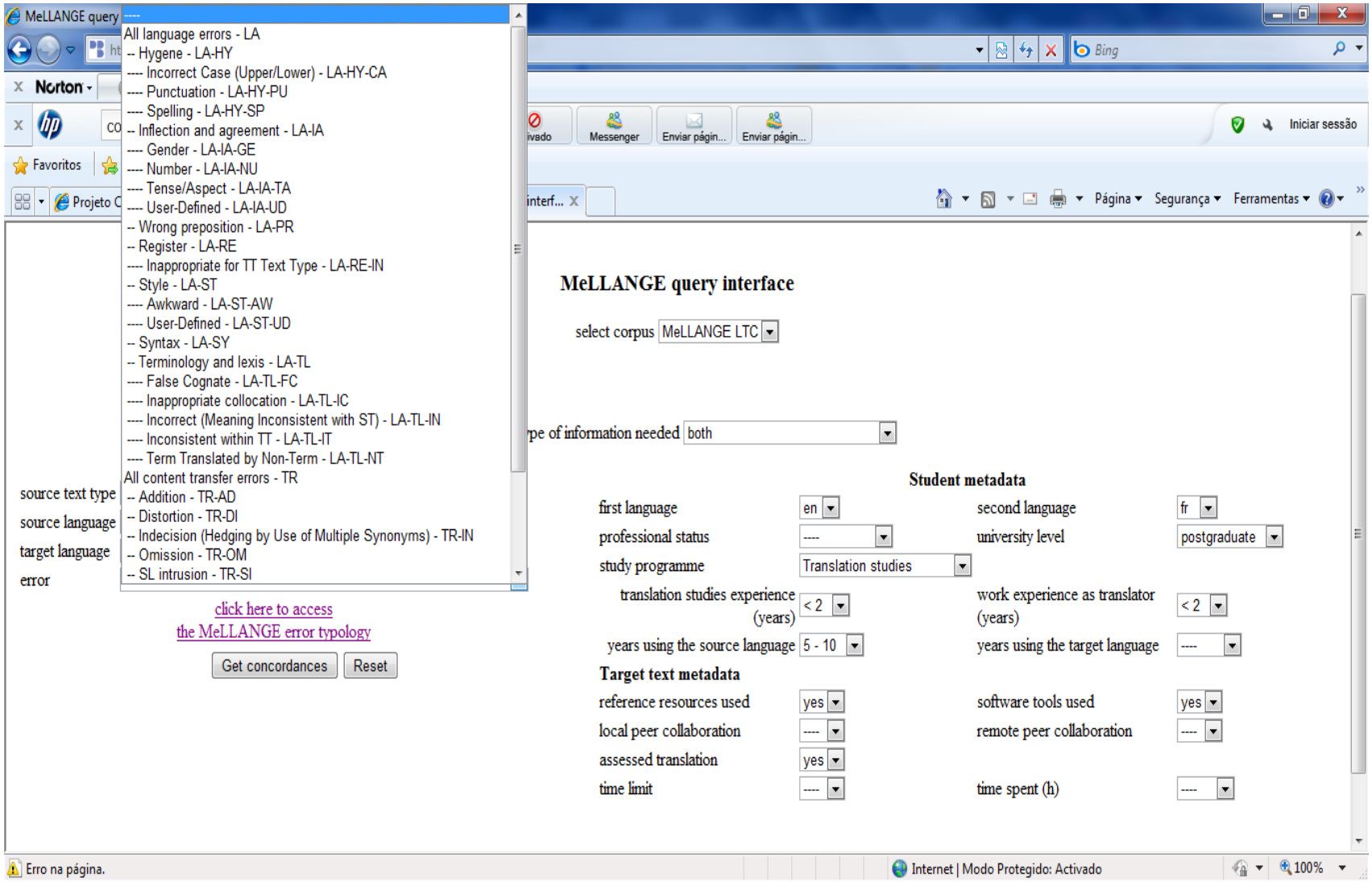
Original	Translation	Error code & name
Les fonctionnaires et agents désireux de passer la visite médicale auprès de l'un des médecins-conseil peuvent s'adresser à l'un des cabinets médicaux.	Civil servants and workers *la-tl-it* wishing to have the check-up with one of the institution's medical officers should contact one of the medical practices.	**la-tl-it** = language - terminology and lexis - inconsistent with TT (target text)
Dans ce cas, les honoraires pour la visite (42€ au maximum) et les factures des examens compris dans le programme de la visite médicale annuelle (annexe 1 et annexe 2) doivent être acquittées directement par la personne examinée et envoyés pour remboursement au service médical dont elle dépend.	In this case, the fees for the visit *la-tl-it* (maximum €42/£29) and the invoices for tests, included in the schedule of the annual medical check-up (annexes *la-hy-ca* 1 and 2), *la-hy-pu* must be settled directly by the employee who *la-hy-pu* can then apply to the responsible medical service for reimbursement.	**la-tl-fc** = language - terminology and lexis - false cognate **la-hy-pu** = language - hygene - punctuation **la-hy-ca** = language - hygene - incorrect case (upper/lower)

Mellange query interface

http://corpus.leeds.ac.uk/mellange/mellange_query_interface.html



Mellange query interface



- Tipologia de erros: http://corpus.leeds.ac.uk/mellange/images/mellange_error_typology_en.jpg

3	language	context	source
Original context	en	Some countries have experienced the most serious economic crisis in history. More than half Argentina's 37m inhabitants now live below the poverty line; the middle classes have been hard hit; and more than a third of the country's active population is either unemployed or under-employed.	Full text
Target sentence	fr	Les classes moyennes ont été [lourdement]$_{LA\text{-}TL\text{-}FC}$ touchées et plus du tiers de la population active est soit au chômage , soit sous-employée .	Full text
Reference context	fr	En Argentine, par exemple, la classe moyenne a été laminée, plus de la moitié des 37 millions d'habitants vivent maintenant dans la pauvreté, et plus du tiers de la population active est sans travail ou sous-employée.	Full text
Alternative student context	fr	Plus de la moitié des 37 millions d'habitants de l'Argentine vit désormais en - dessous du seuil de pauvreté , les classes moyennes ont été durement touchées , et plus du tiers de la population active est soit au chômage , soit exploité .	Full text
Alternative student context	fr	Plus de la moitié des 37 millions d'habitants de l'Argentine vivent désormais sous le seuil de pauvreté , les classes moyennes ont été sérieusement touchées et plus d'un tiers de la population active du pays est au chômage ou sous - employé .	Full text
Alternative student context	fr	Plus de la moitié des 37 millions d'argentins vivent à présent sous le seuil de pauvreté ; les classes moyennes ont été durement touchées ; et plus du tiers de la population active du pays est soit au chômage ou sous - employé .	Full text
Alternative student context	fr	On peut citer l'Argentine où plus de la moitié de ses 37 millions d'habitants vivent actuellement sous le seuil de pauvreté : la classe moyenne a été fortement touchée et plus d'un tiers de la population active est sous - employée ou au chômage .	Full text
Alternative student context	fr	Plus de la moitié des 37 millions d'Argentins vivent désormais sous le seuil de la pauvreté , les classes moyennes ont été durement touchées et plus d'un tiers de la population active du pays est soit au chômage soit sous - employé .	Full text
Alternative student context	fr	Plus de la moitié des habitants de l'Argentine (37 m) vivent maintenant sous la ligne de pauvreté ; les classes moyennes ont été les plus touchées ; et plus d'un tiers de la population active du pays se trouve au chômage ou est sous - employée .	Full text
Alternative student context	fr	La moitié de la population d' Argentine ([[qui dénombre]$_{TR\text{-}AD}$]$_{LA\text{-}ST\text{-}AW}$ 37 millions d' habitants) vit désormais en dessous du seuil de pauvreté ; les classes moyennes ont été fortement [touchées par la crise]$_{TR\text{-}DI}$ et plus d' un tiers de la population active est sans emploi ou sous-employée .	Full text
Alternative student context	fr	Plus de la moitié de la population de l'Argentine , qui compte 37 millions d'habitants , vivent aujourd'hui en dessous du seuil de pauvreté , les classes moyennes ont été les plus touchées , et plus d'un tiers de le population active du pays est soit sans emploi , soit sous - employée .	Full text

Texto original

Viva Brazil! By Ignacio Ramonet BRAZIL'S new president is the head of the Workers' party and former trade union leader, Luiz Inacio "Lula" da Silva. Elected last October, he comes to office at a time when Latin America is in a state of upheaval. For the first time this huge country - which is the world's tenth-ranked industrial nation, with a population of 170m - is about to have a democratic government under a leader with roots in the radical left who rejects liberal globalisation. This is an event of great importance. Although in a very different context, it recalls the Chilean election of the socialist Salvador Allende as president in 1970. So 1 January 2003 marks the beginning of a new historical cycle in Latin America. The preceding cycle began at the end of a dark period of military tyrannies, repression and armed uprisings, and has lasted for two decades, since 1983. Its main characteristics have been the end of the guerrilla movements (apart from those in Colombia, and Subcomandante Marcos's unique and non-violent Zapatista Army in Chiapas); more democratic regimes; and a systematic experimentation with neoliberal economic policies. The application of the free market model has translated into a continuous structural adjustment process, and in all the countries concerned pos: || lemma: quences have been disastrous. The failure has been complete. In 2002 the labour market had the worst conditions in 22 years. Unemployment was rampant and more than half the working population can now find employment only in the informal sector. The number of people living in poverty has been rising continuously. At the same time minimum wages have continued to fall, and growth in the region has declined (-0. 8%). Some countries have experienced the most serious economic crisis in history. More than half Argentina's 37m inhabitants now live below the poverty line; the middle classes have been hard hit; and more than a third of the country's active population is either unemployed or under-employed.

Tradução 1

Vive le Brésil ! Par Ignacio Ramonet LE nouveau président du Brésil , Luiz Inacio " Lula " da Silva , est le chef du parti des travailleurs , et anciennement à la tête d'un syndicat . Elu en octobre dernier , il arrive au gouvernement à un moment où l'Amérique Latine est en plein bouleversement . Pour la première fois , ce pays gigantesque – qui est le dixième pays industriel , avec une population de 170 millions d'habitants – est sur le point d'avoir à sa tête un gouvernement démocratique , avec un dirigeant dont les racines se trouvent dans la gauche radicale , qui rejette la mondialisation . C'est un évènement d'une très grande importance . Dans un contexte certes très différent , cela nous rappelle l'élection du socialiste Salvador Allende à la présidentielle du Chili en 1970 . Le 1er janvier 2003 marque donc le début d'un nouveau chapitre de l'histoire de l'Amérique Latine . Le chapitre précédent avait commencé à la fin de la période sombre de la tyrannie militaire , de la répression , ainsi que des soulèvements armés , et il a duré pendant deux décennies , dès 1983 . Ses caractéristiques principales ont été la fin des guérillas (à part celles en Colombie , celle , unique , du Sous - commandant Marco ainsi que la guérilla non - violente de l'armée zapatiste à Chiapas) ; des régimes plus démocratiques ; et une mise en place systématique des politiques économiques néolibérales . L'application du modèle de marché libre s'est traduit par un processus d'ajustement continuel , et tous les pays concernés ont subi des conséquences sociales désastreuses . pos: || lemma: tal . 2002 a été la pire de ces 22 années pour le marché du travail . Le chômage était très important et par conséquent , plus de la moitié de la population active ne peut désormais s'est accru continuellement . En même temps , le montant des salaires minimum a continué de baisser , et la croissance dans cette région a chuté (- . 8%) . Certains pays ont vécu la crise économique la plus importante de leur histoire . Plus de la moitié des 37 millions d'habitants de l'Argentine vit désormais en - dessous du seuil de pauvreté , les classes moyennes ont été durement touchées , et plus du tiers de la population active est soit au chômage , soit exploité .

Tradução 2

Vive le Brésil Par Ignacio Ramonet Le nouveau président du Brésil est le chef du Parti des travailleurs et un ancien dirigeant syndical , Luiz Inacio " Lula " da Silva . Après avoir été élu en octobre dernier , il prend ses fonctions au moment au l'Amérique latine est en phase de bouleversement . Pour la première fois , cet immense pays – qui se classe au 10ème rang des pays industrialisés , avec une population de 170 millions de personnes – est sur le point d'avoir un gouvernement démocratique conduit par un dirigeant aux origines politiques d'extrême gauche qui rejète la globalisation libérale . Il s'agit d'un évènement de grande importance . Bien que placé dans un contexte très différent , cela rappelle l'élection au Chili du socialiste Salvador Allende en tant que Président en 1970 . Ainsi le 1er janvier 2003 marque le début d'un nouveau cycle historique en Amérique latine . Le cycle précédent commença à la fin de la sombre période des tyrannies militaires , de répression et des soulèvements armés , et a duré deux décennies , depuis 1983 . Ses caractéristiques principales ont été la fin des mouvements de guérilla (à part ceux de Colombie , et celui , unique et non - violent , de l'Armée zapatiste du Sous - commandant Marcos au Chiapas) ; plus de régimes démocratiques ; et l'expérimentation systématique de politiques économiques néolibérales . La mise en application du modèle de marché libre s'est traduit en un processus d'ajustement structurel continu , et dans tous les pays concernés , ses conséquences sociales ont été désastreuses . L'échec a été complet . En 2002 , le marché du travail affichait ses plus mauvaises conditions depuis 22 ans . Le chômage était effréné et plus de la moitié de la population active , à présent , trouve uniquement un emploi dans le secteur informel . Le nombre de personnes vivant dans la pauvreté a continuellement augmenté . Au même moment , les salaires minimum ont continué de baisser , et la croissance dans la région a diminué (, 8%) . Certains pays ont subi la crise économique la plus grave de leur histoire . Plus de la moitié des 37 millions d'argentins vivent à présent sous le seuil de pauvreté ; les classes moyennes ont été durement touchées ; et plus du tiers de la population active du pays est soit au chômage ou sous - employé .

Tradução 3

Vive le Brésil ! Par Ignacio Ramonet . [Le nouveau président du Brésil , Luiz Inacio " Lula " da Silva , est le chef du [[parti travailliste]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IN}$ et c' est aussi un ancien dirigeant syndical . Elu en Octobre dernier , il arrive au pouvoir à une période où l' Amérique latine connaît de grands [troubles]$_{TR\text{-}SI\text{-}TL}$]$_{TR\text{-}DI}$. Pour la première fois , ce pays gigantesque - la dixième nation industrielle du monde , avec une population de 170 millions d' habitants - est sur le point d' être dirigée par un gouvernement démocratique ayant à sa tête un [[ancien militant]$_{TR\text{-}DI}$ d' [[extrême gauche]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IN}$ et anti-mondialiste .]$_{TR\text{-}DI}$ C' est un événement de grande importance . [[Même si]$_{TR\text{-}DI}$]$_{LA\text{-}ST\text{-}AW}$ dans un contexte différent , cela rappelle l' élection du président socialiste Chilien Salvador Allende en 1970 . Le 1er Janvier 2003 marque donc le début d' un nouveau cycle historique en Amérique Latine . Le précédent cycle avait débuté en 1983 , à la fin d' une période sombre marquée par les tyrannies militaires , la répression et les soulèvements armés , et a duré deux décennies . Les principales caractéristiques de [cette période]$_{TR\text{-}AD}$ furent la fin des mouvements de guérilla (excepté le mouvement guérillero de Colombie et le [[mouvement]$_{TR\text{-}AD}$ non-violent]$_{TR\text{-}OM}$ de l' armée zapatiste mené par le sous-commandant [Marco]$_{LA\text{-}HY\text{-}SP}$]$_{LA\text{-}TL\text{-}IG}$ au Chiapas) , des [régimes plus démocratiques]$_{TR\text{-}DI}$, et une expérimentation systématique [des]$_{LA\text{-}PR}$ politiques économiques néolibérales . L' application du [[modèle économique fondé sur le libre marché]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IN}$ [[a entraîné]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IN}$ des ajustements structurels continuels , [et dans]$_{LA\text{-}ST\text{-}AW}$ tous les pays concernés , les conséquences sociales de ces changements furent désastreuses . L' Echec fut retentissant . En 2002 , les [[conditions du marché du travail]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IC}$ furent les pires en 22 ans . [Les chiffres]$_{TR\text{-}AD}$ du chômage [étaient]$_{LA\text{-}IA\text{-}TA}$ [[en augmentation]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IN}$, et désormais plus de la moitié de la population active travaille dans le secteur informel . Le nombre de [[personnes vivant dans un état de pauvreté]$_{TR\text{-}SI\text{-}TL}$]$_{LA\text{-}}$ er alors que le montant des revenus minimum [[n' a cessé de baisser]$_{TR\text{-}DI}$]$_{LA\text{-}ST\text{-}AW}$ et que la [[croissance]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IN}$ dans la région a décliné (-0,8 %) . Certains pays ont été confrontés à la pire crise économique [[qu' ai connu]$_{TR\text{-}AD}$]$_{LA\text{-}ST\text{-}AW}$ l' histoire . La moitié de la population d' Argentine ([[qui dénombre]$_{TR\text{-}AD}$]$_{LA\text{-}ST\text{-}AW}$ 37 millions d' habitants) vit désormais en dessous du seuil de pauvreté ; les classes moyennes ont été fortement [touchées par la crise]$_{TR\text{-}DI}$ et plus d' un tiers de la population active est sans emploi ou sous-employée .

pos: NomC || lemma: chômage

3	language	context	source
Original context	en	Some countries have experienced the most serious economic crisis in history. More than half Argentina's 37m inhabitants now live below the poverty line; the middle classes have been hard hit; and more than a third of the country's active population is either unemployed or under-employed.	Full text
Target sentence	fr	Les classes moyennes ont été [lourdement]$_{LA\text{-}TL\text{-}FC}$ touchées et plus du tiers de la population active est soit au chômage , soit sous-employée .	Full text
Reference context	fr	En Argentine, par exemple, la classe moyenne a été laminée, plus de la moitié des 37 millions d'habitants vivent maintenant dans la pauvreté, et plus du tiers de la population active est sans travail ou sous-employée.	Full text
Alternative student context	fr	Plus de la moitié des 37 millions d'habitants de l'Argentine vit désormais en - dessous du seuil de pauvreté , les classes moyennes ont été durement touchées , et plus du tiers de la population active est soit au chômage , soit exploité .	Full text
Alternative student context	fr	Plus de la moitié des 37 millions d'habitants de l'Argentine vivent désormais sous le seuil de pauvreté , les classes moyennes ont été sérieusement touchées et plus d'un tiers de la population active du pays est au chômage ou sous - employé .	Full text
Alternative student context	fr	Plus de la moitié des 37 millions d'argentins vivent à présent sous le seuil de pauvreté ; les classes moyennes ont été durement touchées ; et plus du tiers de la population active du pays est soit au chômage ou sous - employé .	Full text
Alternative student context	fr	On peut citer l'Argentine où plus de la moitié de ses 37 millions d'habitants vivent actuellement sous le seuil de pauvreté : la classe moyenne a été fortement touchée et plus d'un tiers de la population active est sous - employée ou au chômage .	Full text
Alternative student context	fr	Plus de la moitié des 37 millions d'Argentins vivent désormais sous le seuil de la pauvreté , les classes moyennes ont été durement touchées et plus d'un tiers de la population active du pays est soit au chômage soit sous - employé .	Full text
Alternative student context	fr	Plus de la moitié des habitants de l'Argentine (37 m) vivent maintenant sous la ligne de pauvreté ; les classes moyennes ont été les plus touchées ; et plus d'un tiers de la population active du pays se trouve au chômage ou est sous - employée .	Full text
Alternative student context	fr	La moitié de la population d' Argentine ([[qui dénombre]$_{TR\text{-}AD}$]$_{LA\text{-}ST\text{-}AW}$ 37 millions d' habitants) vit désormais en dessous du seuil de pauvreté ; les classes moyennes ont été fortement [touchées par la crise]$_{TR\text{-}DI}$ et plus d' un tiers de la population active est sans emploi ou sous-employée .	Full text
Alternative student context	fr	Plus de la moitié de la population de l'Argentine , qui compte 37 millions d'habitants , vivent aujourd'hui en dessous du seuil de pauvreté , les classes moyennes ont été les plus touchées , et plus d'un tiers de le population active du pays est soit sans emploi , soit sous - employée .	Full text

Vive le Brésil ! Par Ignacio Ramonet . [Le nouveau président du Brésil , Luiz Inacio " Lula " da Silva , est le chef du [[parti travailliste]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IN}$ et c' est aussi un ancien dirigeant syndical . Elu en Octobre dernier , il arrive au pouvoir à une période où l' Amérique latine connaît de grands [troubles]$_{TR\text{-}SI\text{-}TL}$]$_{TR\text{-}DI}$. Pour la première fois , ce pays gigantesque - la dixième nation industrielle du monde , avec une population de 170 millions d' habitants - est sur le point d' être dirigée par un gouvernement démocratique ayant à sa tête un [[ancien militant]$_{TR\text{-}DI}$ d' [[extrême gauche]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IN}$ et anti-mondialiste .]$_{TR\text{-}DI}$ C' est un événement de grande importance . [[Même si]$_{TR\text{-}DI}$]$_{LA\text{-}ST\text{-}AW}$ dans un contexte différent , cela rappelle l' élection du président socialiste Chilien Salvador Allende en 1970 . Le 1er Janvier 2003 marque donc le début d' un nouveau cycle historique en Amérique Latine . Le précédent cycle avait débuté en 1983 , à la fin d' une période sombre marquée par les tyrannies militaires , la répression et les soulèvements armés , et a duré deux décennies . Les principales caractéristiques de [cette période]$_{TR\text{-}AD}$ furent la fin des mouvements de guérilla (excepté le mouvement guérillero de Colombie et le [[mouvement]$_{TR\text{-}AD}$ non-violent]$_{TR\text{-}OM}$ de l' armée zapatiste mené par le sous-commandant [Marco]$_{LA\text{-}HY\text{-}SP}$]$_{LA\text{-}TL\text{-}IG}$ au Chiapas) , des [régimes plus démocratiques]$_{TR\text{-}DI}$, et une expérimentation systématique [des]$_{LA\text{-}PR}$ politiques économiques néolibérales . L' application du [[modèle économique fondé sur le libre marché]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IN}$ [[a entraîné]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IN}$ des ajustements structurels continuels , [et dans]$_{LA\text{-}ST\text{-}AW}$ tous les pays concernés , les conséquences sociales de ces changements furent désastreuses . L' Echec fut retentissant . En 2002 , les [[conditions du marché du travail]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IC}$ furent les pires en 22 ans . [Les chiffres]$_{TR\text{-}AD}$ du chômage [étaient]$_{LA\text{-}IA\text{-}TA}$ [[en augmentation]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IN}$; et désormais plus de la moitié de la population active travaille dans le secteur informel . Le nombre de [[personnes vivant dans un état de pauvreté]$_{TR\text{-}SI\text{-}TL}$]$_{LA\text{-}}$ [illegible] pos: NomC || lemma: chômage [illegible]er alors que le montant des revenus minimum [[n' a cessé de baisser]$_{TR\text{-}DI}$]$_{LA\text{-}ST\text{-}AW}$ et que la [[croissance]$_{TR\text{-}DI}$]$_{LA\text{-}TL\text{-}IN}$ dans la région a décliné (-0,8 %) . Certains pays ont été confrontés à la pire crise économique [[qu' ai connu]$_{TR\text{-}AD}$]$_{LA\text{-}ST\text{-}AW}$ l' histoire . La moitié de la population d' Argentine ([[qui dénombre]$_{TR\text{-}AD}$]$_{LA\text{-}ST\text{-}AW}$ 37 millions d' habitants) vit désormais en dessous du seuil de pauvreté ; les classes moyennes ont été fortement [touchées par la crise]$_{TR\text{-}DI}$ et plus d' un tiers de la population active est sans emploi ou sous-employée .

Per-fide

Primeiro nível de anotação:

Formatação: paragráfos, secções, títulos, etc.

Metadados: data de publicação, autor, tipo de texto, registro, etc. (ex. **Compara**)

Anotação de carácter propriamente linguístico:

Etiquetação morfossintática (tagging)

Lematização

Anotação prosódica dos corpora orais transcritos (CINTIL)

Nota: Mesmo se os corpora não anotados constituem recursos interessantes, a presença de anotações torna-os mais úteis para a pesquisa linguística.

Per-fide

Para que servem os corpora?

- por um lado por motivos de eficácia técnica: podemos armazenar e distribuir volumes de dados consideráveis que podemos analisar rapidamente (interrogar os corpus, construir listas de frequência ou de co-ocorrências)
- por outro lado, porque constituem um recurso interessante que complementa a pesquisa realizada a partir de dicionários, de glossários ou mesmo na Internet.

Alguns exemplos típicos de utilização de corpora:

- Em linguística, para estudar a competência ou a performance linguística baseando-se em dados autênticos. Para o estudo do vocabulário, da gramática ou, nos estudos diacrónicos, para observar a evolução da língua ou do sentido das palavras

- Em linguística informática (Tratamento automático da linguagem natural: TALN), para treinar e testar ferramentas de análise textual ou construir ontologias, dicionários para domínios específicos.
- Em Linguística Aplicada: na aprendizagem e ensino de uma língua materna ou estrangeira para que se possa emitir hipóteses e testá-las em vez de criar exemplos.

Per-fide

Linguateca http:/
www.linguateca.pt/

(centro de recursos para o processamento computacional da língua portuguesa)

Corpus CETEMPúblico

'interface de pesquisa' :

http://www.linguateca.pt/ACDC/

CONCORDÂNCIA: formato para visualização de resultados que apresenta todas as ocorrências de uma determinada pesquisa em contexto.

Per-fide

Pesquisa de padrões regulares

É possível fazer pesquisas no corpus através da utilização de expressões regulares, desde que colocadas entre aspas.

- **Alternância**
  - Pesquisas alternativas são marcadas com uma barra vertical |
    - **"gostaria|gostava"** dá como resultado todas as ocorrências de *gostaria* e todas as ocorrências de *gostava*

- **Conjuntos de caracteres**
  - Um conjunto de caracteres entre parênteses rectos dá como resultado ocorrências de qualquer um desses caracteres:
    - **"ministr[ao]"** dá como resultado ocorrências de *ministro* e de *ministra*

Per-fide

Pesquisa de padrões regulares

É possível fazer pesquisas no corpus através da utilização de expressões regulares, desde que colocadas entre aspas.

- **Opcionalidade**
  - O "?" (ponto de interrogação) permite assinalar a opcionalidade de qualquer carácter ou expressão que o precede:
    - **"gatos?"** dá como resultado gato e gatos.

- **Ponto**
  - O "." (ponto final) equivale a qualquer ocorrência de um só carácter (letra,dígito ou símbolo):
    - **"pens.“** irá dar como resultado ocorrências de *penso*, *pensa*, *pense*, etc.

Per-fide

Pesquisa de padrões regulares

Iteração

Há três formas diferentes de expressar a iteração:

– O operador * (asterisco) faz com que o carácter ou expressão que o precede seja realizado **zero ou mais vezes**:
  - **"euro.*"** dá como resultado qualquer palavra que começa por *euro*, incluindo a própria forma *euro*
  - **".*euro.*"** dá como resultado qualquer palavra que contém a sequência *euro*

– O operador **+** (mais) é semelhante, mas requer que haja pelo menos uma ocorrência do carácter ou expressão que o precede:
  - **"neuro.+"** dá como resultado qualquer palavra que começa por *neuro* mas excluindo *neuro*

Per-fide

Pesquisa de padrões regulares

Iteração

– Finalmente, {l,n} permite que o número de iterações esteja limitado por um valor mínimo (l) e um valor máximo (n).

- **"psi.{2,8}"** dá como resultado palavras que começam por *psi* e que têm entre 2 e 8 caracteres adicionais

Agrupamento

– Os parêntesis são usados para agrupar expressões. Os operadores descritos acima podem assim aplicar-se ao conjunto da expressão entre parêntesis como se fosse um único carácter:

- **"lind(íssim)?o“** dá como resultado *lindo* e *lindíssimo* (i.e. a sequência *íssimo* a seguir ao *d* é opcional)
- **"ga(to)*"** equivale a *ga*, *gato*, *gatoto*, *gatototo*, etc. (i.e. *to* pode ocorrer zero ou mais vezes)

Per-fide

Pesquisar informação linguística

Pesquisa simples:

Ao pesquisar tradução, obtêm-se ocorrências de *tradução*

A pesquisa apenas dá como resultado formas exactamente iguais à pesquisada.

Esta palavra isolada pode ser pesquisada através do atributo *word*:

[word="tradução"]

Procura formas com a forma ortográfica *tradução*. Os resultados são idênticos aos da pesquisa simples por tradução.

- **A opção por este tipo de pesquisa pode ser útil quando se combina *word* com outros atributos** (ver exemplos abaixo)

Pesquisa de uma sequência de palavras:

"tradução" "literal" dá como resultado ocorrências de *tradução literal*

Per-fide

Pesquisar informação linguística

Pos: Categorias morfossintácticas

- A pesquisa de uma palavra com determinada categoria morfossintáctica é feita através do atributo **pos** (*part-of-speech*):
- **[pos="N"]** encontra formas com a etiqueta POS + nome comum
- **[word=".*fobia" & pos="N"]** devolve formas que são nomes comuns e que contêm a sequência *fobia*
- **[word="saia" & pos="N"]** devolve formas que correspondem ao nome comum *saia* (e não ao verbo *saia:* **[word="saia" & pos="V"]**)
- **[word="situação"] [pos="ADJ"]** devolve formas que correspondem à palavra *situação seguida de adjetivos*

Per-fide

Pesquisar informação linguística

Lema: forma base

Para pesquisar palavras pelo seu lema (sua **forma base**), deve usar-se o atributo **lema**:

[lema=traduzir]

pesquisa formas que têm *traduzir* como lema, tais como *traduz, traduziu*, *traduzido* ou *traduziram, etc.*

[lema="pôr\+se"][word="a"]@[pos="V.*"]
(*põem-se a bailar, pôs-se a gritar , puseram-se a contar , põem-se a chorar, põe-se a rosnar, pôs-se a andar , …)*

Per-fide

Pesquisar informação linguística

FLEXÃO VERBAL

Para pesquisar formas de acordo com os seus traços de flexão verbal, devem usar-se os atributos **temcagr** e **pessnum** (pessoa e número):

[lema="ter" & temcagr="PS_IND" & pessnum="1P"]
pesquisa todas as formas de primeira pessoa plural do lema *ter* no Presente do Indicativo

FLEXÃO NOMINAL

Os atributos género e número têm, respectivamente, os valores **F** (feminino) ou **M** (masculino), e os valores **S** (singular) ou **P** (plural). Estes valores permitem pesquisar ocorrências com valores de flexão específicos:

Nomes no feminino: **[pos="N.*" & gen="F"]**

Adjetivos no feminino plural: **[pos="ADJ" & pessnum="P" & gen="F"]**

Per-fide

Flexão nominal

Alguns tokens possuem anotação de grau, acessível através do atributo **temcagr**:

[pos="N.*" & temcagr="DIM"] pesquisa todas os nomes com o grau diminutivo *(bolsinho)*

deriv - [deriv="DER.*"] (**palavras derivadas por sufixação ou prefixação**)

[pos="N.*" & deriv="DER.*"] ***(queijinho, filinha, estalinismo, churrasquinho, fidelização, branqueamento, fotogaleria, moedita, antiterror, inconsolo, buzinadelas, autojulgamento, multipontos, …)***

Flexão nominal

Alguns tokens possuem anotação de grau, acessível através do atributo **temcagr**:

[pos="N.*" & temcagr="DIM"] pesquisa todas os nomes com o grau diminutivo *(bolsinho)*

deriv - [deriv="DER.*"] (**palavras derivadas por sufixação ou prefixação**)

[pos="N.*" & deriv="DER.*"] *(queijinho, filinha, estalinismo, churrasquinho, fidelização, branqueamento, fotogaleria, moedita, antiterror, inconsolo, buzinadelas, autojulgamento, multipontos, …)*

Per-fide

Pesquisa avançada

- **Através da combinação das diferentes opções de pesquisa descritas acima, é possível construir pesquisas avançadas como as que são de seguida exemplificadas:**

[word="casa"] [pos="ADJ"] pesquisa ocorrências da forma *casa* seguida de adjectivo (*casa típica/própria/comercial/senhorial/ burguesa/romana/londrina /nipónica/mortuária/rural/portuguesa/ agrícola/térrea/mãe/real/*, …)

[lema="tomar"] [pos="N"] pesquisa ocorrências do lema *tomar* seguido de nome *(tomar conta, conhecimento, consciência, parte, partido, posse, posição, nota, medidas, decisões, contacto, …)*

[pos="DET_artd"][pos="N"]
pesquisa ocorrências de formas com a categoria morfossintática artigo definido seguidas de um nome comum

Per-fide

Pesquisa avançada

[pos="DET_artd"][pos="ADJ"]?[pos="N"]
semelhante à pesquisa anterior, mas permite a ocorrência opcional de um adjectivo (indicado pela etiqueta *ADJ* e pelo operador de opcionalidade "?") entre o artigo definido e o nome comum

[pos="DET_artd"][pos!="N"]{2,3}[pos="ADJ"]
dá como resultado sequências com um artigo definido seguido por 2 ou 3 formas que não sejam nomes comuns, seguidas por um adjectivo
... etc.

[pos="N|ADJ"]{3,}
dá como resultado sequências que tenham pelo menos 3 adjectivos e nomes comuns consecutivos (em qualquer uma das ordens possíveis)

- *ajuda financeira maciça, defesa central brasileiro, principais produtores mundiais , campeã olímpica cubana , força militar multinacional , direitos constitucionais básicos ,*

Per-fide

RESUMO SINTAXE DE PESQUISA

Pesquisa simples

uma palavra/
expressão devolve as
suas ocorrências

**Miausculas/
minusculas**

**Expressões para
caracteres**

. qualquer carácter
único

Operadores de repetição

? opcional

* zero ou mais vezes

+ uma ou mais vezes

{n} exactamente n vezes

{n,} n ou mais vezes

{,n} até n vezes

{m,n} de m a n vezes

Expressões combinadas

| alternância

( ) junção

! negação

Pesquisa pela anotação

[atributp="valor"]

[atributo!="valor"]

[atributo="valor" & atributo="valor"]

[atributo="valor" | "atributo"=valor"]

Per-fide

- **Distribuição** : quantificação ou número de ocorrências resultante de uma pesquisa num corpus.

das **formas**

(pesquisar por lemas → ex: **[lema="casa"] "de" @ [ ]** frequência das várias formas do lema *casa* e da forma que se segue) *(ex. banho, campo, habitação, espectáculos, férias, fado, madeira, repouso, pasto, saúde, família, chá…)*

dos **lemas**

[pos="ADJ.*"] [pos="N.*"] [pos="ADJ.*"]

(*poderoso império otomano, grande poder comercial, primeiro cônsul português, fortes raízes ameríndias, …)*

Nesse caso são os lemas do primeiro adjectivo que aparecem.Se quisermos por exemplo os dos nomes, basta anteceder esta unidade do caracter @:

[pos="ADJ.*"] @[pos="N.*"] [pos="ADJ.*"]

Distribuição dos lemas

Procura: **[pos="ADJ.*"] [pos="N.*"] [pos="ADJ.*"]**
Distribuição de **lema**
Corpo: CETEMPúblico 1.7 v. 4.0

283832 casos.

Distribuição

Houve **4386** valores diferentes de **lema**. .

grande	33864
novo	28065
primeiro	16149
bom	11105
único	8359
principal	8331
antigo	7838
actual	7259
pequeno	6540
último	6483
diverso	5574
próximo	5565

Procura: **[pos="ADJ.*"] @[pos="N.*"] [pos="ADJ.*"]**
Distribuição de **lema**
Corpo: CETEMPúblico 1.7 v. 4.0

283832 casos.

Distribuição

Houve **9756** valores diferentes de **lema**.

eleição	4251
ano	2849
grupo	2342
empresa	1877
centro	1803
partido	1740
país	1629
situação	1572
crise	1509
força	1422
condição	1422
problema	1417
sistema	1413

Outras formas de distribuição

- da categoria gramatical (pos) – ex: "casa" lista de frequência das várias categorias gramaticais de uma forma.
- do tempo verbal e/ou do caso pronominal
- de pessoa e/ou número
- de género morfológico
- da função sintática

por secção

- por campo semântico

Per-fide

Corpus COMPARA

Corpus paralelo literário bidirecional (pt-en)

Interface de pesquisa: http://www.linguateca.pt/COMPARA/

Sintaxe de pesquisa:

• formulação ligeiramente diferente:

Ex: CETEMPúblico – DET_artd (artigo definido)

COMPARA – DETardt (artigo definido)

• anotação morfossintática mais detalhada:

Ex: CETEMPúblico

N (nome comum)

N_PROP (nome comum iniciado por maiúscula)

PROP (nome próprio)

PROP_kc (nome próprio composto ligado por &)

Per-fide

COMPARA

N = nomes comuns (ex: *Disse que queria o carro*.);

Nprop = nomes comuns iniciados por maiúscula – personificados ou individualizados (ex:

recebendo o troféu das mãos da Rainha);

N_Nprop = nomes comuns iniciados por maiúscula (ex (início de frase): *Dor no joelho*.);

PROP = nomes próprios (ex: *Teria preferido ir de carro a Londres*);

N_PROP = nomes comuns que podem ser também nomes próprios (ex: *Inverno*)

Nprop_PROP = nomes comuns começados por maiúscula que podem ser também nomes

próprios (ex: *ainda é para ele uma espécie de Paraíso*);

N_Nprop_PROP = nomes comuns que podem ser também nomes comuns começados por

maiúscula e nomes próprios (ex: *Céu e Inferno são concepções sociais para uso da plebe*);

N_V = nomes comuns que podem ser verbos (ex: *Caminhámos até à orla do parque como*

condenados do destino.)

N_Vn = nomes comuns que podem ser formas verbais nominalizadas (ex: *E se eu estivesse*

realmente a contar o sucedido)

Nnumfract = numerais fraccionários que funcionam como nomes (ex: *um terço de todas as*

empresas de engenharia);

Nnummult = numerais multiplicativos que funcionam como nomes (ex: *a terra deve valer hoje*

Per-fide

COMPARA

word: palavra (forma) → [word="…"]
lema: forma base da palavra → [lema="…"]
pos: (*part-of-speech*) categoria grammatical → [pos="…"]
temcagr: tempo caso e grau → [temcagr="…"]
pessnum: pessoa e número → [pessnum="…"]
gen: género → [gen="…"]
emp: locuções → [emp="…"]

↓

N = nominais (ex: *o aquecimento central*)
PRP = prepositivas (ex: *a partir de terça-feira*)
etc

Ex: [word="de" & emp="PRP"]
(*apesar de*; *em vez de*; *por causa de*; *à beira de*; etc)

Per-fide

Exemplos de pesquisa:

a) Procurar *phrasal verbs* flexionados:

[lema="go**"] "**off**"**
[pos="V.*"] "off**"**

b) Procurar verbos flexionados seguidos de uma preposição (de pt para en), exemplo verbo *ir*:

[lema="ir**"] [pos="PRP.*"] [pos="V.*"]**
[lema="ir**" & temcagr="IMPF_IND"] [pos="PRP.*"] [pos="V.*"]**
[lema="ir**"] [word="a"] [pos="V.*"]**

c) Procurar colocações com o verbo *make*:

[lema="make**"] [pos="N.*"]**

d) Pesquisar expressões que integram palavras específicas (de pt para en); por exemplo '*por [adjetivo] que*' :

"por**" [pos="ADJ.*"] "**que**"**

Per-fide

Pesquisa avançada e ultra-avançada:

Pesquisa avançada:

(1) Permite definir a direção de pesquisa: pt para o en ou vice-versa;

(2) Permite introduzir a expressão de pesquisa na língua de partida e

restringir a correspondência da mesma na língua alvo, exemplo (*pt* com

restrição de alinhamento em *en*):

“bonito”

[lema=“nice”]

(3) Permite escolher partes específicas do corpus (variantes, datas, originais ou traduções, textos específicos e autores);

(4) Permite especificar os resultados por concordâncias ou distribuições.

Pesquisa ultra-avançada (formulário expandido de pesquisa avançada)

Per-fide

Corpus OPUS

Corpus paralelo multilingue com vários corpora de vários domínios
- 14 sub-corpus (domínios: administrativo, legislativo, médico, técnico)

Interface de pesquisa: http://opus.lingfil.uu.se/bin/opuscqp.pl

1. Devemos selecionar um corpus e ao clicar sobre o mesmo aparecem as
   várias línguas de partida.
2. Ao selecionar uma das línguas, aparece uma janela onde podemos proceder à nossa pesquisa e à seleção da língua ou das línguas-alvo.

Sintaxe de pesquisa (CWB):

Atributos:
word → **[word="…"]**
lem (lema) → **[lem="…"]**
pos → **[pos="…"]**

Per-fide

Exemplos de pesquisa

! Nem todas as línguas estão anotadas. Não é possível a pesquisa de elementos morfossintáticos para o português, mas podemos partir de uma língua anotada e fazer uma análise contrastiva.

! Etiquetagem em inglês.

Palavra simples: **[word="**folheto**"]** ou **"**folheto**"**

Pesquisar ocorrências com o verbo *take* (flexionado): **[lem="**take**" & pos="V.*"]**

! Ao selecionar uma ou mais opções ('word' , 'id' , 'lem' , 'pos' , 'tree') podemos obter e visualizar a informação selecionada.

Exemplos:

[lem="take**"]** → ocorrências com as formas do lema *take*

[lem="take**"]** → mais opção 'word' → ocorrências com as formas do lema *take*

[lem="take**"]** → mais opção 'lem' → ocorrências apenas com a forma *take*

[lem="take**"]** → mais opção 'word' e 'lem' → ocorrências com as formas do lema *take*/lema

[lem="take**"]** → mais opção 'word' e 'pos' → ocorrências com as formas do lema *take*/anotação morfossintática

Per-fide

Visualização de resultados:

- vertical
- KWIC (*key word in context*)
- horizontal
- (número de ocorrências a visualizar)

Pesquisa avançada (*advanced search*):

- Restrições de alinhamento: permite introduzir a expressão de pesquisa na língua de partida e restringir a correspondência da mesma na língua-alvo, exemplo: (pt) **"**folheto**"** (en) **"**leaflet **"**.

- Restrições de contexto: permite definir o número de elementos – fragmento, ficheiro, cabeçalho, parágrafo, frase - à direita e à esquerda da expressão.

Per-fide

Corpus Per-Fide

Corpus multilingue (pt-es-ru-fr-it-de-en) em vários domínios (literário, técnico, jurídico-legislativo, jornalístico e religioso)

Interface experimental: www.per-fide.ilch.uminho.pt/query

O que já permite?

Pesquisa monolingue:

1. Selecionar uma língua;
2. Selecionar corpus (um ou mais);
3. Inserir palavra de pesquisa, exemplo **carta (Vatican)**;

Pesquisa bilingue:

1. Selecionar um par de línguas (PT-EN);
2. Selecionar corpus (um ou mais); (Europarl)
3. Inserir palavra de pesquisa numa língua de partida e/ou língua-alvo;
   - **PT** fé **(Vatican)**
   - **PT** [word="escrev.*"] **(Vatican)**
   - [word="escrev.*"] "(a|ao|aos)" @ [] {2,3} **(Vatican)**
   - **PT-EN** "dar" "a" "palavra"– "give" "the" "floor" **(Europarl+JRC+Eurlex)**
4. PTD (Dicionário Probabilístico de Tradução):
   a) Ao clicar numa palavra mostra as traduções possíveis na língua-alvo;
   b) → acesso à concordância bilingue;
   c) Nas concordâncias: podemos clicar (duas vezes) em qualquer palavra e obtemos o PTD dessa palavra.

Per-fide

Exercício - CETEMPÚBLICO

[word="dar"][pos="N.*" & pessnum="S"] [pos="PRP"]
[lema="dar" & temcagr="INF"][pos="N.*" & pessnum="S"] [pos="PRP"]
[word="dar"][pos="N.*" & pessnum="S"] [word="a"]

Pesquisa do verbo *dar* seguido de nome comum no singular seguido de preposição

[word="dar"][pos="N.*" & pessnum="S"] [pos="PRP"]

[lema="estar"] [word="em"] [pos="N"]

Pesquisa ocorrências do verbo *estar* flexionado seguido da preposição *em* e de nome comum

(*estar/estou/estavam … em causa/vantagem/risco/perigo/greve/segurança/vigor/ greve/palco/foco/guerra*
/retiro/curso/campo/destaque/alerta/jogo/paz/leilão/chamas…,)

[lema="estar"] [pos="PRP"] [pos="N"]

(estar de acordo/regresso/pé/férias, sem trabalho, a cargo/caminho, para brincadeiras, sob pressão…)