Origines des erreurs en Traduction Spécialisée : différentiation textométrique grâce aux corpus de textes cibles annotés

Natalie Kübler[1], Maria Zimina[1], Serge Fleury[2]

(1) CLILLAC-ARP EA 3967, Université Paris Diderot-Paris 7, Paris

(2) CLESTHIA EA 7345, Sorbonne Nouvelle-Paris 3, Paris, France

Plan

- Présentation du projet
- Typologie d'erreurs **MeLLANGE**
- Corpus ***ER-TRAD-SP1*** et ***ER-TRAD-SP2***
- Base textométrique alignée multiannotée
- Indices quantitatifs sur les constructions potentiellement complexes qui sont à l'origine des difficultés (*Le Trameur*)
- Exploitation des résultats
- Conclusions, perspectives

Projet, corpus de travail, typologie d'erreurs **MeLLANGE**

Corpus et Traduction Spécialisée (TS)

Translation Studies → aspects spécifiques des textes traduits (Baker 1999; Puurtinen 2003, Olohan & Baker 2002; Olohan 2004; Mauranen 2007; Frankenberg-Garcia 2009; Loock 2013)

Enseignement de la TS et recherche expérimentale →

- Bowker & Bennison (2003) : Student translation Tracking System.
- Pearson (2003) : petit corpus parallèle ; traductions des étudiants /vs/ trad. Professionnelle
- Castagnoli et al. (2011) : sensibiliser les étudiants aux différentes stratégies adoptés dans la traduction
- Frankenberg-Garcia (2016) Are translations longer than source texts?

Objectifs & Méthodologie

Paris 7 depuis 15 ans en Master ILTS

- 1999: photographie numérique ; TA HOWTO
- 2004 : collaboration étroite avec les experts du domaine abordé : sciences de la terre et des planètes (STEP) + corpus
- 2004-07: projet européen MeLLANGE
- 2013-2015 : évaluation de la méthodologie et des résultats obtenus
- 2015-2016 : modifications => typologie et enseignement

Création de corpus

Méthodologie de travail étudiant : traduire en langue de spécialité

→ articles en anglais: sciences de la terre

→ compilation de corpus comparable EN/FR

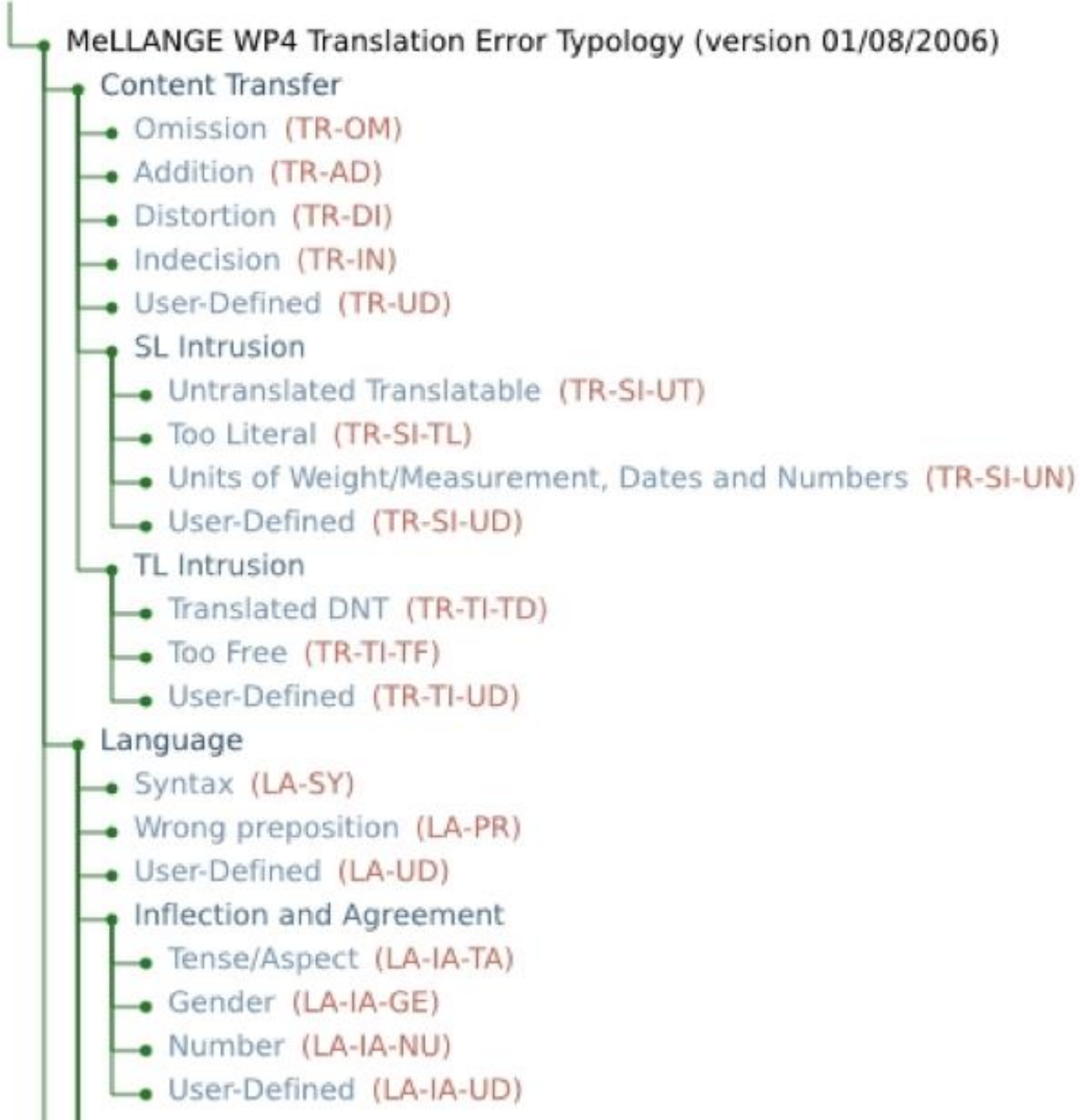
→ analyse terminologique et phraséologique

→ collaboration avec les experts

→ BD terminologique et phraséologique

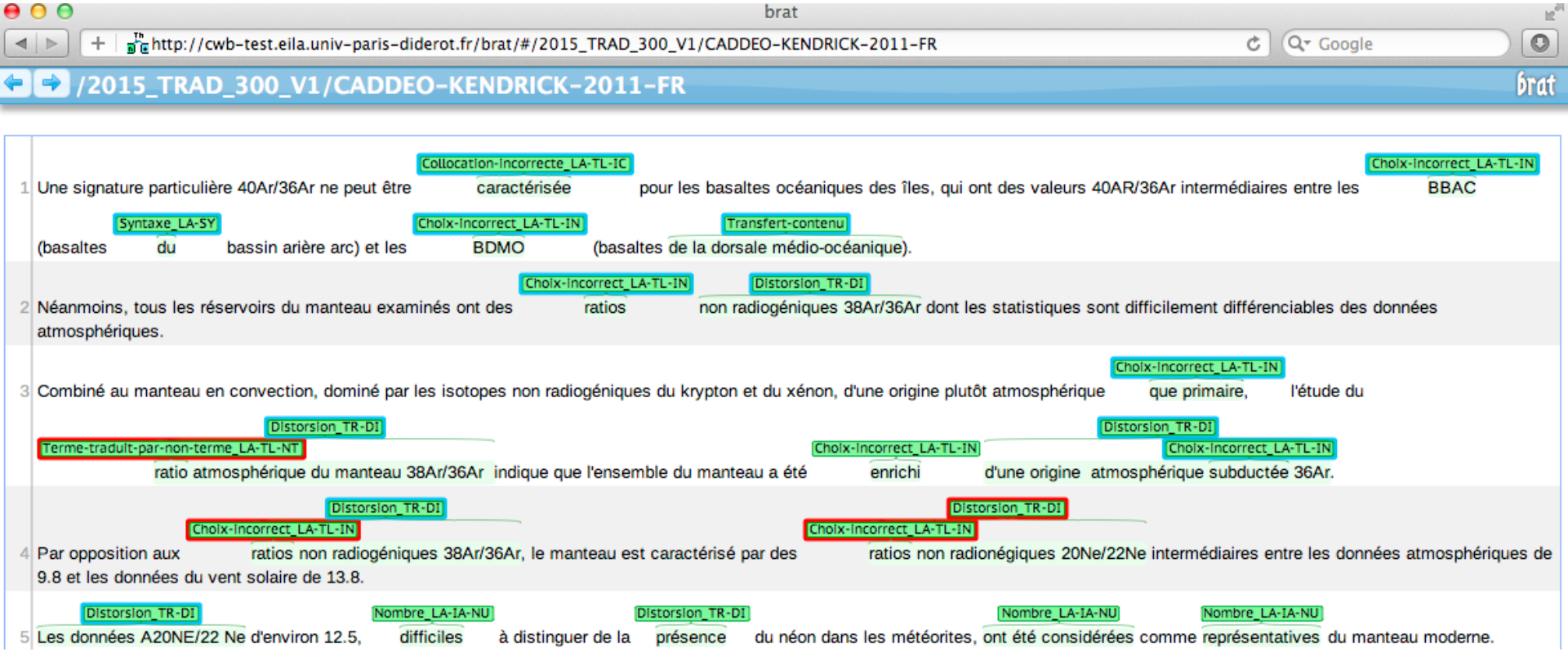
→ traduction avec et sans corpus

Production et annotation de traductions


MeLLANGE WP4 Translation Error Typology (version 01/08/2006)
Content Transfer
Omission (TR-OM)
Addition (TR-AD)
Distortion (TR-DI)
Indecision (TR-IN)
User-Defined (TR-UD)
SL Intrusion
Untranslated Translatable (TR-SI-UT)
Too Literal (TR-SI-TL)
Units of Weight/Measurement, Dates and Numbers (TR-SI-UN)
User-Defined (TR-SI-UD)
TL Intrusion
Translated DNT (TR-TI-TD)
Too Free (TR-TI-TF)
User-Defined (TR-TI-UD)
Language
Syntax (LA-SY)
Wrong preposition (LA-PR)
User-Defined (LA-UD)
Inflection and Agreement
Tense/Aspect (LA-IA-TA)
Gender (LA-IA-GE)
Number (LA-IA-NU)
User-Defined (LA-IA-UD)

- Terminology and Lexis
  - Incorrect (LA-TL-IN)
  - False Cognate (LA-TL-FC)
  - Term Translated by Non-Term (LA-TL-NT)
  - Inconsistent with Glossary (LA-TL-IG)
  - Inconsistent within TT (LA-TL-IT)
  - Inappropriate collocation (LA-TL-IC)
  - User-Defined (LA-TL-UD)
- Hygiene
  - Spelling (LA-HY-SP)
  - Accents or Diacritics (LA-HY-AC)
  - Incorrect Case (Upper/Lower) (LA-HY-CA)
  - Punctuation (LA-HY-PU)
  - User-Defined (LA-HY-UD)
- Register
  - Inconsistent with ST (E.g. Form of Address) (LA-RE-IS)
  - Inappropriate for TT Text Type (LA-RE-IN)
  - Inconsistent within TT (LA-RE-IT)
  - User-Defined (LA-RE-UD)
- Style
  - Awkward (LA-ST-AW)
  - Tautology (LA-ST-TA)
  - User-Defined (LA-ST-UD)
- User-Defined (UD)

Traduction annotée sur Brat



Corpus ER-TRAD-SP1 et ER-TRAD-SP2

53 extraits de 14 articles scientifiques : 32 815 occurrences :

- ER-TRAD_SP1: 14 189 occurrences: **sans accès au corpus**
- ER-TRAD-SP2: 18 626 occurrences: **avec accès au corpus**

Bi-texte: analyses quantitatives sur annotations multiples

BRADT => base textométrique annotée

Format propriétaire d'annotation BRAT

```
T1      Transfert-contenu 17 21      aura
#1      AnnotatorNotes T1 Indicatif ou
        conditionnel, pour "would have".
```

Conversion

```
48  <item type="forme" pos="6"><f>aura</f><c>VER_futu</c><l>avoir</l><a>Transfert-contenu</a><a>
    Indicatif ou conditionnel, pour "would have".</a></item>
49
50  <item type="delim" pos="7"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a></item>
51
52  <item type="forme" pos="8"><f>peu</f><c>ADV</c><l>peu</l><a>-</a><a>-</a></item>
53
54  <item type="delim" pos="9"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a></item>
55
56  <item type="forme" pos="10"><f>de</f><c>PRP</c><l>de</l><a>-</a><a>-</a></item>
57
58  <item type="delim" pos="11"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a></item>
59
60  <item type="forme" pos="12"><f>conséquences</f><c>NOM</c><l>conséquence</l><a>-</a><a>-
    </a></item>
```

Base textométrique alignée

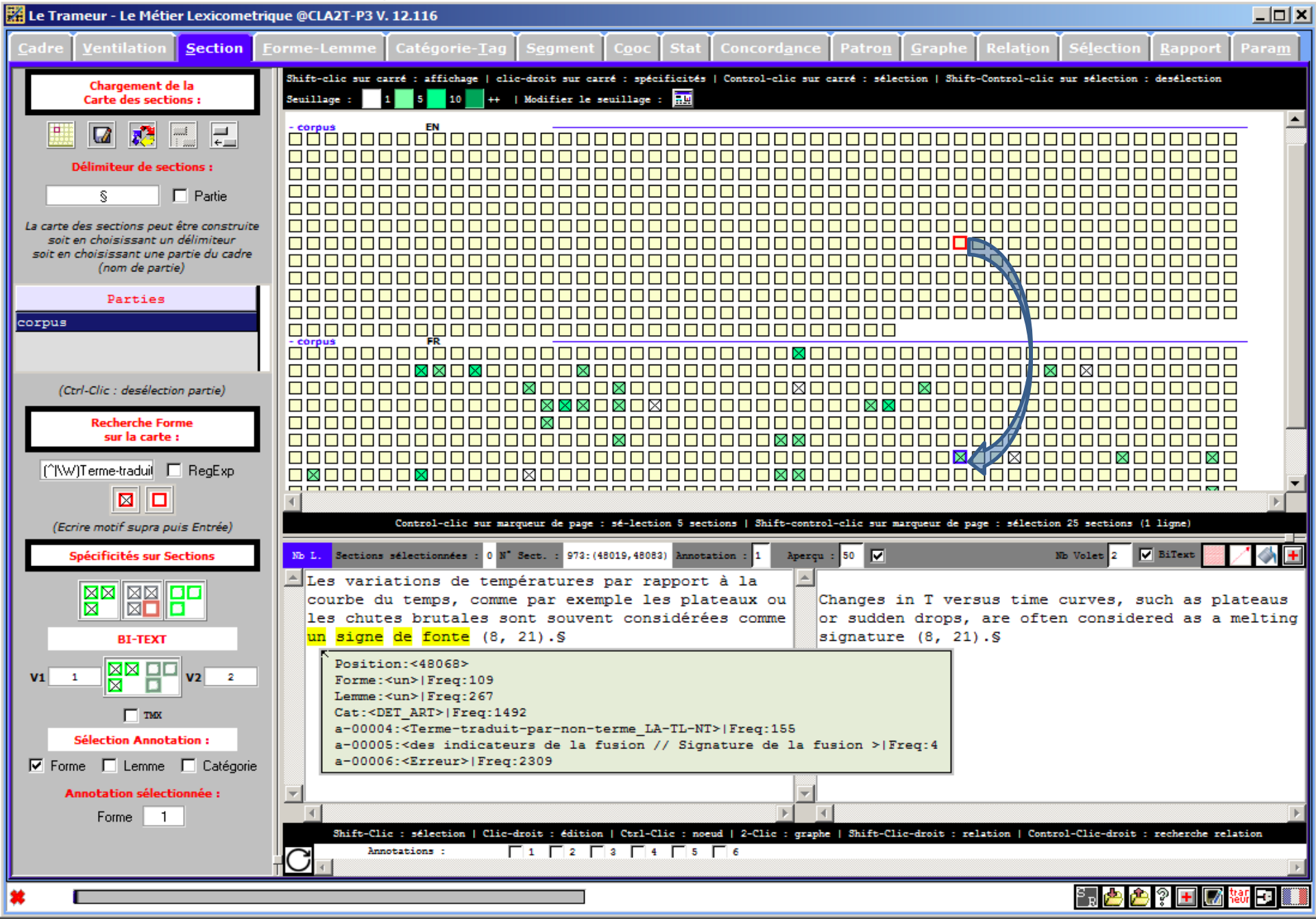
```
<item type="forme" pos="6"><f>would</f><c>MD</c><l>would</l><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="7"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>-
</a></item>
<item type="forme" pos="8"><f>have</f><c>VH</c><l>have</l><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="9"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>-
</a></item>
<item type="forme" pos="10"><f>only</f><c>JJ</c><l>only</l><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="11"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>-
</a></item>
<item type="forme" pos="12"><f>a</f><c>DT</c><l>a</l><a>-</a><a>-</a><a>-</a></item>
<item type="delim" pos="13"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>-
</a></item>
<item type="forme" pos="14"><f>small</f><c>JJ</c><l>small</l><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="15"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>-
</a></item>
<item type="forme" pos="16"><f>effect</f><c>NN</c><l>effect</l><a>-</a><a>-</a><a>-
</a></item>
```

original

traduction

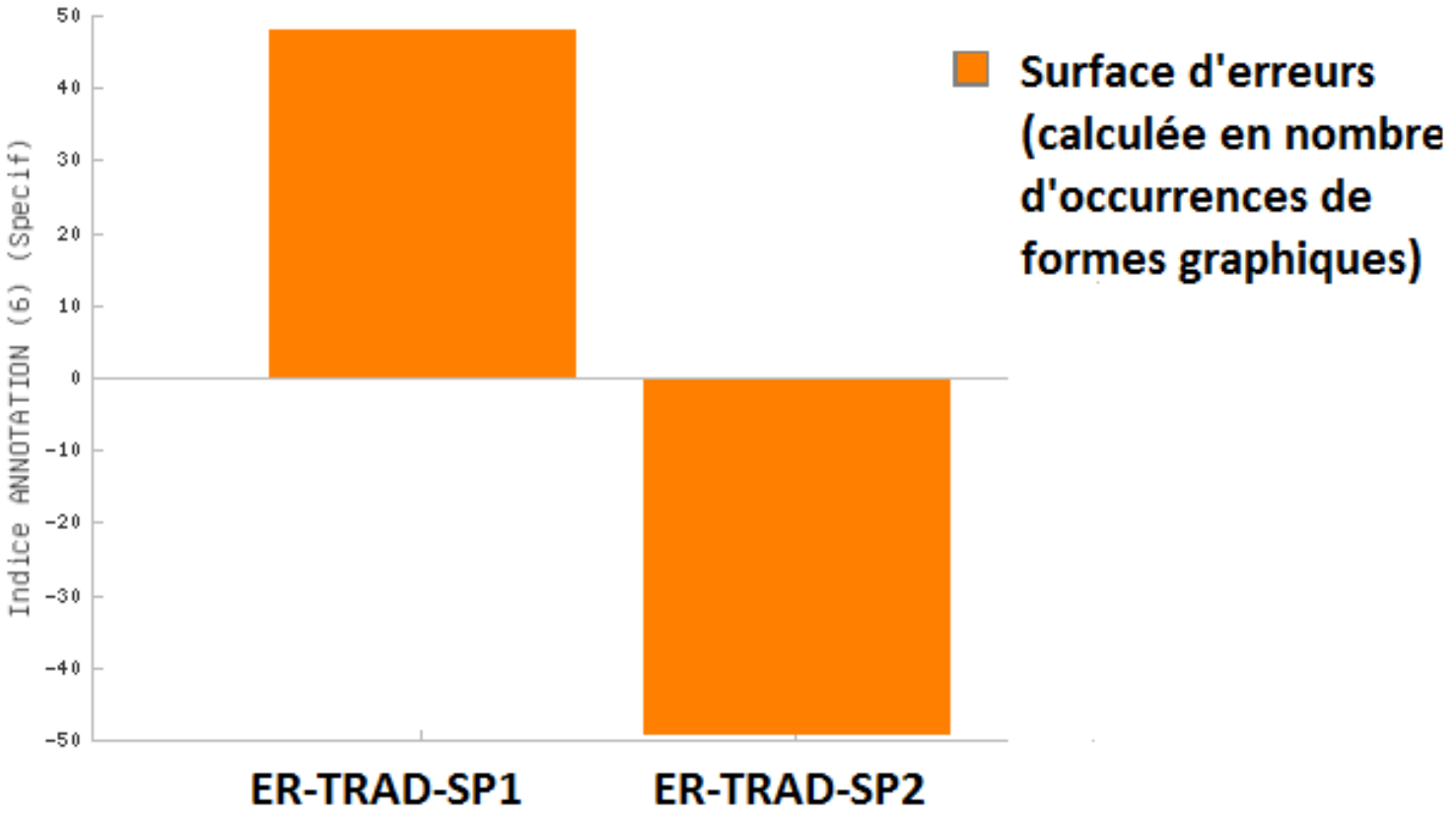
```
<item type="forme" pos="26884"><f>aura</f><c>VER_futu</c><l>avoir</l><a>
Transfert-contenu</a><a>Indicatif ou conditionnel, pour "would have".</a><a>Erreur
</a></item>
<item type="delim" pos="26885"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>
-</a></item>
<item type="forme" pos="26886"><f>peu</f><c>ADV</c><l>peu</l><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="26887"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>
-</a></item>
<item type="forme" pos="26888"><f>de</f><c>PRP</c><l>de</l><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="26889"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>
-</a></item>
<item type="forme" pos="26890"><f>conséquences</f><c>NOM</c><l>conséquence</l><a>-
</a><a>-</a><a>-</a></item>
```

Import dans ***Le Trameur***



Comparaison des surfaces d'erreurs ***ER-TRAD-SP1 & ER-TRAD-SP2*** (1/2)

Analyse des **spécificités sur annotations**



La méthode des spécificités (Lafon 1984; Lebart et Salem, 1994)

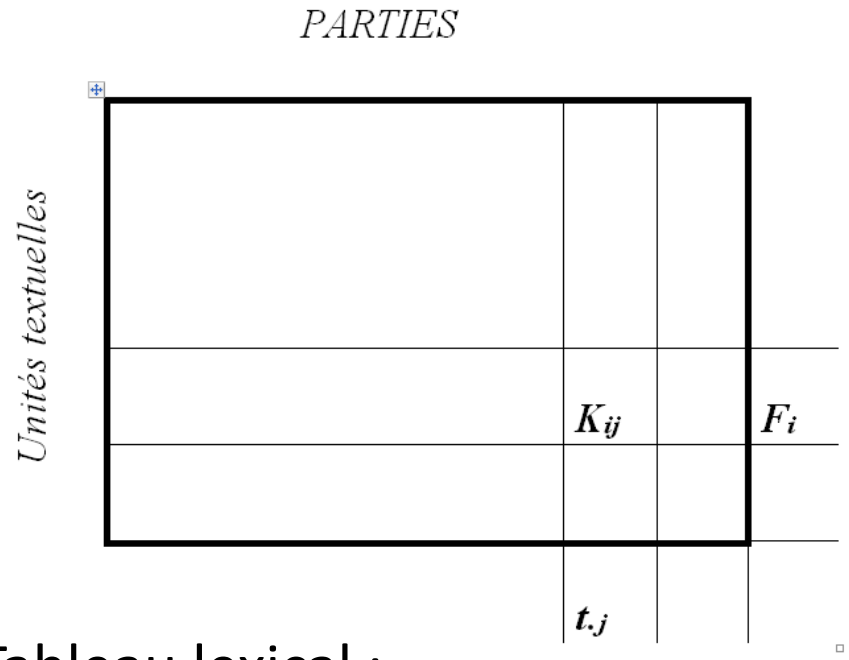
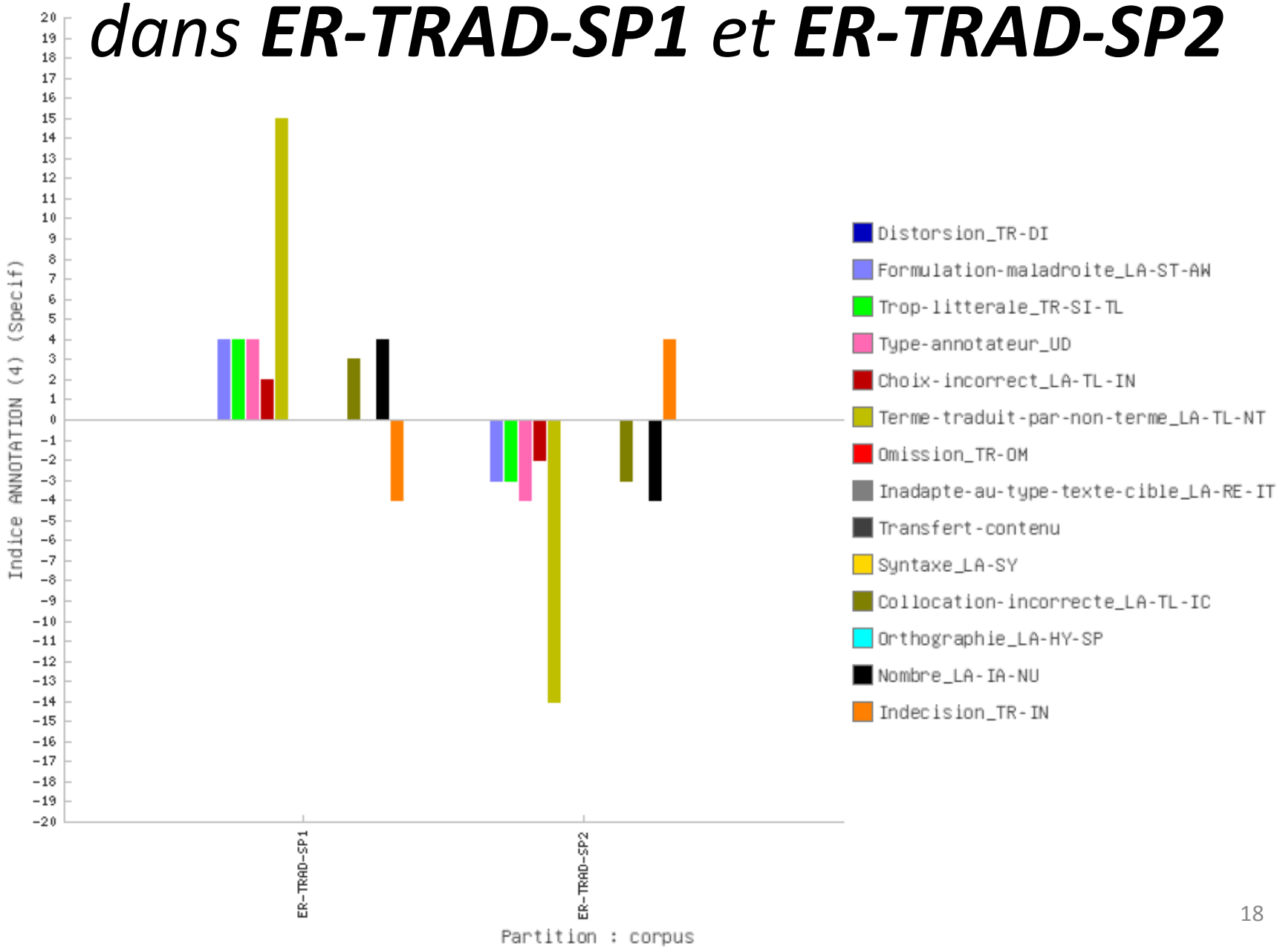


Tableau lexical :

K_{ij} : fréquence de l'unité ***j*** dans la partie ***i***
F_i : fréquence de ***i*** dans le corpus
t_j : taille de la partie ***j***

Si l'effectif K_{ij} ne se situe pas dans les limites de ce que le **calcul probabiliste** permettait d'espérer, on calcule un *indice de spécificité* : *sur-emploi* ou *sous-emploi* de l'unité (spec. ***±xx***)

*Erreurs fréquentes affectant les Noms dans **ER-TRAD-SP1** et **ER-TRAD-SP2***



Erreurs type « Indécision » affectant les Noms dans ***ER-TRAD-SP2***

---------- corpus=ER-TRAD-SP2 ----------

) Les couleurs symbolisent les milliers de comptage/numération pour la fluorescence du Fe et

Les couleurs symbolisent les milliers de comptage/numération pour la fluorescence du Fe et qui

une concentration maximale en Fe ne se superposent/chevauchent pas avec les zones comportant une

concentration maximale en Fe ne se superposent/chevauchent pas avec les zones comportant une concentration

est à son maximum au centre du laser hotspot (?), là où la

à son maximum au centre du laser hotspot (?), là où la phase

Image. 4) Ces valeurs sont également en excellent/parfait accord avec celles des deux précédentes

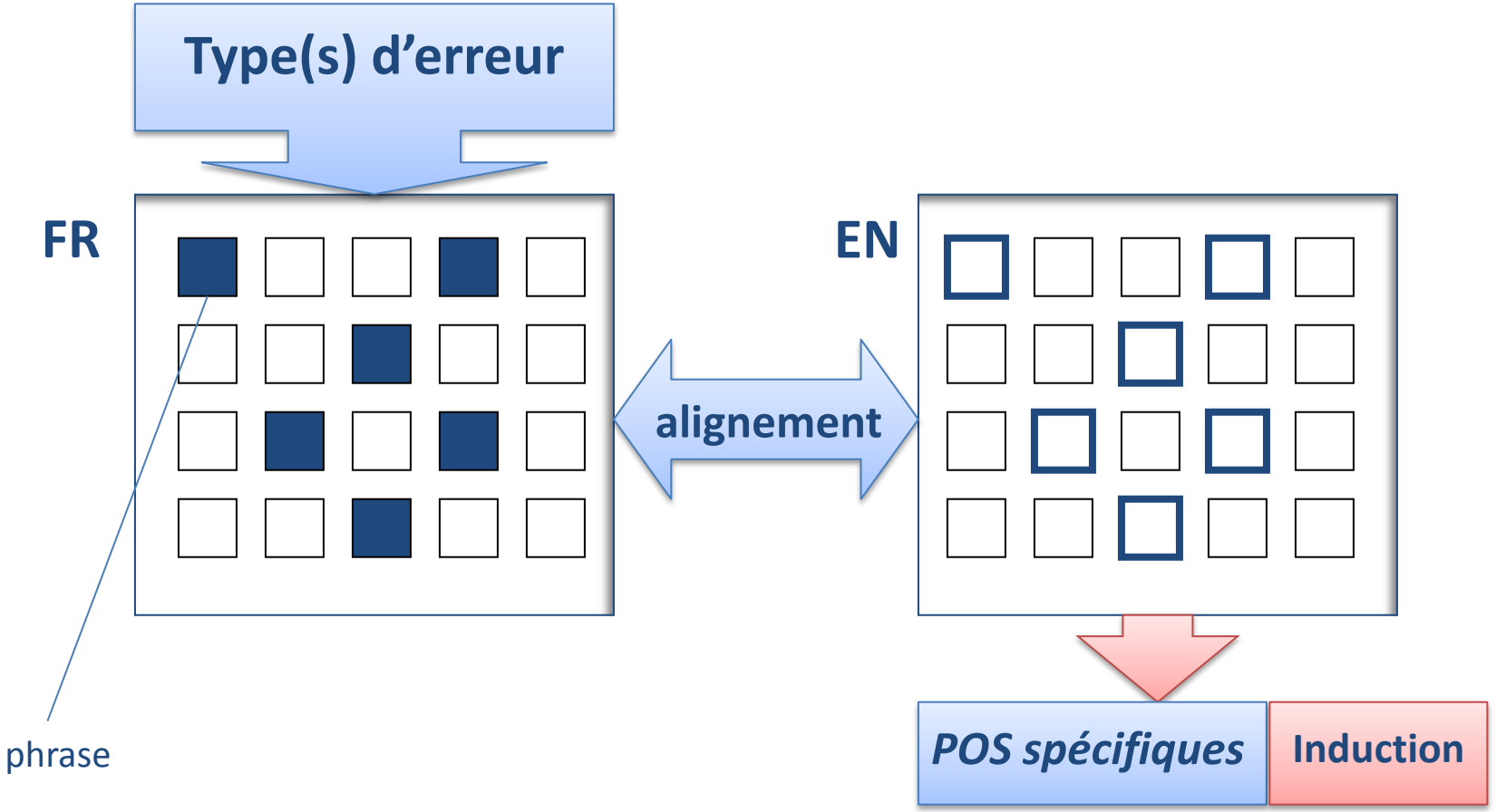
4) Ces valeurs sont également en excellent/parfait accord avec celles des deux précédentes études

principale différence provient de la composition des matériaux de base/matières premières : L'olivine

différence provient de la composition des matériaux de base/matières premières : L'olivine dans

Indices quantitatifs sur les constructions potentiellement complexes de l'original (***ER-TRAD-SP1***)

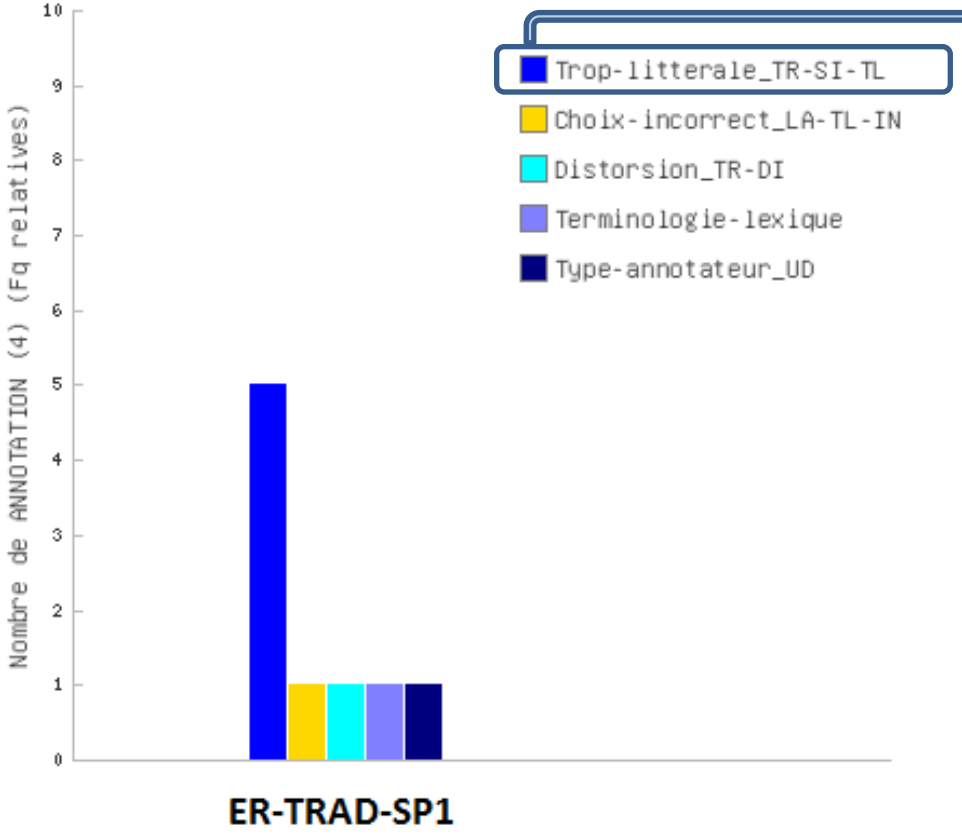
Spécificités sur contextes alignés



Noms affectés par les erreurs (exemples)

Surface d'erreurs en Nb occ. dans ER-TRAD-SP1	**Surface d'erreurs en Nb occ. dans ER-TRAD-SP2**	**Terme (lemme)**
11	**9**	***manteau***
5	9	échantillon
9	3	pression
8	4	fer
7	5	réaction
3	8	rayon
7	4	condition
6	4	rapport
7	3	température
3	6	ratio

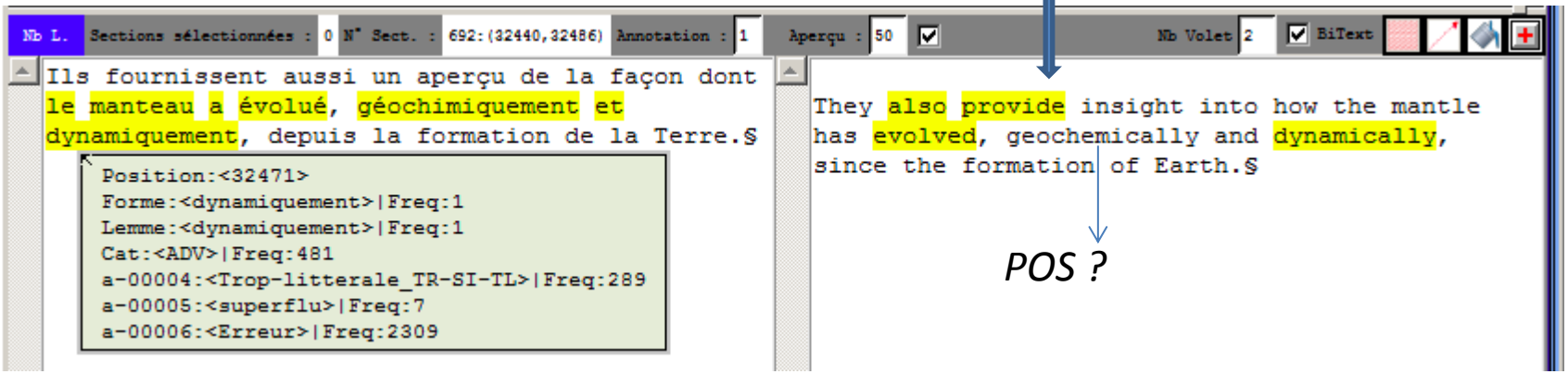
Erreurs affectant « *manteau* »



Verb, past participle
List marker
Modal verb
Adverb, comparative
Adverb
Verb *be*, base form
Verb, base form

Complexité à l'origine des erreurs

Lecture Textométrique Différentielle (LTD)



Constats (***ER-TRAD-SP1***)

- Spécificités des contextes sources (seuil 10) : 2 699 occurrences
- Erreurs relevées dans les productions : 2 277 occurrences
- Certaines erreurs co-occurrent dans les mêmes contextes (*Formulation maladroite* et *Distorsion* : 15 contextes partagés)

Eléments caractéristiques des phrases sources d'erreurs en ***ER-TRAD-SP1*** :

- les ***participes passés*** (spec.+4, F_i=322, K_{ij}=277)
- l'auxiliaire ***be*** au passé (spec.+3, F_i=64, K_{ij}=59)
- les ***modaux*** (spec.+2, F_i=88, K_{ij}=77)
- les ***prépositions*** (spec.+2, F_i=1 620, K_{ij}=1 316)
- les ***noms au pluriel*** (spec.+2, F_i=915, K_{ij}=746)
- la préposition ***to*** (spec.+2, F_i=238, K_{ij}=198)
- les ***gérondifs*** et ***participes présents*** (spec.+2, F_i=175, K_{ij}=149).

Spécificités sources (complexité) et erreurs de traduction ne sont pas des alignements

- **(Original)** Fluid inclusions in the vein minerals representative of first breakdown fluids and fluid inclusions in olivine-enstatite representing final breakdown fluids, were analysed by crushing the samples in vacuum. §
- **(Traduction)** Les échantillons d'inclusions fluides dans les veines minérales représentatives de la[Type-annotateur_UD] première rupture des fluides, et de celles représentant la_rupture[Terme-traduit-par-non-terme] finale, ont été brisés dans_la_chambre_pour_analyse[Distorsion]. §

Structure complexe des GN en anglais :

↓ nombreux termes composés

stratégies de traduction qui rétablissent les relations explicites entre les constituants de la phrase au niveau micro-syntaxique

Spécificités sur catégories morpho-syntaxiques

Erreurs de **syntaxe**

VOLET : EN	VOLET : FR
Thus, H 2 generation may be episodic depending on the rates of formation and destabilization of mineral surface layers during progressive waterrock interaction in an open system.§	Conclusion, la production de H 2 peut être réalisée en plusieurs stades, selon les taux de formation et de déstabilisation de couches superficielles minérales, pendant une interaction eau-roche progressive, au sein dun circuit ouvert. §
Therefore, the extraction of continental crust from this already-depleted reservoir (the EDR) cannot have greatly increased the Sm/Nd ratio of the MORB source; otherwise its 143Nd/144Nd value would have evolved to higher values than those observed for any terrestrial rock.§	Par conséquent, lextraction de croûte terrestre de ce réservoir, déjà appauvrit, ne peux avoir augmenté le rapport Sm/Nd de la source de basalte de dorsale médio-océanique, de façon conséquente. Sans quoi sa valeur en 143Nd/144Nd aurait évolué de manière plus importante que les valeurs observées sur nimporte quelle roche terrestre.§
the volume of mantle from which the continental crust was extracted must be large.§	le volume de croûte terrestre extrait du manteau se doit être inférieur à celui-ci.§
These mechanisms require a positive volume change of dehydration; otherwise, elevated pore pressures will not occur.§	Ces mécanismes requièrent un changement de volume positif de la déshydration; sans quoi il ne pourra pas y avoir de pressions interstitielles élevées.§
Above 6.5 GPa, the antigorite dehydration reaction had a negative Clapeyron slope and volume change of reaction.§	Au-dessus de 6,5 GPa, la réaction de la déshydratation de l'antigorite avait une pente de Clapeyron négative et un changement de volume de réaction négatif. §
Owing to the near-edge spectral characteristics (peak position, structure) for potential candidate (hydr)oxides (for example, ferrihydrite, haematite and goethite), we cannot uniquely identify the Fe(ii)bearing phase and henceforth use the term Fe(iii)-(hydr)oxides to indicate Fe bound to O and/or OH in a variety of crystal structures.§	En raison des caractéristiques spectrales près du seuil d'absorption (position maximale, structure) pour les candidats potentiels d'(hydr-) oxydes (tels que le ferrihydrite, l'hématite et la goethite), nous ne pouvons pas identifier définitivement la phase contenant du Fe(III) et utiliserons donc le terme (hydr-) oxydes de Fe(III) pour signaler l'existance d'un lien entre Fe et O et/ ou HO dans des diverses structures de cristaux. §
Hydrothermal organic reactions affect petroleum formation, degradation, and composition (2, 3), provide energy and carbon sources for deep microbial communities (4, 5), and may be important in the origin of life (6, 7).§	Les réactions organiques hydrothermales ont un effet sur la formation, la dégradation et la composition du pétrole, elles pourvoient de l'énergie et des sources en carbone pour des communautés microbiennes profondes et peuvent avoir leur importance dans l'origine de la vie. §

Conclusions

- Amélioration de la TS et des méthodologies de son enseignement à l'aide d'analyses quantitatives des corpus de traductions annotées
- Pertinence de l'utilisation du schéma d'annotation d'erreurs MeLLANGE
- Propositions pour le profilage des erreurs de traduction en contextes alignés
- Indices quantitatifs sur les constructions potentiellement complexes qui sont à l'origine des difficultés des apprenants (transfert de contenu, erreurs de langue, etc.).

Perspectives

- Supports de cours informatisés qui allient la typologie d'erreurs et l'exploration dynamique des contextes alignés et profilés par type d'erreurs.
- Profils d'erreurs dans les productions réalisées avec et sans apport de corpus
- Poursuite de la réflexion sur les méthodes d'enseignement qui amènent la diminution du nombre d'erreurs en TS

Références

Fleury S., Zimina M. (2014). Trameur: A Framework for Annotated Text Corpora Exploration. Proc. of COLING 2014 (the 25th International Conference on Computational Linguistics: System Demonstrations), August 2014, Dublin, Ireland.

Kübler N., Mestivier A., Pecman M., Zimina M. (2016). "Exploitation quantitative de corpus de traductions annotés selon la typologie d'erreurs pour améliorer les méthodes d'enseignement de la traduction spécialisée." Actes des 13es Journées internationales d'Analyse statistique des Données Textuelles, Nice, juin 2016.

Kübler, N. (2011). Working with different corpora in translation teaching. In Frankenberg-Garcia A., Flowerdew L., and Aston G. editors, New Trends in Corpora and Language Learning. London: Continuum.