

REGISTRE FIAS DES SPÉCIFICATIONS DE RECHERCHE

Notation compacte en couches des auto-réparations et marqueurs discursifs du français oral aligné

IDENTIFIANT DU REGISTRE

FOAR-CON2X-3CDR9PBR6

VERSION

0.9

TYPE D'OBJET

Système d'encodage, d'écriture, de translittération ou de notation

DOMAINE

Linguistique et philologie

MATURITÉ DU DOCUMENT

Candidat pour un usage stable

SOUMISE

29 août 2014, 12:00 UTC

PUBLIÉE

20 octobre 2014, 12:00 UTC

LICENCE

Creative Commons Attribution 4.0

VERSION FAISANT AUTORITÉ

L'enregistrement consigne le résultat de l'examen du Registre pour cette version : la soumission a été suffisamment identifiée, documentée et déterminée en interne pour la procédure indiquée. La portée, l'adoption et les conventions concurrentes sont documentées séparément lorsqu'il y a lieu.

STATUT AU REGISTRE

Acceptée pour enregistrement

20 octobre 2014, 12:00 UTC

Description principale

TYPE DE SOUMISSION

Révision d'une soumission non enregistrée

SOUMISSION ANTÉRIEURE OU IDENTIFIANT DU REGISTRE

FIAS-SUB-LT-001-A

TITRE COMPLET

Notation compacte en couches des auto-réparations et marqueurs discursifs du français oral aligné

TITRE COURT OU ACRONYME

NCL-RMDF

TYPE D'OBJET

Système d'encodage, d'écriture, de translittération ou de notation

STATUT REVENDIQUÉ PAR LE DÉPOSANT

Spécification

MATURITÉ DE CETTE VERSION

Candidat pour un usage stable

VERSION

0.9

DOMAINE PRINCIPAL

Linguistique et philologie

MOTS-CLÉS

- français oral
- auto-réparation
- disfluence
- marqueur discursif
- annotation en couches
- Rhapsodie

PROBLÈME, PORTÉE ET EXCLUSIONS

Cette révision répond aux objections faites à la version 0.7. Elle décrit une notation légère mais composée de trois couches indépendantes : forme transcrite, structure de réparation et fonction discursive. Elle couvre les pauses remplies, amorces, répétitions, substitutions, insertions, interrègles et marqueurs discursifs sur des tours de parole français déjà alignés au signal. Son périmètre exclut segmentation prosodique générale, syntaxe de dépendance, reconnaissance automatique et interprétation psychologique de l'intention du locuteur.

UTILISATEURS ET CONTEXTES VISÉS

- Annotateurs de petits corpus de français oral spontané
- Ateliers universitaires comparant annotation manuelle et DisMo
- Développeurs exportant des annotations vers XML ou tableau relationnel

RELATION AVEC LA PRATIQUE EXISTANTE

Étend une norme ou convention existante

CONTEXTE SCIENTIFIQUE, SOURCES ET ÉCHANGES SCIENTIFIQUES

Cette section décrit les fondements scientifiques, les travaux sous-jacents et les sources de la spécification, ainsi que les groupes de recherche et les cadres scientifiques dans lesquels la solution proposée ou sa formalisation a été présentée ou discutée. Elle documente le contexte et la nature de ces échanges scientifiques.

L'annotation du français oral doit pouvoir représenter plusieurs analyses d'un même segment : sa forme transcrite, sa place dans une réparation et sa fonction discursive. Les travaux de K. Gerdes et al. (2012), puis d'A. Lacheret et al. (2014), organisent ces relations dans les couches alignées de Rhapsodie. J. Belião et A. Lacheret (2013) étudient les rapports entre disfluence prosodique et marqueurs discursifs ; DisMo, présenté par G. Christodoulides, M. Avanzi et J.-P. Goldman (2014), propose également des niveaux distincts pour ces phénomènes.

Ces résultats ont circulé dans le cadre du projet ANR Rhapsodie et des communications à DiSS 2013 et LREC 2014. Les travaux associent notamment MoDyCo, LPP, IRCAM-STMS et LLL, à l'interface de la prosodie, de la syntaxe et de l'ingénierie des corpus. La possibilité de comparer les annotations entre ces approches motive une représentation d'échange dont les unités et les relations soient explicites.

La version 0.9 de NCL-RMDF reprend cette organisation en trois couches et utilise des annotations stand-off identifiées. Les pauses remplies, amorces, répétitions, substitutions, insertions et interrègles peuvent être reliés aux marqueurs discursifs et aux tours de parole, y compris dans les cas de recouvrement ou d'imbrication. La méthodologie de R. Artstein et M. Poesio (2008) conduit à examiner séparément l'accord sur les segments et l'accord sur les catégories. La notation fournit ainsi une base pour la double annotation et la discussion des désaccords dans les groupes de travail sur corpus oraux.

Références:

- Gerdes et al. (2012), Intonosyntactic Data Structures, Texte: <https://aclanthology.org/W12-3611/>
- Belião & Lacheret (2013), Disfluency and discursive markers, Texte: https://www.isca-archive.org/diss_2013/beliao13_diss.html
- Lacheret et al. (2014), Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French, Texte: <https://aclanthology.org/L14-1329/>
- Christodoulides, Avanzi & Goldman (2014), DisMo, Texte: <https://aclanthology.org/L14-1707/>
- Artstein & Poesio (2008), Inter-Coder Agreement, Texte: <https://aclanthology.org/J08-4004/>

STATUT DE NORMALISATION EXTERNE

Aucune prévue

LICENCE DE LA SPÉCIFICATION SOUMISE

Creative Commons Attribution 4.0

DÉCLARATION DE PUBLICATION

Oui

Système d'encodage, d'écriture, de translittération ou de notation

SYSTÈME SOURCE OU RÉPERTOIRE

Transcriptions UTF-8 de français parlé disposant d'identifiants stables de locuteur, tour et token et, si possible, de temps début/fin. Les objets sources sont pause remplie, amorce, répétition, reparandum, interrègne, réparation, insertion, marqueur discursif et chevauchement de tours.

SYSTÈME CIBLE OU RÉPERTOIRE

Trois tables stand-off UTF-8 partageant les identifiants de tokens : FORM pour les événements de surface, REPAIR pour les relations reparandum-interrègne-réparation, et DISCOURSE pour la fonction DM. Une vue imprimable peut employer les symboles de 0.7, mais les tables sont normatives.

UNITÉ DE CORRESPONDANCE

Unités mixtes

RÈGLES OU TABLEAU DE CORRESPONDANCE

Source | Cible | Condition | Règle | Note

Pause remplie | FORM(type=filled_pause,target=tN) | diagnostic acoustique/transcriptif | E1 | cumulable avec DISCOURSE

Amorce | FORM(type=cutoff,target=tN,char_end=k) | interruption intra-token | E2 | index de caractère obligatoire

Répétition | REPAIR(type=repetition,old=tA..tB,new=tC..tD) | matériel repris | E3 | pas d'interrègne obligatoire

Substitution | REPAIR(type=substitution,old=...,new=...) | matériel remplacé | E4 | interregnum facultatif

Insertion | REPAIR(type=insertion,old=∅,new=...) | ajout au redémarrage | E5 | zéro explicite

Interrègne | REPAIR(interregnum=tX..tY) | entre old et new | E6 | span séparé de old et new

Marqueur discursif | DISCOURSE(type=dm,target=...) | fonction discursive justifiée | E7 | couche indépendante

Chevauchement | FORM(type=overlap,target=...,with=speaker-turn-id) | paroles simultanées | E8 | couche séparée de REPAIR

RÉVERSIBILITÉ

Entièrement réversible

POLITIQUE D'AMBIGUÏTÉ ET DE COLLISIONS

Chaque annotation porte un identifiant et vise des identifiants de tokens; plusieurs couches peuvent viser le même token. Deux analyses incompatibles sont conservées comme alternatives avec `certainty` et `ana-set`, jamais fusionnées. Les réparations imbriquées sont permises si leurs identifiants diffèrent. Un span discontinu énumère ses tokens. L'étiquette UNC exige la documentation préalable des alternatives.

RÉPERTOIRE UNICODE ET NORMALISATION

UTF-8 et NFC sont obligatoires. Les valeurs de catégories et noms de colonnes restent en ASCII. Le texte français conserve ses diacritiques. Les symboles [], → et ~ servent au rendu; les exports restent indépendants de leur disponibilité. Une occurrence littérale de ces symboles dans la transcription est échappée dans la vue par une barre oblique inverse, sans modifier le token source.

RÈGLES CONTEXTUELLES

Les frontières de mots et la ponctuation viennent du corpus source. Une réparation reste dans le tour du même locuteur; une reprise par un autre locuteur est une relation discursive distincte. Casse, noms propres et graphies historiques conservent leur forme source. Le statut DM est assigné par fonction en contexte et peut coexister avec FP. Les chevauchements utilisent les temps du corpus, jamais la position typographique seule.

CAS DE TEST

Entrée | Sortie attendue | Règle

t1=je t2=euh t3=pars | FORM f1 filled_pause target=t2 | E1

t1=une t2=mai- t3=maison | FORM f2 cutoff target=t2 char_end=3 | E2

t1=je t2=je t3=pars | REPAIR r1 repetition old=t1 new=t2 | E3

t1=lundi t2=enfin t3=mardi | REPAIR r2 substitution old=t1 interregnum=t2 new=t3 | E4+E6

t1=bon t2=on t3=commence | DISCOURSE d1 dm target=t1 | E7

euh à la fois pause et DM | FORM f3 target=t1 + DISCOURSE d2 target=t1 | E1+E7

réparation imbriquée | deux REPAIR avec identifiants et spans indépendants | politique d'ambiguïté

Historique d'examen du Registre

Soumission reçue

29 août 2014, 12:00 UTC

Soumission reçue.

Acceptée pour enregistrement

20 octobre 2014, 12:00 UTC

Les motifs de révision de 2012 sont traités : les fonctions prosodiques et discursives sont indépendantes, l'interrègne est explicite, les réparations imbriquées restent représentables et la vue Unicode devient un rendu non normatif. Les cas tests sont déterministes. La proposition est acceptée pour enregistrement sous réserve de livrer, avant publication finale, un jeu d'essai à double annotation et un tableau séparant accord de détection des spans et accord de catégorie.