Traitement Automatique des Langues

De la Traduction Automatique à la Recherche d'Information, quels apports pour l'industrie?

Antonio BALVET
Université Lille 3
Dir. adjt. UMR STL

antonio.balvet@univ-lille3.fr

Un peu d'histoire

- 1945 : naissance du Traitement Automatique des Langues
  - contexte de guerre froide
    - besoins de veille stratégique
  - émergence de l'Intelligence Artificielle
    - imiter la compréhension naturelle du langage humain
    - "décoder" des messages en russe vers l'anglais

De la TA à la TAO

- TA = Traduction Automatique
  - texte en langue A => texte en langue B
  - énormes difficultés car interprétation en contexte
    - exemple (Bar-Hillel) :

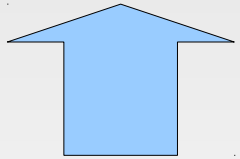
      the pen is in the box / the box is in the pen

      le stylo est dans la boîte / la boîte est dans le stylo? (l'enclos)
    - identification correcte des segments pertinents

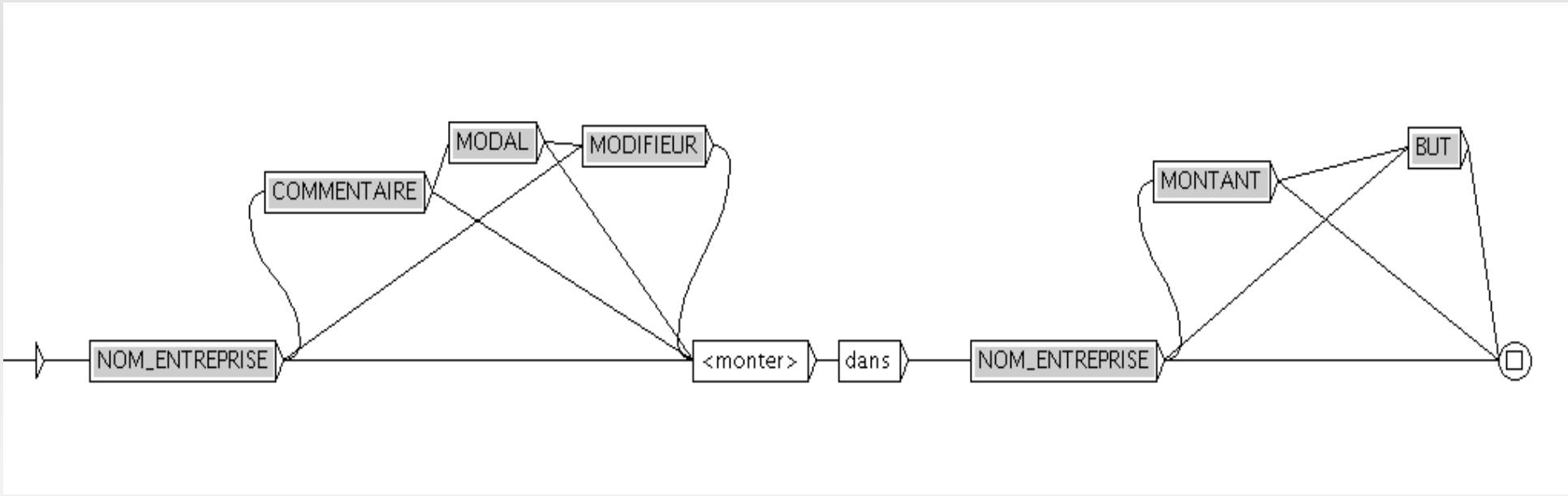
      [pomme de terre] cuite ou **pomme de [terre cuite]**?

      [Thales], après bien des hésitations, s'apprête enfin à **[monter dans] [Ingenico] [à hauteur de 60%]**

Nom_Entreprise COMMENTAIRE MODAL MODIFIEUR monter dans Nom_Entreprise MONTANT, BUT

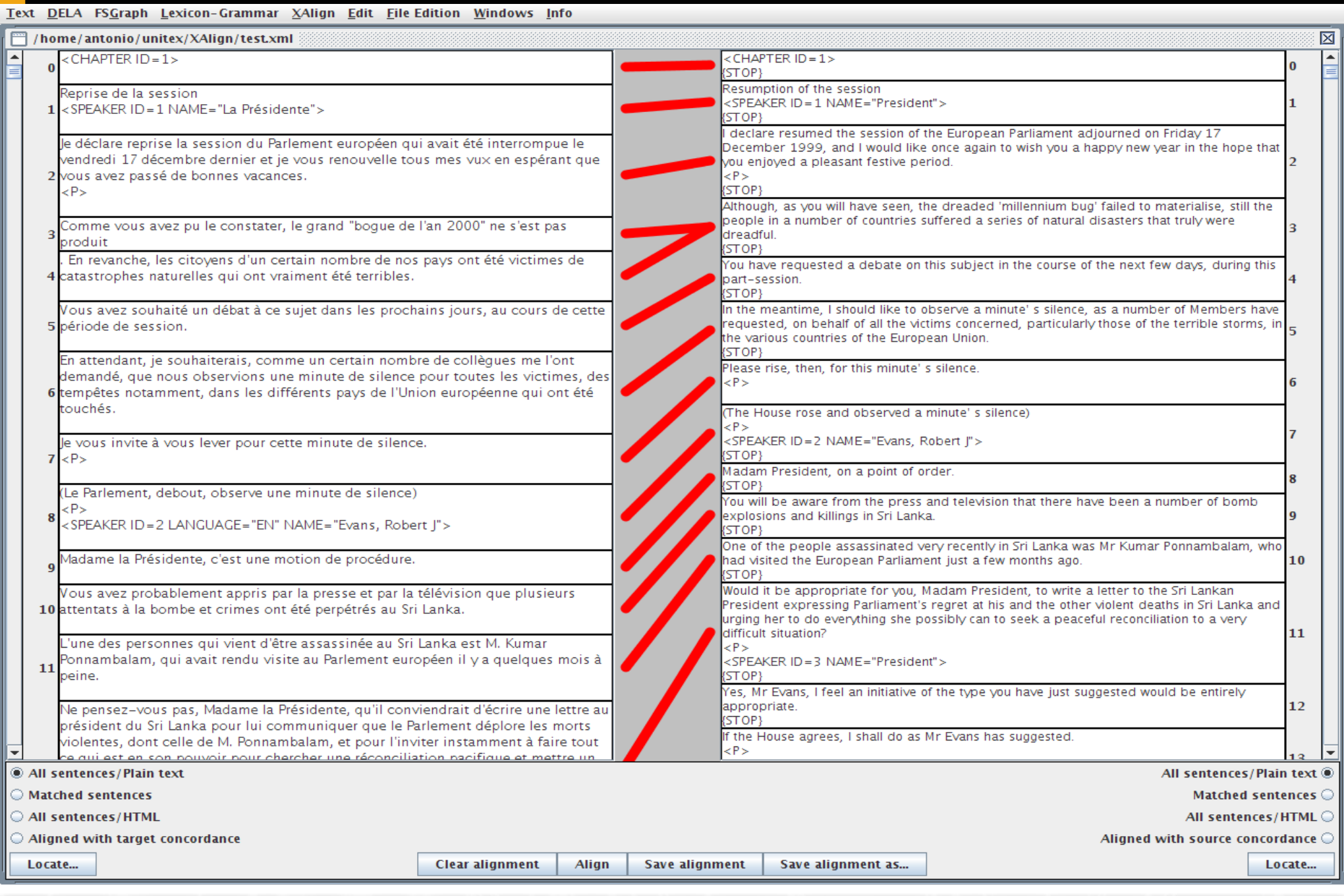


Thales, après bien des hésitations, s'apprête enfin à monter dans Ingenico à hauteur de 60%, pour renforcer sa position


MODAL
MODIFIEUR
COMMENTAIRE
NOM_ENTREPRISE
<monter>
dans
NOM_ENTREPRISE
MONTANT
BUT

De la TA à la TAO

- TAO = Traduction Assistée par Ordinateur
  - textes parallèles langue A/langue B => segments pertinents
    - calcul d'un alignement langue A/langue B
    - analyse de la structure des phrases alignées
    - identification de segments déjà traduits par consultation d'une mémoire de traduction (BDD)
    - détection de nouveaux mots-clés, termes, schémas de phrase => nouveaux segments


Text DELA FSGraph Lexicon-Grammar XAlign Edit File Edition Windows Info
/home/antonio/unitex/XAlign/test.xml
0 <CHAPTER ID=1>
1 Reprise de la session
<SPEAKER ID=1 NAME="La Présidente">
2 Je déclare reprise la session du Parlement européen qui avait été interrompue le vendredi 17 décembre dernier et je vous renouvelle tous mes vux en espérant que vous avez passé de bonnes vacances.
<P>
3 Comme vous avez pu le constater, le grand "bogue de l'an 2000" ne s'est pas produit
4 . En revanche, les citoyens d'un certain nombre de nos pays ont été victimes de catastrophes naturelles qui ont vraiment été terribles.
5 Vous avez souhaité un débat à ce sujet dans les prochains jours, au cours de cette période de session.
6 En attendant, je souhaiterais, comme un certain nombre de collègues me l'ont demandé, que nous observions une minute de silence pour toutes les victimes, des tempêtes notamment, dans les différents pays de l'Union européenne qui ont été touchés.
7 Je vous invite à vous lever pour cette minute de silence.
<P>
8 (Le Parlement, debout, observe une minute de silence)
<P>
<SPEAKER ID=2 LANGUAGE="EN" NAME="Evans, Robert J">
9 Madame la Présidente, c'est une motion de procédure.
10 Vous avez probablement appris par la presse et par la télévision que plusieurs attentats à la bombe et crimes ont été perpétrés au Sri Lanka.
11 L'une des personnes qui vient d'être assassinée au Sri Lanka est M. Kumar Ponnambalam, qui avait rendu visite au Parlement européen il y a quelques mois à peine.
Ne pensez-vous pas, Madame la Présidente, qu'il conviendrait d'écrire une lettre au président du Sri Lanka pour lui communiquer que le Parlement déplore les morts violentes, dont celle de M. Ponnambalam, et pour l'inviter instamment à faire tout ce qui est en son pouvoir pour chercher une réconciliation pacifique et mettre un
<CHAPTER ID=1>
{STOP}
0
Resumption of the session
<SPEAKER ID=1 NAME="President">
{STOP}
1
I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period.
<P>
{STOP}
2
Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the people in a number of countries suffered a series of natural disasters that truly were dreadful.
{STOP}
3
You have requested a debate on this subject in the course of the next few days, during this part-session.
{STOP}
4
In the meantime, I should like to observe a minute' s silence, as a number of Members have requested, on behalf of all the victims concerned, particularly those of the terrible storms, in the various countries of the European Union.
{STOP}
5
Please rise, then, for this minute' s silence.
<P>
6
(The House rose and observed a minute' s silence)
<P>
<SPEAKER ID=2 NAME="Evans, Robert J">
{STOP}
7
Madam President, on a point of order.
{STOP}
8
You will be aware from the press and television that there have been a number of bomb explosions and killings in Sri Lanka.
{STOP}
9
One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam, who had visited the European Parliament just a few months ago.
{STOP}
10
Would it be appropriate for you, Madam President, to write a letter to the Sri Lankan President expressing Parliament's regret at his and the other violent deaths in Sri Lanka and urging her to do everything she possibly can to seek a peaceful reconciliation to a very difficult situation?
<P>
<SPEAKER ID=3 NAME="President">
{STOP}
11
Yes, Mr Evans, I feel an initiative of the type you have just suggested would be entirely appropriate.
{STOP}
12
If the House agrees, I shall do as Mr Evans has suggested.
<P>
13
All sentences/Plain text
Matched sentences
All sentences/HTML
Aligned with target concordance
Locate...
Clear alignment
Align
Save alignment
Save alignment as...
All sentences/Plain text
Matched sentences
All sentences/HTML
Aligned with source concordance
Locate...

TAL et Recherche d'Information

- Approches linguistiques et statistiques, deux approches complémentaires
  - approches statistiques :
    - robustes, standardisées, composants de qualité industrielle (Lucene)
  - approches linguistiques :
    - non sensibles aux signaux faibles, ébauche de compréhension automatique de textes, composants de qualité industrielle

Fichier Édition Affichage Historique Marque-pages Outils Aide

http://www.google.fr/search?hl=fr&q=Poutine&btnG=Recherche+Google&meta=&aq=f&oq=

Getting Started Latest BBC Headlines

Web Images Maps Actualités Vidéo Gmail plus ▼ Connexion

Google

Poutine Rechercher Recherche avancée Préférences

Rechercher dans : Web Pages francophones Pages : France

Web Résultats **1** - **10** sur un total d'environ **1 620 000** pour **Poutine** (**0,26** secondes)

Vladimir Poutine - Wikipédia
Vladimir Vladimirovitch **Poutine**ru-Putin.ogg écouter (en russe : Влади́мир Влади́мирович Пу́тин) est né le 7 octobre 1952 à Léningrad (aujourd'hui ...
fr.wikipedia.org/wiki/Vladimir_**Poutine** - 174k - En cache - Pages similaires

Poutine (plat) - Wikipédia
La **poutine** désigne communément un mets d'origine québécoise traditionnellement constitué de frites et de fromage en grains de cheddar frais que l'on ...
fr.wikipedia.org/wiki/**Poutine**_(plat) - 49k - En cache - Pages similaires

Vladimir Poutine
2 déc 2007 ... Bibliographie / Biographie. Qui est Vladimir **Poutine** ?
www.republique-des-lettres.fr/10198-vladimir-**poutine**.php - 13k -
En cache - Pages similaires

Résultats de recherche de vidéos pour Poutine

Omnikrom avec TTC - Danse la **poutine**
4 min 21 s
www.youtube.com

Chirac décore **Poutine**, Arrêt sur image
3 min 50 s
www.dailymotion.com

Résultats dans l'Actualité pour Poutine

Le Point

Mireille Mathieu séduit **Poutine** et Kadhafi - Il y a 13 heures
Vladimir **Poutine** et Mouammar Kadhafi ont assisté samedi soir à un concert donné au Kremlin par la chanteuse française Mireille Mathieu à Moscou. ...
Europe1 - 242 autres articles »
Poutine: le Kazakhstan est l'un des plus proches alliés de la Russie - XINHUA - 5 autres articles »
La thérapie du groupe de Shanghaï - Courrier International - 118 autres articles »

La poutine : la recette
La recette 'La **poutine**' illustrée du Journal des Femmes : Faire des frites de taille moyenne et de préférence en utilisant le mode belge des deux cuissons.
www.linternaute.com/femmes/cuisine/recette/241405/5033532978/la_**poutine**.shtml -

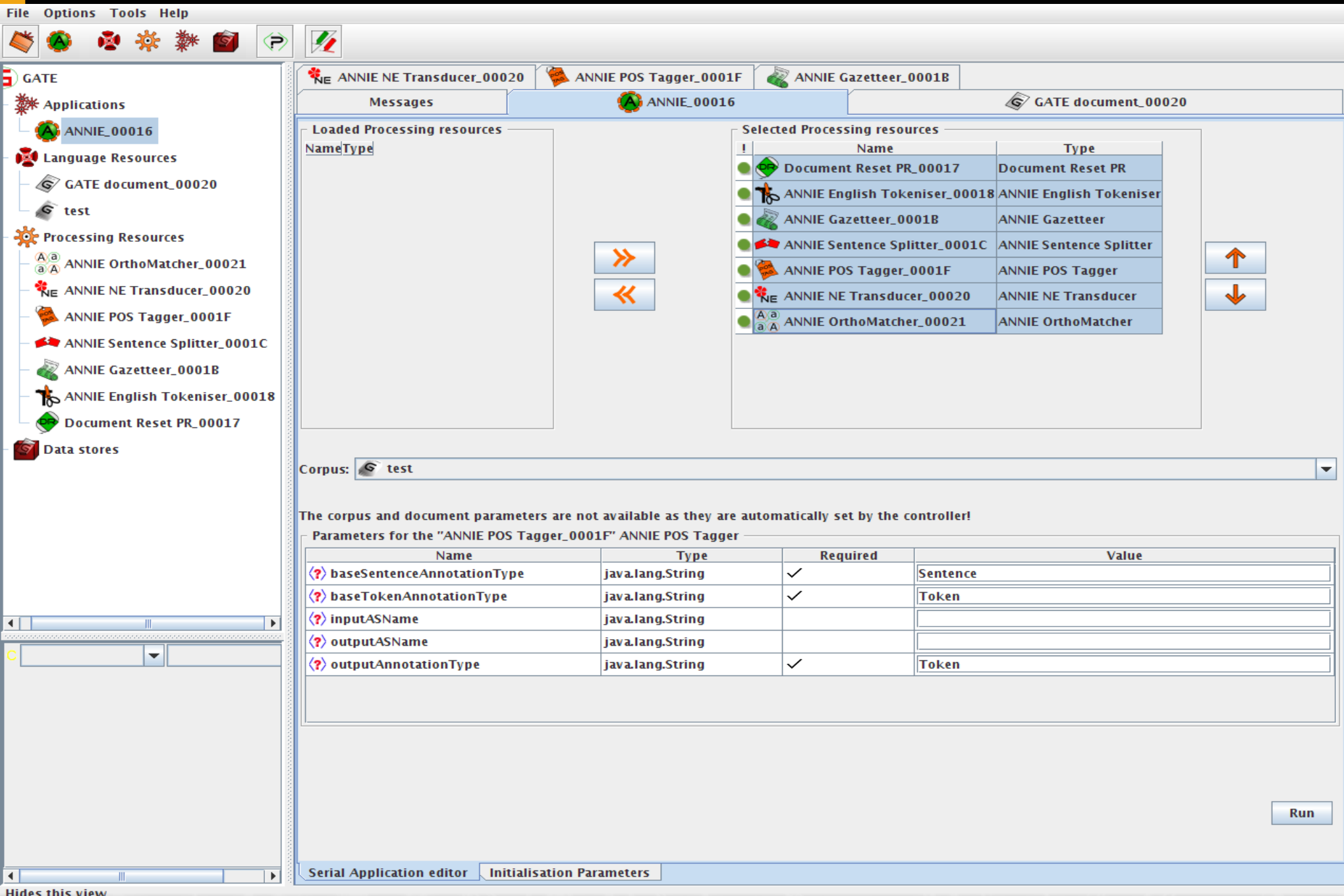
Terminé

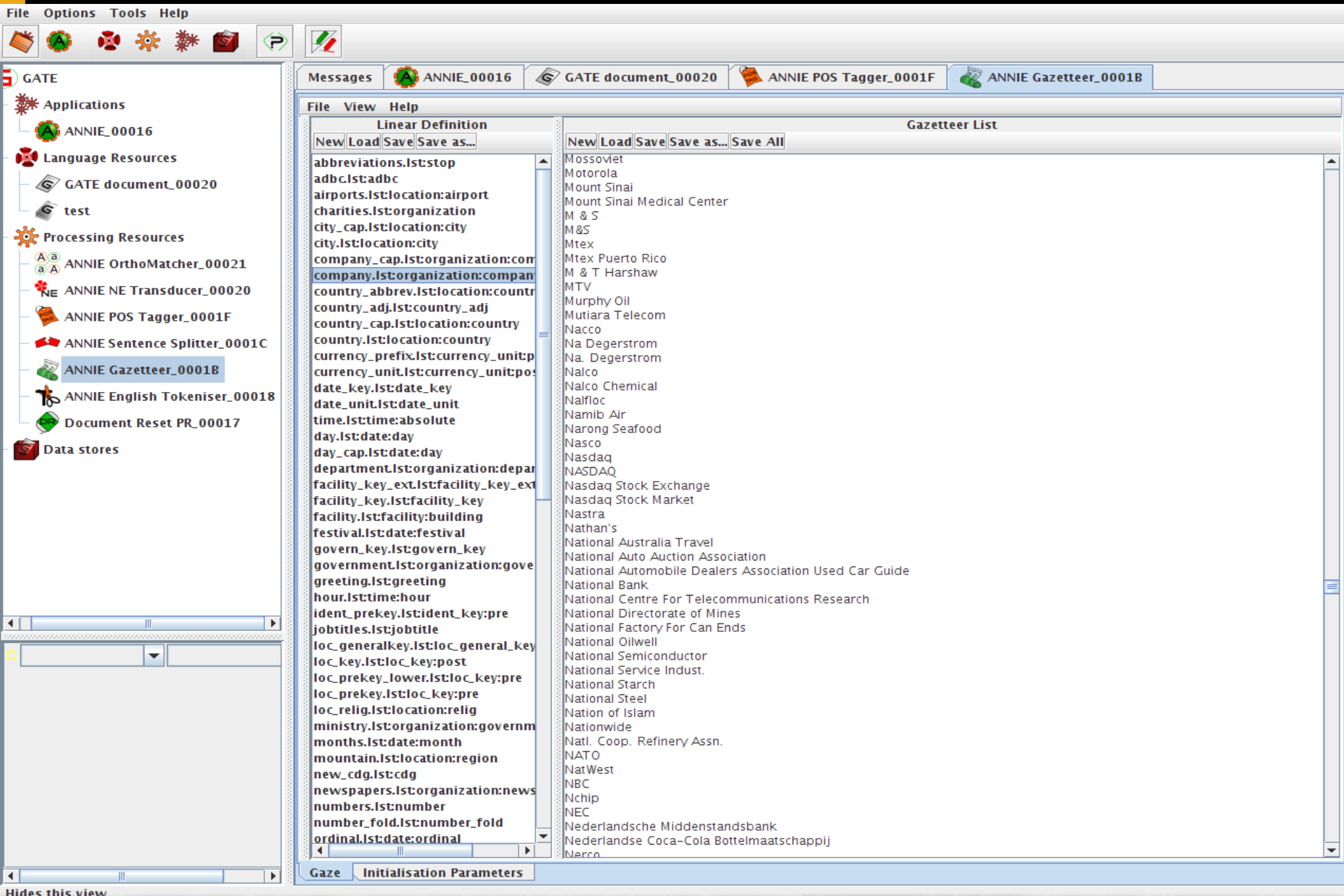
Des tâches fondamentales

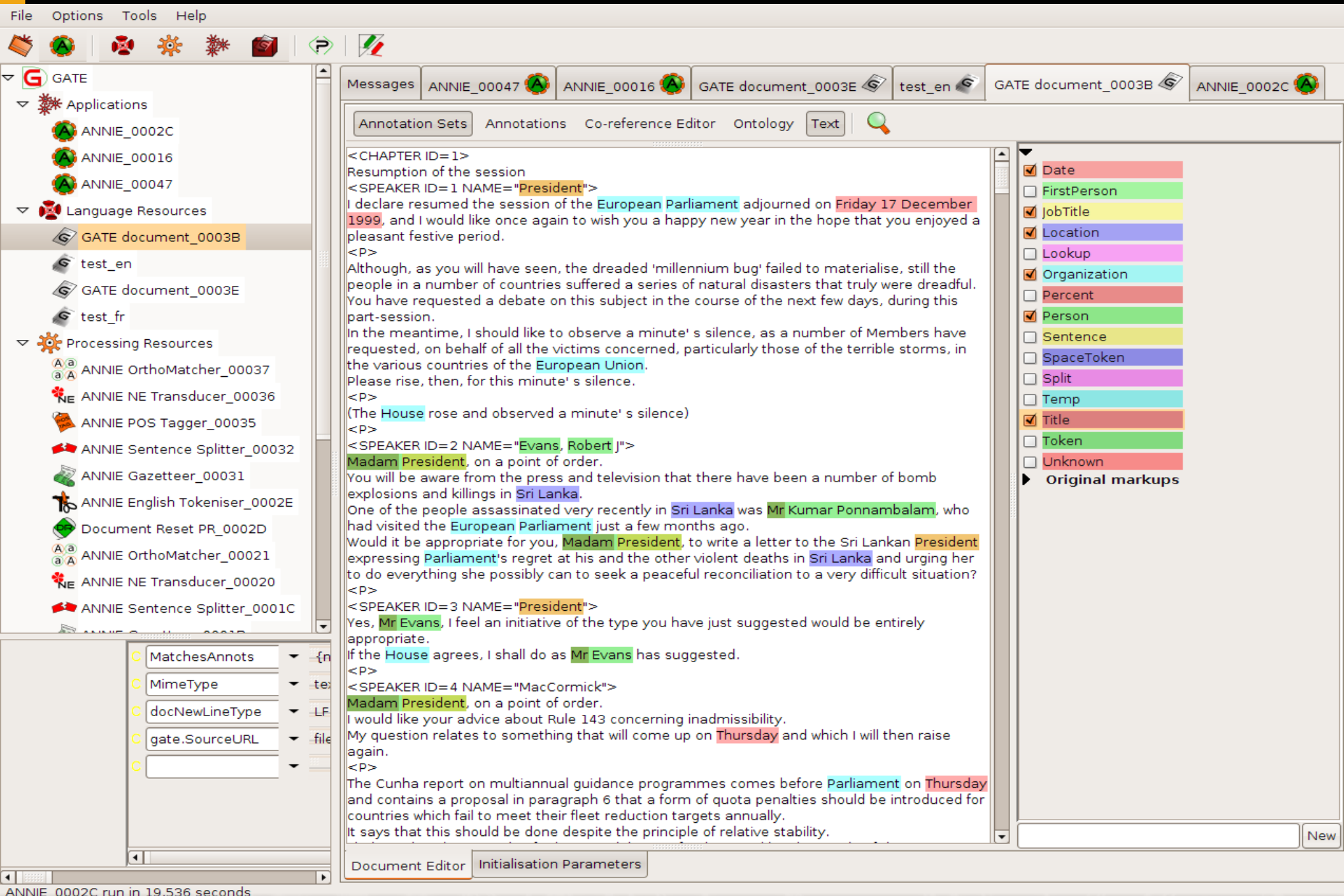
- Prétraitements
  - découpage en unités linguistiques
    - phrases,
    - mots : mots simples/composés
    - unités polylexicales : expressions figées, collocations, patrons de phrase
- Détection d'Entités Nommées
  - noms de lieux, de personnes, de sociétés etc.

Des tâches fondamentales

- L'étiquetage morpho-syntaxique
  - une tâche essentielle à des traitements de plus haut niveau
    - nombreuses approches bien documentées
    - programmes d'évaluation de la qualité des étiqueteurs
    - disponibilité de composants de qualité quasi-industrielle
    - disponibilité de données de paramétrage de référence


File
Options
Tools
Help
GATE
Applications
ANNIE_00016
Language Resources
GATE document_00020
test
Processing Resources
ANNIE OrthoMatcher_00021
ANNIE NE Transducer_00020
ANNIE POS Tagger_0001F
ANNIE Sentence Splitter_0001C
ANNIE Gazetteer_0001B
ANNIE English Tokeniser_00018
Document Reset PR_00017
Data stores
ANNIE NE Transducer_00020
ANNIE POS Tagger_0001F
ANNIE Gazetteer_0001B
Messages
ANNIE_00016
GATE document_00020
Loaded Processing resources
Name
Type
Selected Processing resources
!
Name
Type
Document Reset PR_00017
Document Reset PR
ANNIE English Tokeniser_00018
ANNIE English Tokeniser
ANNIE Gazetteer_0001B
ANNIE Gazetteer
ANNIE Sentence Splitter_0001C
ANNIE Sentence Splitter
ANNIE POS Tagger_0001F
ANNIE POS Tagger
ANNIE NE Transducer_00020
ANNIE NE Transducer
ANNIE OrthoMatcher_00021
ANNIE OrthoMatcher
Corpus:
test
The corpus and document parameters are not available as they are automatically set by the controller!
Parameters for the "ANNIE POS Tagger_0001F" ANNIE POS Tagger
Name
Type
Required
Value
baseSentenceAnnotationType
java.lang.String
Sentence
baseTokenAnnotationType
java.lang.String
Token
inputASName
java.lang.String
outputASName
java.lang.String
outputAnnotationType
java.lang.String
Token
Run
Serial Application editor
Initialisation Parameters
Hides this view


File Options Tools Help
GATE
Applications
ANNIE_00016
Language Resources
GATE document_00020
test
Processing Resources
ANNIE OrthoMatcher_00021
ANNIE NE Transducer_00020
ANNIE POS Tagger_0001F
ANNIE Sentence Splitter_0001C
ANNIE Gazetteer_0001B
ANNIE English Tokeniser_00018
Document Reset PR_00017
Data stores
Messages
ANNIE_00016
GATE document_00020
ANNIE POS Tagger_0001F
ANNIE Gazetteer_0001B
File View Help
Linear Definition
New Load Save Save as...
abbreviations.lst:stop
adbc.lst:adbc
airports.lst:location:airport
charities.lst:organization
city_cap.lst:location:city
city.lst:location:city
company_cap.lst:organization:com
company.lst:organization:compan
country_abbrev.lst:location:countr
country_adj.lst:country_adj
country_cap.lst:location:country
country.lst:location:country
currency_prefix.lst:currency_unit:p
currency_unit.lst:currency_unit:pos
date_key.lst:date_key
date_unit.lst:date_unit
time.lst:time:absolute
day.lst:date:day
day_cap.lst:date:day
department.lst:organization:depar
facility_key_ext.lst:facility_key_ext
facility_key.lst:facility_key
facility.lst:facility:building
festival.lst:date:festival
govern_key.lst:govern_key
government.lst:organization:gove
greeting.lst:greeting
hour.lst:time:hour
ident_prekey.lst:ident_key:pre
jobtitles.lst:jobtitle
loc_generalkey.lst:loc_general_key
loc_key.lst:loc_key:post
loc_prekey_lower.lst:loc_key:pre
loc_prekey.lst:loc_key:pre
loc_relig.lst:location:relig
ministry.lst:organization:governm
months.lst:date:month
mountain.lst:location:region
new_cdg.lst:cdg
newspapers.lst:organization:news
numbers.lst:number
number_fold.lst:number_fold
ordinal.lst:date:ordinal
Gazetteer List
New Load Save Save as... Save All
Mossoviet
Motorola
Mount Sinai
Mount Sinai Medical Center
M & S
M&S
Mtex
Mtex Puerto Rico
M & T Harshaw
MTV
Murphy Oil
Mutiara Telecom
Nacco
Na Degerstrom
Na. Degerstrom
Nalco
Nalco Chemical
Nalfloc
Namib Air
Narong Seafood
Nasco
Nasdaq
NASDAQ
Nasdaq Stock Exchange
Nasdaq Stock Market
Nastra
Nathan's
National Australia Travel
National Auto Auction Association
National Automobile Dealers Association Used Car Guide
National Bank
National Centre For Telecommunications Research
National Directorate of Mines
National Factory For Can Ends
National Oilwell
National Semiconductor
National Service Indust.
National Starch
National Steel
Nation of Islam
Nationwide
Natl. Coop. Refinery Assn.
NATO
NatWest
NBC
Nchip
NEC
Nederlandsche Middenstandsbank
Nederlandse Coca-Cola Bottelmaatschappij
Gaze
Initialisation Parameters
Hides this view


File
Options
Tools
Help
GATE
Applications
ANNIE_0002C
ANNIE_00016
ANNIE_00047
Language Resources
GATE document_0003B
test_en
GATE document_0003E
test_fr
Processing Resources
ANNIE OrthoMatcher_00037
ANNIE NE Transducer_00036
ANNIE POS Tagger_00035
ANNIE Sentence Splitter_00032
ANNIE Gazetteer_00031
ANNIE English Tokeniser_0002E
Document Reset PR_0002D
ANNIE OrthoMatcher_00021
ANNIE NE Transducer_00020
ANNIE Sentence Splitter_0001C
MatchesAnnots
MimeType
docNewLineType
gate.SourceURL
Messages
ANNIE_00047
ANNIE_00016
GATE document_0003E
test_en
GATE document_0003B
ANNIE_0002C
Annotation Sets
Annotations
Co-reference Editor
Ontology
Text
<CHAPTER ID=1>
Resumption of the session
<SPEAKER ID=1 NAME="President">
I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period.
<P>
Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the people in a number of countries suffered a series of natural disasters that truly were dreadful.
You have requested a debate on this subject in the course of the next few days, during this part-session.
In the meantime, I should like to observe a minute' s silence, as a number of Members have requested, on behalf of all the victims concerned, particularly those of the terrible storms, in the various countries of the European Union.
Please rise, then, for this minute' s silence.
<P>
(The House rose and observed a minute' s silence)
<P>
<SPEAKER ID=2 NAME="Evans, Robert J">
Madam President, on a point of order.
You will be aware from the press and television that there have been a number of bomb explosions and killings in Sri Lanka.
One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam, who had visited the European Parliament just a few months ago.
Would it be appropriate for you, Madam President, to write a letter to the Sri Lankan President expressing Parliament's regret at his and the other violent deaths in Sri Lanka and urging her to do everything she possibly can to seek a peaceful reconciliation to a very difficult situation?
<P>
<SPEAKER ID=3 NAME="President">
Yes, Mr Evans, I feel an initiative of the type you have just suggested would be entirely appropriate.
If the House agrees, I shall do as Mr Evans has suggested.
<P>
<SPEAKER ID=4 NAME="MacCormick">
Madam President, on a point of order.
I would like your advice about Rule 143 concerning inadmissibility.
My question relates to something that will come up on Thursday and which I will then raise again.
<P>
The Cunha report on multiannual guidance programmes comes before Parliament on Thursday and contains a proposal in paragraph 6 that a form of quota penalties should be introduced for countries which fail to meet their fleet reduction targets annually.
It says that this should be done despite the principle of relative stability.
Date
FirstPerson
JobTitle
Location
Lookup
Organization
Percent
Person
Sentence
SpaceToken
Split
Temp
Title
Token
Unknown
Original markups
New
Document Editor
Initialisation Parameters
ANNIE_0002C run in 19.536 seconds

Applications

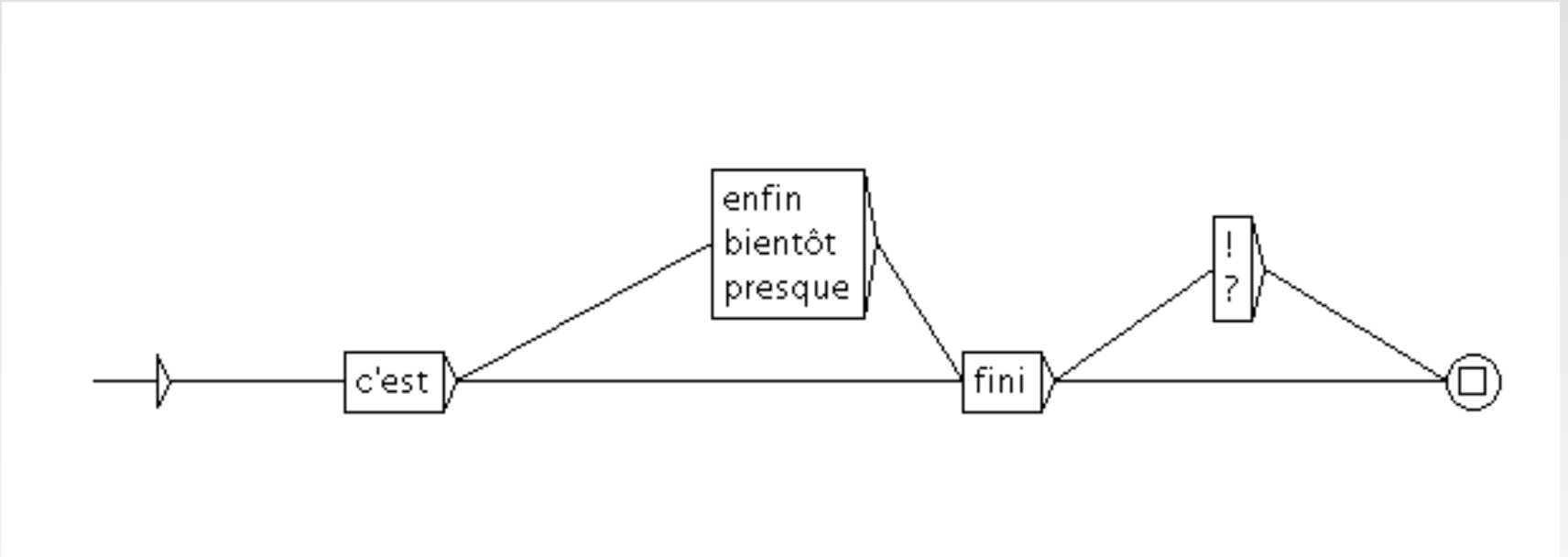
- Thématiques labellisées CREST :
  - Ingénierie Linguistique
    - Traduction Spécialisée Multilingue
      - Ilse Depraetere
    - Lexicographie/Terminographie
      - Pierre Corbin, Nathalie Gasiglia
    - Linguistique de Corpus
      - Antonio Balvet

STL et l'ingénierie linguistique

- Collaborations entreprises/STL
  - formations aux techniques et outils de l'ingénierie linguistique
  - exemple
    - ateliers "détection d'Entités Nommées" avec
      - Unitex
      - GATE

STL et l'ingénierie linguistique

- Autres collaborations
  - accueil d'étudiants :
    - stages de M1/M2
    - thèses CIFRE
  - montage de projets de R&D : MENRT, ANR, ESF
  - accès à des données-métier (texte, oral)
  - taxe d'apprentissage


enfin
bientôt
presque
!
?
c'est
fini