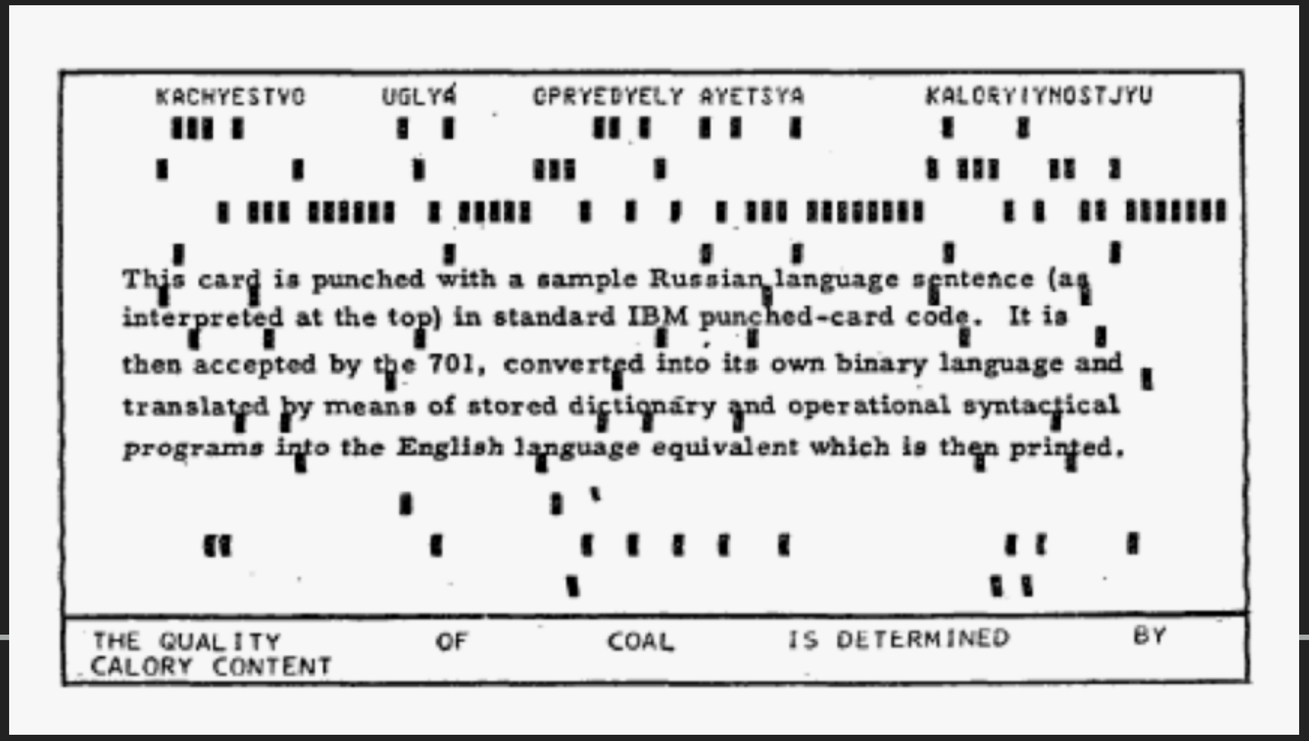
MACHINE TRANSLATION FOR OPEN SCIENCE



MACHINES À TRADUIRE POUR OUVRIR LA SCIENCE

cnrs SORBONNE UNIVERSITÉ Inria Université Paris Cité anr

MATOS – LA SCIENCE DANS PLUS D'UNE LANGUE

MATOS – LA SCIENCE DANS PLUS D'UNE LANGUE

- Initiative d'Helsinki sur le multilinguisme dans la communication savante (2019)

« La recherche est internationale. C'est une bonne chose! Le multilinguisme permet de continuer à mener des recherches pertinentes au niveau local. Préservons-le! La diffusion des résultats de la recherche dans notre propre langue crée de l'impact. Soutenons-la! Il est essentiel d'interagir avec la société et de partager les connaissances au-delà des milieux universitaires. Faisons la promotion de cette ouverture! Les infrastructures disponibles pour communiquer la recherche en langues nationales sont fragiles. Ne les laissons pas disparaître. »

- Plan « science ouverte » MESR + Projets OPERAS
  - Rapport « Traduction et science ouverte »
  - Etudes préparatoires pour un service de traduction de documents scientifiques (2019-2020)

MATOS – LA SCIENCE DANS PLUS D'UNE LANGUE

- Initiative d'Helsinki sur le multilinguisme dans la communication savante (2019)

« La recherche est internationale. C'est une bonne chose! Le multilinguisme permet de continuer à mener des recherches pertinentes au niveau local. Préservons-le! La diffusion des résultats de la recherche dans notre propre langue crée de l'impact. Soutenons-la! Il est essentiel d'interagir avec la société et de partager les connaissances au-delà des milieux universitaires. Faisons la promotion de cette ouverture! Les infrastructures disponibles pour communiquer la recherche en langues nationales sont fragiles. Ne les laissons pas disparaître. »

- Plan « science ouverte » MESR + Projets OPERAS
  - Rapport « Traduction et science ouverte »
  - Etudes préparatoires pour un service de traduction de documents scientifiques (2019-2020)

- Appel à Projet Chist-ERA « *Science in your own language* » (2025-)
- LLM4Eu - « Science WP » (2025-)
- Projets franco-canadiens (2025-)

MATOS – VERS LA TRADUCTION INTÉGRALE DE DOCUMENTS ACADÉMIQUES

- Défis scientifiques
  - Traduire avec des ressources et des dictionnaires
  - Identification des termes et de leur variation
  - Traduire les phénomènes discursifs
  - Utilisation de la structure du document
  - Evaluation de la traduction de documents

MATOS – VERS LA TRADUCTION INTÉGRALE DE DOCUMENTS ACADÉMIQUES

- Défis scientifiques
  - Traduire avec des ressources et des dictionnaires
  - Identification des termes et de leur variation
  - Traduire les phénomènes discursifs
  - Utilisation de la structure du document
  - Evaluation de la traduction de documents
- Résultats attendus
  - De nouvelles ressources
  - Nouvelles méthodes pour la traduction de documents longs
  - Métriques d'évaluation
  - Illustrer les nouvelles possibilités du TAL pour traiter les textes scientifiques

MATOS – ORGANISATION

- Un consortium interdisciplinaire
  - ALTAE / UPC (N. Kübler)
  - ALMAnaCH / Inria (R. Bawden)
  - INIST / CNRS (J.-F. Nominé, M. Huguin)
  - ISIR-MLIA / SU & CNRS (F. Yvon, PI)
- Durée
  - 4 ans, début 01/01/2023
- Personnels non permanents:
  - 3 étudiant-e-s en thèse
  - 39 hm ingénieur
  - 12 hm post-doc

CORPUS PARALLÈLES ET MONOLINGUES

- Traitement des Langues
  - Monolingues
    - ACL anthology, ISCA archive
    - these.fr, ISTEX
  - Parallèles
    - Résumés de these.fr
    - Résumés d'articles (TAL, ISTEX)
    - TA + post-édition de résumés
    - Articles complets - quasi-traductions

CORPUS PARALLÈLES ET MONOLINGUES

- Traitement des Langues
  - Monolingues
    - ACL anthology, ISCA archive
    - these.fr, ISTEX
  - Parallèles
    - Résumés de these.fr
    - Résumés d'articles (TAL, ISTEX)
    - TA + post-édition de résumés
    - Articles complets - quasi-traductions
- Sciences de la Terre
  - Monolingues
    - these.fr, ISTEX
  - Parallèles
    - Résumés de these.fr
    - Résumés d'articles (CRAS, etc)
    - TA + post-édition de résumés
    - Articles complets - centre Mersenne

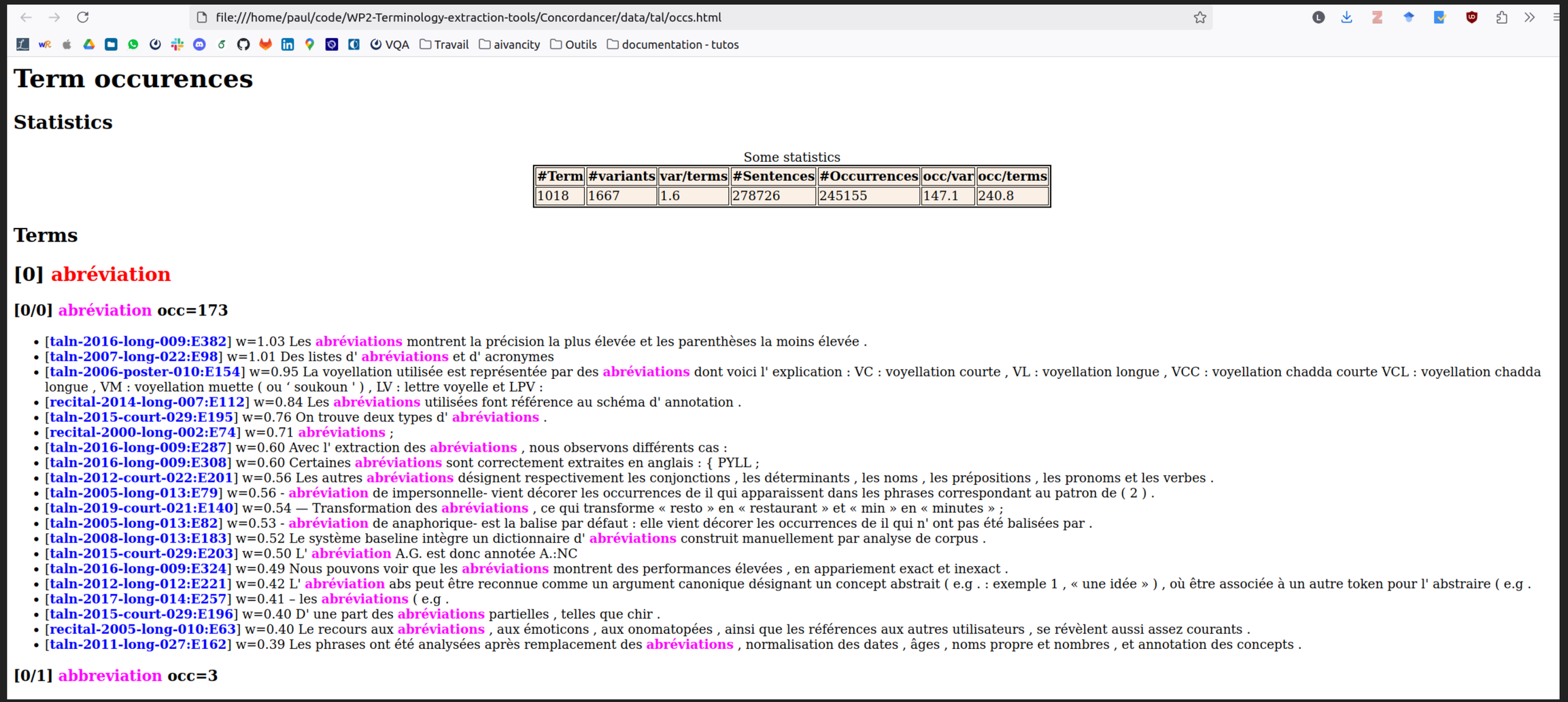
CORPUS PARALLÈLES ET MONOLINGUES

- Traitement des Langues
  - Monolingues
    - ACL anthology, ISCA archive
    - these.fr, ISTEX
  - Parallèles
    - Résumés de these.fr
    - Résumés d'articles (TAL, ISTEX)
    - TA + post-édition de résumés
    - Articles complets - quasi-traductions
- Sciences de la Terre
  - Monolingues
    - these.fr, ISTEX
  - Parallèles
    - Résumés de these.fr
    - Résumés d'articles (CRAS, etc)
    - TA + post-édition de résumés
    - Articles complets - centre Mersenne
- Outils
  - Conversion pdf - txt
  - Alignement de phrases

NOUVELLES TERMINOLOGIES SPÉCIALISÉES

https://loterre.istex.fr/8LP/fr/

REPÉRAGE ET MESURE DE LA VARIATION TERMINOLOGIQUE



Les oeuvres du « Concordanceur »

REPÉRAGE ET MESURE DE LA VARIATION TERMINOLOGIQUE

[820] apprentissage automatique

[820/0] machine learning occ=4

- [taln-2004-long-027:E75] w=0.38 • L' approche machine learning se fonde sur l' apprentissage automatique .
- [taln-1999-poster-009:E13] w=0.04 Une telle hypothèse de travail , qui s' inspire du courant Machine Learning en IA , est encore assez peu exploitée en TALN [Gorin et al .
- [taln-2018-DEFT-002:E19] w=0.04 La volumétrie importante des tweets mis à disposition pour ce concours nous permettait d' envisager des méthodes de type Machine Learning .
- [taln-2004-long-027:E70] w=0.02 Il existe deux grandes approches pour détecter le thème d' un document sachant les représentations vectorielles , l' approche statistique et l' approche machine learning , nous présentons maintenant les principes de quelques unes des méthodes de détection de thème .

[820/1] apprentissage machine occ=3

- [taln-2011-court-025:E33] w=0.03 Nous avons ainsi restreint la liste des rôles à prendre en compte dans l' apprentissage machine .
- [taln-2016-long-018:E39] w=0.02 L' approche la plus commune et efficace pour la tâche de classification d' énoncés en termes d' actes de dialogue est d' employer des techniques supervisées d' apprentissage machine (Tavafi et al .
- [recital-2010-long-001:E257] w=0.01 Soulignons que Luyckx & Daelemans (2008) , qui se sont penchés sur l' influence du nombre d' auteurs , obtiennent un rappel semblable au nôtre (76 %) pour 20 auteurs (avec une approche basée sur l' apprentissage machine) , alors que notre étude porte sur 53 auteurs .

[820/2] apprentissage automatique occ=353

- [taln-2007-poster-008:E57] w=1.41 ensuite un apprentissage automatique permet un filtrage .
- [taln-2010-long-020:E64] w=1.33 Algorithmes d' apprentissage automatique
- [taln-2013-court-026:E151] w=1.11 Mais l' apprentissage automatique nécessite une grande quantité de documents préalablement étiquetés .
- [taln-2015-long-017:E49] w=1.07 L' apprentissage automatique est la base de cette méthodologie d' exploration de corpus .
- [taln-2011-long-038:E222] w=1.03 Leur apprentissage automatique peut améliorer la couverture , mais au détriment de la précision .
- [taln-2019-court-004:E91] w=0.92 L' apprentissage automatique vise à indiquer des frontières de chaque chunk , mais aussi à déterminer son type .
- [taln-2007-long-033:E148] w=0.80 Une méthode d' apprentissage automatique pour évaluer les stades de développement
- [taln-2015-court-019:E38] w=0.78 L' apprentissage automatique n' avait pu encore être mis en oeuvre dans ce contexte , faute jusqu' à présent de corpus annotés et disponibles librement .
- [taln-2013-long-031:E78] w=0.71 Traditionnellement , l' apprentissage automatique se rapproche plutôt d' une classification (attribution d' une classe à un mot) que d' une annotation (délimitation d' une expression linguistique) .
- [taln-2009-long-022:E132] w=0.71 Apprentissage automatique
- [taln-2001-poster-004:E10] w=0.71 Apprentissage Automatique
- [taln-2019-court-004:E89] w=0.71 Apprentissage automatique
- [taln-2013-court-033:E71] w=0.67 SECTION 3 : Classification d' opinion par apprentissage automatique
- [taln-2014-long-019:E42] w=0.63 Ces systèmes utilisent tous des méthodes d' apprentissage automatique .
- [taln-2019-court-021:E227] w=0.63 L' apprentissage automatique proposé tient compte de ces observations et des caractéristiques linguistiques de chaque catégorie retenue .
- [taln-2011-long-014:E80] w=0.60 L' apprentissage automatique se base sur des tests sémantiques , qui permettent de mesurer le degré de subjectivité des termes , ainsi que leur valence s' il s' agit d' adjectifs , et qui sont effectués à l' aide du moteur de recherche Yahoo!.
- [recital-2014-long-005:E15] w=0.58 , Apprentissage Automatique
- [taln-2010-long-020:E201] w=0.57 En apprentissage automatique (AA) , la quantité de données d' apprentissage est une notion cruciale .
- [taln-2018-DEFT-008:E85] w=0.53 L' algorithme d' apprentissage automatique supervisé retenu est liblinear6 (Fan et al .
- [taln-2014-long-028:E319] w=0.52 Ce logiciel fournit des algorithmes d' apprentissage automatique et donne leurs résultats de classification .

Les oeuvres du « Concordanceur »

REPÉRAGE ET MESURE DE LA VARIATION TERMINOLOGIQUE

[820] apprentissage automatique docc=188 (occ=1050)

- ~~[**1540**] apprentissage reduction docc=38 (occ=2423) taln-2011-long-018:1:E4:E5 taln-2009-long-026:0:E36:E36 taln-2015-long-006:3:E195:E198 taln-2017-court-025:0:E112:E112 taln-2019-court-014:0:E17:E17 taln-2009-long-019:2:E24:E26 recital-2017-long-005:1:E227:E228 recital-2017-long-008:1:E339:E340 taln-2009-long-022:1:E118:E119 taln-2012-court-016:1:E70:E71

[823] traduction automatique docc=242 (occ=1406)

- [**823**] TA acronym docc=27 (occ=368) taln-2005-court-014:0:E2:E2 taln-2013-court-019:0:E24:E24 taln-2014-court-001:0:E2:E2 taln-2017-court-026:0:E50:E53 recital-2006-long-002:0:E10:E10 taln-1999-poster-011:0:E2:E2 taln-2005-long-023:0:E163:E163 taln-2009-long-014:0:E15:E15 taln-2011-long-005:0:E22:E22 taln-2014-long-025:0:E0:E2
- [**823**] traductions automatiques inflection docc=6 (occ=48) taln-2014-court-001:1:E2:E3 taln-2013-court-028:4:E7:E36 taln-2008-long-022:4:E4:E8 taln-2013-long-006:5:E0:E5 taln-2009-court-028:1:E7:E29 taln-2019-court-010:2:E31:E48

[823] TA docc=27 (occ=368)

- [**823**] traduction automatique acronym_expansion docc=9 (occ=1406) taln-2019-long-003:0:E212:E212 taln-2015-long-021:1:E6:E8 taln-2008-long-024:9:E196:E232 taln-2011-long-005:3:E90:E93 taln-2013-demo-008:4:E0:E4 taln-2017-court-026:1:E3:E20 recital-2006-long-002:6:E202:E208 taln-2017-long-005:5:E222:E242 taln-2009-long-014:7:E49:E71

[830] annotation manuelle docc=104 (occ=215)

- ~~[**76**] annotation reduction docc=21 (occ=2883) recital-2013-long-005:0:E226:E226 taln-2007-long-032:0:E69:E69 taln-2010-court-034:0:E45:E45 taln-2012-court-003:1:E23:E25 recital-2011-long-002:1:E7:E8 taln-2010-long-016:1:E250:E251 taln-2014-long-032:1:E83:E84 taln-2014-long-027:1:E253:E256 recital-2017-long-010:2:E189:E191 taln-2011-long-012:0:E321:E324
- ~~[**76**] annotations reduction docc=4 (occ=1229) taln-2015-long-026:3:E201:E204 taln-2010-court-013:3:E69:E72 taln-2016-long-016:3:E21:E24 taln-2007-long-024:4:E188:E192

[835] modèles de markov docc=32 (occ=84)

- ~~[**870**] modèle reduction docc=8 (occ=4394) recital-2005-court-003:2:E11:E13 taln-2012-long-003:1:E57:E58 taln-2014-court-010:2:E11:E13 taln-2014-long-028:1:E75:E76 taln-2007-poster-030:3:E2:E5 taln-2014-court-030:3:E36:E39 taln-2004-long-010:1:E24:E39 taln-2005-court-003:3:E11:E67
- ~~[**870**] modèles reduction docc=3 (occ=2105) taln-2016-court-012:5:E23:E31 taln-2010-long-019:6:E171:E177 taln-2011-long-002:6:E50:E56
- [**835**] modèle de markov inflection docc=2 (occ=26) taln-2014-long-028:1:E75:E85 recital-2005-court-003:6:E62:E68

[835] modèle de markov docc=7 (occ=26)

- ~~[**870**] modèle reduction docc=3 (occ=4394) taln-2005-court-003:1:E110:E111 taln-2019-court-014:6:E156:E162 recital-2010-long-003:9:E77:E86

[836] masculin docc=48 (occ=89)

Les oeuvres du « Concordanceur »

REPÉRAGE ET MESURE DE LA VARIATION TERMINOLOGIQUE

[820] apprentissage automatique docc=188 (occ=1050)

- ~~[**1540**] apprentissage reduction docc=38 (occ=2423) taln-2011-long-018:1:E4:E5 taln-2009-long-026:0:E36:E36 taln-2015-long-006:3:E195:E198 taln-2017-court-025:0:E112:E112 taln-2019-court-014:0:E17:E17 taln-2009-long-019:2:E24:E26 recital-2017-long-005:1:E227:E228 recital-2017-long-008:1:E339:E340 taln-2009-long-022:1:E118:E119 taln-2012-court-016:1:E70:E71

[823] traduction automatique docc=242 (occ=1406)

- [**823**] TA acronym docc=27 (occ=368) taln-2005-court-014:0:E2:E2 taln-2013-court-019:0:E24:E24 taln-2014-court-001:0:E2:E2 taln-2017-court-026:0:E50:E53 recital-2006-long-002:0:E10:E10 taln-1999-poster-011:0:E2:E2 taln-2005-long-023:0:E163:E163 taln-2009-long-014:0:E15:E15 taln-2011-long-005:0:E22:E22 taln-2014-long-025:0:E0:E2
- [**823**] traductions automatiques inflection docc=6 (occ=48) taln-2014-court-001:1:E2:E3 taln-2013-court-028:4:E7:E36 taln-2008-long-022:4:E4:E8 taln-2013-long-006:5:E0:E5 taln-2009-court-028:1:E7:E29 taln-2019-court-010:2:E31:E48

[823] TA docc=27 (occ=368)

- [**823**] traduction automatique acronym_expansion docc=9 (occ=1406) taln-2019-long-003:0:E212:E212 taln-2015-long-021:1:E6:E8 taln-2008-long-024:9:E196:E232 taln-2011-long-005:3:E90:E93 taln-2013-demo-008:4:E0:E4 taln-2017-court-026:1:E3:E20 recital-2006-long-002:6:E202:E208 taln-2017-long-005:5:E222:E242 taln-2009-long-014:7:E49:E71

[830] annotation manuelle docc=104 (

- ~~[**76**] annotation reduction docc=21 (occ=2883) recit … 003:1:E23:E25 recital-2011-long-002:1:E7:E8 taln-2010-long-016:1:E250:E251 taln-2014-long-032:1:E83:E84
- ~~[**76**] annotations reduction docc=4 (occ=1229) taln- … :4:E188:E192

[835] modèles de markov docc=32 (occ

- ~~[**870**] modèle reduction docc=8 (occ=4394) recital-… :E75:E76 taln-2007-poster-030:3:E2:E5 taln-2014-court-030:3:E36:E39 taln-2004-long-010:1:E24:E39 ta
- ~~[**870**] modèles reduction docc=3 (occ=2105) taln-2016-court-012:5:E23:E31 taln-2010-long-019:6:E171:E177 taln-2011-long-002:6:E50:E56
- [**835**] modèle de markov inflection docc=2 (occ=26) taln-2014-long-028:1:E75:E85 recital-2005-court-003:6:E62:E68

[835] modèle de markov docc=7 (occ=26)

- ~~[**870**] modèle reduction docc=3 (occ=4394) taln-2005-court-003:1:E110:E111 taln-2019-court-014:6:E156:E162 recital-2010-long-003:9:E77:E86

[836] masculin docc=48 (occ=89)

- Sélection des termes vedettes
- Filtrage des variants
- Extraction de contextes définitoires

Les oeuvres du « Concordanceur »

EXTRACTION DE CONTEXTES DÉFINITOIRES

Qu’est-ce qu’un alignement de mots ?

allauzen-wisniewski-2009-modeles: « *Un alignement mot à mot entre une phrase et sa traduction consiste à extraire des relations d'appariement entre les mots de la phrase source et les mots de sa traduction.* »

allauzen-wisniewski-2009-modeles: « *L'alignement mot à mot est une tâche intermédiaire dont le seul objectif est d'extraire des ressources pour une tâche « de plus haut niveau » (système de traduction automatique de recherche d’information…).* »

mdhaffar-etal-2019-apport: alignement mot à mot « *Un alignement mot à mot utilisant la distance de Levenshtein est réalisé entre la transcription manuelle (référence) et la transcription automatique (hypothèse).* »

tomeh-etal-2011-estimation: alignement mot à mot « *Un alignement mot à mot entre une phrase source et sa traduction (la phrase cible) regroupe un ensemble de liens décrivant une relation de traduction entre mots.* »

lardilleux-etal-2011-generalisation: « *L'alignement sous-phrastique consiste à extraire des traductions d'unités textuelles de grain inférieur à la phrase à partir de textes multilingues parallèles alignés au niveau de la phrase.* »

lardilleux-lepage-2009-anymalign: « *L'alignement sous-phrastique consiste à extraire des traductions d'unités textuelles de grain inférieur à la phrase à partir de textes multilingues dont les phrases ont préalablement été mises en correspondance* ».

ozdowska-2007-trois: « *ALIBI est un système d'alignement sous-phrastique qui vise à mettre en correspondance des unités textuelles de taille inférieure à la phrase qui sont potentiellement en relation de traduction (Ozdowska 2006)* ».

Les oeuvres du « Concordanceur »

NÉONYMIE COMPUTATIONNELLE

Comment traduire des néologismes ? En exploitant des définitions …

zero-shot learning + « Façon d'apprendre une tâche sans avoir eu de démonstration ou d'exemple » => apprentissage sans exemple

et des giga modèles de langue (GML).

Questions de recherche:

- Comment utiliser des définitions pour prédire des termes inconnus ?
- Comment retrouver et include des exemples de termes pertinents ?
- Les connaissances morphologiques des grands modèles de langue ?

NÉONYMIE COMPUTATIONNELLE ET GML MULTILINGUES

- Conditions et *prompts*
  - TERM: traduction directe du terme anglais

    « le terme anglais {src_term} peut se traduire en français par : »
  - DEF: génération depuis une définition en français

    « {def} définit le terme : "{def} defines the term : »
  - TERM+DEF : traduction et génération depuis une définition

    « {def} définit le terme anglais {src_term} qui peut se traduire en français par : »

NÉONYMIE COMPUTATIONNELLE ET GML MULTILINGUES

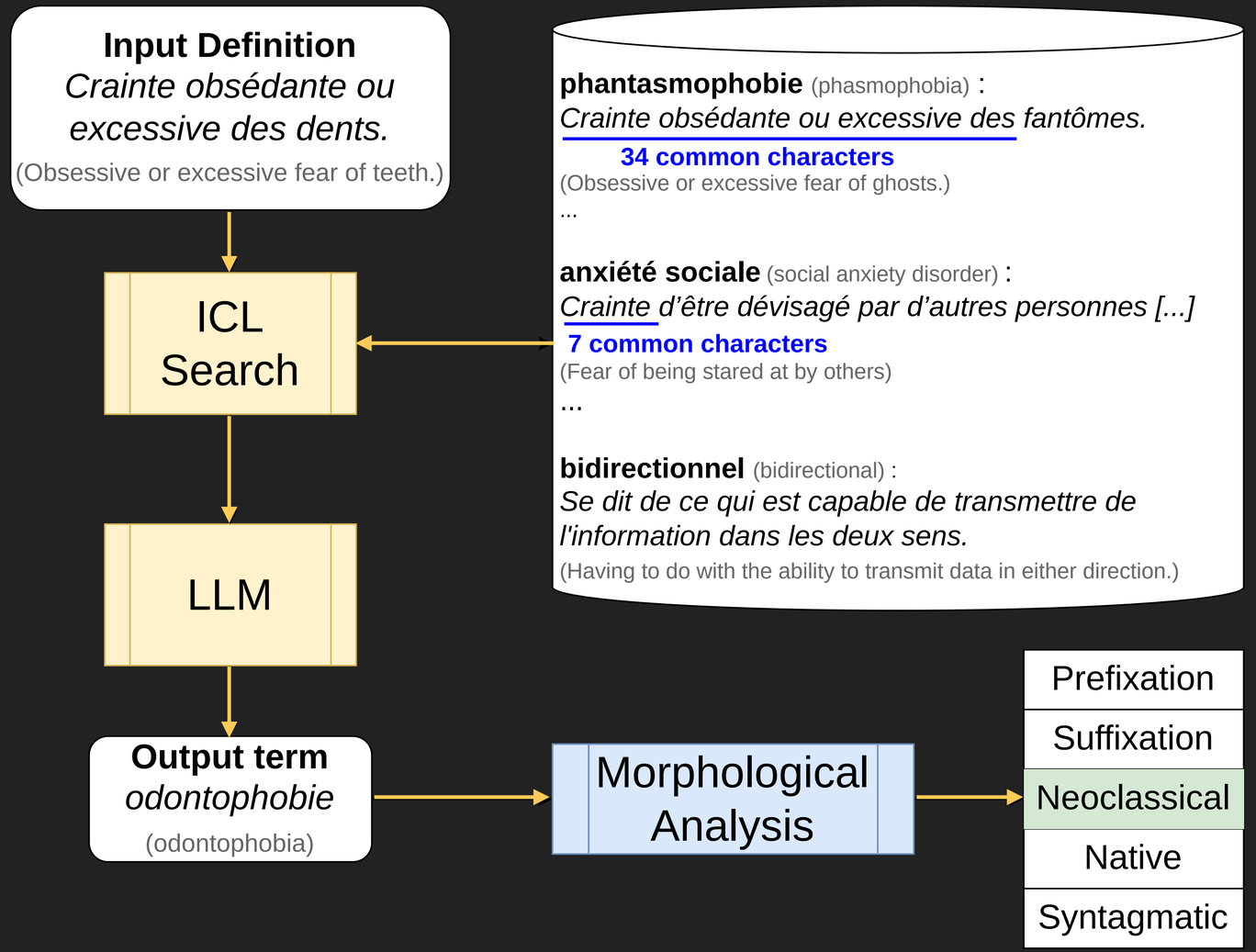
- Conditions et *prompts*
  - TERM: traduction directe du terme anglais
    « le terme anglais {src_term} peut se traduire en français par : »
  - DEF: génération depuis une définition en français
    « {def} définit le terme : "{def} defines the term : »
  - TERM+DEF : traduction et génération depuis une définition
    « {def} définit le terme anglais {src_term} qui peut se traduire en français par : »

- La génération utilise des giga modèles de langue multilingues + des exemples (RAGs)

NÉONYMIE COMPUTATIONNELLE – GENERATION PAR ANALOGIE

Sélection d'exemples

1. Aléatoire
2. Recherche de « co-hyponymes »: exploite les définitions en français
3. Recherche des familles morphologiques: exploite les termes anglais
4. Combine termes et définitions



NÉONYMIE COMPUTATIONNELLE: RÉSULTATS PRINCIPAUX

- Pour les GLM considérés, DEF est très inférieur à TERM
- Utiliser une définition (DEF+TERM) améliore TERM
- Combiner les deux stratégies de recherche « linguistiques » fournit les meilleurs résultats (~ 35 EM)
- Le procédé morphologique utilisé est souvent correct
- La segmentation sous-lexicale obscurcit les relations morphologiques

Towards the Machine Translation of Scientific Neologisms (Lerner & Yvon, Proc. COLING 2025)

VERS LA TRADUCTION DE DOCUMENTS

Phase par phrase

I like this color.	⟹	J'aime cette couleur.
It is very elegant.	⟹	Elle est très élégante.
It represents ...	⟹	Elle représente ...

VERS LA TRADUCTION DE DOCUMENTS

Phase par phrase

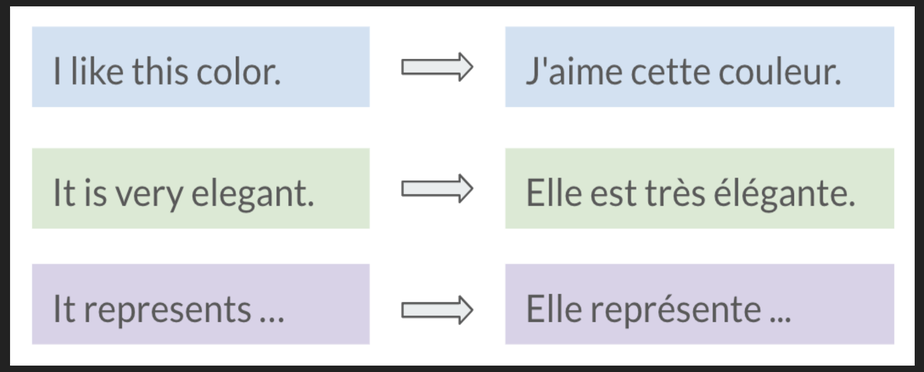
I like this color.	⇨	J'aime cette couleur.
It is very elegant.	⇨	Elle est très élégante.
It represents ...	⇨	Elle représente ...

Phrases en contexte (fenêtre = 1)

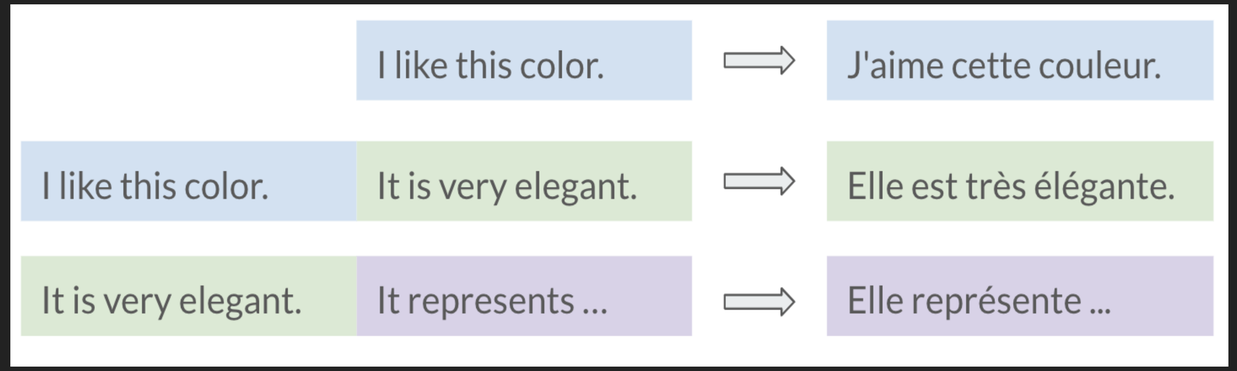
	I like this color.	⇨	J'aime cette couleur.
I like this color.	It is very elegant.	⇨	Elle est très élégante.
It is very elegant.	It represents ...	⇨	Elle représente ...

VERS LA TRADUCTION DE DOCUMENTS

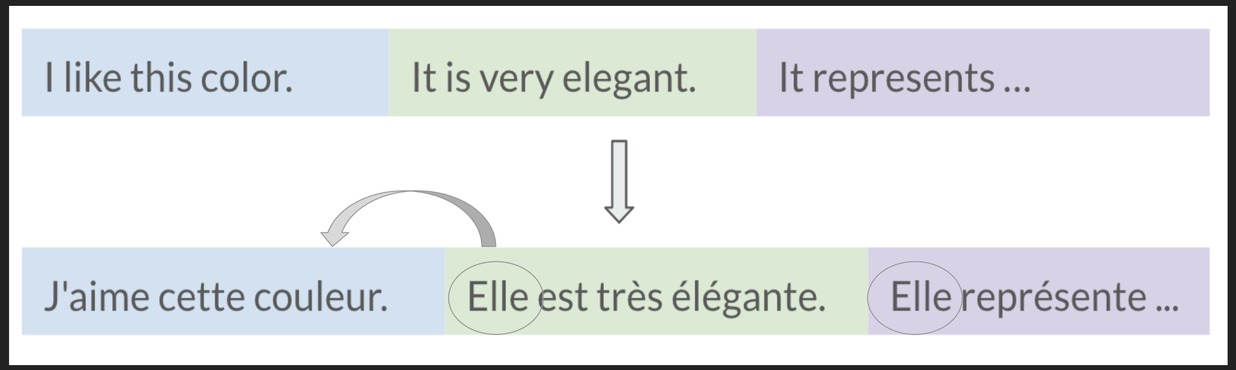
Phase par phrase



Phrases en contexte (fenêtre = 1)



« Holistique »



- Intègre toute l'information disponible
- Permet des restructurations du texte
- Meilleure cohérence ?

VERS LA TRADUCTION DE DOCUMENTS

Phase par phrase

I like this color.	⇨	J'aime cette couleur.
It is very elegant.	⇨	Elle est très élégante.
It represents …	⇨	Elle représente ...

Phrases en contexte (fenêtre = 1)

	I like this color.	⇨	J'aime cette couleur.
I like this color.	It is very elegant.	⇨	Elle est très élégante.
It is very elegant.	It represents …	⇨	Elle représente ...

« Holistique »

I like this color.	It is very elegant.	It represents …
J'aime cette couleur.	Elle est très élégante.	Elle représente ...

- Intègre toute l'information disponible
- Permet des restructurations du texte
- Meilleure cohérence ?

Nouveaux défis :

- Information éparse
- Dilution de l'attention
- Complexité des calculs
- Impact des erreurs de recherche

LA TRADUCTION HOLISTIQUE EST SOUS-OPTIMALE

- Systèmes de base
  - NLLB, TA multlingue
  - TowerBase, LLM multilingue
- Allonger la longueur des segments dégrade fortement les performances de traduction automatique

	l_{max}	2014	2015	2016	2017
NLLB	sent	45.1 (0.97)	43.9 (0.98)	41.7 (1.00)	41.8 (1.00)
	256	33.9 (0.82)	35.4 (0.84)	33.3 (0.86)	33.5 (0.87)
	512	14.6 (0.44)	16.0 (0.56)	15.2 (0.52)	13.8 (0.49)
	768	7.3 (0.27)	7.9 (0.32)	10.0 (0.46)	6.7 (0.27)
	1024	8.8 (0.56)	7.4 (0.51)	7.5 (0.50)	6.5 (0.48)
TOWERBASE	sent	43.4 (0.98)	42.9 (0.99)	39.7 (1.00)	38.7 (1.00)
	256	44.0 (0.96)	42.8 (0.98)	40.9 (1.00)	39.4 (1.00)
	512	42.9 (0.96)	39.8 (0.98)	39.9 (1.00)	40.6 (1.00)
	768	39.6 (0.98)	39.0 (0.97)	38.1 (0.99)	39.9 (1.00)
	1024	38.5 (0.98)	33.1 (0.99)	35.4 (1.00)	35.4 (0.98)
	1200	37.4 (0.92)	35.5 (0.98)	36.2 (1.00)	35.6 (0.98)
	1600	33.3 (0.96)	34.9 (0.96)	26.7 (0.94)	31.0 (0.97)
	2048	24.0 (0.97)	27.7 (0.95)	27.2 (0.96)	23.5 (0.87)

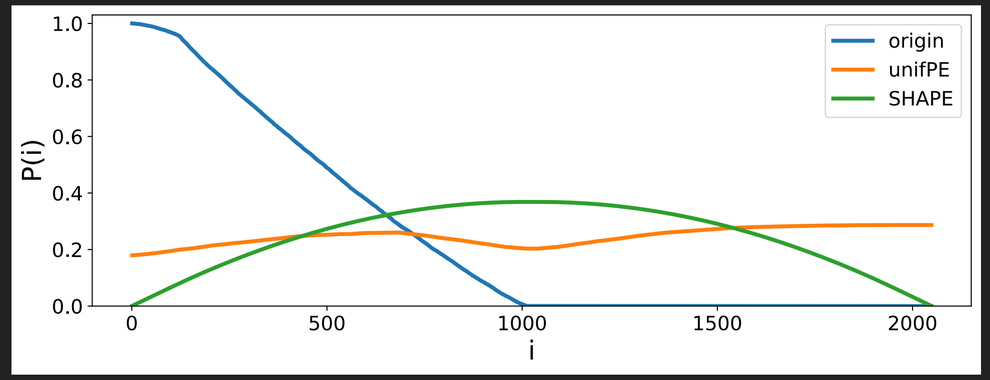
Scores BLEU, jeux de tests IWSLT, segmentations variables

AMÉLIORER LA TRADUCTION HOLISTIQUE

Adaptations:

A. Affinage avec des documents de taille variable

B. Manipulation des plongements positionnels

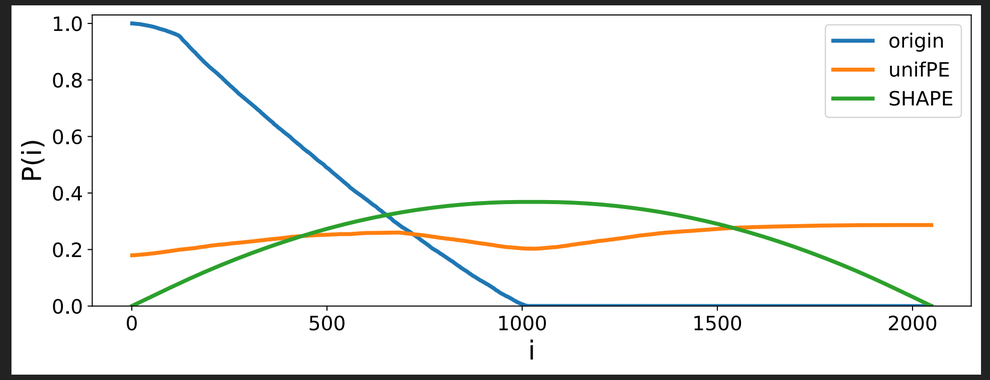


Investigating Length Issues in Document-Level Machine Translation (Peng et al., Proc.MT-Summit 2025)

AMÉLIORER LA TRADUCTION HOLISTIQUE

Adaptations:

A. Affinage avec des documents de taille variable
B. Manipulation des plongements positionnels



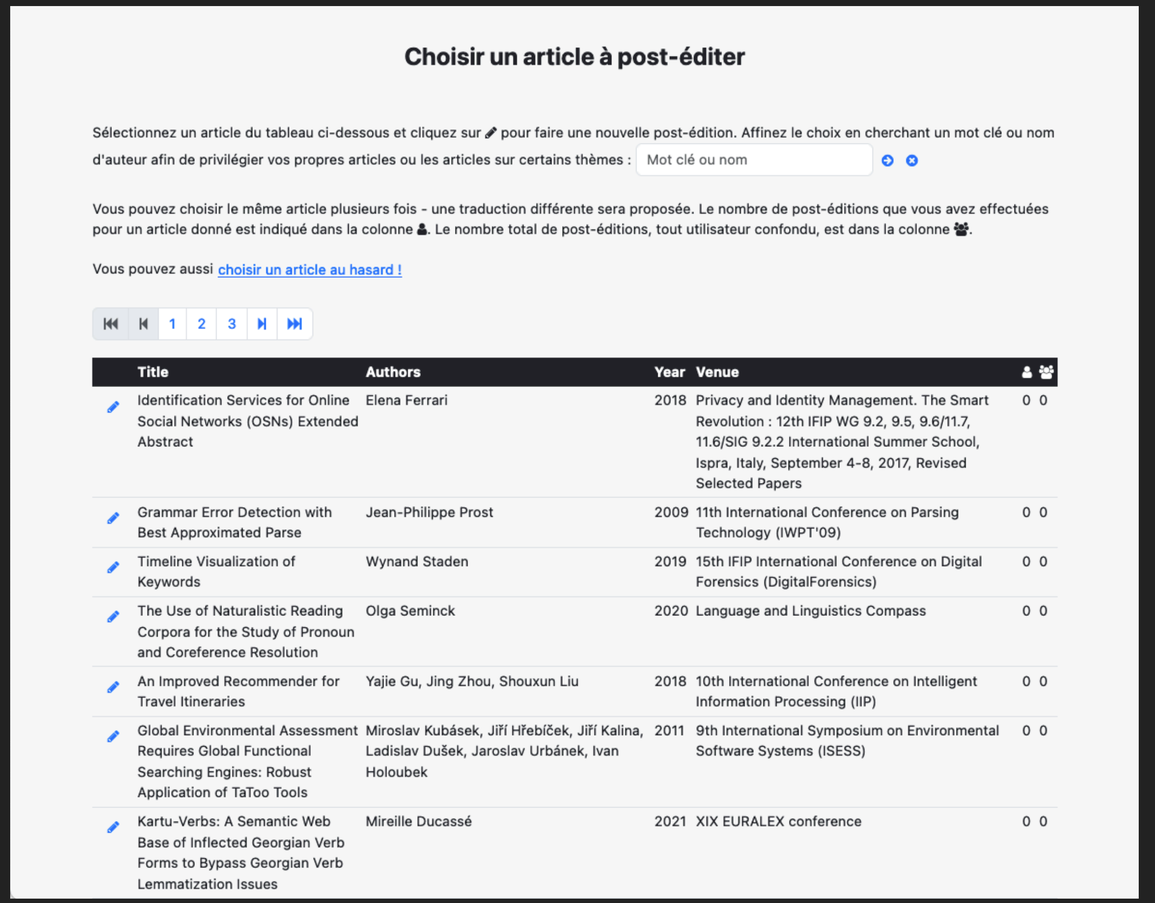
Conclusions (provisoires)

1. (A) et (B) améliorent les systèmes à base de PP
2. Les problèmes de longueur subsistent
3. La traduction de phrases en contexte donne les meilleurs résultats

Investigating Length Issues in Document-Level Machine Translation (Peng et al., Proc.MT-Summit 2025)

POST-ÉDITION PARTICIPATIVE

- Evaluer la TA en conditions réelles
  - Mesurer la qualité de la TA
  - Acceptabilité de la PE
  - Identifier les principales erreurs
  - Collecter des données de référence (termes & corpus)



Choisir un article à post-éditer

Sélectionnez un article du tableau ci-dessous et cliquez sur ✎ pour faire une nouvelle post-édition. Affinez le choix en cherchant un mot clé ou nom d'auteur afin de privilégier vos propres articles ou les articles sur certains thèmes : Mot clé ou nom

Vous pouvez choisir le même article plusieurs fois - une traduction différente sera proposée. Le nombre de post-éditions que vous avez effectuées pour un article donné est indiqué dans la colonne 👤. Le nombre total de post-éditions, tout utilisateur confondu, est dans la colonne 🐾.

Vous pouvez aussi choisir un article au hasard !

⏮ ⏪ 1 2 3 ⏩ ⏭

	Title	Authors	Year	Venue	👤	🐾
✎	Identification Services for Online Social Networks (OSNs) Extended Abstract	Elena Ferrari	2018	Privacy and Identity Management. The Smart Revolution : 12th IFIP WG 9.2, 9.5, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Ispra, Italy, September 4-8, 2017, Revised Selected Papers	0	0
✎	Grammar Error Detection with Best Approximated Parse	Jean-Philippe Prost	2009	11th International Conference on Parsing Technology (IWPT'09)	0	0
✎	Timeline Visualization of Keywords	Wynand Staden	2019	15th IFIP International Conference on Digital Forensics (DigitalForensics)	0	0
✎	The Use of Naturalistic Reading Corpora for the Study of Pronoun and Coreference Resolution	Olga Seminck	2020	Language and Linguistics Compass	0	0
✎	An Improved Recommender for Travel Itineraries	Yajie Gu, Jing Zhou, Shouxun Liu	2018	10th International Conference on Intelligent Information Processing (IIP)	0	0
✎	Global Environmental Assessment Requires Global Functional Searching Engines: Robust Application of TaToo Tools	Miroslav Kubásek, Jiří Hřebíček, Jiří Kalina, Ladislav Dušek, Jaroslav Urbánek, Ivan Holoubek	2011	9th International Symposium on Environmental Software Systems (ISESS)	0	0
✎	Kartu-Verbs: A Semantic Web Base of Inflected Georgian Verb Forms to Bypass Georgian Verb Lemmatization Issues	Mireille Ducassé	2021	XIX EURALEX conference	0	0

POST-ÉDITION PARTICIPATIVE

- Evaluer la TA en conditions réelles
  - Mesurer la qualité de la TA
  - Acceptabilité de la PE
  - Identifier les principales erreurs
  - Collecter des données de référence (termes & corpus)

Post-éditez la traduction d'un titre et d'un résumé dans le domaine du TAL

Instructions :

Modifiez le texte (titre et résumé) pour qu'il soit clair, compréhensible et acceptable, comme vous le feriez pour une publication dans un journal en français (p. ex. la revue TAL). Pour ce faire, merci de ne pas vous servir d'outils de traduction automatique. Dans la mesure du possible, merci de faire cette révision sans vous interrompre pour que la durée enregistrée corresponde au temps effectif de post-édition.

Attention : Si vous quittez cette page (en fermant la fenêtre ou en revenant sur la page précédente, vous perdrez les modifications apportées).

Titre :	**A Comparison of Various Methods for Concept Tagging for Spoken Language Understanding**
Publié dans :	LREC
Auteurs :	Stefan Hahn, Patrick Lehnen, Christian Raymond, Hermann Ney
Année :	2008
ID Hal :	1321122

Résumé d'origine :

A Comparison of Various Methods for Concept Tagging for Spoken Language Understanding

The extraction of flat concepts out of a given word sequence is usually one of the first steps in building a spoken language understanding (SLU) or dialogue system. This paper explores five different modelling approaches for this task and presents results on a French state-of-the-art corpus, MEDIA. Additionally, two log-linear modelling approaches could be further improved by adding morphologic knowledge. This paper goes beyond what has been reported in the literature, e.g. in (Raymond & Riccardi 07). We applied the models on the same training and testing data and used the NIST scoring toolkit to evaluate the experimental results to ensure identical conditions for each of the experiments and the comparability of the results. Using a model based on conditional random fields, we achieve a concept error rate of 11.8% on the MEDIA evaluation corpus.

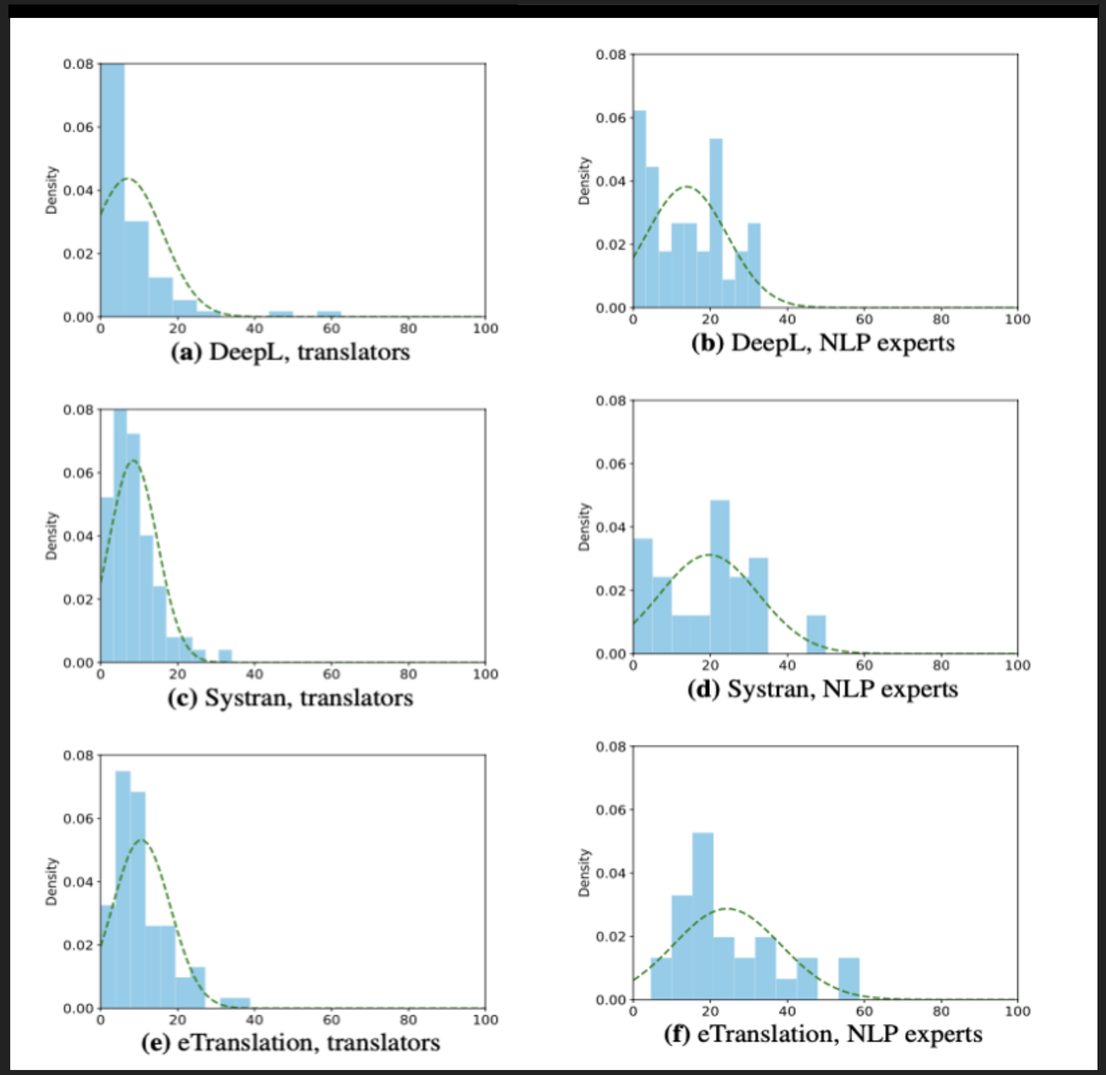
Traduction à post-éditer :

Comparaison de diverses méthodes d'étiquetage de concepts pour la compréhension du langage parlé

L'extraction de concepts plats à partir d'une séquence de mots donnée est généralement l'une des premières étapes de la construction d'un système de compréhension du langage parlé (SLU) ou d'un système de dialogue. Cet article explore cinq approches de modélisation différentes pour cette tâche et présente des résultats sur un corpus français de pointe, MEDIA. En outre, deux approches de modélisation log-linéaire pourraient être améliorées en ajoutant des connaissances morphologiques. Cet article va au-delà de ce qui a été rapporté dans la littérature, par exemple dans (Raymond & Riccardi 07). Nous avons appliqué les modèles sur les mêmes données de formation et de test et utilisé la boîte à outils de notation du NIST pour évaluer les résultats expérimentaux afin de garantir des conditions identiques pour chacune des expériences et la comparabilité des résultats. En utilisant un modèle basé sur des champs aléatoires conditionnels, nous obtenons un taux d'erreur de concept de 11,8% sur le corpus d'évaluation MEDIA.

COLLECTE 2023: RÉSULTATS

- Conclusions principales
  - Les outils de TA sont performants
  - Post-éditer prend peu de temps
  - Différences de stratégie PE « spécialistes » et « traducteurs »
  - ~350 résumés traduits et post-édités
  - Principales source d'erreur: les termes

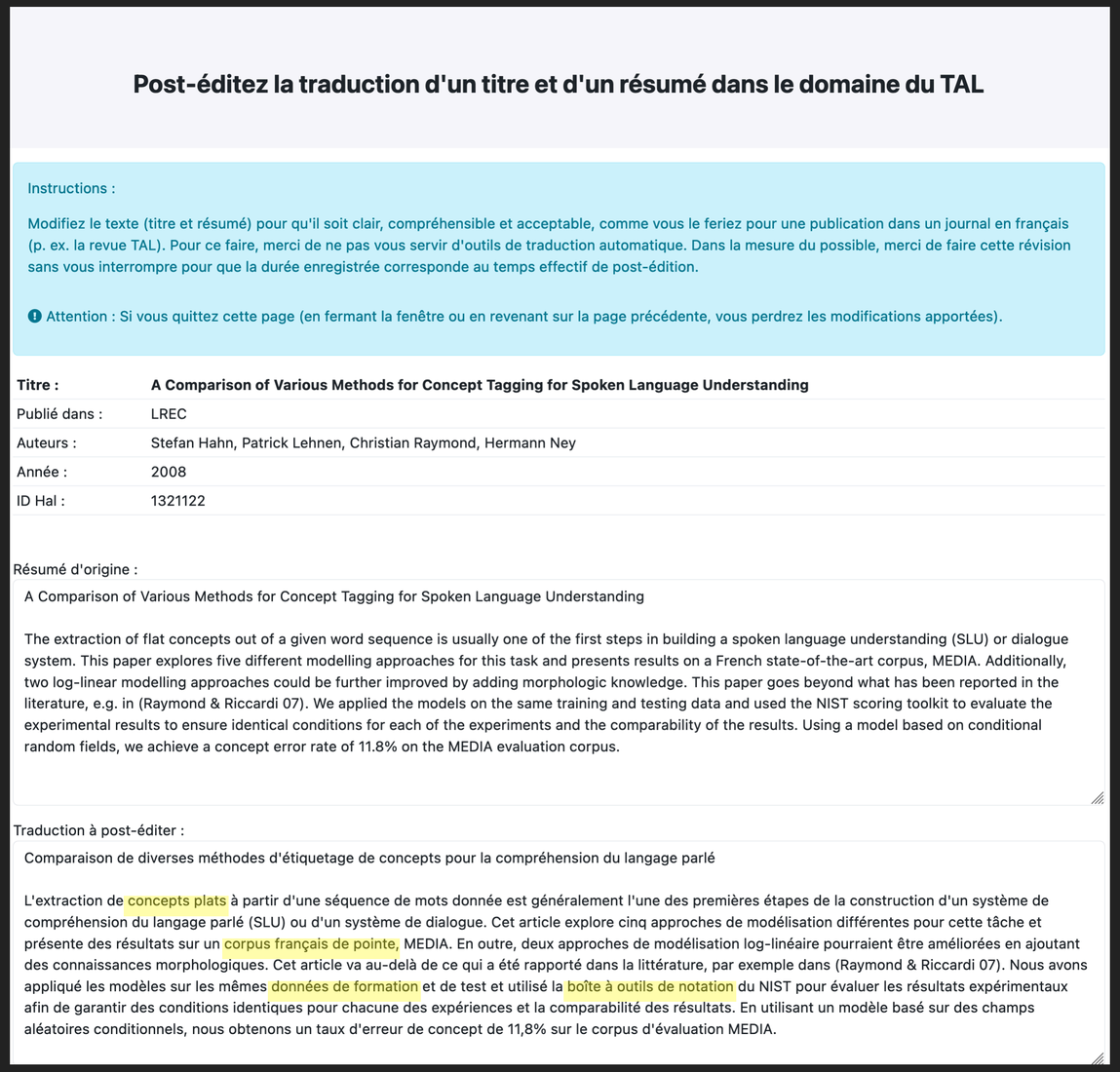


Translate your Own: a Post-Editing Experiment in the NLP domain (Bawden et al., Proc.EAMT 2024)

NOUVELLE COLLECTE 2025

https://postedition.anr-matos.fr/tal

- Objectifs
  - Evaluer la TA à base de GLM
  - Variantes de l’interface de PE
  - Intérêt de la pré-annotation de zones d’erreurs
  - Collecte de nouvelles données



Post-éditez la traduction d'un titre et d'un résumé dans le domaine du TAL

Instructions :

Modifiez le texte (titre et résumé) pour qu'il soit clair, compréhensible et acceptable, comme vous le feriez pour une publication dans un journal en français (p. ex. la revue TAL). Pour ce faire, merci de ne pas vous servir d'outils de traduction automatique. Dans la mesure du possible, merci de faire cette révision sans vous interrompre pour que la durée enregistrée corresponde au temps effectif de post-édition.

Attention : Si vous quittez cette page (en fermant la fenêtre ou en revenant sur la page précédente, vous perdrez les modifications apportées).

Titre :	**A Comparison of Various Methods for Concept Tagging for Spoken Language Understanding**
Publié dans :	LREC
Auteurs :	Stefan Hahn, Patrick Lehnen, Christian Raymond, Hermann Ney
Année :	2008
ID Hal :	1321122

Résumé d'origine :

A Comparison of Various Methods for Concept Tagging for Spoken Language Understanding

The extraction of flat concepts out of a given word sequence is usually one of the first steps in building a spoken language understanding (SLU) or dialogue system. This paper explores five different modelling approaches for this task and presents results on a French state-of-the-art corpus, MEDIA. Additionally, two log-linear modelling approaches could be further improved by adding morphologic knowledge. This paper goes beyond what has been reported in the literature, e.g. in (Raymond & Riccardi 07). We applied the models on the same training and testing data and used the NIST scoring toolkit to evaluate the experimental results to ensure identical conditions for each of the experiments and the comparability of the results. Using a model based on conditional random fields, we achieve a concept error rate of 11.8% on the MEDIA evaluation corpus.

Traduction à post-éditer :

Comparaison de diverses méthodes d'étiquetage de concepts pour la compréhension du langage parlé

L'extraction de concepts plats à partir d'une séquence de mots donnée est généralement l'une des premières étapes de la construction d'un système de compréhension du langage parlé (SLU) ou d'un système de dialogue. Cet article explore cinq approches de modélisation différentes pour cette tâche et présente des résultats sur un corpus français de pointe, MEDIA. En outre, deux approches de modélisation log-linéaire pourraient être améliorées en ajoutant des connaissances morphologiques. Cet article va au-delà de ce qui a été rapporté dans la littérature, par exemple dans (Raymond & Riccardi 07). Nous avons appliqué les modèles sur les mêmes données de formation et de test et utilisé la boîte à outils de notation du NIST pour évaluer les résultats expérimentaux afin de garantir des conditions identiques pour chacune des expériences et la comparabilité des résultats. En utilisant un modèle basé sur des champs aléatoires conditionnels, nous obtenons un taux d'erreur de concept de 11,8% sur le corpus d'évaluation MEDIA.

NOUVELLE COLLECTE 2025

https://postedition.anr-matos.fr/tal

- Objectifs
  - Evaluer la TA à base de GLM
  - Variantes de l'interface de PE
  - Intérêt de la pré-annotation d'erreurs
  - Collecte de nouvelles donné

DÉVELOPPEMENTS EN COURS OU À VENIR

- Ressources
  - Terminologie STEP
  - Poursuite de l’acquisition d’articles traduits
  - Traitements automatiques de HAL

- Annotation, évaluation et modélisation de la variation terminologique

- Traduction Automatique
  - Traduire avec RAG (intra et inter doc)
  - Traduire avec des termes
  - Intégration de la structure

- Evaluation de la TA
  - Evaluer avec des GLM
  - Evaluer la traduction de termes
  - Méta-évaluation de la traduction niveau document
  - Post édition: passage à l’échelle

POUR EN SAVOIR PLUS

https://anr-matos.github.io