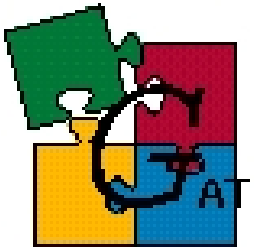
L'ingénierie de la langue avec GATE

Luc Plamondon

RALI/DIRO

Université de Montréal, 2004

Qu'est-ce que GATE



ATE — General Architecture for Text Engineering

Environnement pour faire de l'ingénierie linguistique

1996 : version 1
2003 : version 2
2004 : version 3 prévue

Ingénierie linguistique ?

- Développement d'applications de TAL (expérimentales/commerciales)

- Résultats mesurables et répétables
  - outils de mesure
  - portabilité
  - documentation
  - facilité d'utilisation

- Processus de conception mesurable et répétable
  - coûts

Environnement de développement ?

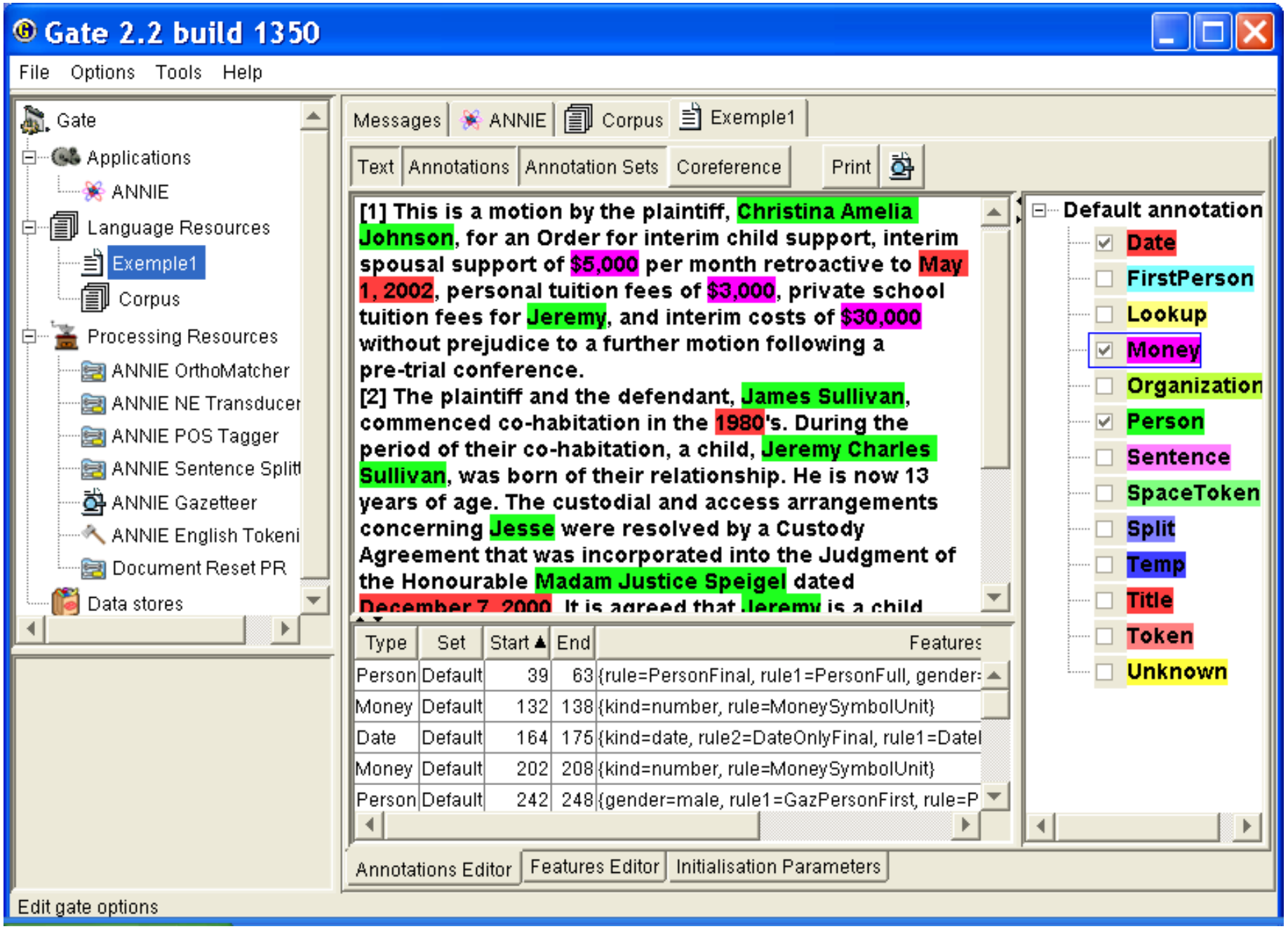
- Architecture
  - une application = pipeline de modules autonomes
  - passages des infos entre modules : annotations

- Bibliothèque de classes Java
  - objets de base en TAL : doc, corpus, annotation, tâche
  - ... et les fonctions courantes s'y rattachant

- Interface graphique
  - étiqueter manuellement un corpus de référence
  - jongler avec les modules
  - visualiser et comparer les résultats de l'application

Possibilités

- Extraction d'information (anglais, bengali)
- Recherche d'information (Lucene)
- Bases de données (Oracle, PostgresQL)
- Ontologies (DAML+OIL)
- Bibliothèques numériques (MUMIS, Perseus)
- Web sémantique (SEKT)
- Apprentissage automatique pour classification (WEKA)
- Constitution de corpus (EMILLE : 65 millions de mots de langues sud-asiatiques)
- Anonymisation (RALI)
- Question-réponse (RALI)
- Résumé automatique (RALI)


Gate 2.2 build 1350
File Options Tools Help
Gate
Applications
ANNIE
Language Resources
Exemple1
Corpus
Processing Resources
ANNIE OrthoMatcher
ANNIE NE Transducer
ANNIE POS Tagger
ANNIE Sentence Splitt
ANNIE Gazetteer
ANNIE English Tokeni
Document Reset PR
Data stores
Messages
ANNIE
Corpus
Exemple1
Text
Annotations
Annotation Sets
Coreference
Print
[1] This is a motion by the plaintiff, Christina Amelia Johnson, for an Order for interim child support, interim spousal support of $5,000 per month retroactive to May 1, 2002, personal tuition fees of $3,000, private school tuition fees for Jeremy, and interim costs of $30,000 without prejudice to a further motion following a pre-trial conference.
[2] The plaintiff and the defendant, James Sullivan, commenced co-habitation in the 1980's. During the period of their co-habitation, a child, Jeremy Charles Sullivan, was born of their relationship. He is now 13 years of age. The custodial and access arrangements concerning Jesse were resolved by a Custody Agreement that was incorporated into the Judgment of the Honourable Madam Justice Speigel dated
Type Set Start End Features
Person Default 39 63 {rule=PersonFinal, rule1=PersonFull, gender=
Money Default 132 138 {kind=number, rule=MoneySymbolUnit}
Date Default 164 175 {kind=date, rule2=DateOnlyFinal, rule1=Datel
Money Default 202 208 {kind=number, rule=MoneySymbolUnit}
Person Default 242 248 {gender=male, rule1=GazPersonFirst, rule=P
Default annotation
Date
FirstPerson
Lookup
Money
Organization
Person
Sentence
SpaceToken
Split
Temp
Title
Token
Unknown
Annotations Editor
Features Editor
Initialisation Parameters
Edit gate options

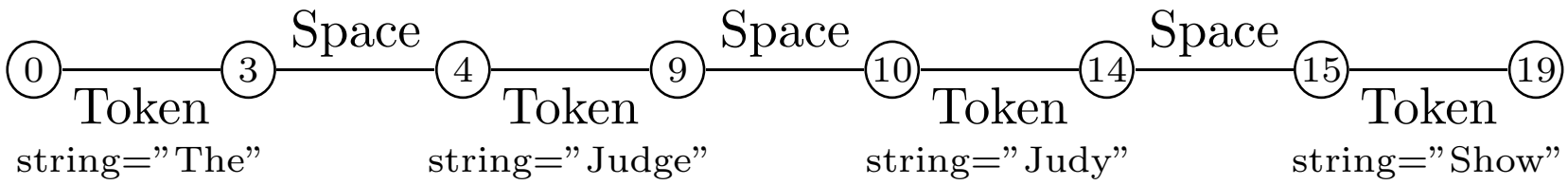
Graphe des annotations

3 façons de créer des annotations :

- Balises du fichier d'origine
- Accès direct aux caractères du texte (tokeniseur, gazetteer)
- Manipulation des annotations existantes : grammaire JAPE

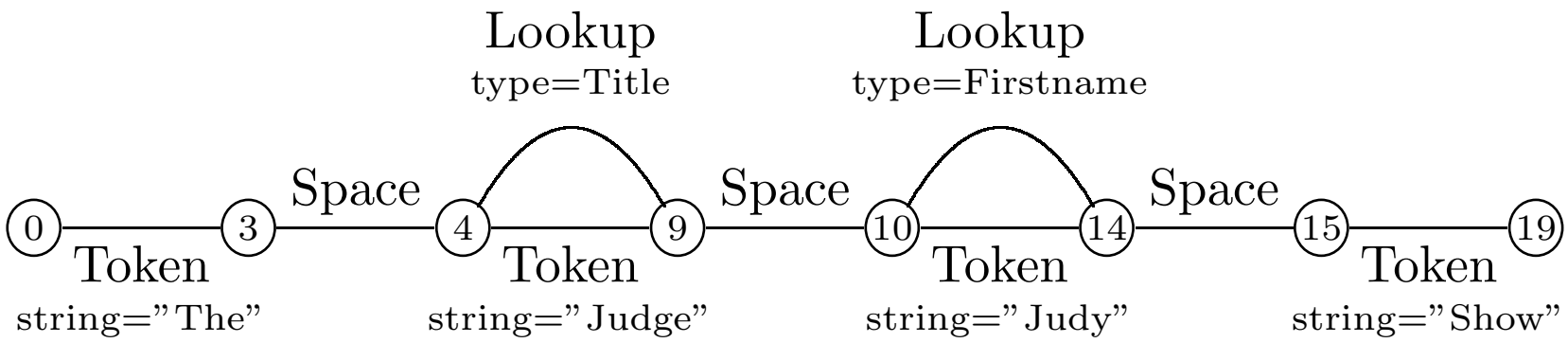
Graphe des annotations
Accès direct au texte

The *Judge* *Judy* *Show*



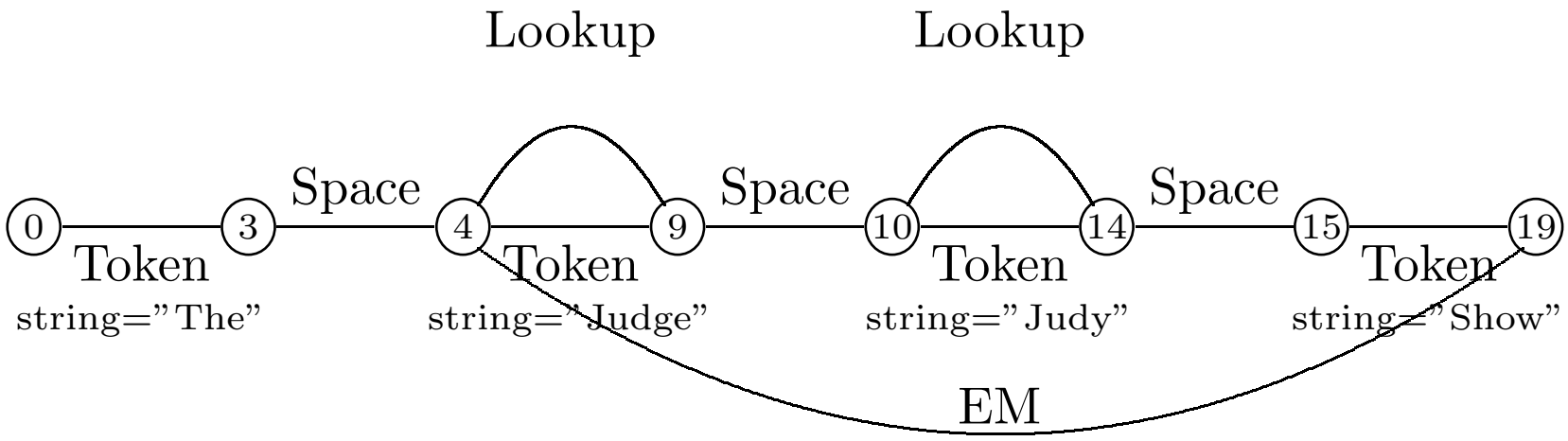
Graphe des annotations
Accès direct au texte

The *Judge* *Judy* *Show*



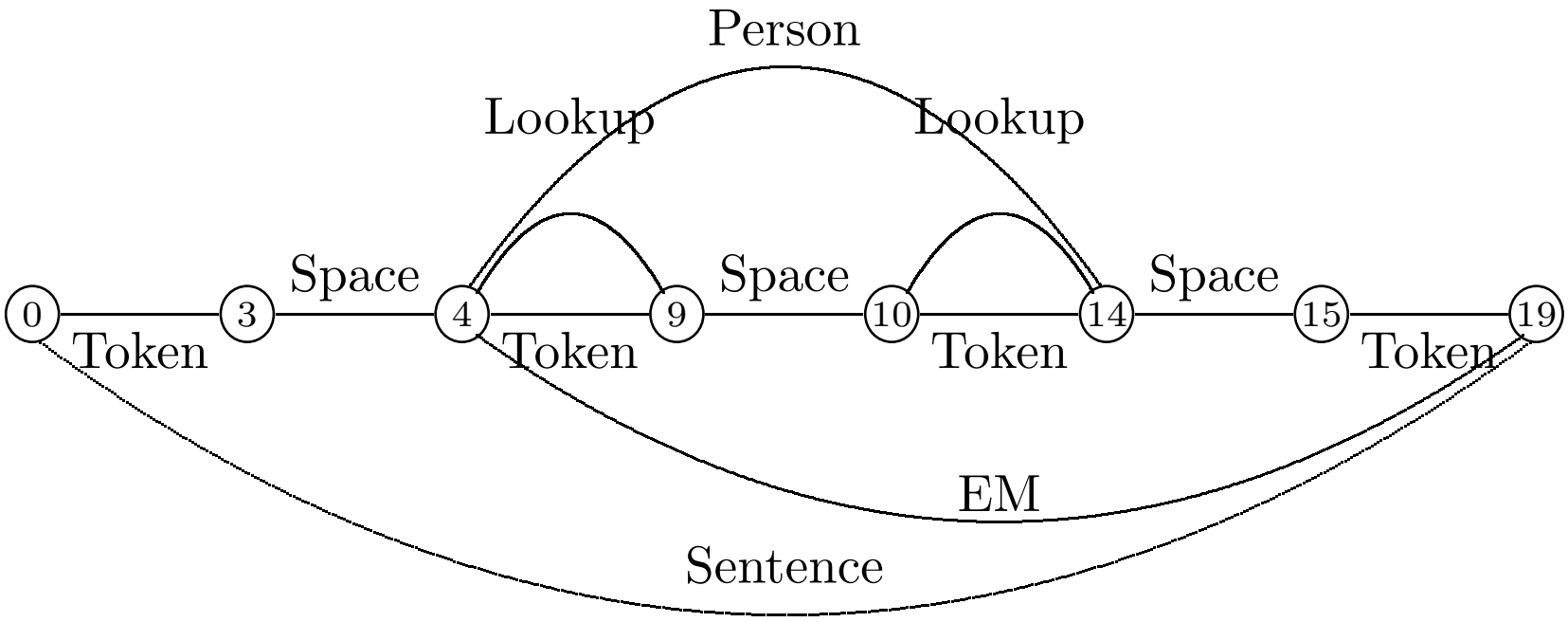
Graphe des annotations
Balises d'origine

The *Judge* *Judy* *Show*



Graphe des annotations
JAPE sur Annotations existantes

The *Judge* *Judy* *Show*



Langage JAPE

- Dérivé de CPSL (Common Pattern Specification Language), issu de Tipster, Doug Appelt et coll.

- Règles portent sur les annotations, pas sur le texte

- 3 types de déclenchement des règles :
  - first : 1re du fichier qui s'applique
  - brill : chacune qui s'applique
  - appelt : couvre la plus longue portion du doc, mécanisme de priorité pour les ambiguïtés

- Code Java dans la partie droite

Grammaire

```
Phase: Name
Input: Token Lookup
Options: control = appelt debug = false

Rule: PersonAndPerson
Priority: 30
// Anne and Kenton

( {Lookup.majorType == firstname} ) :person1
( {Token.string == "and"} )
( {Token.orth == upperInitial} ) :person2
-->
:person1.Person = {rule = "PersonAndPerson"},
:person2.Person = {rule = "PersonAndPerson"}
```

Grammaire avec code Java

```
Rule: TitleGender
Priority: 50
(
 {Lookup.majorType == title, Lookup.minorType == male}|
 {Lookup.majorType == title, Lookup.minorType == female}
):person
-->
{
 gate.AnnotationSet person = (gate.AnnotationSet)bindings.get("person");
 gate.Annotation personAnn = (gate.Annotation)person.iterator().next();
 gate.FeatureMap features = Factory.newFeatureMap();
 features.put("gender", personAnn.getFeatures().get("minorType"));
 annotations.add(person.firstNode(), person.lastNode(), "Title",
 features);
}
```

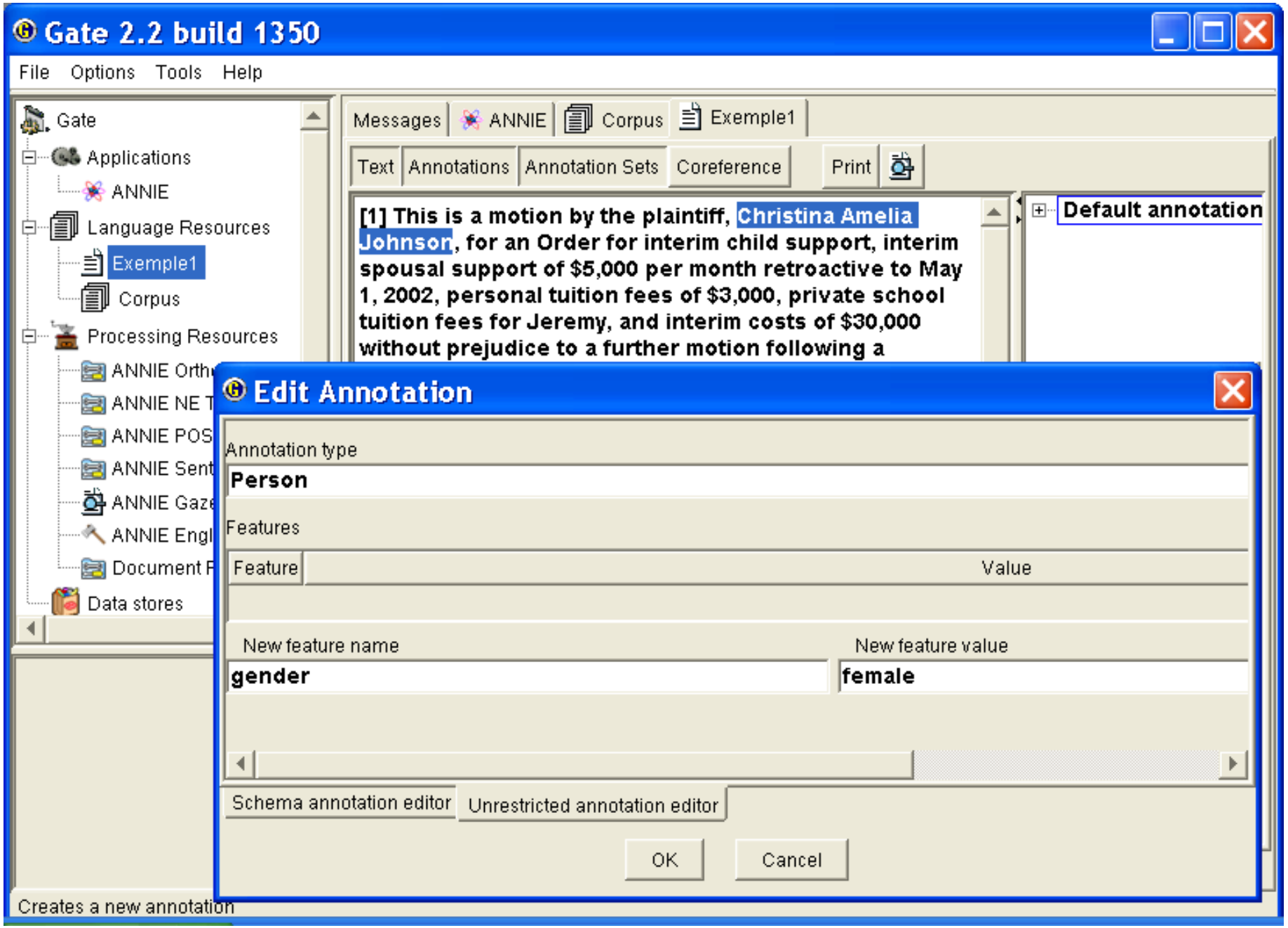
Cycle de développement d'une application de TAL

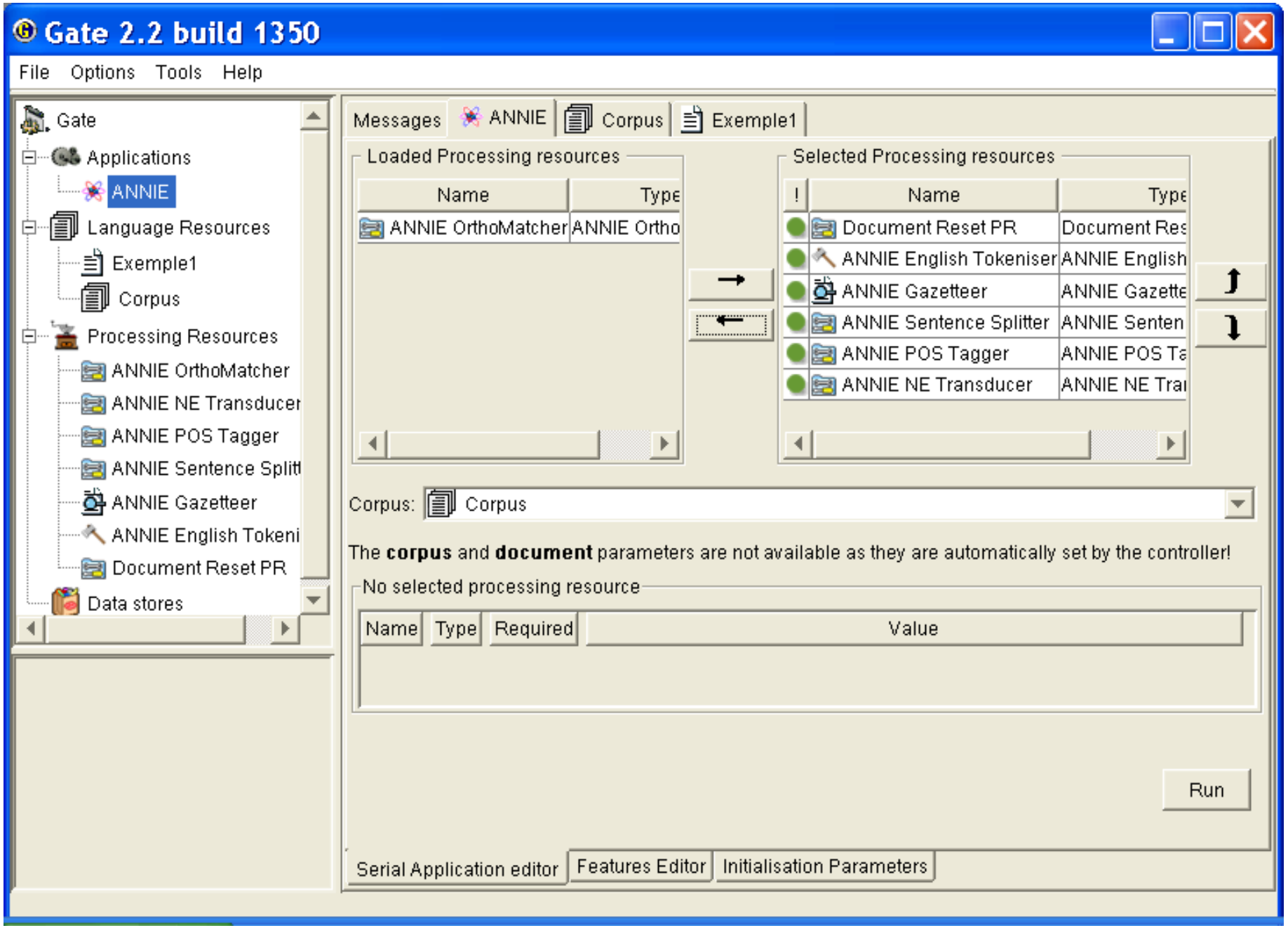
Exemple : extraction d'entités nommées

- Lecture des documents
- Annotation manuelle du corpus de référence
- Traitement du corpus de test
  - Tokenisation
  - Segmentation en phrases
  - Étiquetage grammatical
  - Étiquetage des entités nommées
- Visualisation des résultats
- Évaluation

Lecture des documents

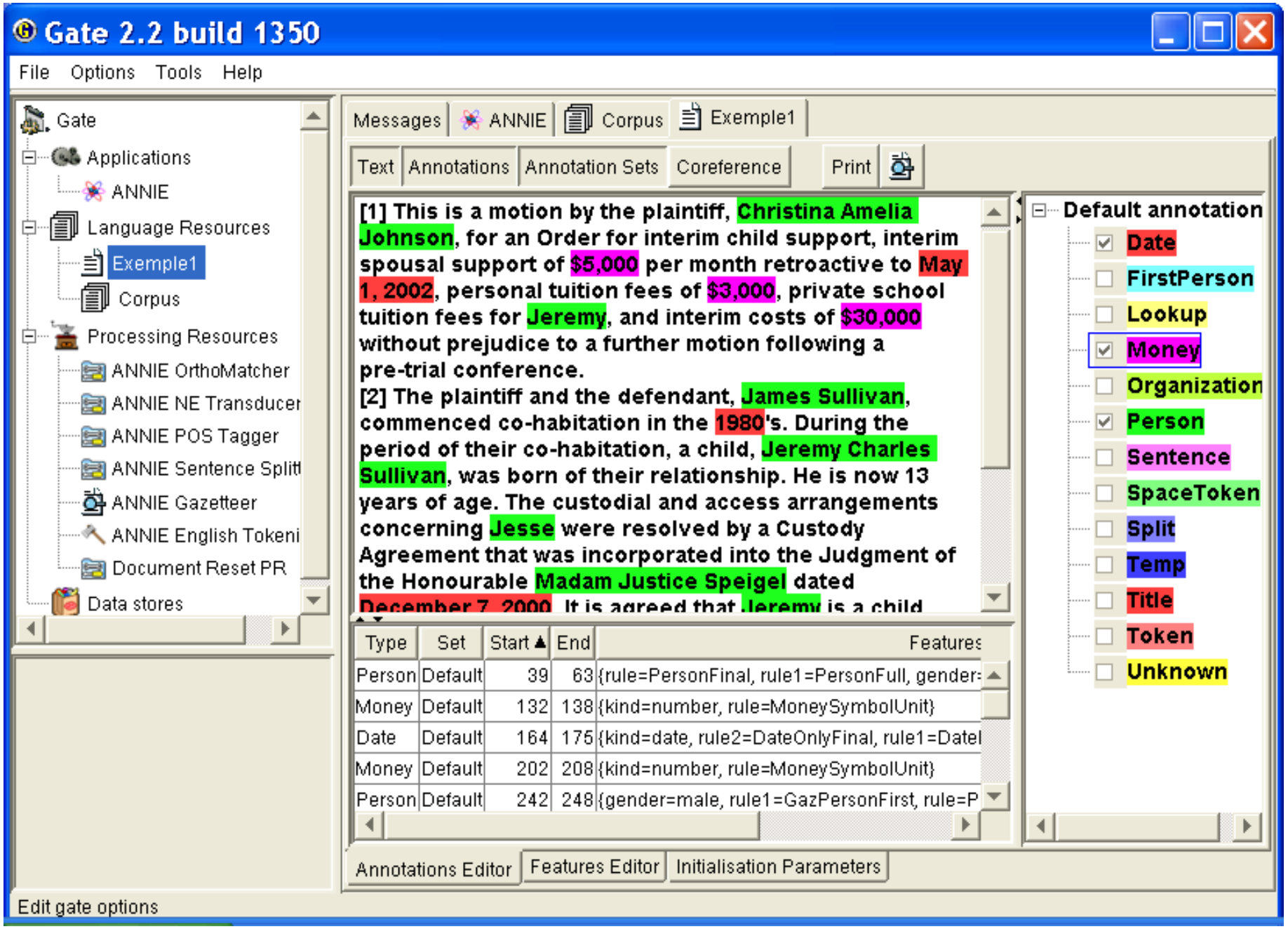
- Unicode
- texte régulier, XML, HTML, SGML, RTF, email
- Possibilité de garder les balises originales et d'en faire des annotations
- Écriture du document en XML seulement

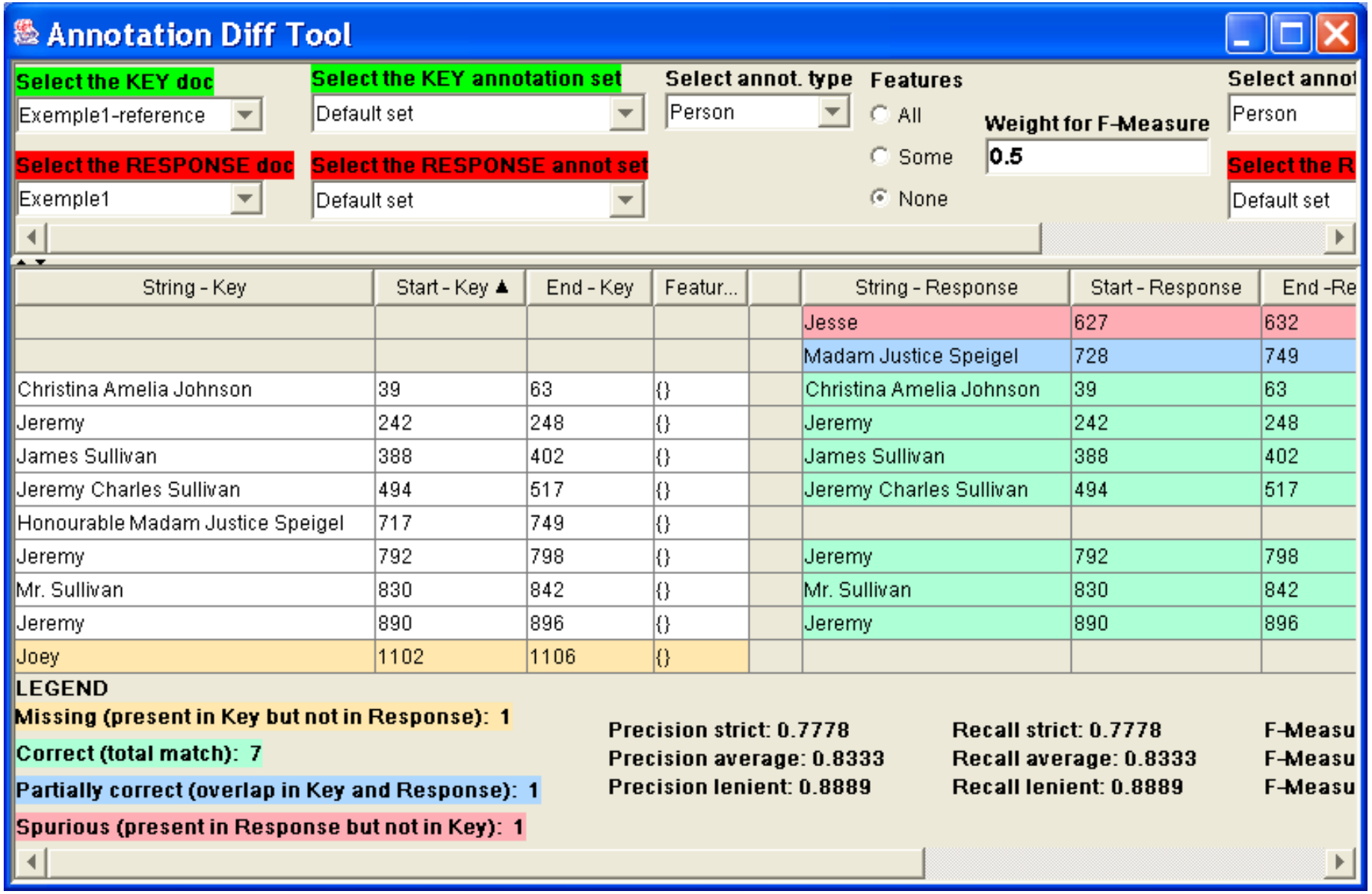

Gate 2.2 build 1350
File Options Tools Help
Gate
Applications
ANNIE
Language Resources
Exemple1
Corpus
Processing Resources
Data stores
Messages
ANNIE
Corpus
Exemple1
Text
Annotations
Annotation Sets
Coreference
Print
[1] This is a motion by the plaintiff, Christina Amelia Johnson, for an Order for interim child support, interim spousal support of $5,000 per month retroactive to May 1, 2002, personal tuition fees of $3,000, private school tuition fees for Jeremy, and interim costs of $30,000 without prejudice to a further motion following a
Default annotation
Edit Annotation
Annotation type
Person
Features
Feature
Value
New feature name
gender
New feature value
female
Schema annotation editor
Unrestricted annotation editor
OK
Cancel
Creates a new annotation


Gate 2.2 build 1350
File Options Tools Help
Gate
Applications
ANNIE
Language Resources
Exemple1
Corpus
Processing Resources
ANNIE OrthoMatcher
ANNIE NE Transducer
ANNIE POS Tagger
ANNIE Sentence Splitt
ANNIE Gazetteer
ANNIE English Tokeni
Document Reset PR
Data stores
Messages
ANNIE
Corpus
Exemple1
Loaded Processing resources
Name
Type
ANNIE OrthoMatcher
ANNIE Ortho
Selected Processing resources
!
Name
Type
Document Reset PR
Document Res
ANNIE English Tokeniser
ANNIE English
ANNIE Gazetteer
ANNIE Gazette
ANNIE Sentence Splitter
ANNIE Senten
ANNIE POS Tagger
ANNIE POS Ta
ANNIE NE Transducer
ANNIE NE Trar
Corpus:
Corpus
The corpus and document parameters are not available as they are automatically set by the controller!
No selected processing resource
Name
Type
Required
Value
Run
Serial Application editor
Features Editor
Initialisation Parameters

ANNIE

1. Tokenisation
   - Grammaire maison + grammaire JAPE
2. Segmentation en phrases
   - Lexique d'abréviations + grammaire JAPE
3. Étiquetage grammatical
   - Brill modifié (Hepple 2000)
4. Étiquetage sémantique
   - Lexique de villes, prénoms, etc.
   - 21 grammaires JAPE


Gate 2.2 build 1350
File Options Tools Help
Gate
Applications
ANNIE
Language Resources
Exemple1
Corpus
Processing Resources
ANNIE OrthoMatcher
ANNIE NE Transducer
ANNIE POS Tagger
ANNIE Sentence Splitt
ANNIE Gazetteer
ANNIE English Tokeni
Document Reset PR
Data stores
Messages
ANNIE
Corpus
Exemple1
Text
Annotations
Annotation Sets
Coreference
Print
[1] This is a motion by the plaintiff, Christina Amelia Johnson, for an Order for interim child support, interim spousal support of $5,000 per month retroactive to May 1, 2002, personal tuition fees of $3,000, private school tuition fees for Jeremy, and interim costs of $30,000 without prejudice to a further motion following a pre-trial conference.
[2] The plaintiff and the defendant, James Sullivan, commenced co-habitation in the 1980's. During the period of their co-habitation, a child, Jeremy Charles Sullivan, was born of their relationship. He is now 13 years of age. The custodial and access arrangements concerning Jesse were resolved by a Custody Agreement that was incorporated into the Judgment of the Honourable Madam Justice Speigel dated
Type Set Start End Features
Person Default 39 63 {rule=PersonFinal, rule1=PersonFull, gender=
Money Default 132 138 {kind=number, rule=MoneySymbolUnit}
Date Default 164 175 {kind=date, rule2=DateOnlyFinal, rule1=Datel
Money Default 202 208 {kind=number, rule=MoneySymbolUnit}
Person Default 242 248 {gender=male, rule1=GazPersonFirst, rule=P
Default annotation
Date
FirstPerson
Lookup
Money
Organization
Person
Sentence
SpaceToken
Split
Temp
Title
Token
Unknown
Annotations Editor
Features Editor
Initialisation Parameters
Edit gate options


Annotation Diff Tool
Select the KEY doc
Exemple1-reference
Select the KEY annotation set
Default set
Select annot. type
Person
Features
All
Some
None
Weight for F-Measure
0.5
Select annot
Person
Select the RESPONSE doc
Exemple1
Select the RESPONSE annot set
Default set
Select the R
Default set
String - Key
Start - Key
End - Key
Featur...
String - Response
Start - Response
End -Re
Jesse 627 632
Madam Justice Speigel 728 749
Christina Amelia Johnson 39 63 {} Christina Amelia Johnson 39 63
Jeremy 242 248 {} Jeremy 242 248
James Sullivan 388 402 {} James Sullivan 388 402
Jeremy Charles Sullivan 494 517 {} Jeremy Charles Sullivan 494 517
Honourable Madam Justice Speigel 717 749 {}
Jeremy 792 798 {} Jeremy 792 798
Mr. Sullivan 830 842 {} Mr. Sullivan 830 842
Jeremy 890 896 {} Jeremy 890 896
Joey 1102 1106 {}
LEGEND
Missing (present in Key but not in Response): 1
Correct (total match): 7
Partially correct (overlap in Key and Response): 1
Spurious (present in Response but not in Key): 1
Precision strict: 0.7778
Precision average: 0.8333
Precision lenient: 0.8889
Recall strict: 0.7778
Recall average: 0.8333
Recall lenient: 0.8889
F-Measu
F-Measu
F-Measu

Autres modules

- Coréférences orthographiques/pronominales/nominales
- Éditeur d'ontologies
- Segmenteur de groupes verbaux
- Interface graphique pour Wordnet
- Interface pour moteur de recherche Lucene
- Interface pour WEKA (apprentissage automatique de classificateurs)

Points forts

- Portabilité : Java, Unicode, XML
- Modularité : plug-n-play, favorise la réutilisation, facilite la mise en commun
- Un seul environnement graphique, un seul langage, une seule bibliothèque de classes, un seul format d'échange de données
- Bibliothèque de classes permet l'incorporation à d'autres programmes
- Code source libre et gratuit

Points faibles

- Java portable ?
- Langage de haut niveau : lent
- Vorace en mémoire
- Langage JAPE est verbeux, syntaxe peu flexible, pas de négation
- Bogues

Conclusion

- Enfin !
- Emphase sur la modularité, la flexibilité et la maintenabilité au détriment de l'efficacité
- Excellent pour les applications expérimentales, mais la lenteur et les besoins en mémoire limitent les perspectives commerciales
- Excellent outil pour l'apprentissage de l'informatique linguistique