UNIVERSITÉ
PARIS-EST
MARNE-LA-VALLÉE

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Extraction d'information dans l'environnement Unitex/Gramlab

Claude Martineau

Plan de présentation

1. Notion d'Extraction d'information
2. Environnement Unitex/Gramlab
3. Un exemple d'extraction d'opinion
4. Perspectives

Extraction d'information

Conversion du texte en données structurées répondant à des questions factuelles :

QUI FAIT/PENSE QUOI QUAND OÙ COMMENT …

- Reconnaissance d'entités nommées (noms de personnes, lieu, dates, valeurs monétaires)
- Normalisation : 4 juillet 2014 → 2014-07-04
- Polysémie : ***Washington*** *(**Lieu** VS **Personne**)*
- Traitement de gros corpus

Extraction d'information

- Reconnaître des séquences de un ou plusieurs mots porteuses d'information
- Délimiter et baliser ces séquences
  - <XXX> 20 novembre 2013</XXX>
- Annoter ces séquences
  - <Date>20 novembre 2013</Date>
  - <Date type=absolue norm=2013/11/20>20 novembre 2013</Date>

→ **Corpus annoté**

L’environnement Unitex/Gramlab

- Environnement open-source développé principalement par Sébastien Paumier (écrit en C et interface en Java)
- Multiplateforme : Windows, Mac, Linux,IOS,Android
- Multilingue : allemand, anglais, arabe, coréen, finnois, français, espagnol, géorgien ancien, grec, grec ancien, italien, norvégien bokmal, norvégien nynorsk, polonais, portugais du Portugal, portugais du Brésil, russe, serbo-croate latin, serbo-croate cyrillique, thaï. (une langue à la fois)
- Unicode 3.0 (UTF16, UTF8)

Unitex/Gramlab

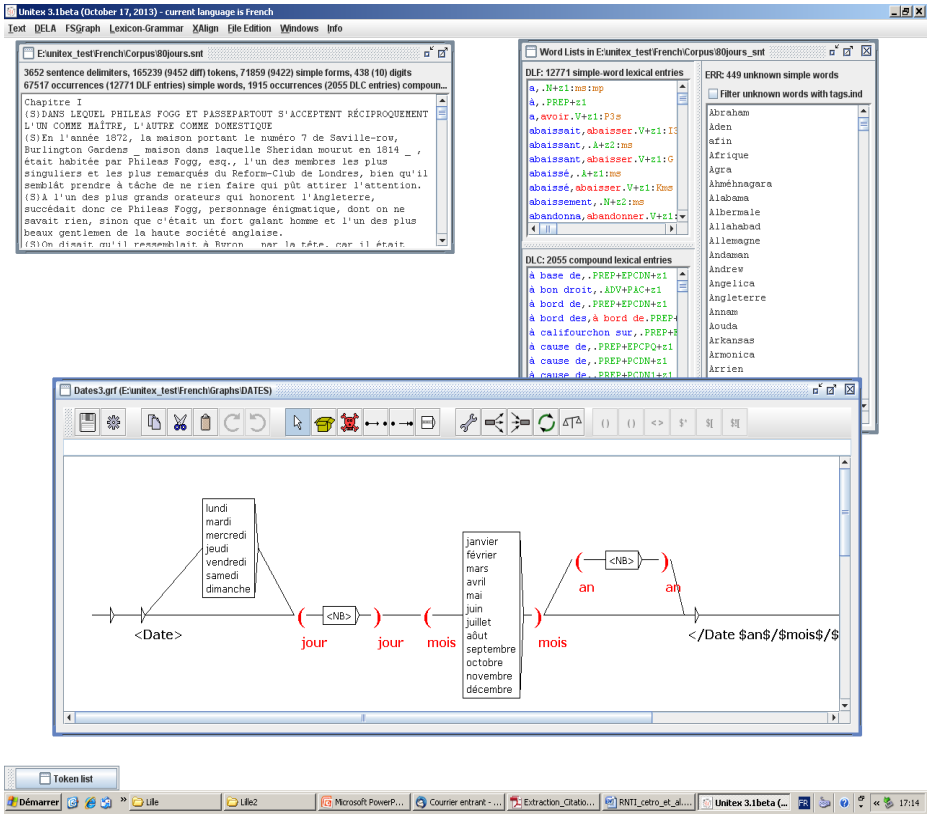
- Environnement d'analyse automatique de corpus et création de ressources linguistiques.
- Réseau d'automates à états finis (RTN)
- Gramlab (version industrielle, travail collaboratif avec SVN, MAVEN)
- Utilisable sous forme d'une librairie en C ou JAVA
- Système de fichiers virtuels
  - ➡ Traitements en mémoire

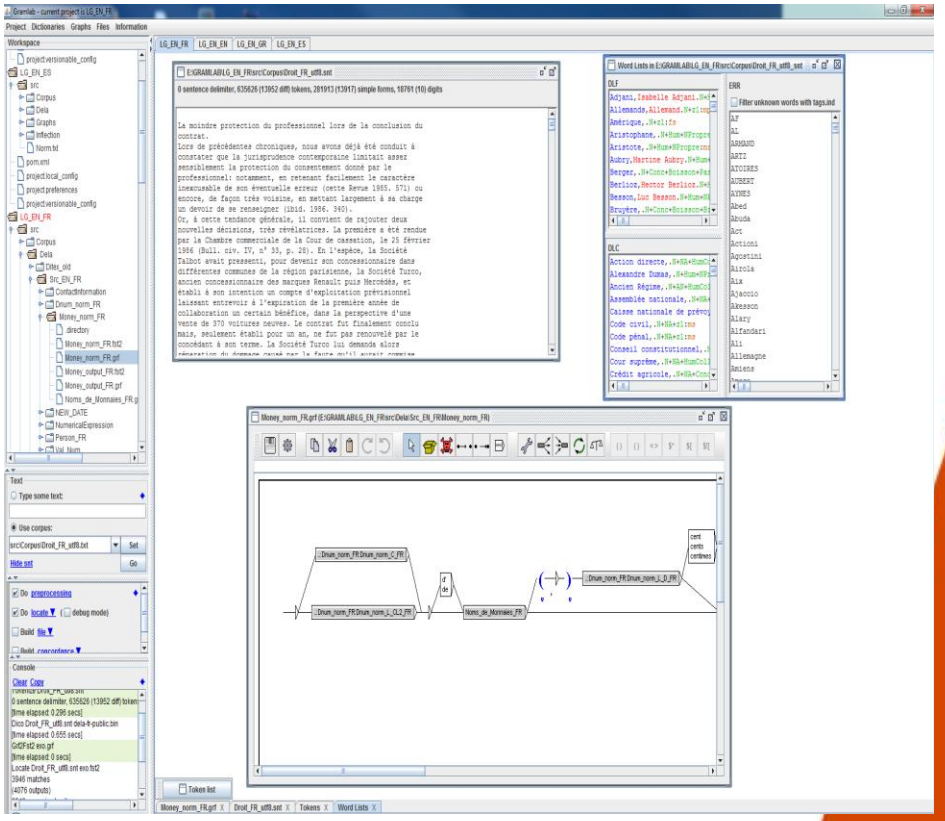
Environnement Unitex/Gramlab

Unitextool « Commande »

Interface Unitex

Interface Gramlab





Système DELA

Dictionnaire électronique du LADL

f_fléchie*,*f_canonique*.*cat_gram+infos syntax-sém*: *infos flexionnelles

- mots simples (DELAF): 984 723 entrées

  praesidium,*praesidium*.*N+HumColl*:*ms*
  praesidia,*praesidium*.*N*+*HumColl*:*mp*
  présidium,*présidium*.*N*+*HumColl*+*praesidium*:*ms*
  présidiums,*présidium*.*N*+*HumColl*+*praesidium*:*mp*

- mots composés (DELACF): 276 000 entrées

  week-end,*week-end*.*N*+*Tps*+*weekend*:*ms*
  week-ends,*week-end*.*N*+*Tps*+*weekend*:*mp*

Système DELA

Code	Description
m	masculin
f	féminin
n	neutre
s	singulier
p	pluriel
1,2,3	1ère, 2e,3e personnes

Code	Description
P	indicatif présent
I	indicatif imparfait
J	passé simple
F	futur
S	subjonctif présent
T	subjonctif imparfait
C	Conditionnel présent
W	infinitif
G	participe présent
K	participe passé
Y	impératif

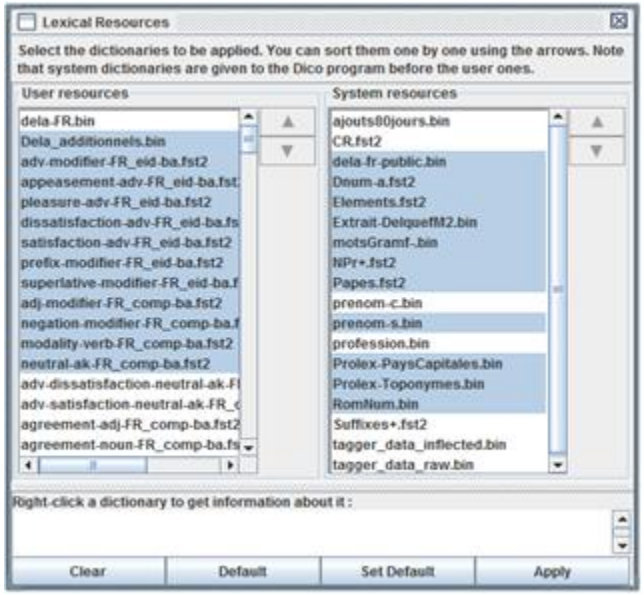
Système DELA

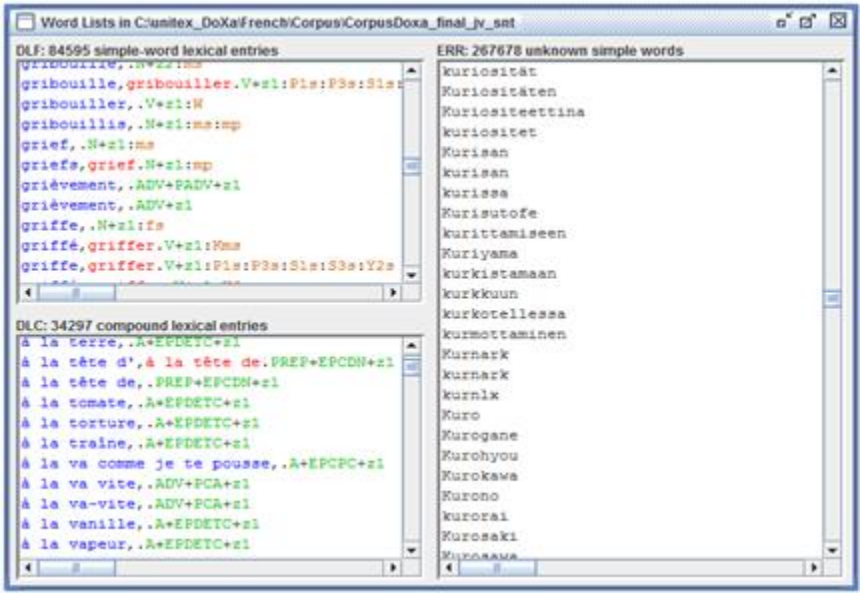
Code	Description
A	adjectif
ADV	adverbe
CONJC	conjonction de coordination
CONJS	conjonction de subordination
DET	déterminant
INTJ	interjection
N	nom
PREP	préposition
PRO	pronom
V	verbe

Code	Description
Abst	abstrait
Anl	animal
AnlColl	animal collectif
Conc	concret
ConcColl	concret collectif
Hum	humain
HumColl	humain collectif
i	intransitif
t	transitif
z1	langue générale
z2	langage spécialisé
z3	langage très spécialisé, rare

Unitex : utilisation des ressources

- Dictionnaires électroniques au format DELA
- Graphes dictionnaires : grammaires locales qui produisent dynamiquement des entrées au format DELA





Sélection des ressources

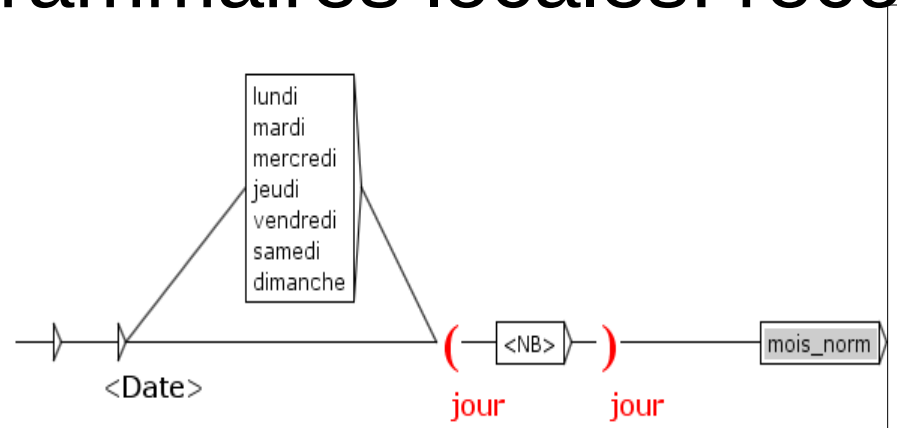
Dictionnaire du texte

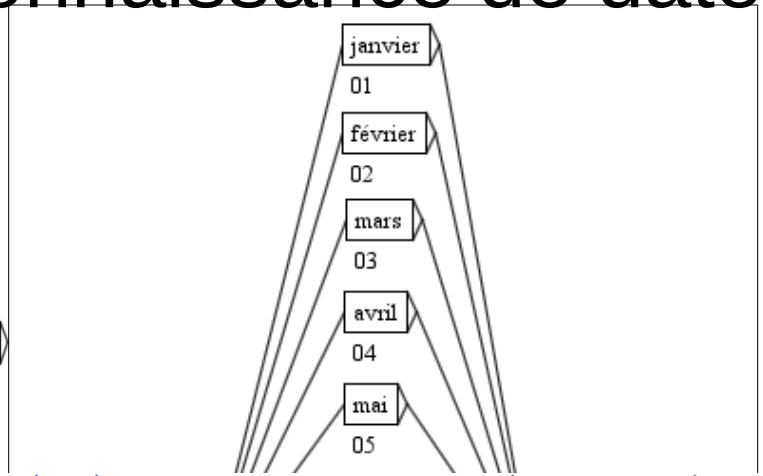
Expressions Régulières

<ADV><<ment$>><V:K>

Guinée-Bissau, du Niger et a été récemment élargi au Sénégal.{S} Le Sommet, à l
.les du nord par les mutins, sont actuellement détenus par ces derniers à Korhog
.armée française, a toutefois été formellement démentie par le porte-parole de l
: Puma, achetés en Roumanie, sont récemment arrivés afin de renforcer larmée.{S}
ı comédien et opposant Camara H., vraisemblablement tué par un «escadron de la m
de comptabilité moderne.{S} Sont également prévus le développement de mécanisme
.paux bailleurs de fonds", a-t-il également indiqué. ayv-jlh/dep tmf ###1103: {S
ı exliqué que cet évènement avait fortement contribué au blocage dans les négoci
)upe de soldats qui devaient être prochainement démobilisés sest soulevé, a indi
ıaitons que ce cessez-le-feu soit pleinement mis en oeuvre et quil soit respecté
ı première période, les Mimos ont littéralement dominé leur sujet, en inscrivant
'iron.{S} Le premier vol (AF 703) normalement prévu dimanche aura quant à lui dé
ıies Kofi Annan pour un entretien essentiellement consacré au réglement de la cr
les voyages en Côte dIvoire" et "fortement pressé ceux qui sy trouvent de parti
:tat Richard Boucher. "Nous avons particulièrement fait part de notre intérêt à
: et le message clair".{S} Elle a également précisé que lambassadeur de France e
:s témoins sur place.{S} Il a été notamment accueilli par le Premier ministre Pa
ıse Moïse Lida Kouassi.{S} Il est immédiatement monté à bord de lhélicoptère pré
.intérieur Emile Boga Doudou, ont notamment trouvé la mort. jlh/ag ###1421: {S}L
.lles des personnes detenues sont également informées par le CICR sur la situati
de ses conseillers.{S} Le sommet initialement prévu ce week-end et finalement r
.nitialement prévu ce week-end et finalement reporté au lundi 16 décembre, se ti

Grammaires locales: reconnaissance de dates





le pont. Le lendemain, <Date>8 novembre</Date /11/8>, au lever du soleil, la goélet
la personne de Fix. Le <Date>dimanche 20 octobre</Date /10/20>, vers midi, on eut c
oir. " Arrivé à Paris, <Date>jeudi 3 octobre</Date /10/3>, 7 heures 20 matin. " Qui
ingt-neuf du matin, ce <Date>mercredi 2 octobre 1872</Date 1872/10/2>, vous êtes à
es : " Quitté Londres, <Date>mercredi 2 octobre</Date /10/2>, 8 heures 45 soir. " A
sque c'est aujourd'hui <Date>mercredi 2 octobre</Date /10/2>, je devrai être de ret
ns se compromettre. Le <Date>mercredi 30 octobre</Date /10/30>, dans l'après-midi,
ivit donc, ce jour-là, <Date>mercredi 9 octobre</Date /10/9>, son arrivée à Suez, q
soir. " Arrivé à Suez, <Date>mercredi 9 octobre</Date /10/9>, 11 heures matin. " To
sieur Phileas Fogg. Le <Date>mercredi 9 octobre</Date /10/9>, on attendait pour onz
répondit Mr. Fogg, le <Date>samedi 21 décembre 1872</Date 1872/12/21>, à huit heur
ême du Reform-Club, le <Date>samedi 21 décembre</Date /12/21>, à huit heures quaran
éraire convenu ? Et le <Date>samedi 21 décembre</Date /12/21>, à huit heures quaran
à cette date fatale du <Date>samedi 21 décembre</Date /12/21>, à huit heures quaran
.. " Arrivé à Brindisi, <Date>samedi 5 octobre</Date /10/5>, 4 heures soir. " Embarq
le Mont-Cenis à Turin, <Date>vendredi 4 octobre</Date /10/4>, 6 heures 35 matin. "

Modèle DoXa: représentation sémantique des O&S

Les 20 catégorie sémantiques du modèle DoXa

Catégorie Sémantique	Pol. Intrinsèque	Etiquette	Etiq. Cat. Ant.	Exemple
Accord	positive	Agreement	Disagreement	approbation
Colère	négative	Anger		exaspération
Apaisement	positive	Appeasement		rassurée
Valorisation	positive	Appraisal	Depreciation	bienveillant
Ennui	négative	Boredom		rébarbatif
Mépris	négative	Contempt		<prendre> en grippe
Dévalorisation	négative	Depreciation	Appraisal	dénigrer
Mésentente	négative	Disagreement	Agreement	<mettre> en doute
Gêne	négative	Discomfort		perturber
Déplaisir	négative	Displeasure		répugnant
Insatisfaction	négative	Dissatisfaction	Satisfaction	incompétent
Crainte	négative	Fear		effroi
Surprise Négative	négative	NegSurprise	PosSurprise	sidéré
Plaisir	positive	Pleasure		divertir
Surprise Positive	positive	PosSurprise	NegSurprise	<couper> le souffle
Tristesse	négative	Sadness		découragement
Satisfaction	positive	Satisfaction	Dissatisfaction	adorable
Connotation méliorative	positive	MelConnot		bravo
Connotation péjorative	négative	PejConnot		problématique
Attente	neutre	Expectation		souhaiterais

Modèle DoXa: calcul de l'intensité

- Echelle d'intensité à 10 valeurs entières [1,…,10]

Intensité de 3 à 7 => constituants de base : A, ADV, N, V, formes (semi-)figées

Ex : <*intéressant*> –> cat_Satisfaction|int3

- Echelle de modification d'intensité :

3 niveaux en intensification et 3 en atténuation

Ex: *peu* –> AdvAtt1 *très* –> AdvInt2 *hyper* –> PrefInt3

- Règles de calcul de l'intensité résultante :

Combinaisons modifieurs et constituants de base

Ex: {AdvInt2} {Satisfaction|Int3} –> cat_Satisfaction|int5

« *très intéressante* » cat_Satisfaction|int5

Modèle DoXa: traitement de la négation

- La présence d'une négation introduit un changement de polarité du segment évaluatif qui s'exprime :
  - soit par passage à la classe antonyme si elle existe

    Ex: <*intéressant*> –> cat_Satisfaction|int3

    « ***pas*** *intéressants* » cat_**Dis**satisfaction|int3
  - soit par l'ajout de l'attribut *neg*

    Ex: <*inquiet*> –> cat_Fear|int3

    « ***pas*** *inquiets* » cat_Fear|int3|**neg**
- Le calcul de polarité se représente également par des règles

  Ex: *pas* {cat_Satisfaction|int3} –> cat_Dissatisfaction|int3

  pas {cat_Fear|int3} –> cat_Fear|int3|neg
- Prise en compte de l'intensité des négations simples et composées

  Ex : pas, Neg guère, LowNeg absolument pas, HighNeg

Annotation de segments complexes
calcul incrémental de l'intensité

unanimement | vraiment | le plus | hyper | intéressant | qu'on connaisse

+1 +2 +2 +3 3

+3 +6

+8

+10

Module Transverse

- Lexique
- Règle

Module Satisfaction

- Lexique
- Règle
- Règle
- Règle

DLF: 48 simple-word lexical entries

```
hyper,.{PrefInt3}+{ModInt3}+{PrefModifier}
intéressant,intéressant.{cat_Satisfaction|int3}+{SatisfactionAdj}+{SatisfactionAdjInt3}
unanimement,.{AdvInt1}+{AdvIntsInt1}+{ModInt1}+{AdvModifier}
vraiment,.{AdvInt2}+{AdvExtqInt2}+{ModInt2}+{AdvModifier}
vraiment,.{AdvInt2}+{AdvPSInt2}+{ModInt2}+{AdvModifier}
```

DLC: 19 compound lexical entries

```
hyper intéressant,.{cat_Satisfaction|int6}+{SatisfactionAdjInt6}+{SatisfactionPref}+{SatisfactionComp}+{SatisfactionCompInt6}
le plus hyper intéressant qu'on connaisse,.{cat_Satisfaction|int8}+{SatisfactionAdjInt8}+{SatisfactionComp}+{SatisfactionCompInt8}
le plus hyper intéressant,.{cat_Satisfaction|int8}+{SatisfactionAdjInt8}+{SatisfactionComp}+{SatisfactionCompInt8}
le plus,.{SupInt2}+{ModInt2}+{SupModifier}
qu'on connaisse,.{SuperlativeEnd}
unanimement vraiment le plus hyper intéressant qu'on connaisse,.{cat_Satisfaction|int10}+{SatisfactionComp}+{SatisfactionCompInt10}
unanimement vraiment,.{AdvModifierCompInt3}
vraiment le plus hyper intéressant qu'on connaisse,.{cat_Satisfaction|int10}+{SatisfactionComp}+{SatisfactionCompInt10}
```

Segments consécutifs reliés par des connecteurs

<Q2>A chaque fois que j'ai téléphoné, je suis tombé sur des interlocuteurs tout à fait aimables mais pas compétents : j'ai l'impression qu'ils sont là seulement pour décrocher mais ils ne nous apportent aucune solution et on s'aperçoit qu'il n'y a aucun suivi de dossier. CT</Q2>

<Q2>J'ai été bien accueilli et tout a été bien expliqué : la personne a été correcte au niveau téléphone, elle était agréable et pas agressive. CT</Q2>

<Q2>Aucun soucis, parfait, tout va très bien. Le sérieux EDF : le moindre renseignement nous est donné de suite, ils font bien leur travail. .</Q2>

<Q2>Bons rapports : rapport commercial normal classique. On me demande quelque chose et j'ai répondu oui j'ai pas à dire que c'était excessivement bien ou excessivement mal.

<O&S>tout à fait aimables mais pas compétents</O&S{Opposition}+Annotation1={cat_Appraisal|int5} + {AppraisalComp} + {AppraisalCompInt5};Annotation2={cat_Depreciation|int4}+{DepreciationMais}+{MaisComp}>

<O&S>agréable et pas agressive </O&S{Coordination}+Annotation1={cat_Satisfaction|int3}+{SatisfactionAdj}+ {SatisfactionAdjInt3}; connecteur=et; Annotation2={cat_Appraisal|int5}+{AppraisalComp}+{AppraisalCompInt5}>

<O&S>Aucun soucis, parfait, tout va très bien </O&S {Enumeration}+Annotation1={cat_Satisfaction|int7}+{SatisfactionAdj} + {SatisfactionAdjInt7};connecteur=virgule;Annotation2={cat_Satisfaction|int6}+{SatisfactionSemiFrozen}+ {SatisfactionSemiFrozen6}>

<O&S>c'était excessivement bien ou excessivement mal </O&S {Disjonction}+Annotation1={cat_Satisfaction|int7}+ {SatisfactionSemiFrozen} +{ SatisfactionSemiFrozen7}; connecteur=ou;Annotation2={cat_Depreciation|int5}+ {DepreciationComp}+ {DepreciationCompInt5}>

Perspectives

- Amélioration des ressources
  - Modularité
  - Adaptabilité
  - Réutilisabilité
- Extractions complexes :
  - Analyseur syntaxique FRMG
    Éric de la Clergerie Alpage INRIA

Unitex/Gramlab

Site : http://igm.univ-mlv.fr/~unitex/

Téléchargement :

Il est préférable de télécharger la version bêta qui bénéficie des dernières corrections

http://igm.univ-mlv.fr/~unitex/index.php?page=3&html=beta.html

Forum :

https://groups.google.com/forum/#!forum/unitex-gramlab

MERCI