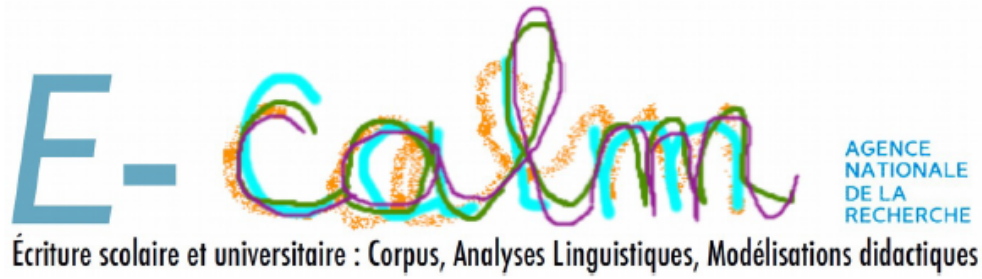


Annotation outillée de corpus scolaires et académiques

Lydia-Mai Ho-Dac

UE TAL

Plan

Plan

Le projet ANR E :Calm (2017-2022, Défi 8)

Écriture scolaire et universitaire : **C**orpus, **A**nalyses **L**inguistiques, **M**odélisations didactiques

Dresser une cartographie des compétences d'écriture depuis l'école primaire jusqu'à l'université basée sur des données attestées écologiques [Doquet et al., 2017]

Questions de départ

- Quelles compétences linguistiques les scripteurs manifestent-ils dans leurs textes et leurs brouillons selon leur âge, les caractéristiques de leur milieu et la situation d'écriture ?
- Quel rôle jouent les interventions des enseignants dans ces manifestations ?
- Comment outiller les enseignants pour qu'ils puissent adapter leurs interventions p.r. aux différents types de problèmes que pose l'acquisition de l'écriture (orthographe, genre textuel, coréhence, organisation du discours, etc.) ?

Le projet ANR E :Calm (2017-2022, Défi 8)

Dresser une cartographie des compétences d'écriture depuis l'école primaire jusqu'à l'université basée sur des données attestées écologiques [Doquet et al., 2017]

Objectifs

- Structurer et mettre à disposition de la communauté scientifique un vaste corpus d'écrits d'élèves et d'étudiants permettant des analyses quantitatives et des traitements automatiques
- Caractériser ces écrits et les attentes des enseignants du point de vue de l'acquisition de l'orthographe et de la cohérence, dans des analyses sociologiquement contextualisées
- Étudier les modalités d'écriture dans les textes et avant-textes (plans, notes, brouillons), notamment à travers l'influence réciproque des écrits remis et des interventions des enseignants sur les copies

Le projet ANR E :Calm (2017-2022, Défi 8)

Dresser une cartographie des compétences d'écriture depuis l'école primaire jusqu'à l'université basée sur des données attestées écologiques [Doquet et al., 2017]

Objectifs

- Structurer et mettre à disposition de la communauté scientifique un vaste corpus d'écrits d'élèves et d'étudiants permettant des analyses quantitatives et des traitements automatiques
- Caractériser ces écrits et les attentes des enseignants du point de vue de l'acquisition de l'orthographe et de la cohérence, dans des analyses sociologiquement contextualisées
- Étudier les modalités d'écriture dans les textes et avant-textes (plans, notes, brouillons), notamment à travers l'influence réciproque des écrits remis et des interventions des enseignants sur les copies

Le projet ANR E :Calm (2017-2022, Défi 8)

Dresser une cartographie des compétences d'écriture depuis l'école primaire jusqu'à l'université basée sur des données attestées écologiques [Doquet et al., 2017]

Objectifs

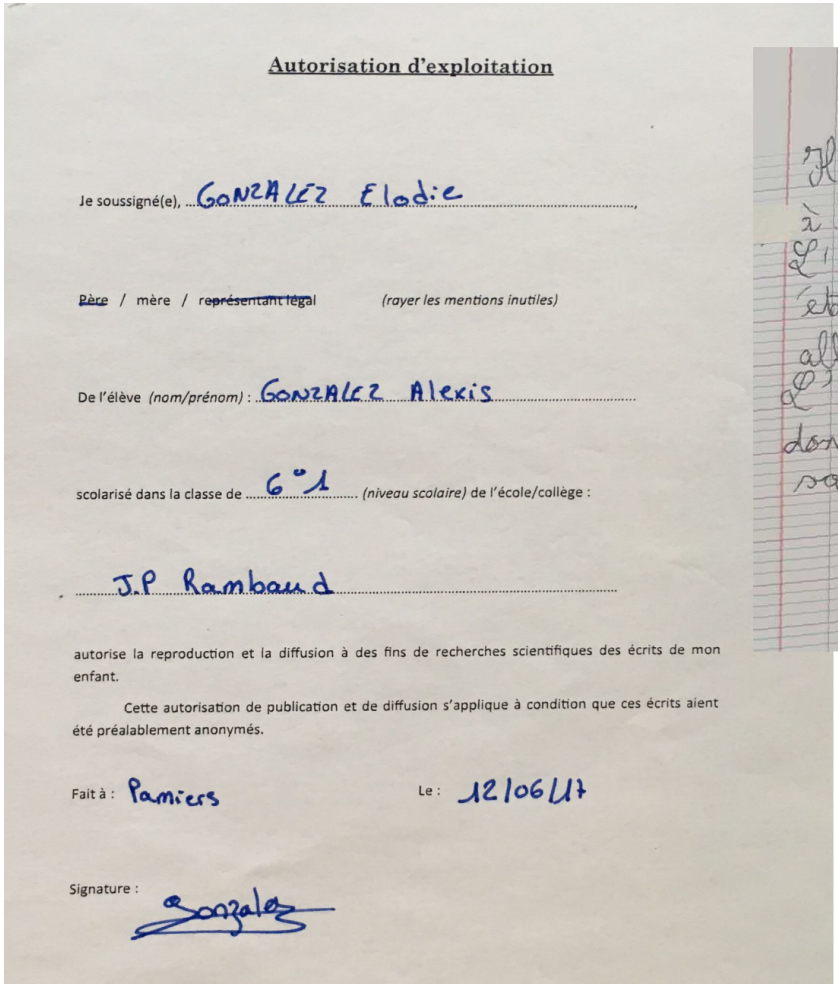
- Structurer et mettre à disposition de la communauté scientifique un vaste corpus d'écrits d'élèves et d'étudiants permettant des analyses quantitatives et des traitements automatiques → Constitution de la ressource E :CALM
- Caractériser ces écrits et les attentes des enseignants du point de vue de l'acquisition de l'orthographe et de la cohérence, dans des analyses sociologiquement contextualisées
- Étudier les modalités d'écriture dans les textes et avant-textes (plans, notes, brouillons), notamment à travers l'influence réciproque des écrits remis et des interventions des enseignants sur les copies

Processus +/− commun de constitution

1. Collecte des copies auprès d'écoles, collèges, lycées et universités
   - Autorisation de diffusion signée par les parents
   - Variété d'activités scolaires (rédaction, écriture libre, devoir, etc.)
2. Scan, rognage et anonymisation des copies
3. Encodage des méta-données (teiHeader TEI-P5)
4. Digitalisation du texte (transcription) au format XML TEI-P5
5. Vérification de la transcription XML via une visualisation html
6. Normalisation orthographique : annotation des segments de texte erronés et correction
7. Vérification de la normalisation orthographique
8. POS-tagging and parsing
9. Annotation de la cohérence et de certains indices de cohésion

Plan

Collecter les données et l'autorisation de les utiliser pour la recherche



Autorisation d'exploitation

Je soussigné(e), GONZALEZ Elodie,

~~Père~~ / mère / ~~représentant légal~~ *(rayer les mentions inutiles)*

De l'élève *(nom/prénom)* : GONZALEZ Alexis

scolarisé dans la classe de 6°1 *(niveau scolaire)* de l'école/collège :

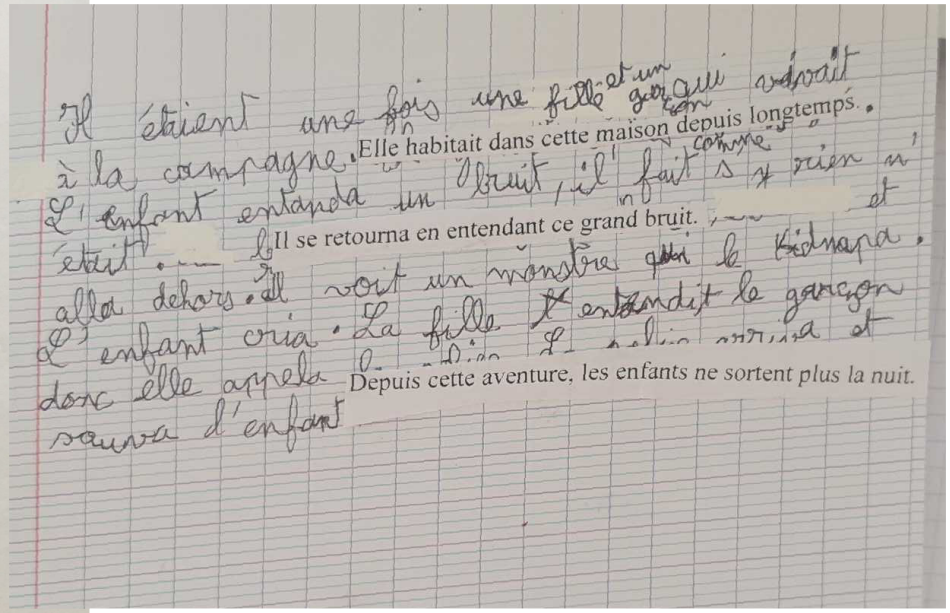
J.P Rambaud

autorise la reproduction et la diffusion à des fins de recherches scientifiques des écrits de mon enfant.

Cette autorisation de publication et de diffusion s'applique à condition que ces écrits aient été préalablement anonymés.

Fait à : Pamiers Le : 12/06/17

Signature : Gonzalez



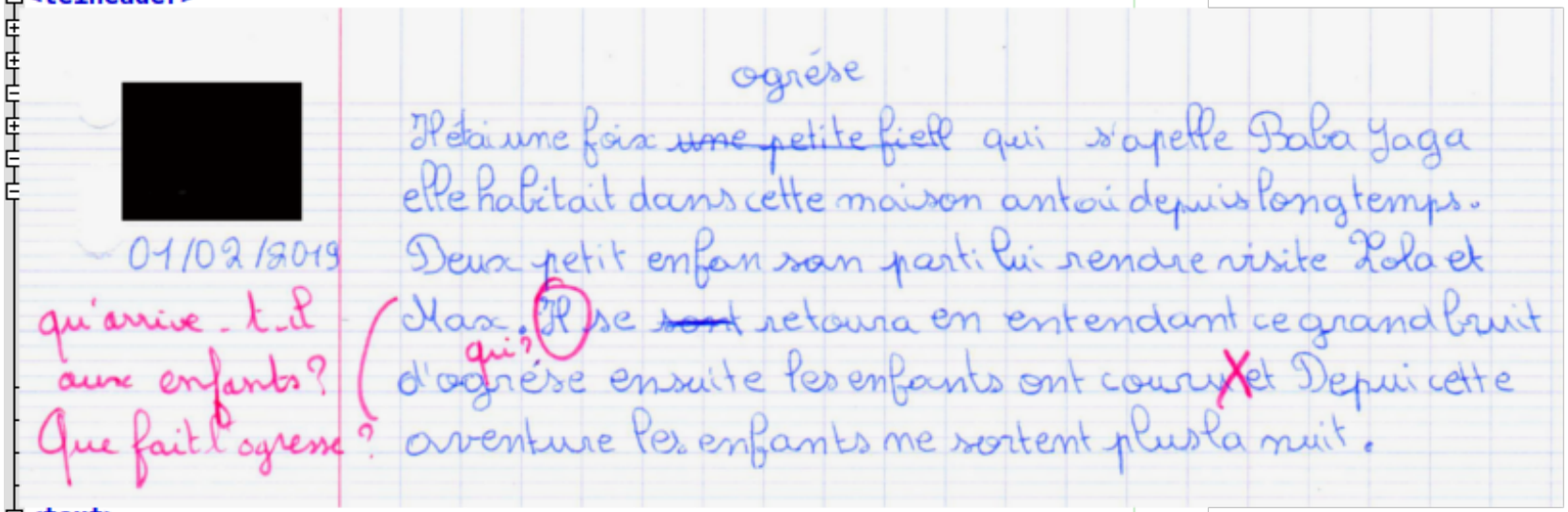
Il étaient une fois une fille et un garçon qui vivait à la campagne. Elle habitait dans cette maison depuis longtemps. L'enfant entenda un bruit, il fait comme si rien n'était. Il se retourna en entendant ce grand bruit. et alla dehors. il voit un monstre qui le kidnapa. L'enfant cria. La fille entendit le garçon donc elle appela Depuis cette aventure, les enfants ne sortent plus la nuit. sauva d'enfant

Encoder les méta-données : teiHeader

```
      <projectDesc>
        Ce fichier a été produit dans le cadre du projet É:calm : "Écriture scolaire et universitaire : Corpus, An
        Linguistiques, Modélisations didactiques" (ANR-17-CE28-0004-04) démarré en 2018 et achevé en 2022.
        http://e-calm.huma-num.fr/
      </projectDesc>
    </encodingDesc>
    <profileDesc>
      <langUsage>
        <language ident="fr">french</language>
      </langUsage>
      <settingDesc>
        <setting>
          <name type="city">Fontenay-sous-Bois (94120)</name>
          <name type="region">Île-de-France</name>
          <date>2019-01-08</date>
          <locale>Collège Joliot Curie</locale>
          <activity>Rédaction</activity>
          <p>
            <!-- indiquer ici toute information complémentaire sur la situation de production -->
          </p>
        </setting>
      </settingDesc>
      <textDesc n="schoolwork">
        <channel mode="w">manuscript</channel>
        <constitution type="single">chaque texte correspond à une copie produite en classe par un élève ou un étud
        certaines copies, des interventions de l'enseignant sont également présentes.</constitution>
        <derivation type="original"/>
        <domain type="education"/>
        <factuality type="fiction">La consigne donnée aux enfants demande de raconter une histoire sans indication
        la factualité.</factuality>
        <interaction type="none"/>
        <preparedness type="none"/>
        <purpose degree="high" type="express">Le texte résulte d'un travail réalisé en classe selon une consigne p
        demandant au locuteur de raconter une histoire.</purpose>
      </textDesc>
      <particDesc>
        <listPerson>
          <person age="na" id="R3" sex="X">élève né.e en XXXX au mois XX</person>
        </listPerson>
      </particDesc>
    </profileDesc>
  </teiHeader>
```

Transcription des traces d'écriture et des interventions

une première couche d'annotation non assistée (mais des possibilités e.g. TACT)



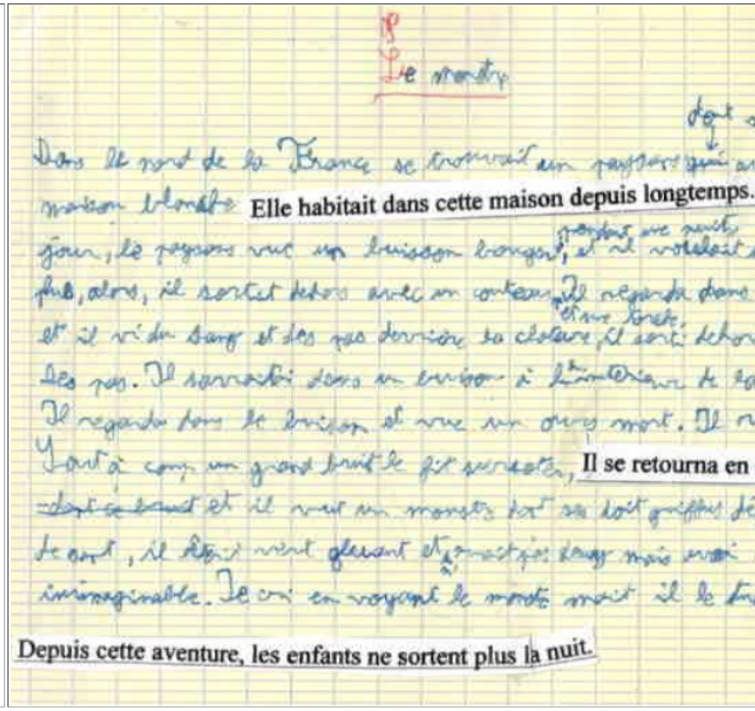

```
<?xml version='1.0' encoding='UTF-8'?>
<TEI>
<teiHeader>
<text>
<front>
  <p>Nom de l'élève</p>
  <p>01/02/2019</p>
</front>
<body>
  <p>
      Il étai une foix <mod type="subst"><del>une petite fiell</del><add>
      ogrése</add></mod> qui s'apelle BabaYaga <lb/>elle habitait dans cette
      maison antai depuis longtemps. <lb/>Deux petit enfan son parti lui
      rendre visite Lola et <lb/>Max. Il se <mod type="subst"><del>sont
      </del></mod> retoura en entendant ce grand bruit <lb/>d'ogrése ensuite
      les enfants ont couru et Depuis cette <lb/>aventure les enfants ne
      sortent plus la nuit.
  </p>
```

Vérification des transcriptions (transformation XSLT)

EC-CM2-2016-SGLEA-D1-R24-V1	Ile de France	2017-04-25	école publique en zone mixte Entre 300 et 400 élèves

Le monstre

Dans le nord de la France se trouvait un pay*san* |dont sa grand-mêre ~~qui~~ avait une maison blanche. Elle habitait dans cette maison depuis longtemps. Un jour, le paysan vue un buisson bouger |pendant une *nuit*, il ne voulait en savoir plus, alors, il sortit dehors avec un couteau |et une torche. Il regarda dans le buisson et il vi du sang et des pas derrière la cloture, il sorti dehors et suivi les pas. Il *sarretai* dans un *buisson* à l'*interieur* de la forêt
Il regarda dans le buisson et vue un ours mort. Il reparti alors.
Tout à coup un grand bruit le fit sursoter, Il se retourna en entendant ce grand bruit.
~~-dont ce bruit~~ et il vue un monstre dont ses doit griffe*s* dégout*xx* *deont*, il *xx vert* gluant et |ª *xx* : dans mais *xx inodaxxt* inimaginable. Je cri en voyant le monstre il le tuer.
Depuis cette aventure, les enfants ne sortent plus la nuit.



- Il est difficile de comprendre l'écriture de l'élève.
- Ligne 1 et 3 : les ajouts ont été effectués au dessus de la ligne à l'aide d'une sorte de flèche.
- Ligne 4 et 10 : les ajouts ont été effectués au dessous de la ligne à l'aide d'une sorte de flèche.

Si vous voyez une erreur dans la transcription, merci d'envo
hodac@univ-tlse2.fr

Plan

Normalisation de l'orthographe, une première couche d'annotation

- Détection des erreurs uniquement (pas de caractérisation)
- Caractérisation des erreurs pensée dans un deuxième temps, étape par étape (en cours, les temps verbaux non composés)
- Guide de normalisation orthographique (ScolEdit, Ecriscol) :

  Ortho. lexicale Il était une foie ⇒ fois

  Ponctuation point / majuscule, virgule dans les énumérations, non-fermeture/ouverture des citations, dialogues ou parenthèses

  Temps verbaux on ne corrige pas la cohérence des temps verbaux
- Possibilité de proposer plusieurs corrections

Digression : Annoter en linguistique

Handbook of linguistic annotation [Ide, 2017]

Linguistic annotation of language data was originally performed in order to provide information for the development and testing of linguistic theories, or, as it is known today, corpus linguistics. At the time, considerable time and effort was required to annotate data with even the ***simplest linguistic phenomena****, and the annotated corpora available for study were* ***quite small****. Over the past three decades, advances in computing power and storage together with development of robust methods for automatic annotation have made linguistically-annotated data increasingly available in ever-growing quantities. As a result,* ***these resources now serve not only linguistic studies, but also the field of natural language processing (NLP)****, which relies on linguistically-annotated text and speech corpora to evaluate new human language technologies and, crucially, to develop reliable* ***statistical models for training*** *these technologies.*

Rapide historique

- 1970 : Corpus Brown Corpus of Standard American English
  - 1 M de mots
  - Premier projet d'annotation morphosyntaxique
- Années 2000 : développements des corpus arborés
  - Puis développement des annotations sémantiques
- Développement de l'informatique et du TAL
  - TAL probabiliste, mobilisation des méthodes d'approvisionnement par la foule (crowdsourcing) comme Amazon mechanical turk

Ce qu'annoter implique... [Ide, 2017]

Linguistic annotation involves ***the association of descriptive or analytic notations with language data****. The* ***raw data*** *may be textual, drawn from any source or genre, or it may be in the form of time functions (audio, video and/or physiological recordings). The* ***annotations*** *themselves may include transcriptions of all sorts (from phonetic features to discourse structures), part-of-speech and sense tags, syntactic analyses, "named entity" labels, semantic role labels, time and event identification, coreference chains, discourse-level analyses, and many others. Resources vary in the range of annotation types they contain : some resources contain only one or two types, while others contain* ***multiple annotation "layers"*** *or "tiers" of linguistic descriptions.*

Quelles annotations en linguistique ?

Annotation morphosyntaxique ou grammaticale

- Texte à annoter (écrits journalistiques, académiques, web, oral, brouillons, LSF, SMS, interactions en ligne, etc.)
- Unité = token
- Tagset largement stabilisé, aujourd'hui le cadre des UD

Annotation syntaxique

- Texte à annoter (idem)
- Unité = token et constituants (i.e. token consécutifs), difficulté de segmentation pour l'oral
- Tagset largement stabilisé, aujourd'hui le cadre des UD

→ Penn Treebank, French Treebank

https://universaldependencies.org/

open community effort, over 300 contributors, nearly 200 treebanks, over 100 languages

Quelles annotations en linguistique ?

Annotations sémantiques et pragmatiques

- Depuis les années 2000, beaucoup d'efforts sont faits pour automatiser certaines annotations sémantiques et pragmatiques
  - annotation des émotions, des frames, des entités nommées, des actes de parole, du hate speech
- Annotation principalement manuelle
- Textes à annoter : courts et très diversifiés (car besoins liés aux *big data*)
- Tagset peu stabilisés (sauf peut-être pour les EN)
- Évaluation nécessaire avec des scores d'accord inter-annotateurs
- Méthodes d'apprentissage automatique (et d'active learning)
- Des challenges

→ http://www.quaero.org/, annotations des EN dans des documents techniques, e.g.

Quelles annotations en linguistique ?

Annotations discursives

- Textes à annoter : textes longs, experts, normalisés, encore majoritairement écrits (difficulté d'annoter l'oral, rôle important de la prosodie et des disfluences)
- Unité = UDE, marqueurs de discours, portions de texte, etc.
- Tagset peu stabilisés (faible consensus)
  - Relations de discours

  → Penn Discourse Treebank, French Discourse Treebank

  → Réseau TextLink

  unify numerous but scattered linguistic resources on discourse structure.
  - Coréférence
  - Autres objets (structures énumératives, encapsulations, etc.)

Fundamental components of a linguistic annotation project [Ide, 2017]

Définir le modèle d'annotation

The most critical component of a linguistic annotation project is the ***annotation scheme*** *that defines the* ***labels*** *and associated* ***features*** *to be associated with the appropriate annotation unit*

Stabiliser une méthode (un outillage) pour assurer une certaine qualité

[...] modern manual or semi-automatic annotation efforts typically rely on an ***annotation tool with an interface*** *that enables identification of spans of characters and/or links between such spans [...]*

Multi annotation pour évaluer la qualité et fournir un consensus (*gold*)

[...] measure ***inter-annotator agreement*** *(IAA) for two or* ***more annotators*** *using one of several popular metrics, in order to measure* ***consensus****, define a threshold of expected performance by automatic annotation tools, and/or determine if a particular scale is appropriate for measuring the phenomenon in question, etc.*

Ce qu'annoter implique

- Définir le schéma d'annotation et les besoins outils et annotateurs
- Collecter les données primaires (une première étape qui s'apparente parfois à de l'annotation)
- Lancer l'annotation

Ce qu'annoter implique

Procédure typique d'une annotation

- Charger le corpus/texte à annoter
- Délimiter des segments de texte : les unités (**U**)
- Associer des "traits" aux unités selon
- Relier des unités : les relations (**R**)
- Regrouper des relations et des unités dans des structures complexes : les schémas (**S**)

Ce qu'annoter implique

Procédure typique d'une annotation

- Charger le corpus/texte à annoter
  - texte brut
  - document word ou pdf
  - texte structurés voire prémarqués (e.g. XML) $\Rightarrow$ multi-couche
- Délimiter des segments de texte : les unités (**U**)
- Associer des "traits" aux unités selon
- Relier des unités : les relations (**R**)
- Regrouper des relations et des unités dans des structures complexes : les schémas (**S**)

Ce qu'annoter implique

Procédure typique d'une annotation

- Charger le corpus/texte à annoter
- Délimiter des segments de texte : les unités (**U**)
  - **Délimitation dans le texte** (annotation intégrée, *in situ* $\Rightarrow$ le texte source est modifié ce qui empêche l'interopérabilité et le "multicouche".
  - **Délimitation hors le texte** (annotation débarquée, *stand-off*) $\Rightarrow$ le texte source n'est jamais édité. Chaque annotation est indépendante et n'interfère pas sur le texte source. Permet des analyses croisées.
    - Géolocalisation des bornes de début et de fin au token (TXM - URS), au caractère (Glozz) ou au choix (Inception)
- Associer des "traits" aux unités selon
- Relier des unités : les relations (**R**)
- Regrouper des relations et des unités dans des structures complexes : les schémas (**S**)

Ce qu'annoter implique

Procédure typique d'une annotation

- Charger le corpus/texte à annoter
- Délimiter des segments de texte : les unités (**U**)
- Associer des "traits" aux unités selon
- Relier des unités : les relations (**R**)
- Regrouper des relations et des unités dans des structures complexes : les schémas (**S**)

Ce qu'annoter implique

Procédure typique d'une annotation

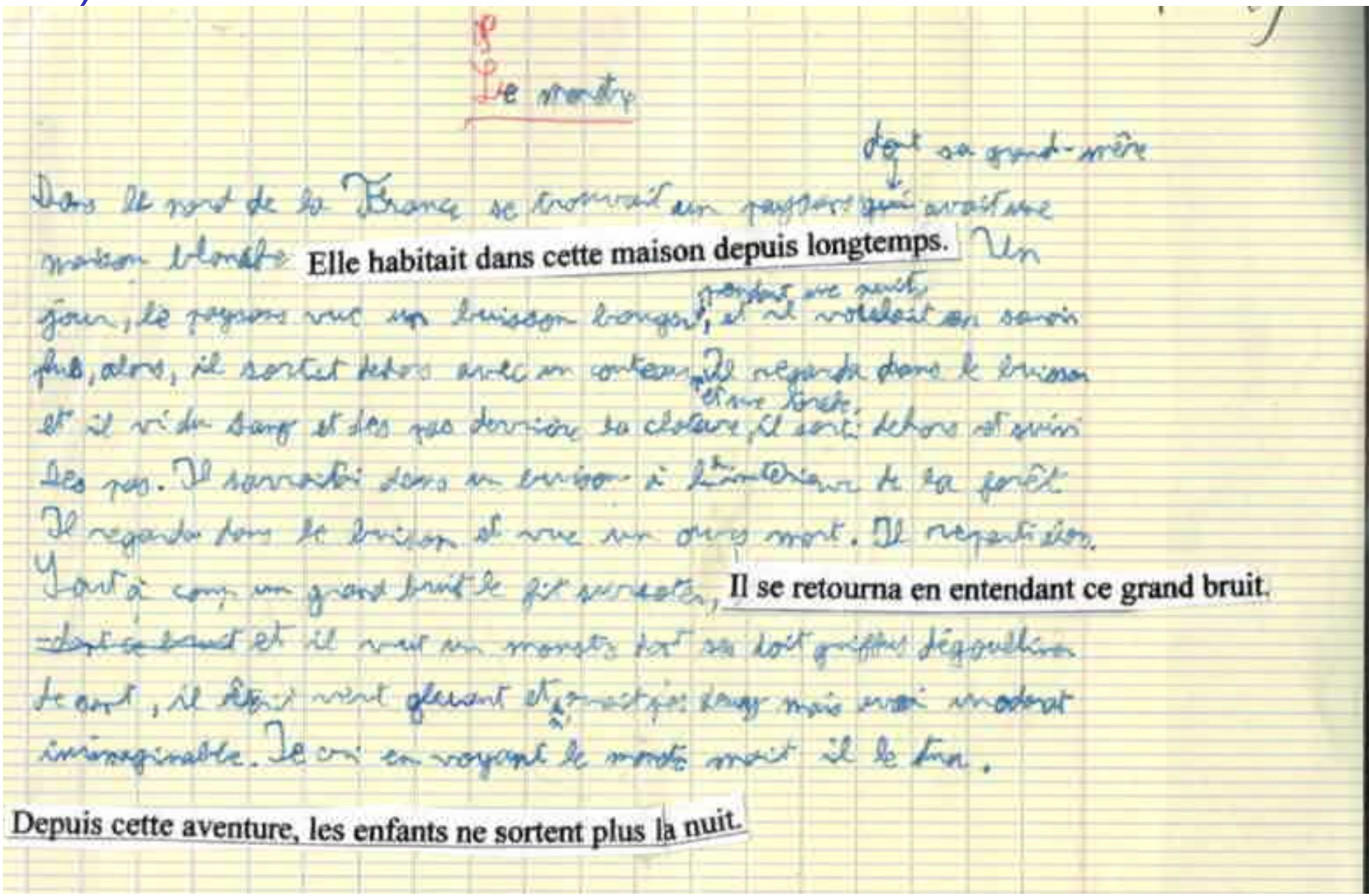
- Charger le corpus/texte à annoter
- Délimiter des segments de texte : les unités (**U**)
- Associer des "traits" aux unités selon
  - un modèle d'annotation prédéfini non modifiable
  - un modèle d'annotation construit dynamiquement
- Relier des unités : les relations (**R**)
- Regrouper des relations et des unités dans des structures complexes : les schémas (**S**)

Des outils pour annoter

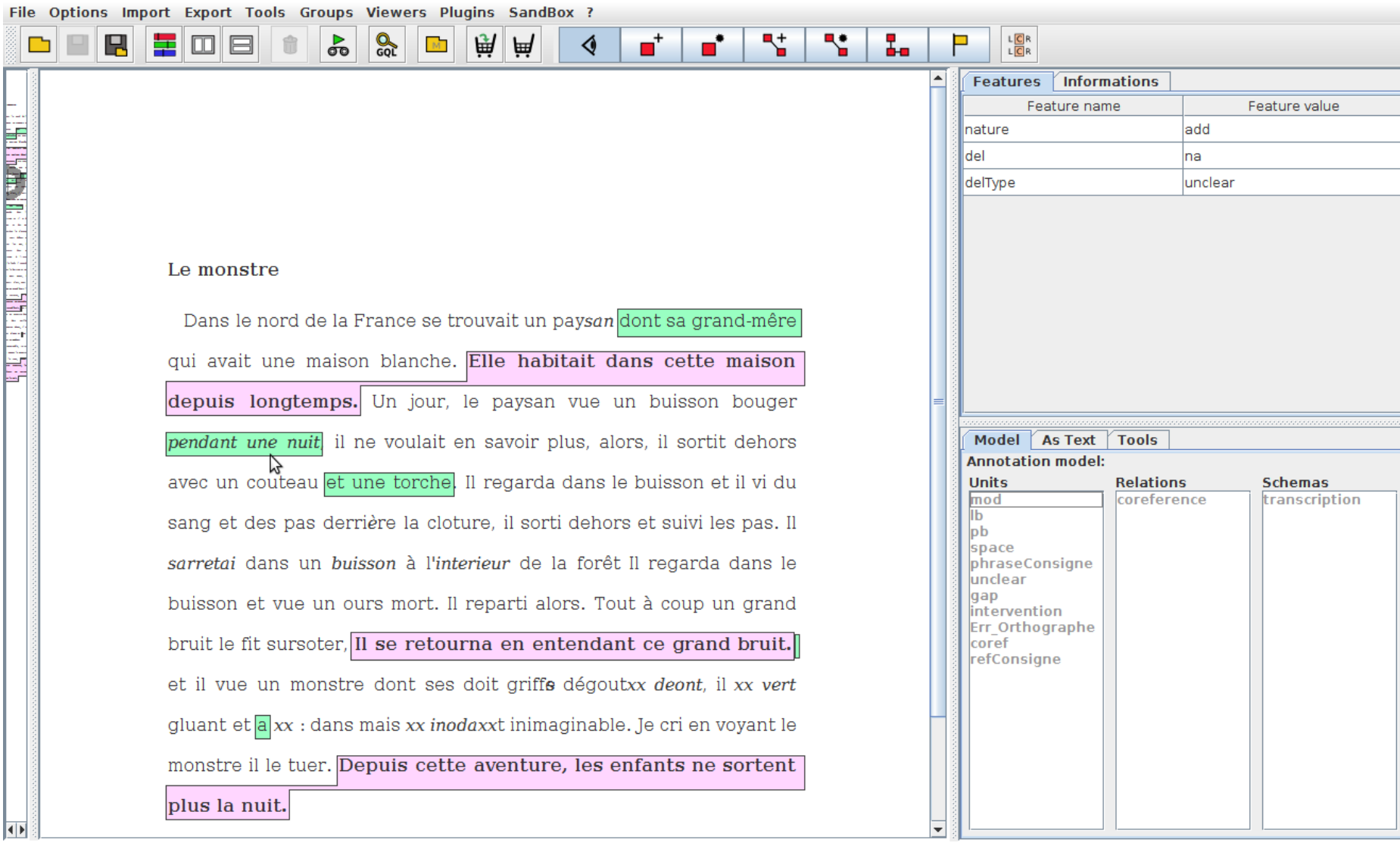
Des outils qui se veulent génériques

- mais qui ont des "restes" issus de leur origine :
  - outil existant d'exploration de corpus
    e.g. TXM et Analec → TXM–URS
    http://textometrie.ens-lyon.fr/
  - un projet visant la constitution d'une ressource
    e.g. AnnoDis → Glozz http://www.glozz.org
  - outil d'annotation enrichi pour de l'active learning
    e.g WebAnno → Inception https://inception-project.github.io
- Des principes communs
  - annotation d'objets linguistiques dans des textes digitalisés
  - permettant une AAO ("analyse assistée par ordinateur") pouvant tirer partie de traitements automatiques e.g. étiquetage morpho-syntaxique, parsing, projection de patrons et de lexiques
  - permettant plusieurs couches d'annotation
  - permettant des analyses croisées entre couches d'annotation
  - permettant la constitution de ressources pour la communauté

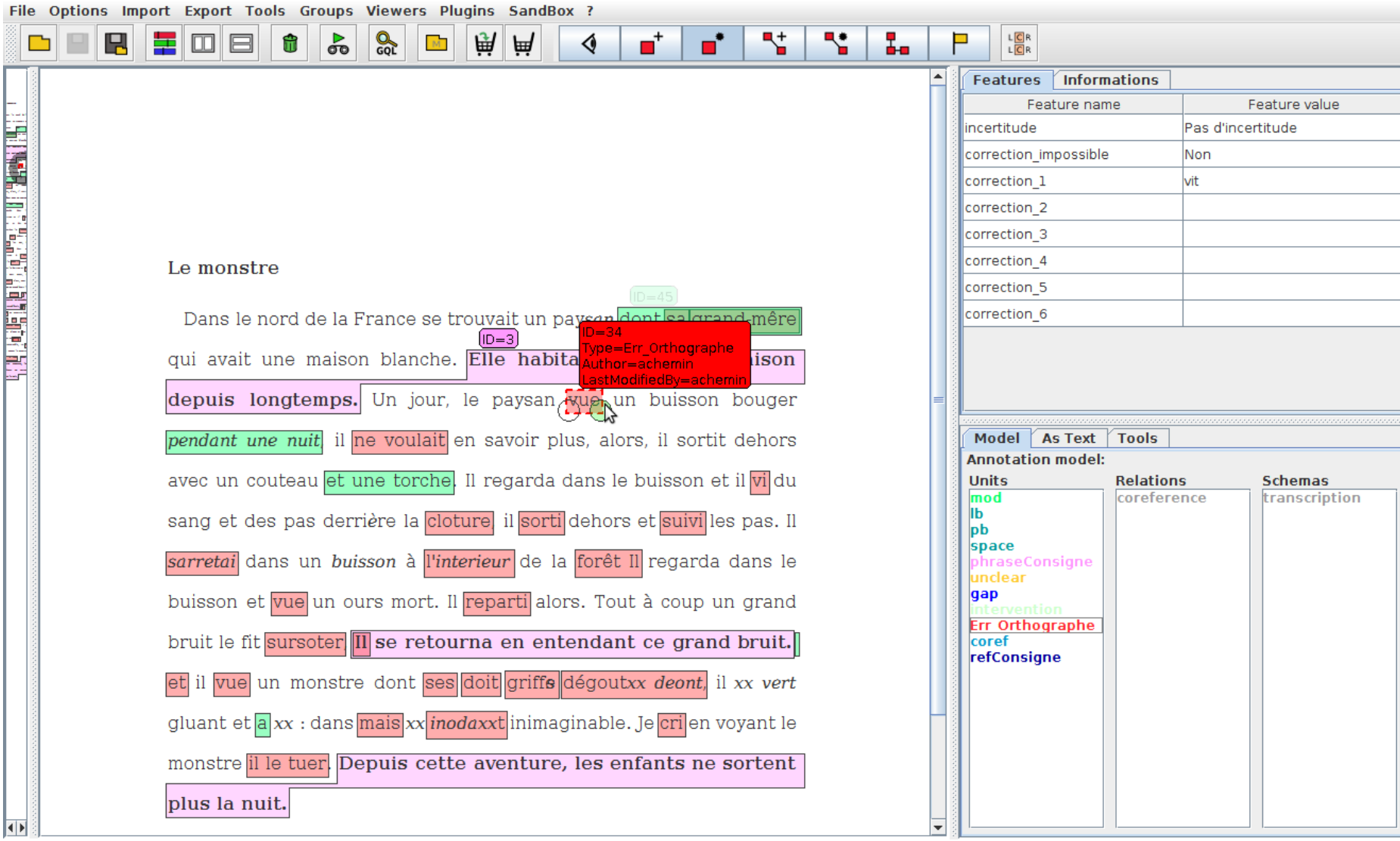
Chargement d'un texte et d'un modèle d'annotation (Glozz)



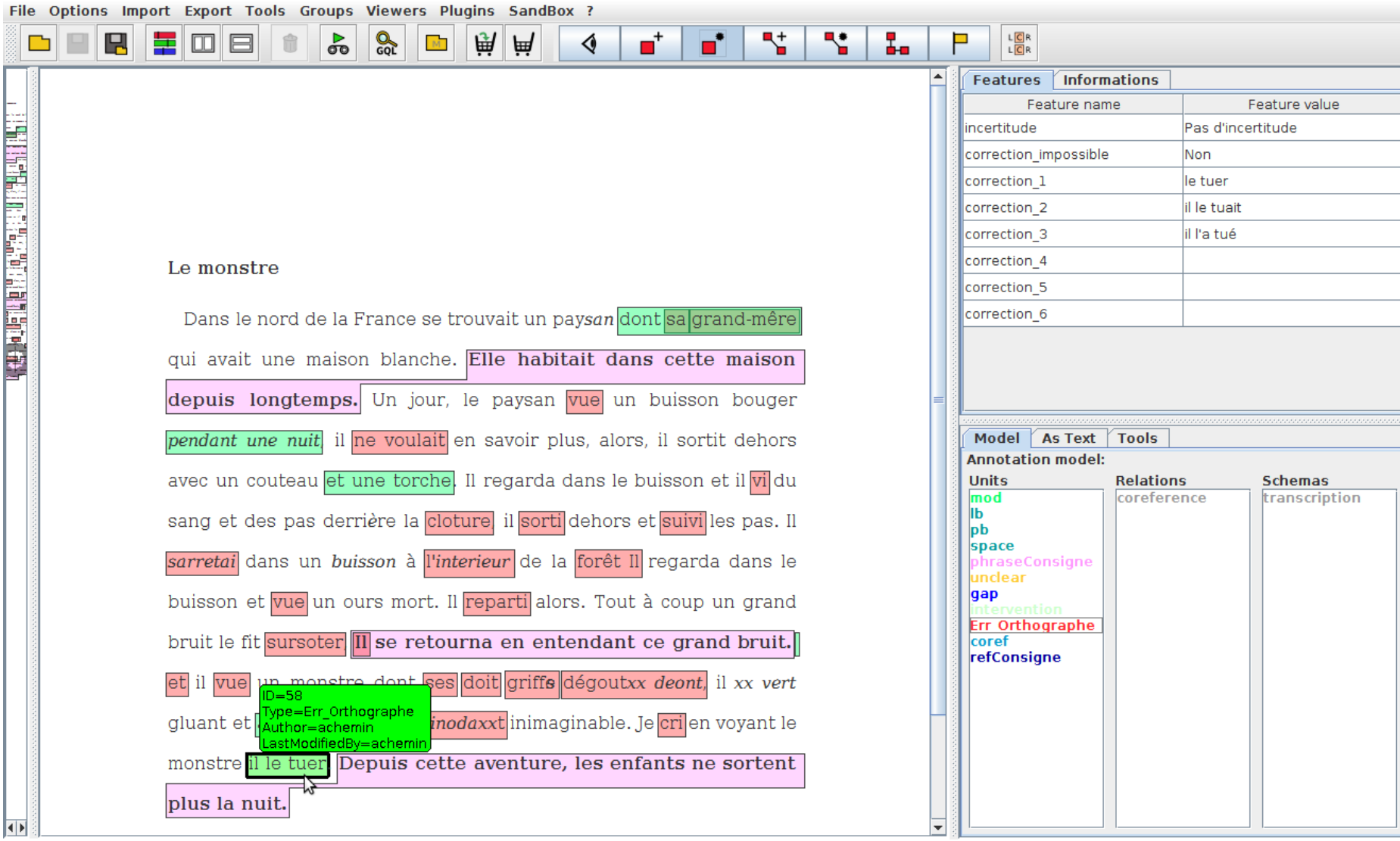
Chargement d'un texte et d'un modèle d'annotation (Glozz)



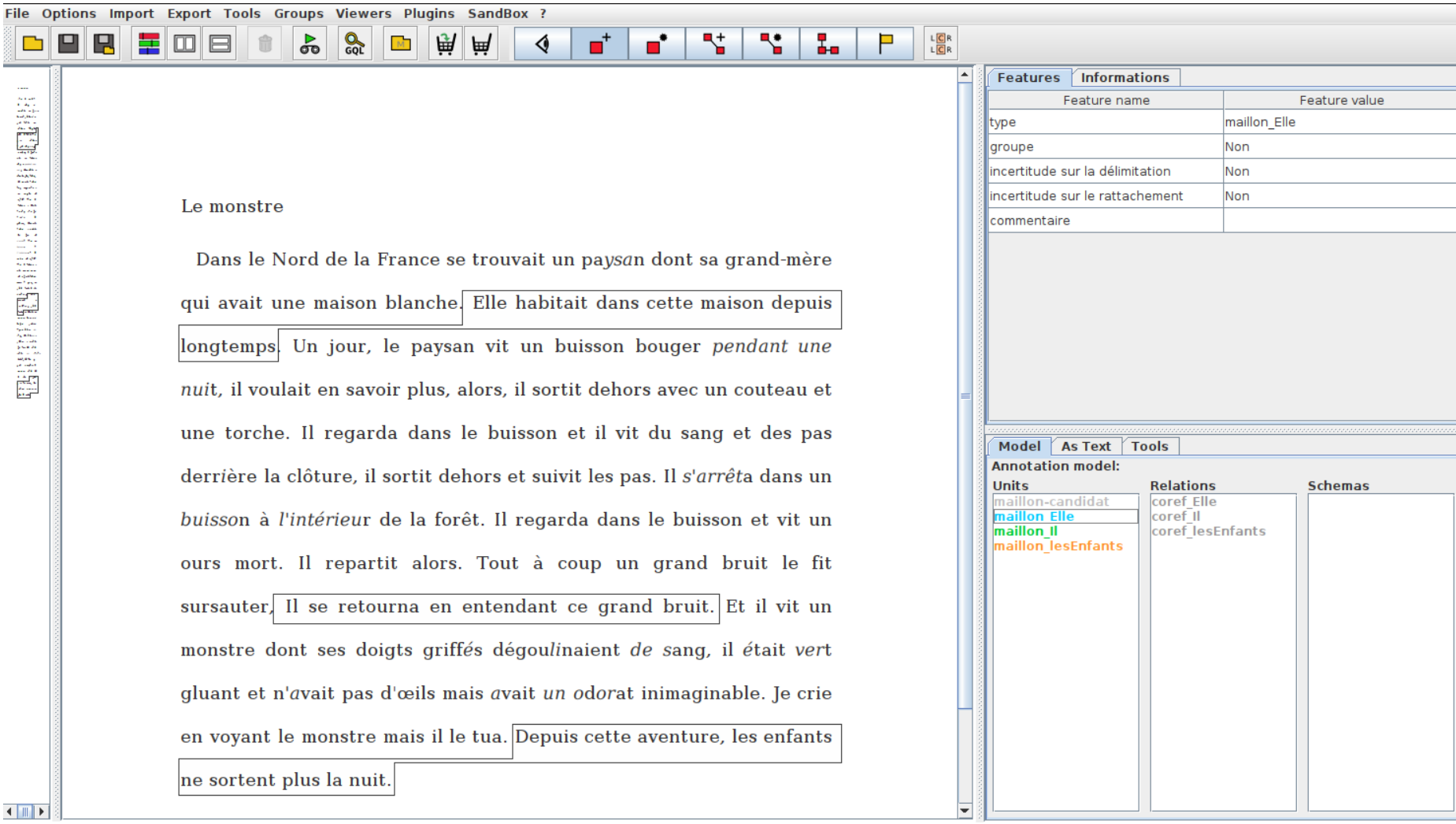
Normalisation de l'orthographe avec Glozz



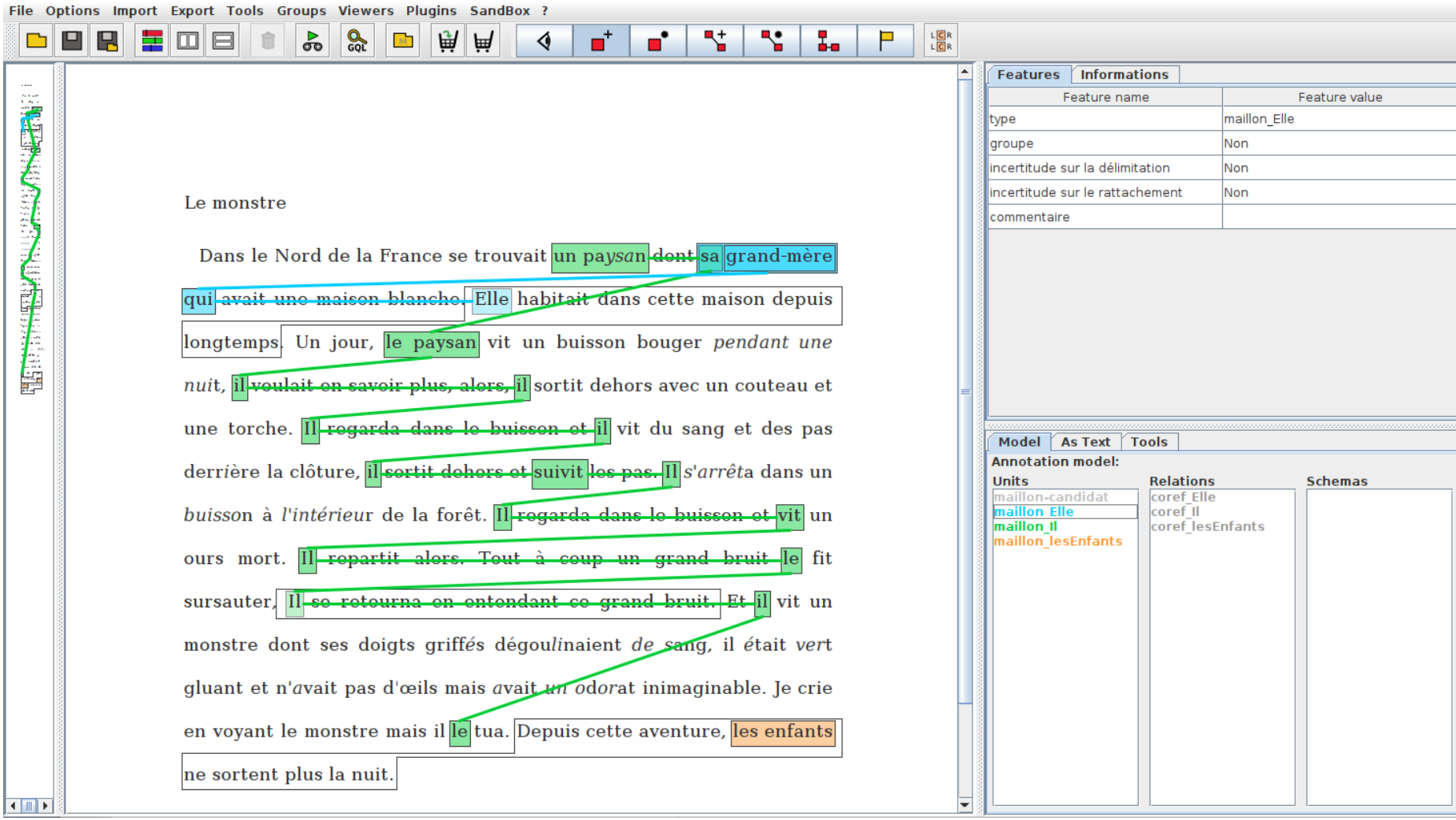
Normalisation de l'orthographe avec Glozz



Charger un autre modèle, ajouter une couche d'annotation



Charger un autre modèle, ajouter une couche d'annotation



Besoins à l'origine de Glozz : projet ANNODIS, annotation discursive

Spécificité des textes et objets à annoter

- Textes longs et structurés (structures de haut niveau, prise en compte de la structure de document)
- Structures complexes, qui peuvent se "chevaucher" et sont influencer par la structure du document

Besoins

- Avoir une vue de haut du texte => ruban
- Avoir une vue "document" des textes à annoter
- Permettre de cibler vers des zones à annoter => prémarquage
- Avoir une vue graphique des annotations

Besoins à l'origine de Glozz : projet ANNODIS, annotation discursive

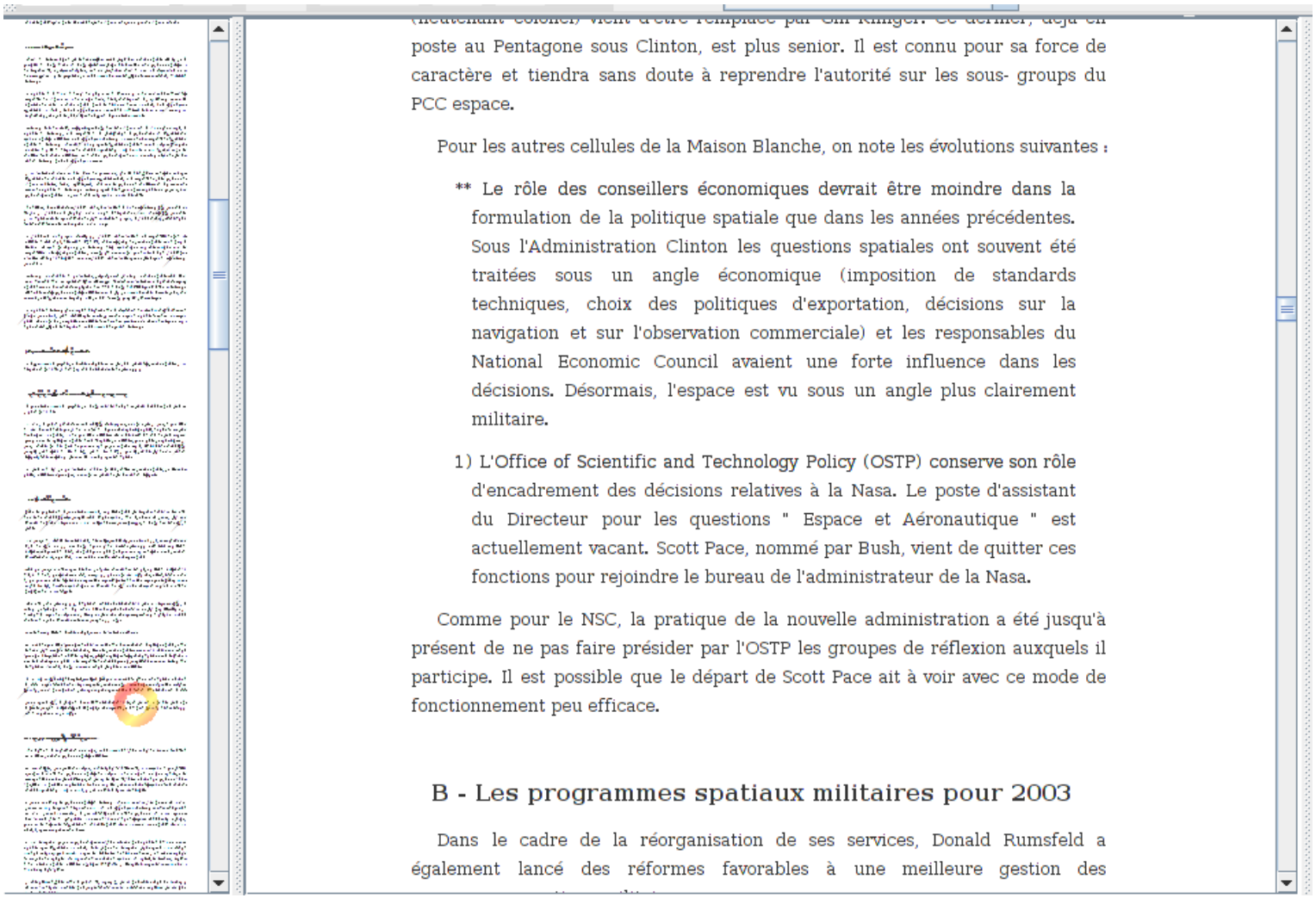
Spécificité des textes et objets à annoter

- Textes longs et structurés (structures de haut niveau, prise en compte de la structure de document)
- Structures complexes, qui peuvent se "chevaucher" et sont influencer par la structure du document

Besoins

- **Avoir une vue de haut du texte => ruban**
- **Avoir une vue "document" des textes à annoter**
- Permettre de cibler vers des zones à annoter => prémarquage
- Avoir une vue graphique des annotations

Vue de haut – ruban – et vue réaliste du "document"



poste au Pentagone sous Clinton, est plus senior. Il est connu pour sa force de caractère et tiendra sans doute à reprendre l'autorité sur les sous- groups du PCC espace.

Pour les autres cellules de la Maison Blanche, on note les évolutions suivantes :

** Le rôle des conseillers économiques devrait être moindre dans la formulation de la politique spatiale que dans les années précédentes. Sous l'Administration Clinton les questions spatiales ont souvent été traitées sous un angle économique (imposition de standards techniques, choix des politiques d'exportation, décisions sur la navigation et sur l'observation commerciale) et les responsables du National Economic Council avaient une forte influence dans les décisions. Désormais, l'espace est vu sous un angle plus clairement militaire.

1) L'Office of Scientific and Technology Policy (OSTP) conserve son rôle d'encadrement des décisions relatives à la Nasa. Le poste d'assistant du Directeur pour les questions " Espace et Aéronautique " est actuellement vacant. Scott Pace, nommé par Bush, vient de quitter ces fonctions pour rejoindre le bureau de l'administrateur de la Nasa.

Comme pour le NSC, la pratique de la nouvelle administration a été jusqu'à présent de ne pas faire présider par l'OSTP les groupes de réflexion auxquels il participe. Il est possible que le départ de Scott Pace ait à voir avec ce mode de fonctionnement peu efficace.

B - Les programmes spatiaux militaires pour 2003

Dans le cadre de la réorganisation de ses services, Donald Rumsfeld a également lancé des réformes favorables à une meilleure gestion des

Besoins à l'origine de Glozz : projet ANNODIS, annotation discursive

Spécificité des textes et objets à annoter

- Textes longs et structurés (structures de haut niveau, prise en compte de la structure de document)
- Structures complexes, qui peuvent se "chevaucher" et sont influencer par la structure du document

Besoins

- Avoir une vue de haut du texte => ruban
- Avoir une vue "document" des textes à annoter
- Permettre de cibler vers des zones à annoter => prémarquage
- Avoir une vue graphique des annotations

Besoins à l'origine de Glozz : projet ANNODIS, annotation discursive

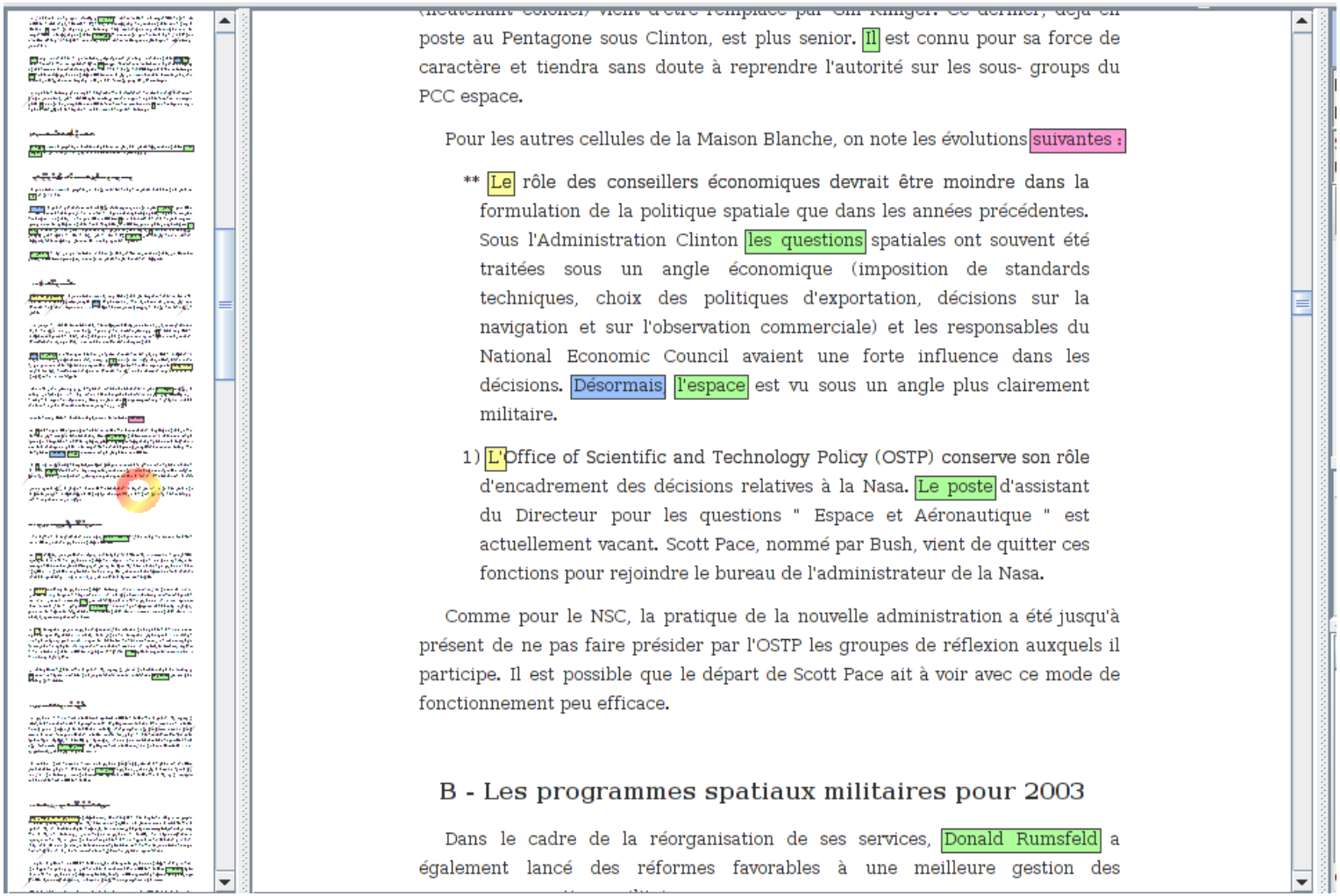
Spécificité des textes et objets à annoter

- Textes longs et structurés (structures de haut niveau, prise en compte de la structure de document)
- Structures complexes, qui peuvent se "chevaucher" et sont influencer par la structure du document

Besoins

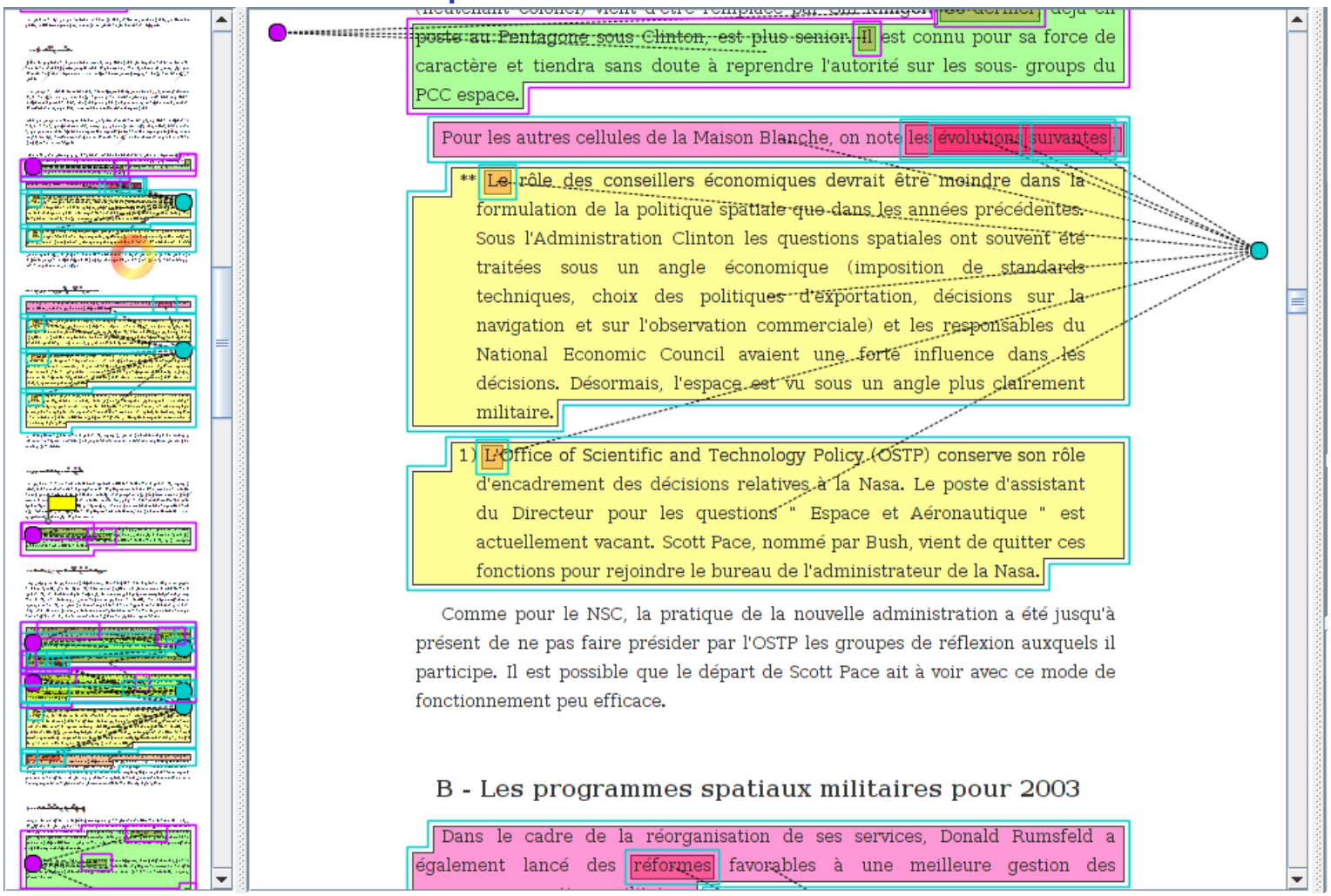
- Avoir une vue de haut du texte => ruban
- Avoir une vue "document" des textes à annoter
- **Permettre de cibler vers des zones à annoter => prémarquage**
- Avoir une vue graphique des annotations

Vue de haut de zones à annoter (concentration de prémarques ou zone non annotée)



Annotation localement et largement, indépendemment de la structure de document

Annoter des structures complexes et permettre des chevauchements <p> et <structure>



Plan

Ce que l'annotation outillée permet

Préatraitements et postraitements

- Prémarquer les textes pour des annotations complexes
- Nettoyer les annotations
- Appliquer des modules TAL : POStagging, Parsing, etc.

Croiser des annotations

- Annotations multiples et mesure de l'accord inter-annotateur
- Annotations relevant de différentes couches d'annotation (syntaxe, discours, ratures, orthographes, etc.)

Analyser les annotations

- Statistiques
- Concordances, textométrie

Ce que l'annotation outillée permet

Préatraitements et postraitements

- Prémarquer les textes pour des annotations complexes
- Nettoyer les annotations
- Appliquer des modules TAL : POStagging, Parsing, etc.

Croiser des annotations

- Annotations multiples et mesure de l'accord inter-annotateur
- Annotations relevant de différentes couches d'annotation (syntaxe, discours, ratures, orthographes, etc.)

Analyser les annotations

- Statistiques
- Concordances, textométrie

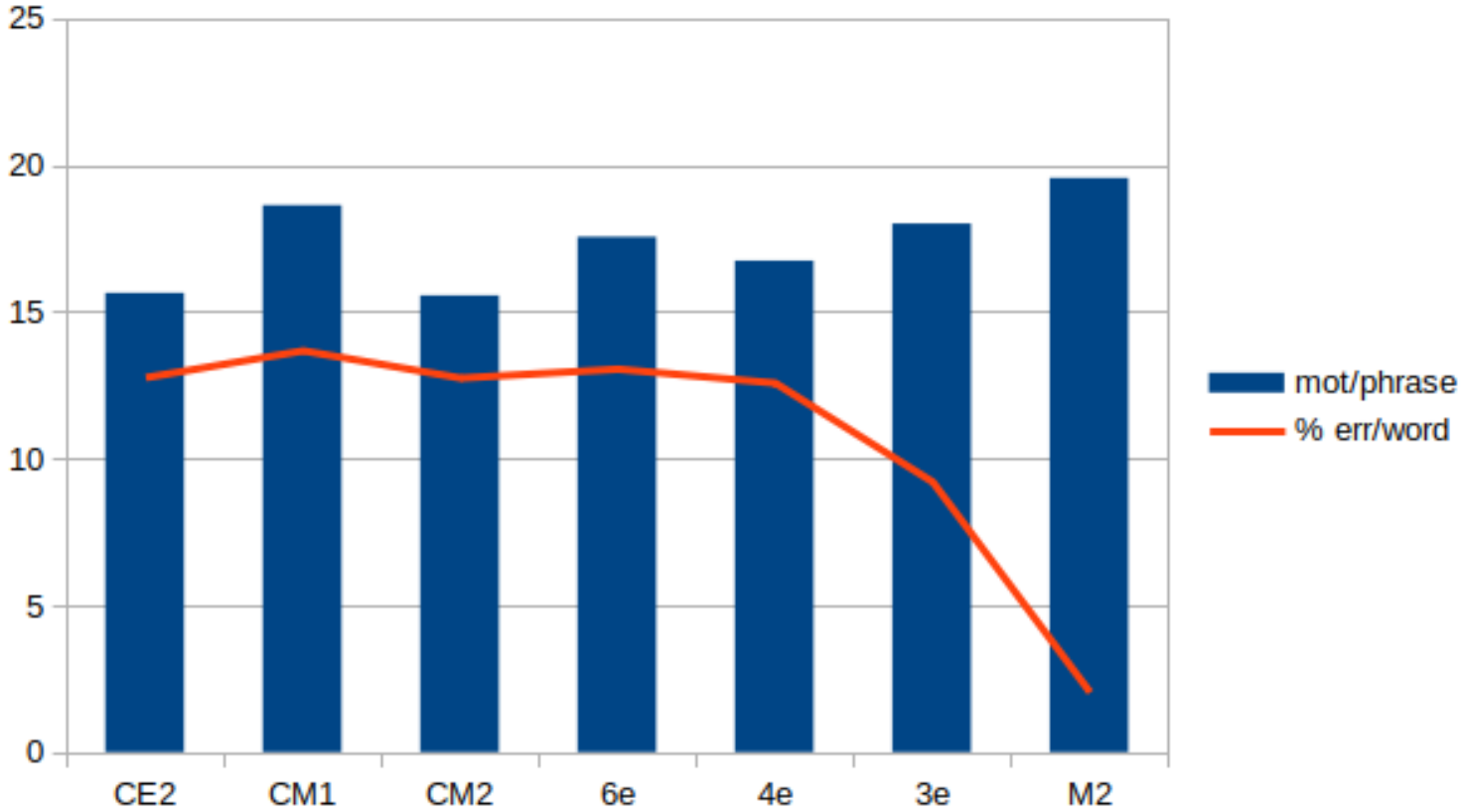
Retour à la ressource E :CALM : : POS-tagging et Parsing

Intérêts

- Analyser les zones de résistance / les interrogations persistantes
- Permettre un prémarquage pour l'annotation de structures discursives [Asher et al., 2017] :
  - Segmentation en unités de discours
  - Connecteurs de discours
  - Continuités co-référentielles

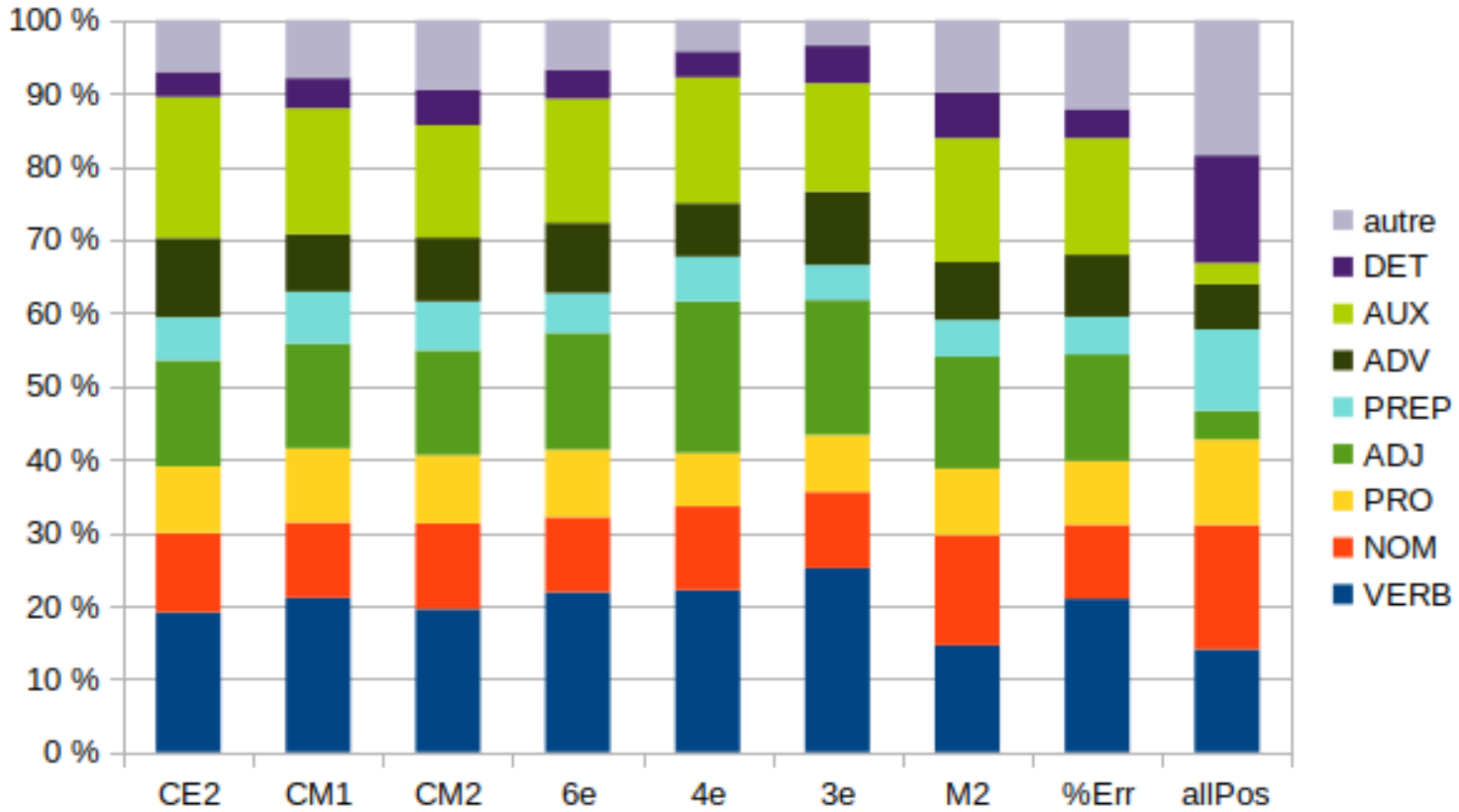
Premières observations issues de la normalisation E :CALM

Longueur des phrases et % erreur par mots



Premières observations issues de la normalisation E :CALM

% erreur par POS et niveaux (Stanza)



L'annotation outillée rend également possible des annotations plus complexes

Annotations complexes

- Annotations coûteuses en temps et en concentration
- Annotations impliquant une interprétation fine du texte
- Annotations faisant appel à la subjectivité de l'annotateur $\Rightarrow$ nécessité de multi-annoter, adjudiquer et mesurer un accord

Plan

Les continuités référentielles dans les textes d'élèves

Des liens de cohésion qui se tissent et donnent au texte une *texture* [Halliday and Hasan, 1976]

- Des liens entre les référents 'principaux' d'un texte (*topics/participants continuity* [Givón, 1983])
- Qui impliquent une variété "procédures" pour introduire et maintenir ces référents dans les discours : *The normal procedure for introducing information is [...] a reference is first established (typically with an indefinite NP), and then is subsequently maintained (with some kind of anaphora). This essentially simple picture is complicated by three things : (i) evoked knowledge differs enormously, both qualitatively and quantitatively, from one person to another ; (ii) if an entity is manifest in the environment, then it is often unnecessary to go through the whole reference procedure ; and (iii) as we have seen, information can be introduced non-standardly, by way of accommodation.* [Werth, 1999]

Les continuités référentielles dans les textes d'élèves

Des liens de cohésion qui se tissent et donnent au texte une *texture*

- Des liens entre les référents 'principaux' d'un texte (*topics/participants continuity* [Givón, 1983])
- Qui impliquent une variété "procédures" pour introduire et maintenir ces référents dans les discours : *The **normal procedure** for introducing information is [...] complicated by three things : (i) evoked knowledge differs enormously [...] from one person to another ; (ii) [...] it is often unnecessary to go through the whole reference procedure ; and (iii) [...] information can be introduced non-standardly [...]* [Werth, 1999]

Questions générales

- Quelles "procédures" adoptées par les élèves ?
- Y a-t-il des **procédures plus fréquentes** selon le niveau scolaire ?

Annotation des continuités référentielles dans E-Calm

Tirer partie de précédents projets d'annotation discursive du Français

ANNODIS [Asher et al., 2017]

- Corpus diversifié annoté en structures discursives dont les "Chaînes Topicales"
- Annotation *top-down* : des structures aux indices
- Annotateurs "naïfs"

DEMOCRAT [Landragin, 2015]

ANR Democrat

- Corpus diversifié annoté en expressions référentielles et chaînes de référence
- Annotation *bottom-up* : des mentions aux chaînes
- Annotateurs experts

Le corpus RÉSOLCO  [Garcia-Debanc et al., 2017]

Une consigne d'écriture qui interpelle la cohérence discursive

Racontez une histoire dans laquelle vous insérerez, séparément et dans l'ordre donné, les trois phrases suivantes :

P1 | Elle habitait dans cette maison depuis longtemps.

P2 | Il se retourna en entendant ce grand bruit.

P3 | Depuis cette aventure, les enfants ne sortent plus la nuit.

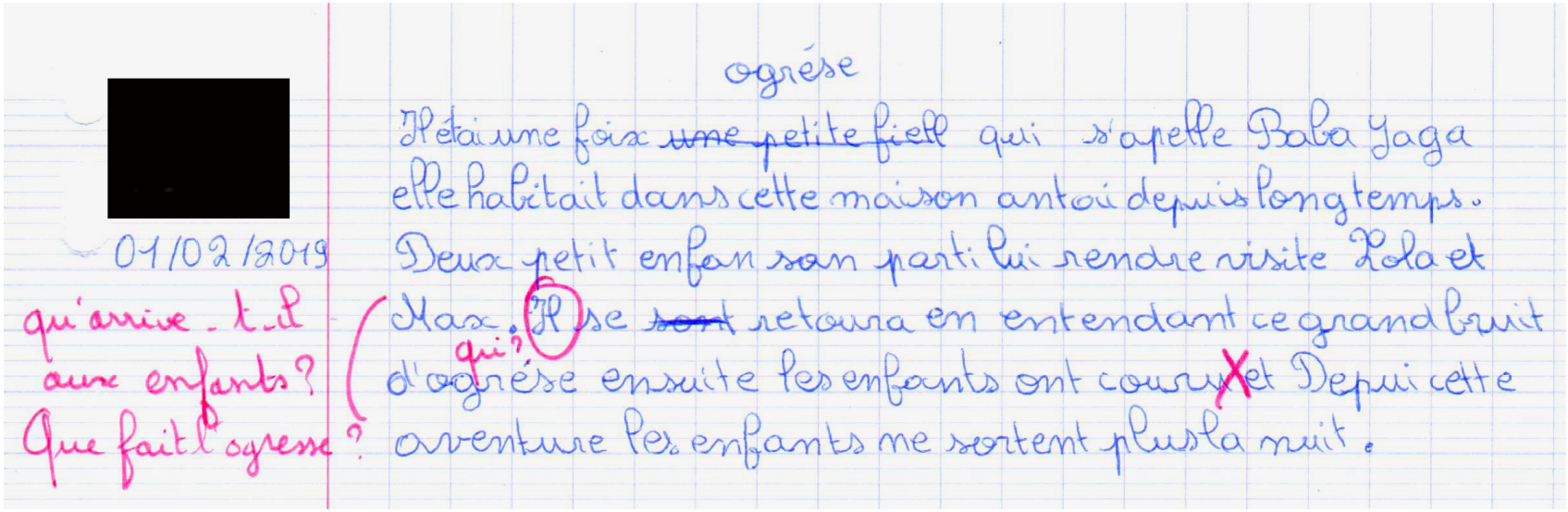
Vous pouvez découper les bandelettes contenant les phrases ci-dessous ou bien recopier chaque phrase avec soin à l'identique de celles qui vous sont données. [Garcia-Debanc, 2016]

Une "tâche-problème" impliquant la résolution de problèmes de cohésion

- Qui est *Elle* et *Il* ? Sont-ce des enfants ? *les enfants* de P3 ?
- Comment rendre ces personnages accessibles au moment des bandelettes ?
- Quel besoin informationnel pour intégrer les démonstratifs *cette maison*, *ce grand bruit*, *cette aventure* ?

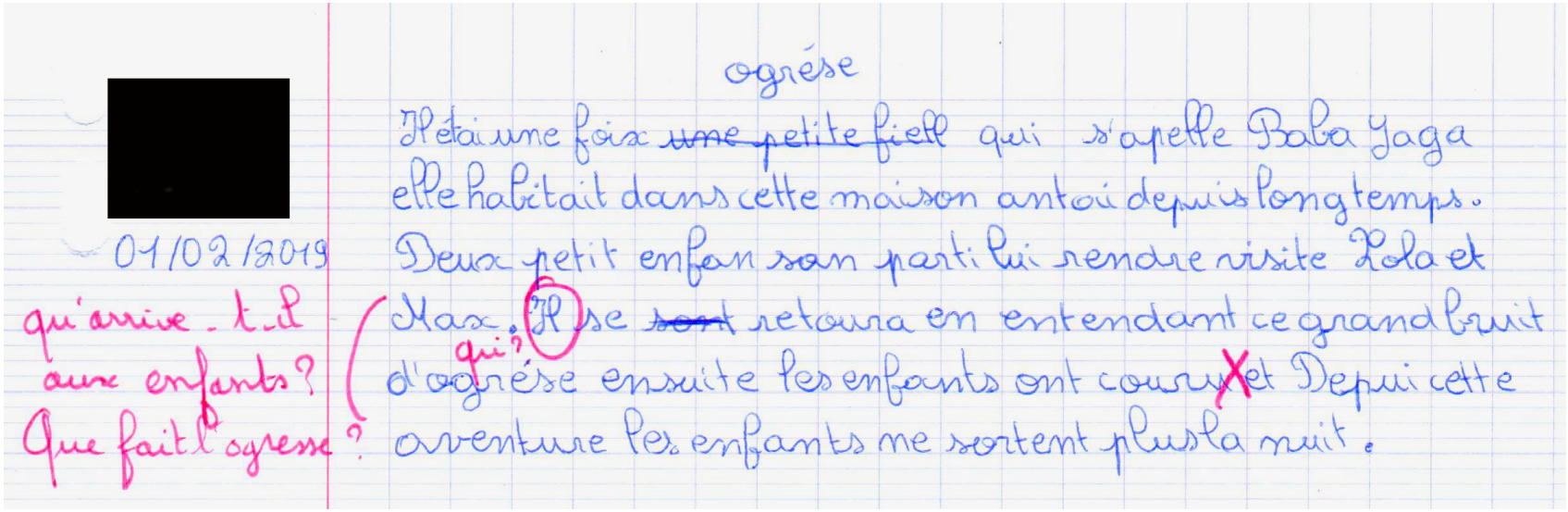
Une consigne provoquant des problèmes de cohérence

copie de CM1



Une consigne provoquant des problèmes de cohérence

copie de CM1

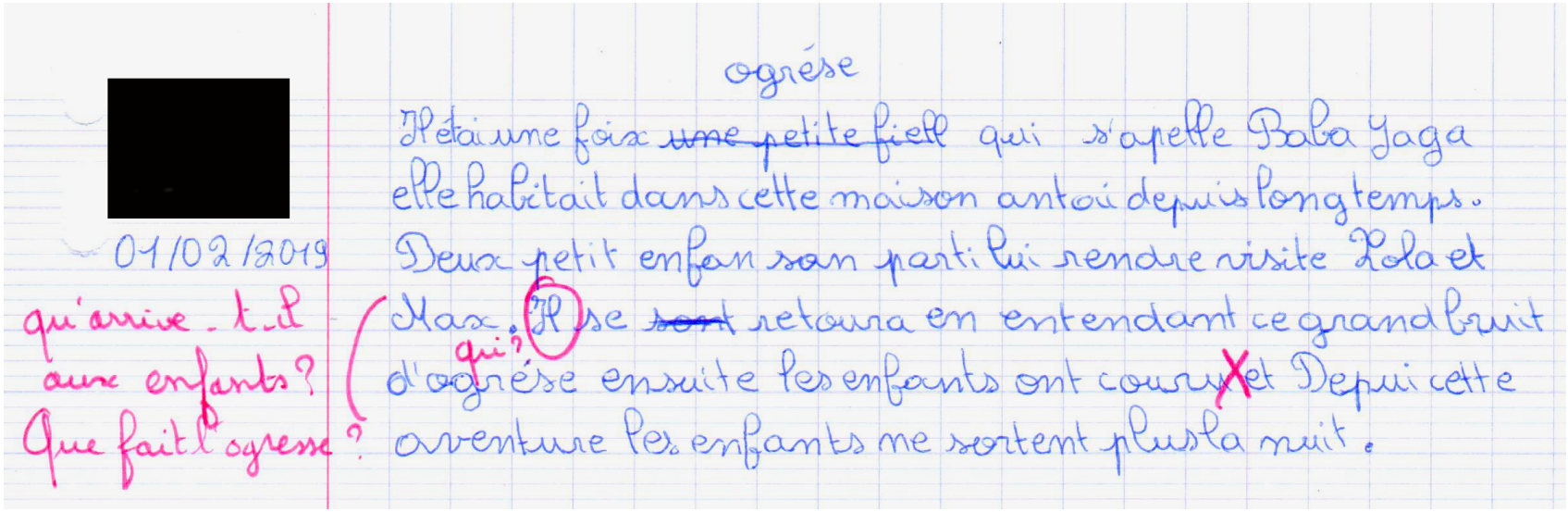


Interventions de l'enseignant.e

- Qui est "Il" ?
- Qu'arrive-t-il aux enfants ?
- Que fait l'ogresse ?

Une consigne provoquant des problèmes de cohérence

copie de CM1



Une annotation pour décrire les procédures employées pour

- introduire les référents
- maintenir le référent actif i.e. la continuité référentielle
- gérer la cohabitation entre les référents

Principes d'annotation

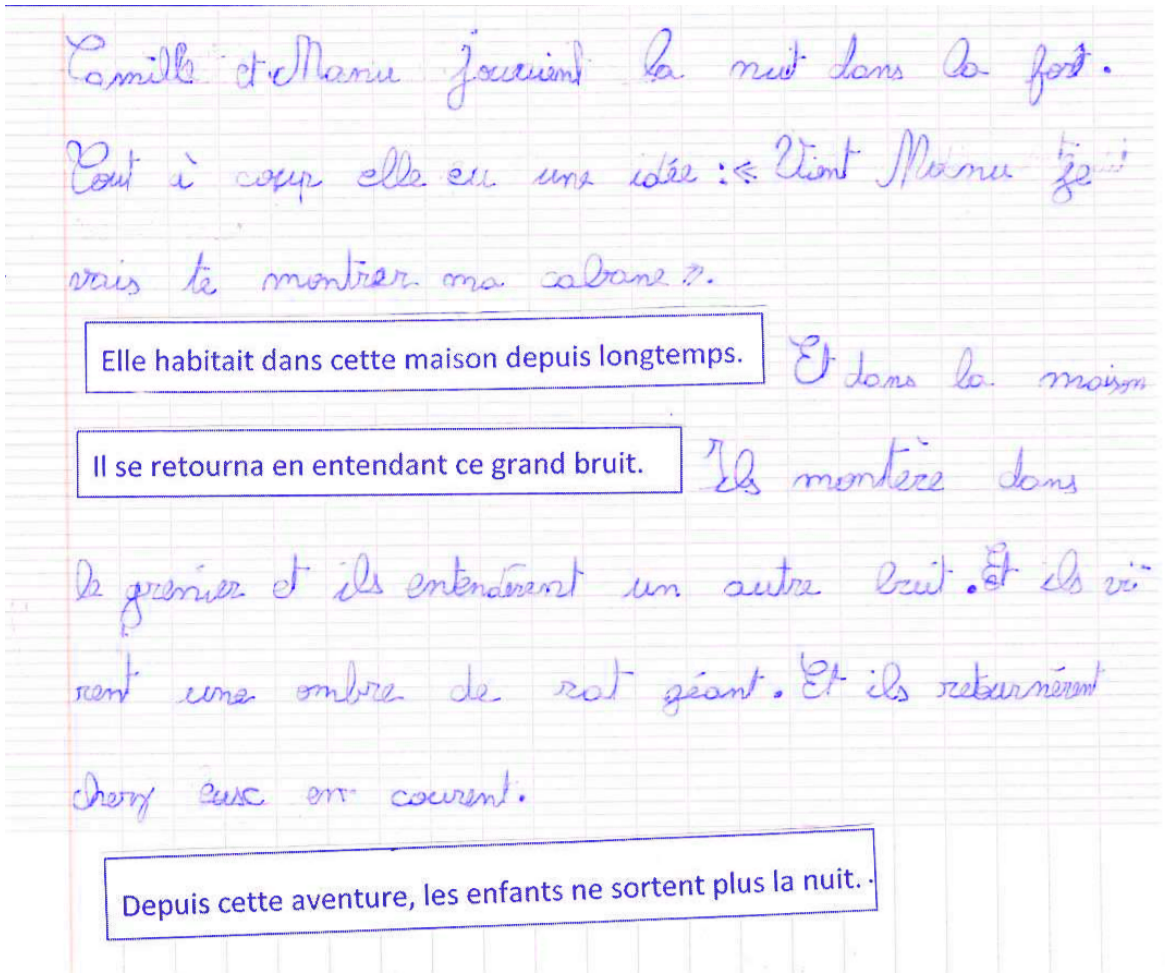
Une annotation *bottom-up* des continuités référentielles

- Se focaliser sur les référents-personnages communs à tous les textes

P1	**Elle** habitait dans cette maison depuis longtemps.
P2	**Il** se retourna en entendant ce grand bruit.
P3	Depuis cette aventure, **les enfants** ne sortent plus la nuit.

- Annoter les mentions ***en lien*** avec ces personnages (pas d'identité référentielle stricte)
  - Inclure les passages au discours direct
  - Inclure les transitions d'un individu au groupe auquel il appartient
  - Indiquer les incertitudes (ambiguïtés ou incohérences)
- Annoter chaque référent-personnage l'un après l'autre
- Outiller l'annotation (prémarquage et interface) [Péry-Woodley et al., 2011]
- Identification des mentions selon le guide DemoCrat [Landragin, 2015]
- Annotateurs experts

Étape -1 : les textes récoltés (CM1)



Camille et Manu jouaient la nuit dans la forêt.
Tout à coup elle eu une idée : « Dit Manu je vais te montrer ma cabane ».

Elle habitait dans cette maison depuis longtemps.

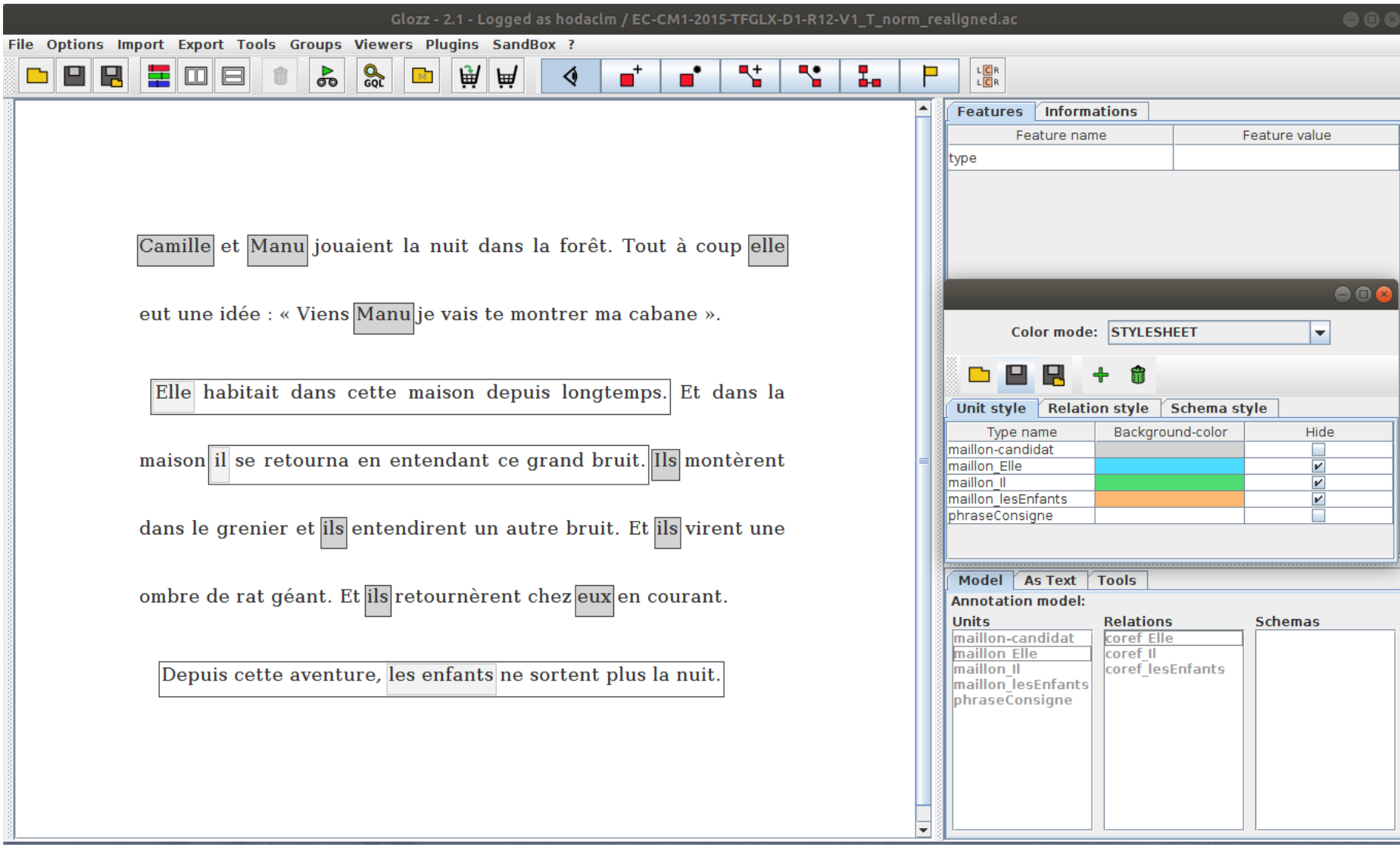
Et dans la maison

Il se retourna en entendant ce grand bruit.

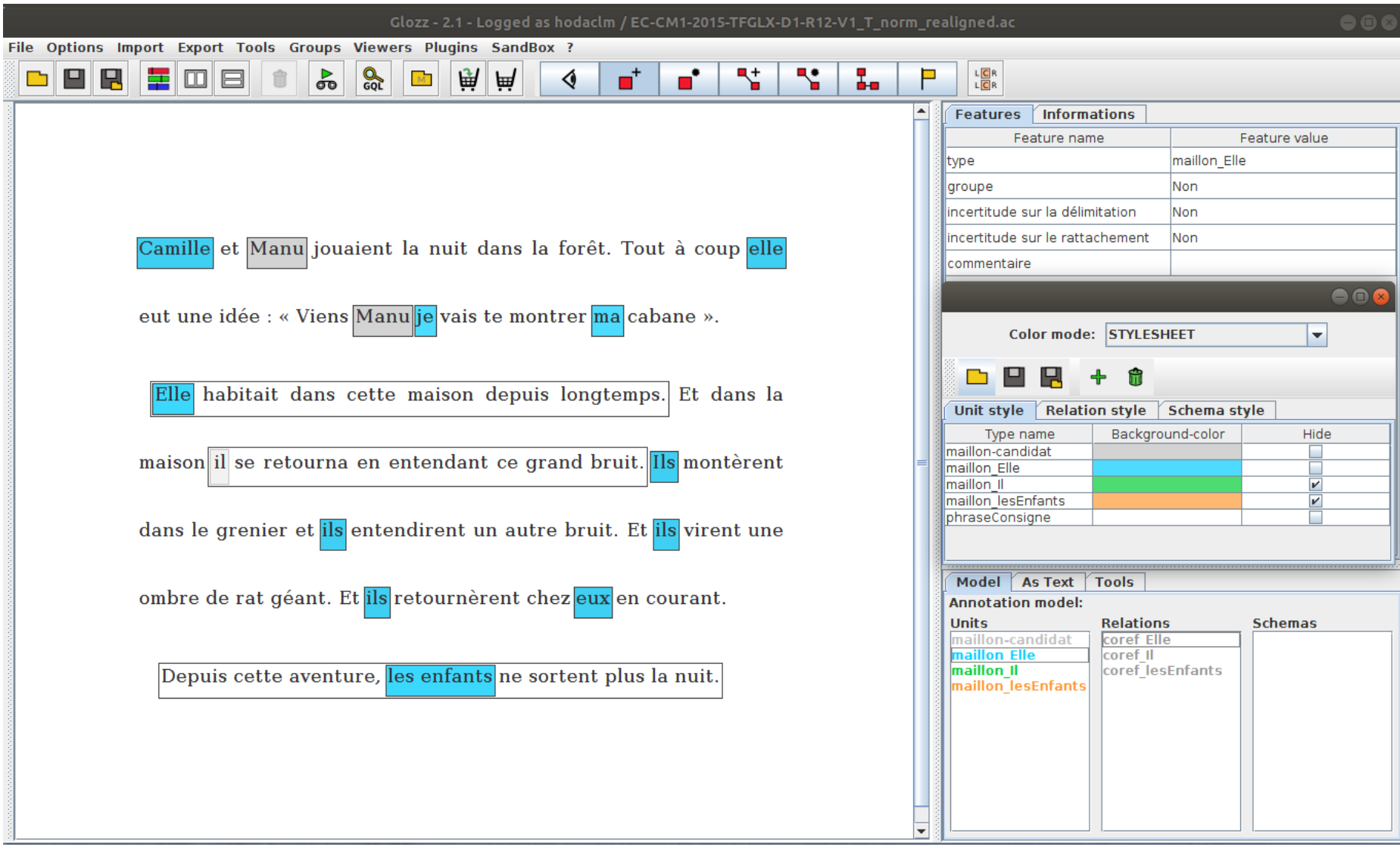
Ils montèrent dans le grenier et ils entendirent un autre bruit. Et ils virent une ombre de rat géant. Et ils retournèrent chez eux en courant.

Depuis cette aventure, les enfants ne sortent plus la nuit.

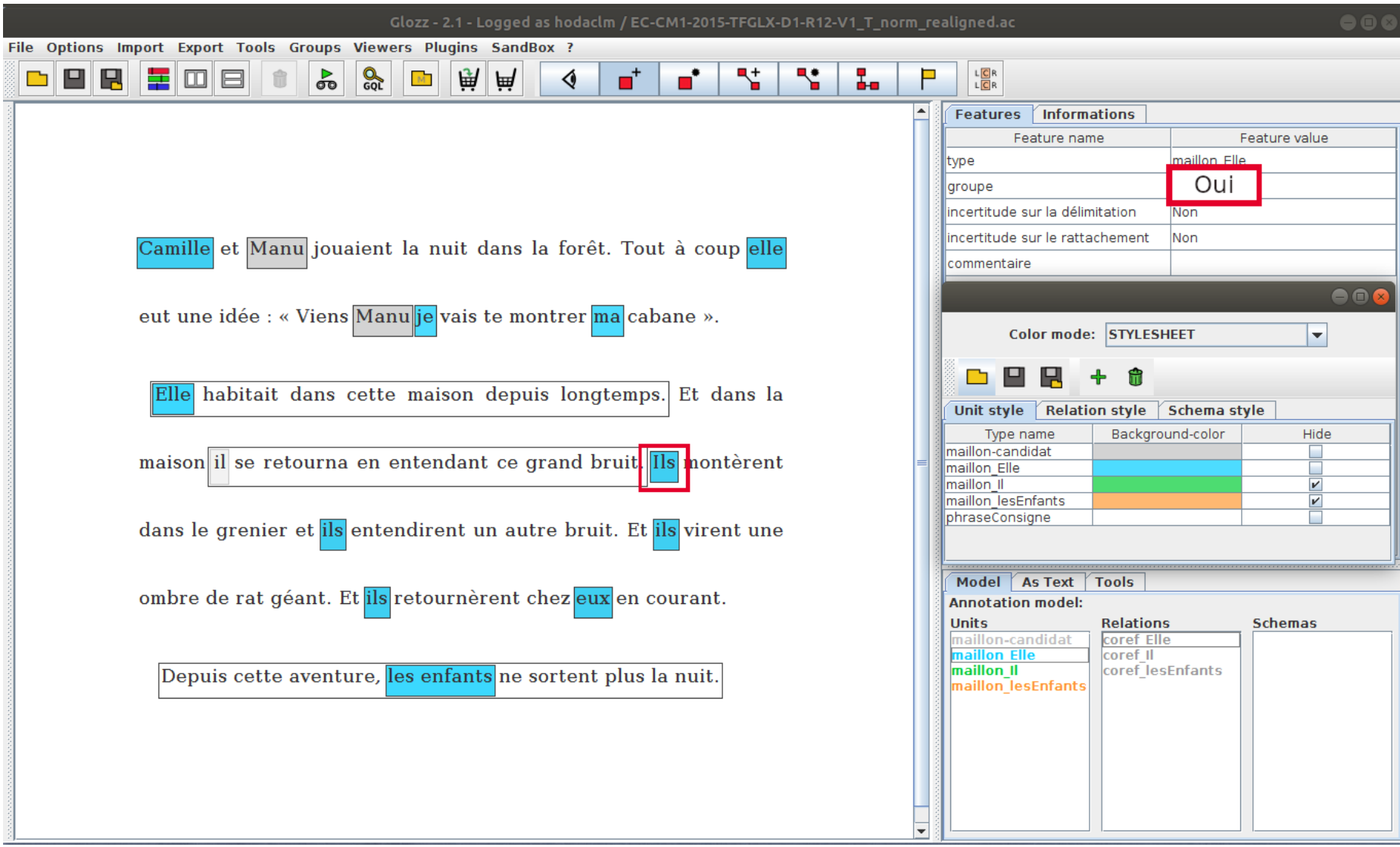
Étape 0 : Annotation outillée



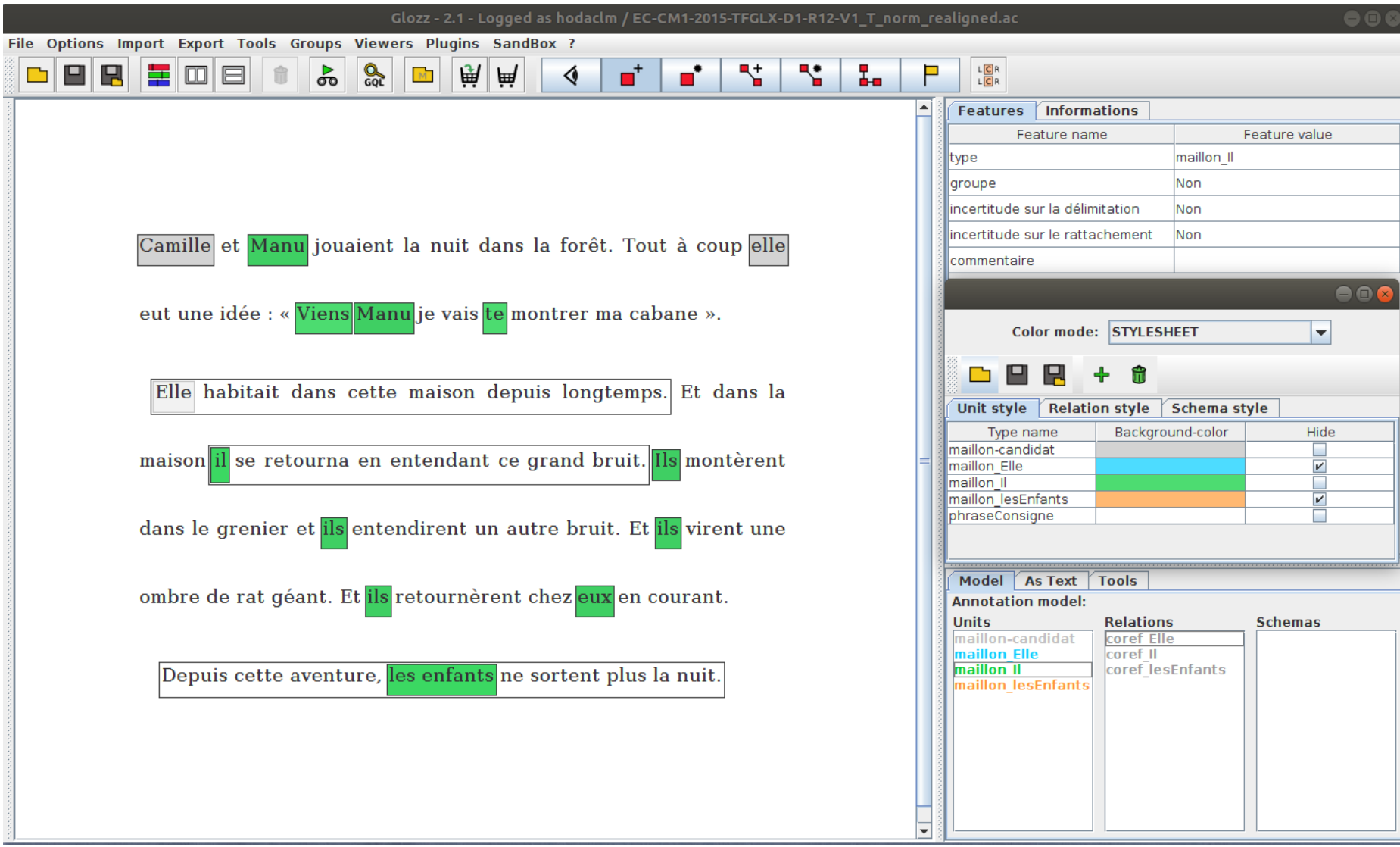
Étape 1.1 : Identification des maillons_Elle (ici, discours direct)



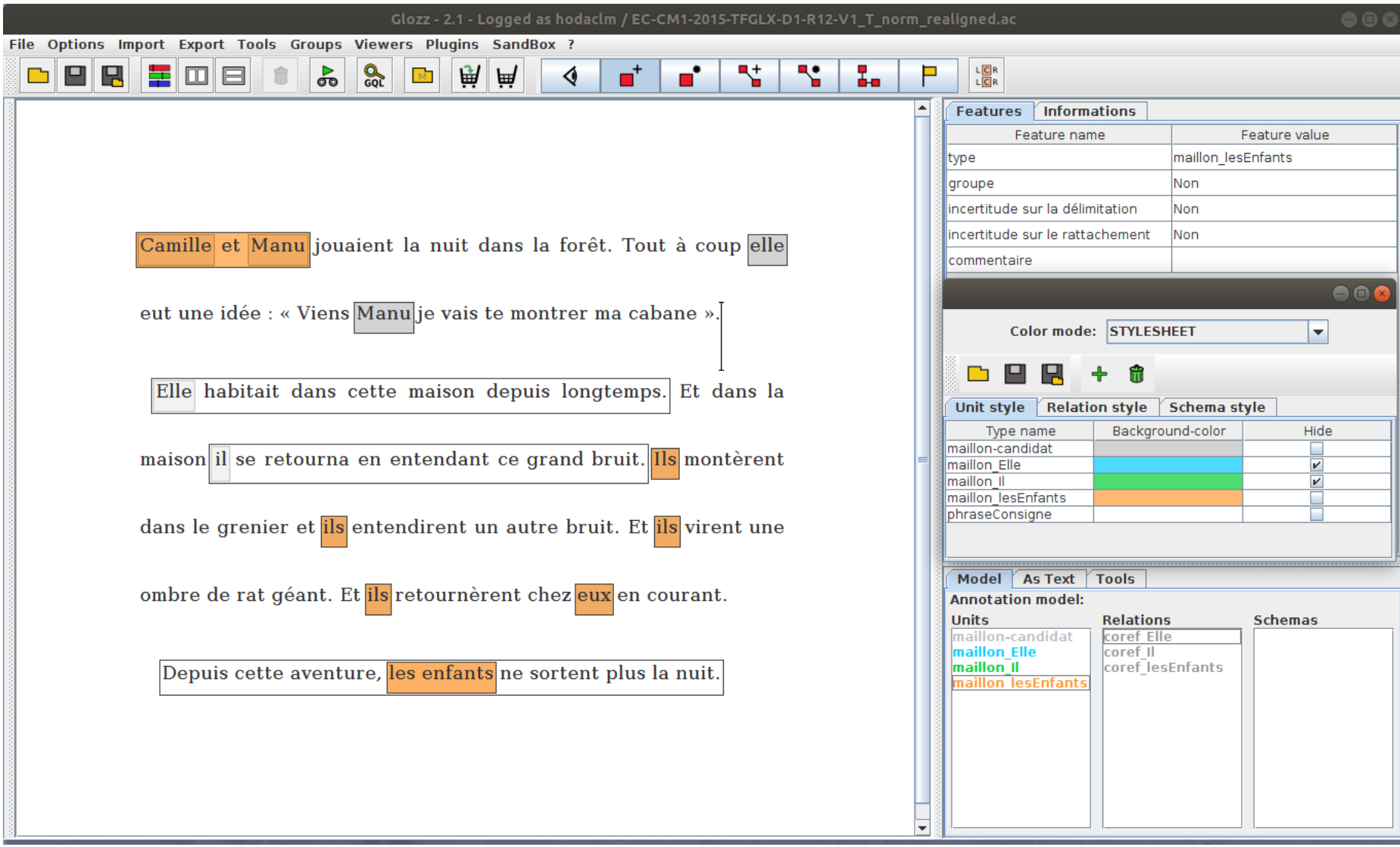
Étape 1.1 : Transition d'un *Elle* individuel à un *Elle* groupe



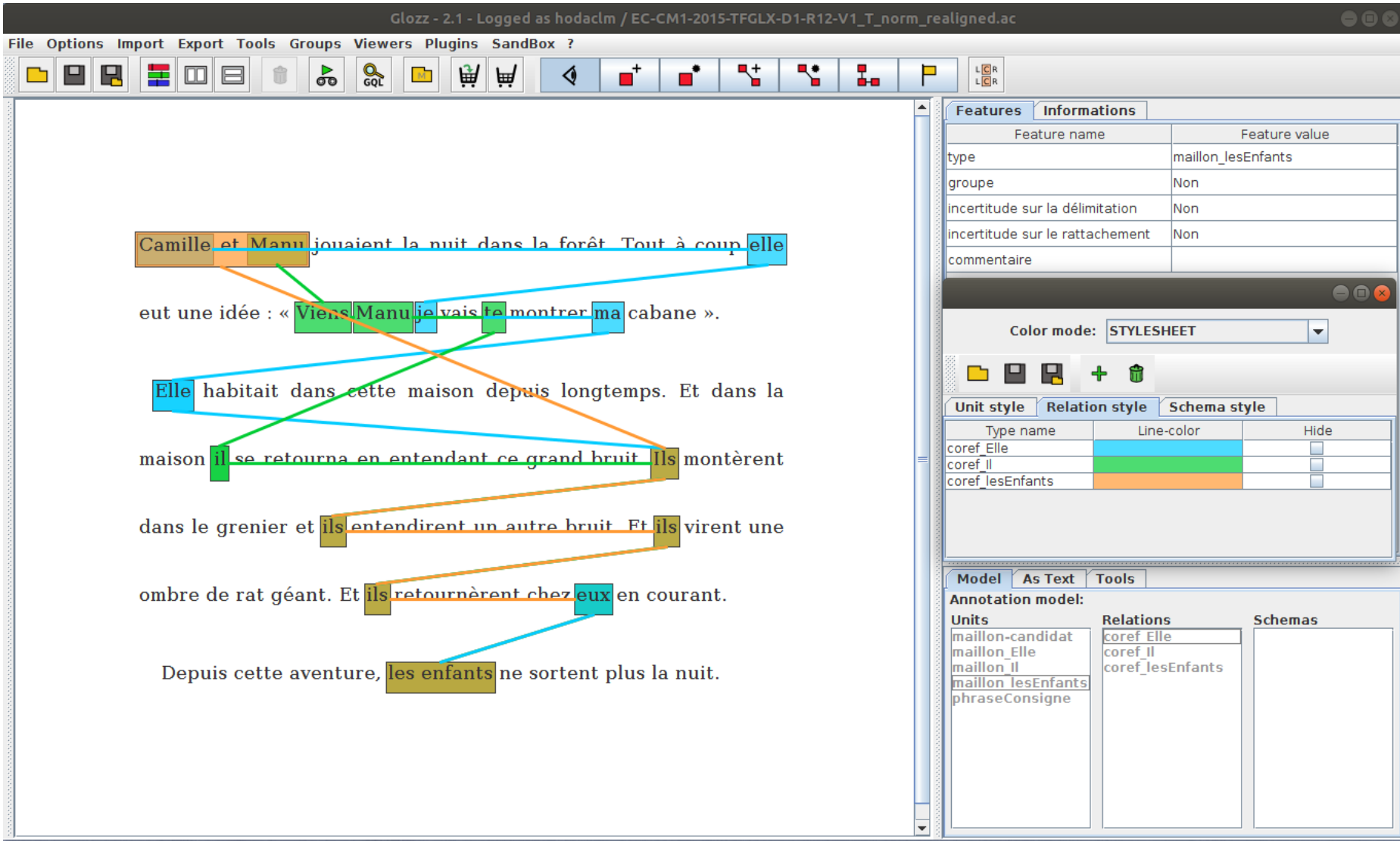
Étape 1.2 : Identification des maillons_Il (indiv. + groupe)



Étape 1.3 : Identification des maillons_lesEnfants



Étape 1.3 : Vérification par matérialisation des continuités



Faire "parler" les annotations obtenues

Avertissement

- Ce n'est pas parce qu'on a tout annoté qu'on a analysé et qu'on a des réponses à nos questions de recherche
- Nécessite de mettre en oeuvre, d'imaginer des méthodes pour interroger les données, les annotations obtenues
- Besoin de compétences en statistique (decrire les données, les préparer, avoir conscience des biais, ...)
- Des premiers résultats souvent décevants car
  - évidents
  - parfois réducteurs

Premier échantillon d'annotations dans le corpus RÉSOLCO

P1 Elle habitait dans cette maison depuis longtemps.

P2 Il se retourna en entendant ce grand bruit.

P3 Depuis cette aventure, les enfants ne sortent plus la nuit.

Longueur moyenne des 3 chaînes en nombre de maillons

		Maillons_Elle/texte			Maillons_Il/texte			Maillons_lesEnf./texte		
	nb. textes	moy	min	max	moy	min	max	moy	min	max
CE2	21	8	1	22	8	1	19	4	1	8
6e	24	12	3	21	11	0	28	5	0	22
3e	21	14	2	30	13	1	35	7	1	21
Master	40	23	4	76	22	6	63	10	1	22
	106	**16**	**1**	**76**	**15**	**0**	**63**	**7**	**0**	**22**

- Une augmentation du nombre de maillons avec le niveau scolaire (et du nombre de mots par texte)
- 2x plus de maillons pour les chaînes *Elle Il* que pour les chaînes *les enfants*. Jusqu'à **76** maillons
- **X** Très peu de textes sans introduction et/ou maintien des personnages (**1** seule mention i.e. la bandelette –) sauf pour *les enfants* (3% *Elle*, 8% *Il* vs. 23%) → incompréhension de la consigne ?

Annotations dans le corpus RÉSOLCO

Longueur moyenne des 3 chaînes en nombre de maillons

	nb. textes	Maillons_Elle/texte	Maillons_Il/texte	Maillons_lesEnf./texte
CE2	53	9	7	6
CM1	50	7	8	5
CM2	57	12	13	7
6e	106	13	13	8
4e	45	15	18	13
3e	54	17	18	10
M2	42	19	20	14
TOTAL	407	13	14	8

Des référents-personnages seuls et en groupe

Annotation des transitions d'un individu au groupe

> *Les pluriels et les groupes d'individus impliqués construisent des entités du discours au même titre qu'une expression référant à un individu unique. [...] Annoter un pluriel permet[tant] ainsi de* ***modéliser de manière exhaustive les transitions référentielles d'un individu au groupe auquel il appartient et inversement****.*
>
> *projet MC4 [Landragin, 2011]*

Une diversité de maillons de type "groupe"

maillon_Elle	maillon_ElleGroupe	
Camille	*Camille et Manu*	coordination
elle	*la bande de copines*	collectif
maillon_Il	maillon_IlGroupe	
il	*les deux frères*	pluriel
Pierre	*l'un d'eux*	indéfini + pronom

Des référents-personnages seuls et en groupe

Une consigne qui amène à imaginer *Elle* et/ou *Il* faisant partie d'un groupe
Plusieurs interprétations (et donc procédures) possibles :

- *Camille* et *Manu* sont *les enfants* qui ne sortent plus la nuit
- Beaucoup de possibilités e.g. *Elle* est parent et *Il* est un des enfants :

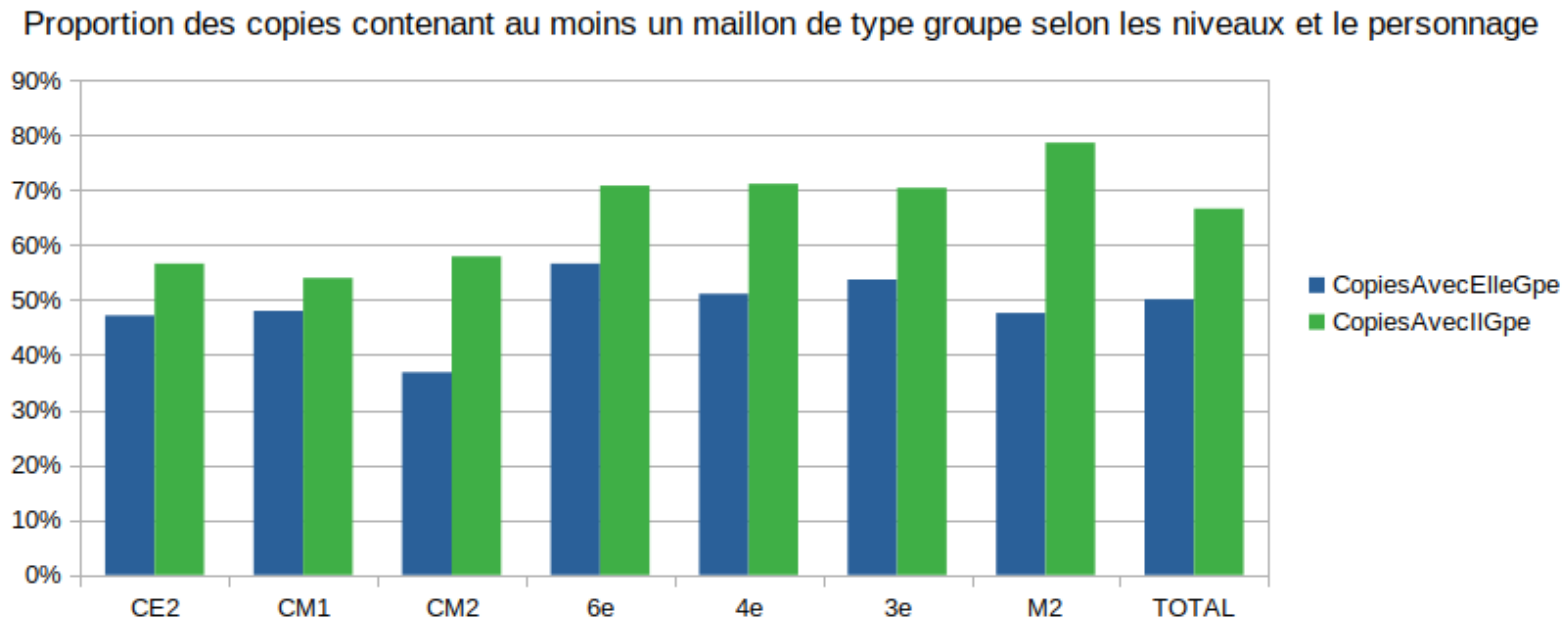
Il y a une mère qui sortait souvent le chien. Elle habitait dans cette maison depuis longtemps. C'était une grande maison avec une multitude de couleurs. On voyait aussi souvent les enfants jouer dehors, l'un s'appelait Pierre et l'autre Gile. Une nuit Pierre se réveilla et il avait faim. Il descendit donc pour aller à la cuisine. Il se retourna en

Copie de 3e

Fréquence des continuités avec groupes selon le niveau

Un recours à des maillons "groupe" qui varie plus selon le personnage que selon le niveau scolaire

- $>$ 50% des copies montrent une intégration de *Elle* ou *Il* dans un groupe
- Plus pour les chaînes *Il* que pour les chaînes *Elle*
  (e.g. exemple *Il y a une mère qui... les enfants ... Pierre*)



Fréquence des continuités avec groupes selon le niveau

Une proportion de maillons "groupe" qui varie plus selon le personnage que selon le niveau scolaire (malgré la différence de longueur des chaînes)

- $>$ 15% des maillons *Elle* sont de type groupe
- $>$ 25% des maillons *Il* sont de type groupe

	nbTextes	% *Elle*_groupe	(Max)	% *Il*_groupe	(Max)
CE2	21	17	(50)	34	(83)
6e	24	17	(50)	25	(85)
3e	21	14	(80)	30	(83)
M2	40	15	(68)	28	(67)
TOTAL	106	16	(80)	29	(85)

- Des annotations fréquemment associées à des incertitudes de rattachement (12 % des maillons de type groupe chez les 6e)

Deuxième échantillon et premières esquisses des stratégies de gestion des référents

- Quels indices pour étudier les stratégies mises en place par les élèves pour gérer la continuité référentielle (et la consigne RésolCo) ?
- La place de la première mention du référent de *elle*, *il* et *les enfants* comme indice des stratégies rédactionnelles dominantes
- Deux configurations à l'étude :

Elle<P1<Il<P2 Stratégie de type "pas à pas" : l'élève instancie un personnage susceptible d'être repris par le pronom *elle* dans P1. Lorsque l'élève abandonne ce personnage pour introduire un candidat à une reprise par *il* dans P2, il se trouve en situation de gérer chaque référent-personnage l'un après l'autre, ce qui évite la gestion en simultanée de 2 référents.

Elle&Il<P1 Stratégie de type "planification" : l'élève instancie les deux personnages susceptibles d'être repris par les pronoms *elle* et *il* avant P1. Cette configuration signifie que l'élève a anticipé les 3 phrases consignes et a choisi d'introduire dès le début tous les personnages du récit, ce qui peut compliquer la gestion des continuités référentielles.

TABLE – Répartition des textes annotés selon les niveaux scolaires et en fonction du point d'insertion de la première mention du référent de *elle* et de *il*

	nb textes	[Elle<P1<Il<P2]		[Elle&Il<P1]	
		nb textes	%	nb textes	%
CE2	21	7	33	6	29
6e	41	18	44	11	27
3e	41	10	24	16	39
Total	103	35	34	33	32

- Préférence des 6e pour une stratégie pas à pas
- Préférence des 3e pour une stratégie de planification
- Les CE2 hors jeu (dans 38% de textes de CE2 Elle=P1, contre 20% en 6e et 22% en 3e)

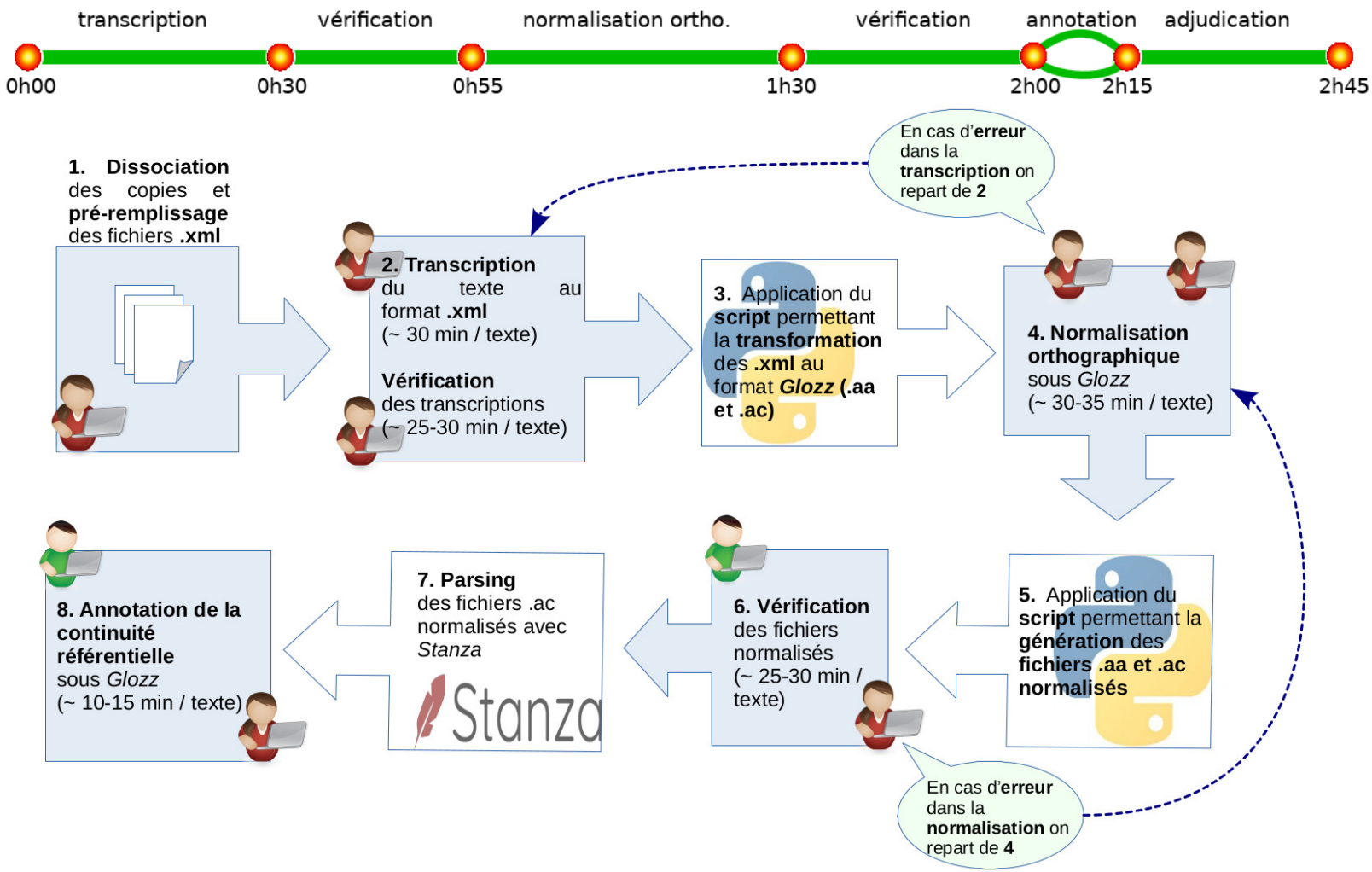
Plan

Pour conclure

Une usine à gaz qui en vaut la peine

- Permet d'être rigoureux et objectif (ne pas juger la copie, d'où la nécessité de travailler sur la version normée)
- Permet une analyse contrastive
  - Par niveaux scolaires, profil d'élèves, profil de copies selon d'autres critères (nb. de ratures, d'erreurs d'orthographe, longueur des phrases, etc.)
  - Pour d'autres textes (textes d'experts, narrations selon d'autres consignes, textes rédigés hors classe, etc.)
- Permet la reproductibilité : bonnes pratiques, guides, et nouvelles données
- Assure la diffusion et pérennité de la ressource
- Assure l'inter-opérabilité de la ressource et permettant ainsi l'application d'autres modèles d'annotation, de modules TAL et de méthodes statistiques partagées

Mais tout de même un processus de constitution chronophage



Un modèle d'annotation de la continuité référentielle qui semble fonctionner

- Annotation ascendante de la continuité référentielle
  - Focalisée sur les référents principaux
  - Permet de tisser des liens entre des mentions sans référence stricte
- Transposable à d'autres productions avec des consignes d'écriture différentes
- Encore beaucoup d'analyses à imaginer et réaliser

Une ressource à valoriser (dernière étape du projet E :CALM

- Valoriser auprès des chercheurs... démo du site
- Faire connaître au grand public, aux enseignants et formateurs d'enseignants
  - erreurs fréquentes (persistantes) selon le niveau scolaire
  - construction de la phrase graphique au fil de la scolarité
  - outil d'évaluation de la cohérence et notamment de la gestion de plusieurs référents dans un même récit
- http://redac.univ-tlse2.fr/corpus/resolco/


Asher, N., Muller, P., Bras, M., Ho-Dac, L. M., Benamara, F., Afantenos, S., and Vieu, L. (2017).
Annodis and related projects : Case studies on the annotation of discourse structure.
In Ide, N. and Pustejovsky, J., editors, *Handbook of Linguistic Annotation*, pages 1241–1264. Springer Netherlands, Dordrecht.


Doquet, C., David, J., Fleury, S., and (Eds) (2017).
Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement.
In *Corpus [Online]*, volume 16 (Special Issue). OpenEdition.


Garcia-Debanc, C. (2016).
Une tâche problème pour analyser les compétences d'élèves de sixième en matière de cohésion textuelle.
In L., S., D., V., and (coord)., C. B., editors, *Connexion et indexation. Ces liens qui tissent le texte*, Connexion et indexation. Ces liens qui tissent le texte, pages 263–278. ENS Editions.


Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., and Rebeyrolle, J. (2017).
Vers l'annotation discursive de textes d'élèves.
Corpus, 16.


Givón, T. (1983).
Topic continuity in discourse : an introduction.
In T.Givon, editor, *Topic continuity in discourse : a quantitative cross-language study*, pages 1–42. John Benjamins : Amsterdam/Philadelphia.


Halliday, M. and Hasan, R. (1976).
Cohesion in English.
Longman : London.


Ide, N. (2017).
Introduction to the handbook of linguistic annotation.
In Ide, N. and Pustejovsky, J., editors, *Handbook of Linguistic Annotation*, pages 1–18. Springer Netherlands, Dordrecht.


Landragin, F. (2011).
Une procédure d'analyse et d'annotationdes chaînes de coréférence dans des textes écrits.
Corpus, (10) :61–80.


Landragin, F. (2015).
Description, modélisation et détection automatique des chaînes de référence (democrat).
Bulletin de l'AFIA, 92 :11–15.


Péry-Woodley, M.-P., Afantenos, S. D., Ho-Dac, L.-M., and Asher, N. (2011).
La ressource annodis, un corpus enrichi d'annotations discursives.
Traitement Automatique des Langues, 52(3) :71–101.


Werth, P. (1999).
Text worlds : Representing conceptual space in discourse.
Longman : London.