

FIAS REGISTRY OF RESEARCH SPECIFICATIONS

Compact Layered Notation for Repair and Discourse Events in Time-Aligned Spoken French

REGISTRY IDENTIFIER

FOAR-CON2X-3CDR9PBR6

VERSION

1.0

OBJECT TYPE

Encoding, writing, transliteration, or notation scheme

SUBJECT AREA

Linguistics and philology

DOCUMENT MATURITY

Stable

SUBMITTED

16 April 2018, 12:00 UTC

PUBLISHED

11 June 2018, 12:00 UTC

LICENCE

Creative Commons Attribution 4.0

AUTHORITATIVE VERSION

Registration records the outcome of the Registry review for this version: the submission was sufficiently identified, documented, and internally determinate for the stated procedure. Scope, adoption, and competing conventions are documented separately where relevant.

REGISTRY STATUS**Registered**

11 June 2018, 12:00 UTC

Core description

SUBMISSION KIND

Revision of an unregistered submission

EARLIER SUBMISSION OR REGISTRY IDENTIFIER

FIAS-SUB-LT-001-B

FULL TITLE

Compact Layered Notation for Repair and Discourse Events in Time-Aligned Spoken French

SHORT TITLE OR ACRONYM

CLN-RDE

OBJECT TYPE

Encoding, writing, transliteration, or notation scheme

STATUS CLAIMED BY THE SUBMITTER

Project standard

MATURITY OF THIS VERSION

Stable

VERSION

1.0

PRINCIPAL SUBJECT AREA

Linguistics and philology

KEYWORDS

- spoken French
- disfluency annotation
- repair structure
- discourse markers
- stand-off annotation
- inter-annotator agreement

PROBLEM, SCOPE, AND EXCLUSIONS

Version 1.0 freezes a small interoperable annotation for observable repair and discourse events in already transcribed, time-aligned spoken French. It represents surface events,

structured repairs, interregna, discourse-marker functions, speaker overlap, uncertainty and alternative analyses without modifying the transcript. It is suitable for manual annotation, conversion to XML or relational tables, and evaluation of an automatic pre-annotator. It excludes full prosodic phrasing, dependency syntax, dialogue acts, clinical fluency diagnosis, causal claims about production, and automatic-transcription accuracy.

INTENDED USERS AND CONTEXTS

- Corpus linguists annotating contemporary spoken French
- Research-methods seminars teaching reliability and adjudication
- NLP developers evaluating rule-based or statistical disfluency pre-annotation
- Data stewards exchanging stand-off event layers without rewriting transcripts

RELATION TO EXISTING PRACTICE

Replaces an earlier specification

SCIENTIFIC CONTEXT, SOURCES, AND SCHOLARLY DISCUSSION

This section describes the scientific background, underlying work, and sources of the specification, together with the research groups and scholarly settings in which the proposed solution or its formalisation was presented or discussed. It documents the context and nature of these scholarly exchanges.

Research on spoken French treats repair, hesitation and discourse organisation as interacting levels of analysis. K. Gerdes et al. (2012) and A. Lacheret et al. (2014) develop this approach through Rhapsodie's aligned prosodic and syntactic structures. J. Belião and A. Lacheret (2013) examine the relation between disfluency and discourse markers. DisMo, presented by G. Christodoulides, M. Avanzi and J.-P. Goldman (2014), organises tokenisation, disfluency and discourse annotation in separate levels, while G. Christodoulides and M. Avanzi (2015) refine the treatment of simple and structured disfluencies.

These methods were presented and discussed through the Rhapsodie research network and publications at DiSS 2013, LREC 2014 and Interspeech 2015. The Rhapsodie work connects MoDyCo, LPP, IRCAM-STMS and LLL. This shared setting brings corpus construction, linguistic analysis and annotation tools into contact, making explicit exchange conventions important for comparing analyses across teams.

Version 1.0 consolidates a three-layer representation of observable forms, repair structure and discourse function. Stand-off annotations identify events, overlapping spans, interregna and alternative analyses while preserving the transcript. Defined offsets and deterministic examples make the representation reproducible across tools. The evaluation distinguishes span detection from category agreement, following the methodological concerns developed by R. Artstein and M. Poesio (2008). The specification therefore supports both the exchange of manually annotated data and the examination of automatic pre-annotation results.

References:

- Gerdes et al. (2012), Intonosyntactic Data Structures, Source: <https://aclanthology.org/W12-3611/>
- Belião & Lacheret (2013), Disfluency and discursive markers, Source: https://www.isca-archive.org/diss_2013/beliaio13_diss.html
- Lacheret et al. (2014), Rhapsodie, Source: <https://aclanthology.org/L14-1329/>
- Christodoulides, Avanzi & Goldman (2014), DisMo, Source: <https://aclanthology.org/L14-1707/>
- Christodoulides & Avanzi (2015), Automatic detection and annotation of disfluencies in spoken French corpora, Source: https://www.isca-archive.org/interspeech_2015/christodoulides15_interspeech.html
- Artstein & Poesio (2008), Inter-Coder Agreement, Source: <https://aclanthology.org/J08-4004/>

EXTERNAL STANDARDISATION STATUS

None intended

LICENCE FOR THE SUBMITTED SPECIFICATION

Creative Commons Attribution 4.0

PUBLICATION DECLARATION

Yes

Encoding, writing, transliteration, or notation scheme

SOURCE SYSTEM OR REPERTOIRE

A source package contains (1) immutable UTF-8 transcript tokens with speaker, turn and token identifiers; (2) optional millisecond start/end times; and (3) optional audio. Candidate observable events are filled pauses, intra-token cut-offs, lengthenings, repetitions, deletions, substitutions, insertions and speaker overlaps. Functional annotations are discourse-marker uses and unresolved alternatives.

TARGET SYSTEM OR REPERTOIRE

Canonical stand-off annotations serialised as UTF-8 TSV or equivalent XML. FORM rows encode one observable event; REPAIR rows link ordered reparandum, interregnum and repair token lists; DISCOURSE rows attach a functional label independently. Every row has annotation_id, layer, type, targets, annotator_id, certainty and revision.

UNIT OF MAPPING

Mixed units

MAPPING RULES OR TABLE

Source | Canonical target | Condition | Rule | Constraint

Filled pause | FORM:filled_pause targets=[t] | audible conventional hesitation | E1 | token retained

Cut-off | FORM:cutoff targets=[t] offset=n | audible interruption inside token | E2 | Unicode-scalar offset

Lengthening | FORM:lengthening targets=[t] | duration judged non-lexical | E3 | time required

Repetition | REPAIR:repetition old=[...] new=[...] | same material restarted | E4 | lists nonempty

Deletion | REPAIR:deletion old=[...] new=[] | reparandum abandoned | E5 | empty new permitted

Substitution | REPAIR:substitution old=[...] new=[...] | replacement differs | E6 | lists nonempty

Insertion | REPAIR:insertion old=[] new=[...] | new material inserted at restart | E7 | anchor required

Interregnum | REPAIR.interregnum=[...] | material between old and new | E8 | may be empty

Discourse marker | DISCOURSE:dm targets=[...] | discourse-organising function | E9 | independent of FORM

Speaker overlap | FORM:overlap targets=[...] with_turn=turn-id | times overlap | E10 | both turns identified

Alternative analysis | ALT members=[annotation-id...] | evidence underdetermines one analysis | E11 | no member privileged without adjudication

REVERSIBILITY

Fully reversible

AMBIGUITY AND COLLISION POLICY

Stand-off rows never delete, reorder or replace source tokens. Multiple layers may target the same token. Mutually exclusive analyses are members of one ALT set and carry certainty values from {certain, probable, possible}. Nested repairs are legal when each has a distinct identifier; crossing repair spans are legal only when the annotator supplies a note. Adjudication creates a new revision and retains the pre-adjudication rows. Unknown event type is encoded FORM:unclassified, not guessed.

UNICODE REPERTOIRE AND NORMALISATION

UTF-8 and NFC are mandatory for transcript strings and notes. Identifiers, layer names and controlled values use ASCII. Character offsets count Unicode scalar values after NFC, never bytes or display graphemes. Canonical storage uses stand-off offsets rather than bracket markup. The optional reading view renders U+27E6/U+27E7, U+2192 and U+2301 but escapes literal occurrences; lack of a display glyph has no effect on canonical data.

CONTEXTUAL RULES

Source orthography, case, accents, punctuation, proper names and historical spellings are immutable. Token boundaries follow the source package. A structured repair is speaker-internal and stays within a turn; collaborative completion is outside scope. A DM judgement must cite its contextual function and may coexist with a filled-pause annotation. Overlap requires temporal evidence. Lengthening requires audio and is recorded only when that evidence survives.

TEST CASES

Input | Expected canonical rows | Rule

je euh pars | FORM f1 filled_pause targets=t2 | E1

une mai- maison | FORM f2 cutoff targets=t2 offset=3 | E2

je je pars | REPAIR r1 repetition old=t1 new=t2 interregnum=[] | E4

lundi enfin mardi | REPAIR r2 substitution old=t1 interregnum=t2 new=t3 | E6+E8

je voulais enfin je souhaite | REPAIR r3 substitution old=t1,t2 interregnum=t3 new=t4,t5 | E6+E8

bon on commence | DISCOURSE d1 dm target=t1, with contextual note | E9

euh functioning both ways | FORM f3 filled_pause + DISCOURSE d2 dm, same target | E1+E9

Nested repair | two valid REPAIR rows with distinct IDs | ambiguity policy

Competing span boundaries | ALT a1 members=r4,r5 | E11

Cut-off with UTF-8 é before interruption | offset counts Unicode scalars after NFC | Unicode policy

Registry review record

Submission received

16 April 2018, 12:00 UTC

Submission received.

Registered

11 June 2018, 12:00 UTC

Registered as a narrow project specification after documentary, technical and domain-method review. Version 1.0 makes the stand-off form normative, preserves nested and alternative analyses, defines Unicode offsets, separates observable and functional layers, and supplies deterministic conformance fixtures. Registration denotes determinacy and reproducibility, not disciplinary endorsement or adoption by Rhapsodie, DisMo, their authors, or their institutions.