Approches génériques du TAL pour la RI

Patrice Bellot — Aix-Marseille Université (AMU) / LSIS

octobre 2012
patrice.bellot@univ-amu.fr

Aix*Marseille université

Plan

- Linguistique et TAL : objets d'étude, applications...
- Un exemple bien connu : l'analyse syntaxique
- Des ressources linguistiques informatisées
- Extraction d'information
  - Reconnaissance d'entités nommées
  - Annotation automatique
- Recherche d'informations précises
  - Systèmes de questions-réponses

Introduction

Quelques thèmes de la revue TAL

http://atala.org/-Appels-a-soumission-

- Modèles et algorithmes pour la résolution d'anaphores.
- Systèmes de question/réponse.
- Les dictionnaires électroniques : pour les personnes, les machines ou pour les deux ?
- Le résumé automatique de texte
- Fouille de données textuelles : complexité, algorithmique et passage à l'échelle.
- Recherches actuelles en phonologie et en phonétique : interfaces avec le traitement automatique des langues.
- Opinions, sentiments et jugements d'évaluation.
- Traitement automatique des informations temporelles et spatiales en langage naturel.
- Du bruit dans le signal : gestion des erreurs en traitement automatique des langues.
- ..... **la Recherche d'information**

Exemples d'applications du TAL en RI

- Indexation de syntagmes nominaux
- Normalisation d'unités lexicales (mots, entités nommées...) et désambiguïsation sémantique
- Typage (annotation) des mots des documents
- Recherche de paraphrases (identification de variantes)
- RI multilingue (traduction automatique, alignement)
- Analyse de la temporalité, de l'opinion, de sentiments...
- Interactivité et systèmes de dialogue
- Correction orthographique, suggestion de requêtes
- RI multimédia (reconnaissance de la parole)
- RI à partir de documents *bruités* (OCR, tweets, SMS ...)

stemmer et lemmatiseur

Evaluation de diverses stratégies de désambiguïsation lexicale

Claire Fautsch, Jacques Savoy

Institut d'informatique
Université de Neuchâtel, rue Emile Argand 11, 2009 Neuchâtel (Suisse)
Claire.Fautsch@unine.ch, Jacques.Savoy@unine.ch

```
<NUM> C180 </NUM>
<EN-TITLE> Bankruptcy of Barings </EN-TITLE>
<EN-DESC> What was the extent of the losses in the Barings
 </EN-DESC>
<EN-NARR> Relevant documents must quantify in some way
the collapse of the oldest bank in Great Britain </EN-NARR>
...
<NUM> C180 </NUM>
<EN-TITLE>
 <TERM ID="10.2452/180-AH-1" LEMA = "bankruptcy" POS = "NNP">
 <WF> Bankruptcy </WF>
  <SYNSET SCORE = "0.48196658883771086" CODE = "10386276-n"/>
  <SYNSET SCORE = "0.51803341116228914" CODE = "10386165-n"/>
<TERM ID = "10.2452/180-AH-2" LEMA = "of" POS = "IN">
 <WF> of </WF> </TERM>
<TERM ID = "10.2452/180-AH-3" LEMA = "baring" POS = "NNPS">
 <WF> Barings </WF>
 <SYNSET SCORE = "1" CODE = "00819570-n"/> </TERM>
</EN-TITLE>
...
```

Figure 2 : *Exemple d'une requête avec les indications du lemme, de sa discours, et des numéros de classe de WordNet associées*

	Précision moyenne (MAP)					
	aucun	S-stemmer ‡	Porter	Lovins	SMART	lemme
Okapi	**0,3743**	0,4044†	**0,4150**†‡	**0,3930**	**0.4152**†‡	0,3988†
DFR-PL2	0,3703	0,4006†	0,4116†‡	0,3927†	0,4096†‡	**0,3994**†
DFR-I(n_e)C2	0,3731	**0,4054**†	0,4141†‡	0,3894 ‡	0,4139†‡	0,3988†
LM	0,3445	0,3709†	0,3809†‡	0,3522 ‡	0,3760†‡	0,3602
tf idf	0,2230	0,2393†	0,2399†	0,2194 ‡	0,2431†‡	0,2308
Moyenne	0,3370	0,3641	0,3723	0,3493	0,3716	0,3576
% différence		+8,0 %	+10,5 %	+3,6 %	+10,2 %	+6,1 %

Table 1. Précision moyenne (MAP) obtenues avec différents modèles et enracineurs (284 requêtes, format T)

	Précision moyenne (MAP)		
	lemme	lemme & POS	lemme & synset
Okapi	0,3988	**0,4053**†	0.3986
DFR-PL2	**0,3994**	0,4013	0,3918
DFR-I(n_e)C2	0,3988	0,4047	**0,4018**
LM	0,3602	0,3659†	0,3546
tf idf	0,2308	0,2315	0,2325
Moyenne	0,3576	0,3617	0,3559
% différence		+1,2 %	-0,5 %

Table 2. Précision moyenne (MAP) pour différents modèles de RI et variantes d'analyse morphologique (284 requêtes, format T)

Combinaison d'index ?

- Thèse de F. Moreau (IRISA, 2006) «*Revisiter le couplage traitement automatique des langues et recherche d'information*»

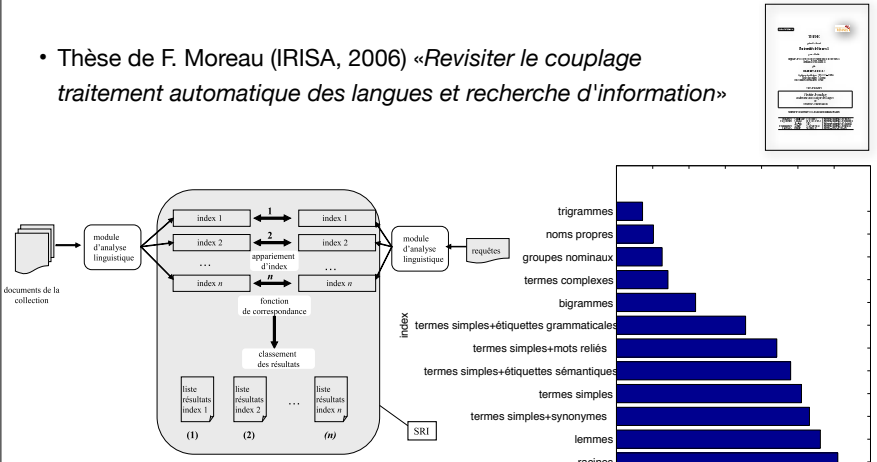


FIG. 3.2 – Intégration au sein du SRI des représentations multi-index

FIG. 3.3 – Performance du SRI pour chaque information linguistique manipulée

paire d'index (index 1 + index 2)	taux de résultats différents (en %)	documents pertinents retrouvés uniquement par le 1er index (en %)	documents pertinents retrouvés uniquement par le 2nd index (en %)
trigrammes + synonymes	72,22	1	99
trigrammes + étiquettes sémantiques	66,72	1	99
trigrammes + mots reliés	66,30	2	98
groupes nominaux + synonymes	61,96	0	100
noms propres + synonymes	59,83	4	96
groupes nominaux + étiquettes sémantiques	57,92	2	98
noms propres + étiquettes sémantiques	56,37	6	94
groupes nominaux + mots reliés	56,30	2	98
noms propres + mots reliés	55,97	8	92
termes complexes + synonymes	54,18	2	98
termes complexes + étiquettes sémantiques	50,50	3	97
termes complexes + mots reliés	49,53	4	96
bigrammes + synonymes	44,86	8	92
racines + trigrammes	43,31	97	3
lemmes + trigrammes	42,66	98	2
bigrammes + mots reliés	41,95	13	87
racines + groupes nominaux	41,54	99	1
bigrammes + étiquettes sémantiques	41,08	10	90
lemmes + groupes nominaux	40,60	99	1
termes simples + trigrammes	40,23	97	3
noms propres + racines	38,43	8	92
termes simples +groupes nominaux	38,22	97	3
racines + termes complexes	38.09	95	5

	Index des racines	Fusion sans infos req. (amélioration %)
Précision	27.90	26.49 (-5.04%)
Rappel	51.32	41.65 (-18.84%)
F-mesure	29.67	30.28 (+2.05%)

	Fusion infos req. + rang Moyenne (écart-type)	Fusion sans infos req. Moyenne (écart-type)
Précision	20.48 (3.81)	26.49 (4)
Rappel	43.88 (5.77)	41.65 (4.72)
F-mesure	34.50 (3.92)	30.28 (3.97)

TAB. 4.10 – Moyennes et écarts-type des valeurs de précision, rappel et F-mesure obtenus sur les 10 jeux de test par la méthode de fusion avec et sans prise en compte des informations sur les requêtes

TAB. 4.6 – Moyennes des précision, rappel et F-mesure obtenues par la méthode de fusion sans prise en compte des informations sur la requête et comparées aux performances d'un SRI exploitant des racines

Exemples de requêtes TREC

- Question-Answering Track
  - What Jeopardy contestant is the biggest money-winner in television game show history?
  - Into how many languages has "Harry Potter and the Goblet of Fire" been translated?

Utilisation de *narrative* ?

```
▼<topic num="68">
 ▼<template id="2">
    What [financial relationships] exist between [DARPA] and [BBN]?
  </template>
 ▼<narrative>
    The analyst would like to know of financial relationships between the Defense Advanced Research Projects
    Agency and BBN Technologies in Cambridge. Specifically, the analyst would like to know of projects
    funded by DARPA at BBN.
  </narrative>
</topic>
```

- Entity Track

```
<query>
<num>109</num>
<entity_name>Airbus</entity_name>
<entity_URL>clueweb09-en0004-69-31337</entity_URL>
<target_entity>airline</target_entity>
<narrative>airlines that fly the Airbus 320 plane</narrative>
</query>
```


Exemples de requêtes TREC (2)

- Medical Track
  - Patients diagnosed with localized prostate cancer and treated with robotic surgery
  - Patients who received methotrexate for cancer treatment while in the hospital
- Chemical IR Track

```
<top>
 <num>TCS-xx</num>
 <query>liquid coating composition</query>
 <narr>
   I need all available data on the production of liquid coating compositions.
 </narr>
</top>
```

- Enterprise Track

```
<top>
<num>CE-053</num>
<query>selenium soil</query>
<narr>
Can you please provide a current e-mail address, or failing that can you please put me in contact with the
group responsible for the research into the use of selenium as an additive to soils, to promote sheep
productivity/health. There were some trials conducted in WA and I am looking for aditional information on
these.
</narr>
</top>
```

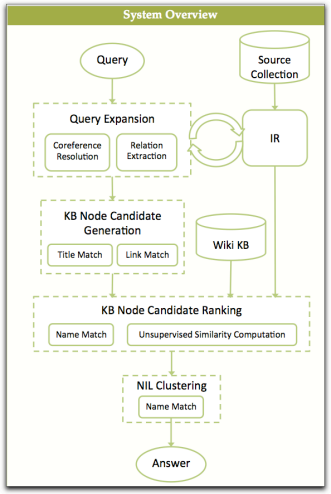

Exemples de tâches DEFT (Défi Fouille de Textes)

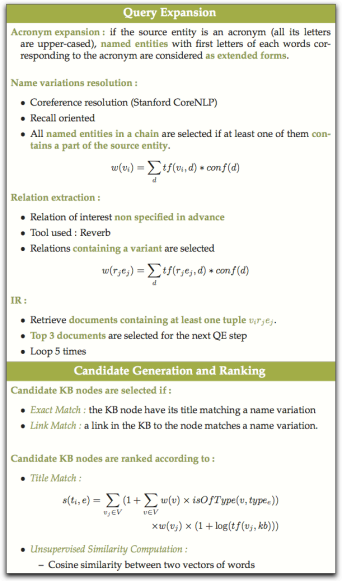
- détection de **l'opinion** (positive, neutre, négative) exprimée dans un texte (débats parlementaires, critiques de livres et de jeux vidéos, relectures d'articles scientifiques)
- **classification** automatique de corpus **en genres** (Le Monde vs. Wikipédia) et en catégories (International, Politique française, Société, Economie, Sciences, Art, Littérature, Sports, Télévision)
- **détermination du parti politique d'appartenance** des parlementaires d'après leurs interventions + détection des articles globalement objectifs (**factuels**) ou **subjectifs** (porteurs d'opinion) dans les corpus de journaux
- **identification de l'année** de publication d'un extrait de journal (de 1801 à 1944) et appariement de résumés avec des articles scientifiques
- **identification des mots-clés** indexant le contenu d'un article scientifique paru en revue de SHS

http://deft.limsi.fr

Exemple de tâche

- TAC KBP (*Knowledge Base Population*)





TAL : quelles unités d'étude ?

- **Mot** ? **Phrase** ? **Paragraphe** ? **Texte** ? **Document** ?
- ***Chunk*** : séquence de mots + catégorie associée
  (groupe nominal, groupe verbal / pas d'inclusion d'autre chunk)
- **Syntagme** : nominal, verbal
- **Clause** : sujet + prédicat + ...
- **Énoncé** : prise de parole
- **Proposition** : valeur de vérité associée

Quelques définitions

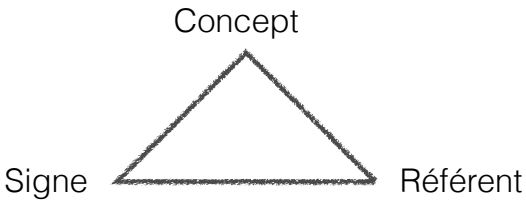
- **Mot** : unité (empirique) du *lexique* — mots dénominatifs, désignant, grammaticaux...
- **Lemme** : forme simple d'un mot permettant d'identifier un groupe de flexions
- **Terme** : nom, entité nommée, mot de spécialité utile pour l'indexation (mot clé)
- **Thesaurus** : base de connaissance lexicale qui permet d'identifier des associations de mots (synonymie, hyperonymie....)
- **Syntaxe** : [étude] ordre des mots / des groupes — micro/macro syntaxe
  vs. **morphologie** (structure interne des mots : flexions, dérivations...) → **morphosyntaxe**
  vs. **sémantique...**
- **Syntagme** : constituant syntaxique, organisé autour d'une tête (nom pour SN, verbe pour SV, adjectif pour SA...) : *le **stylo** que tu m'as offert écrit très bien, je **prendrai** un café serré, très **heureux** d'être ici !*

Phrase, *signification* et *sens*

- *Le cours commence dans dix minutes.*
  - **Sémantique** : le sens indépendamment du contexte = la signification
  - **Pragmatique** : le sens en contexte (environnement, texte qui précède, voire hypothèses posées par l'auditeur) = le sens
  - Lexique : les mots sont porteurs de sens / transforment le sens = liens avec *concepts, catégories* (les entités désignées par un mot)
  - Syntaxe : indique que c'est le *cours* qui va commencer
  - Contexte : aide à l'interprétation, influence le sens des mots
  - Un *énoncé* : une phrase + informations non linguistiques (locuteur, ...)

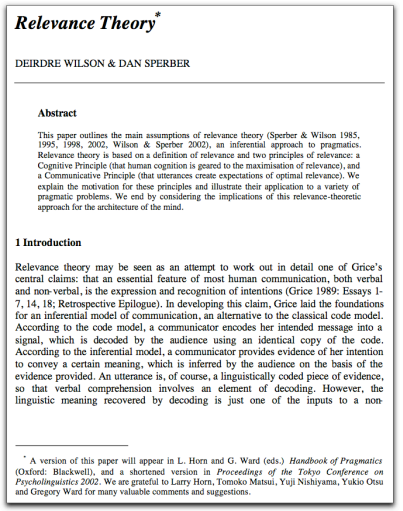
Le triangle sémiotique

- Mot / signe, forme orthographique / signifiant, concept / signifié
- **Triangle d'Odgen et Richards** : les mots (*signes*) désignent des *référents* (entités du monde, l'extension) via des *concepts* (l'intension)
  - Noms propres : lien direct signe / référent
  - Connecteurs, adverbes, pronoms... : pas de sens sans référent (*information procédurale*) --> résolution des références (anaphores...)
  - Expression génériques : peu importe le référent (classe conceptuelle)



Notion de pertinence (*relevance*)

- Pertinence *vs.* Vérité (subjectif/ objectif, vrai/faux...)
- Est pertinent ce qui crée de l'information (dans l'esprit)
- Notion d'effort / coût (adapté au contexte, auditeurs, besoin...)
- RI : constitution de référentiels (qrels), mesures de précision/rappel...



Relevance Theory*

DEIRDRE WILSON & DAN SPERBER

Abstract

This paper outlines the main assumptions of relevance theory (Sperber & Wilson 1985, 1995, 1998, 2002, Wilson & Sperber 2002), an inferential approach to pragmatics. Relevance theory is based on a definition of relevance and two principles of relevance: a Cognitive Principle (that human cognition is geared to the maximisation of relevance), and a Communicative Principle (that utterances create expectations of optimal relevance). We explain the motivation for these principles and illustrate their application to a variety of pragmatic problems. We end by considering the implications of this relevance-theoretic approach for the architecture of the mind.

1 Introduction

Relevance theory may be seen as an attempt to work out in detail one of Grice's central claims: that an essential feature of most human communication, both verbal and non-verbal, is the expression and recognition of intentions (Grice 1989: Essays 1-7, 14, 18; Retrospective Epilogue). In developing this claim, Grice laid the foundations for an inferential model of communication, an alternative to the classical code model. According to the code model, a communicator encodes her intended message into a signal, which is decoded by the audience using an identical copy of the code. According to the inferential model, a communicator provides evidence of her intention to convey a certain meaning, which is inferred by the audience on the basis of the evidence provided. An utterance is, of course, a linguistically coded piece of evidence, so that verbal comprehension involves an element of decoding. However, the linguistic meaning recovered by decoding is just one of the inputs to a non-

* A version of this paper will appear in L. Horn and G. Ward (eds.) *Handbook of Pragmatics* (Oxford: Blackwell), and a shortened version in *Proceedings of the Tokyo Conference on Psycholinguistics 2002*. We are grateful to Larry Horn, Tomoko Matsui, Yuji Nishiyama, Yukio Otsu and Gregory Ward for many valuable comments and suggestions.

http://people.bu.edu/bfraser/Relevance%20Theory%20Oriented/Sperber%20&%20Wilson%20-%20RT%20Revisited.pdf

Analyseurs (syntaxic

Analyseurs en partie du discours (*POS tagger*)

- Etiquetage morpho-syntaxique (la syntaxe influence la forme)
  — nom, verbe, article/déterminant, adjectif, pronom, adverbe, préposition, conjonction
  — masculin / féminin (/ neutre), singulier / pluriel, auxiliaire, temps, participe, voie (active / passive)...
- Méthodes :
  — règles d'étiquetage
  — modèles de Markov cachés (HMM)
- Exemples :
  — *TreeTagger* http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
  anglais, français, allemand, italien, swahili...
  — *HMMTagger* (dans UIMA)

Sample output:

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

http://hal.inria.fr/docs/00/49/38/47/PDF/article-taln-2010.pdf

Un étiqueteur de rôles grammaticaux libre pour le français intégré à Apache UIMA
Charles Dejean Manoel Fortun Clotilde Massot Vincent Pottier Fabien Poulard· Matthieu Vernier

The Stanford Natural Language Processing Group

home · people · teaching · research · publications · software · events · local

http://www-nlp.stanford.edu/software/

Stanford CoreNLP
An integrated suite of natural language processing tools for English in Java, including tokenization, part-of-speech tagging, named entity recognition, parsing, and coreference. Online CoreNLP demo

Stanford Parser
Implementations of probabilistic natural language parsers, both highly optimized PCFG and dependency parsers, and a lexicalized PCFG parser in Java. Includes: Online parser demo, Stanford Dependencies page, and Parser FAQ.

Stanford POS Tagger
A maximum-entropy (CMM) part-of-speech (POS) tagger for English, Arabic, Chinese, French, and German, in Java.

Stanford Named Entity Recognizer
A Conditional Random Field sequence model, together with well-engineered features for Named Entity Recognition in English and German. Online NER demo

Stanford Word Segmenter
A CRF-based word segmenter in Java. Supports Arabic and Chinese.

Stanford Classifier
A machine learning classifier, directed at text categorization. A conditional loglinear classifier (a.k.a. a maximum entropy or multiclass logistic regression model).

Tregex and Tsurgeon
A Tgrep2-style utility for matching patterns in trees, and a tree-transformation utility built on top of this matching language.

Phrasal
A state-of-the-art phrase-based machine translation system.

Stanford Biomedical Event Parser (SBEP)
Biomedical Event Extraction for the BioNLP 2009/2011 shared task.

Stanford EnglishTokenizer
A fast tokenizer for English text (producing Penn Treebank tokenization, roughly)

Stanford TokensRegex
A tool for matching regular expressions over tokens.

Stanford Temporal Tagger (SUTime)
A rule-based temporal tagger for English text. Online SUTime demo

Stanford Parser

http://nlp.stanford.edu/software/lex-parser.shtml

- Anglais, Chinois, Arabe, Allemand (adaptable à d'autres langues)
- *The strongest rain ever recorded in India shut down the financial hub of Mumbai, snapped communication lines, closed airports and forced thousands of people to sleep in their offices or walk home during the night, officials said today.*

```
The/DT strongest/JJS rain/NN ever/RB recorded/VBN in/IN India/NNP
shut/VBD down/RP the/DT financial/JJ hub/NN of/IN Mumbai/NNP ,/,
snapped/VBD communication/NN lines/NNS ,/, closed/VBD airports/NNS
and/CC forced/VBD thousands/NNS of/IN people/NNS to/TO sleep/VB in/IN
their/PRP$ offices/NNS or/CC walk/VB home/NN during/IN the/DT night/NN
,/, officials/NNS said/VBD today/NN ./.
```


Stanford Parser

http://nlp.stanford.edu/software/lex-parser.shtml

- Anglais, Chinois, Arabe, Allemand (adaptable à d'autres langues)
- *The strongest rain ever recorded in India shut down the financial hub of Mumbai, snapped communication lines, closed airports and forced thousands of people to sleep in their offices or walk home during the night, officials said today.*

```
The/DT strongest/JJS
shut/VBD down/RP the
snapped/VBD communic
and/CC forced/VBD th
their/PRP$ offices/N
,/, officials/NNS sa
```

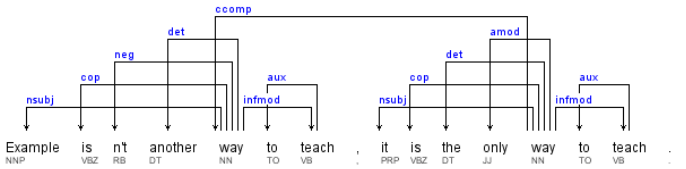
```
(ROOT
  (S
    (S
      (NP
        (NP (DT The) (JJS strongest) (NN rain))
        (VP
          (ADVP (RB ever))
          (VBN recorded)
          (PP (IN in)
            (NP (NNP India)))))
      (VP
        (VP (VBD shut)
          (PRT (RP down))
          (NP
            (NP (DT the) (JJ financial) (NN hub))
            (PP (IN of)
              (NP (NNP Mumbai)))))
        (, ,)
        (VP (VBD snapped)
          (NP (NN communication) (NNS lines)))
        (, ,)
        (VP (VBD closed)
          (NP (NNS airports)))
        (CC and)
        (VP (VBD forced)
```

```
ia/NNP
P ,/,
rts/NN
p/VB i
DT nig
```

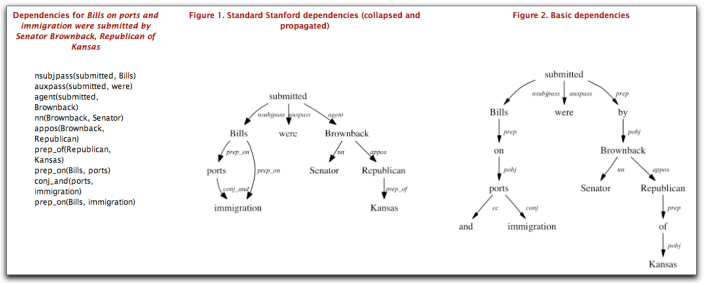
```
det(rain-3, The-1)
amod(rain-3, strongest-2)
nsubj(shut-8, rain-3)
nsubj(snapped-16, rain-3)
nsubj(closed-20, rain-3)
nsubj(forced-23, rain-3)
advmod(recorded-5, ever-4)
partmod(rain-3, recorded-5)
prep_in(recorded-5, India-7)
ccomp(said-40, shut-8)
prt(shut-8, down-9)
det(hub-12, the-10)
amod(hub-12, financial-11)
dobj(shut-8, hub-12)
prep_of(hub-12, Mumbai-14)
conj_and(shut-8, snapped-16)
ccomp(said-40, snapped-16)
nn(lines-18, communication-17)
dobj(snapped-16, lines-18)
conj_and(shut-8, closed-20)
ccomp(said-40, closed-20)
dobj(closed-20, airports-21)
conj_and(shut-8, forced-23)
ccomp(said-40, forced-23)
dobj(forced-23, thousands-24)
```


Stanford Parser : GUI

- http://chaoticity.com/dependensee-a-dependency-parse-visualisation-tool/



- GrammarScope / Stanford dependencies



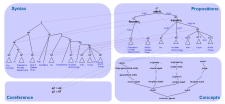
Stanford CoreNLP

http://www-nlp.stanford.edu/software/dcoref.shtml

- Résolution des références (dont anaphores)
- Evaluation sur la tâche CoNLL-2011 (http://conll.cemantix.org/2011/introduction.html)

  She had a good suggestion and it was unanimously accepted

  – données du projet OntoNotes (http://www.bbn.com/ontonotes/)



Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky
Stanford NLP Group

Official; Closed track; Predicted mentions

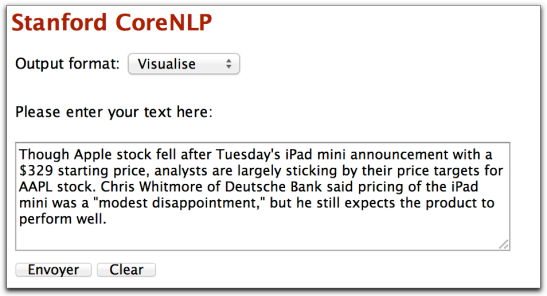
System	MD F	MUC F¹	B-CUBED F²	$CEAF_m$ F	$CEAF_e$ F³	BLANC F	Official $\frac{F^1+F^2+F^3}{3}$
lee	70.70	59.57	68.31	56.37	45.48	73.02	57.79
sapena	43.20	59.55	67.09	53.51	41.32	71.10	55.99
chang	64.28	57.15	68.79	54.40	41.94	73.71	55.96
nugues	68.96	58.61	65.46	51.45	39.52	71.11	54.53
santos	65.45	56.65	65.66	49.54	37.91	69.46	53.41
song	67.26	59.95	63.23	46.29	35.96	61.47	53.05
stoyanov	67.78	58.43	61.44	46.08	35.28	60.28	51.92
sobha	64.23	50.48	64.00	49.48	41.23	63.28	51.90
kobdani	61.03	53.49	65.25	42.70	33.79	62.61	51.04
zhou	62.31	48.96	64.07	47.53	39.74	64.72	50.92
charton	64.30	52.45	62.10	46.22	36.54	64.20	50.36
yang	63.93	52.31	62.32	46.55	35.33	64.63	49.99
hao	64.30	54.47	61.01	45.07	32.67	65.35	49.38
xinxin	61.92	46.62	61.93	44.75	36.23	64.27	48.46
zhang	61.13	47.28	61.14	44.46	35.19	65.21	48.07
kummerfeld	62.72	42.70	60.29	45.35	38.32	59.91	47.10
zhekova	48.29	24.08	61.46	40.43	35.75	53.77	40.43
irwin	26.67	19.98	50.46	31.68	25.21	51.12	31.28

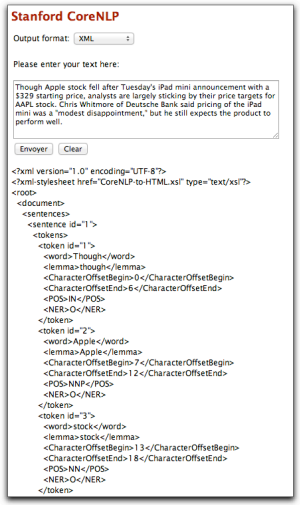
Scores on various corpora described in Raghunathan et al. 2010

	MUC			B cubed			Pairwise		
	P	R	F1	P	R	F1	P	R	F1
ACE2004 dev	86.0	75.5	80.4	89.3	76.5	82.4	81.7	55.2	65.9
ACE2004 test	82.7	70.2	75.9	88.7	74.5	81.0	77.2	44.6	56.6
ACE2004 nwire	84.6	75.1	79.6	87.3	74.1	80.2	79.4	50.1	61.4
MUC6 test	90.6	69.1	78.4	90.6	63.1	74.4	89.7	57.0	69.7

Stanford CoreNLP (2) - démo

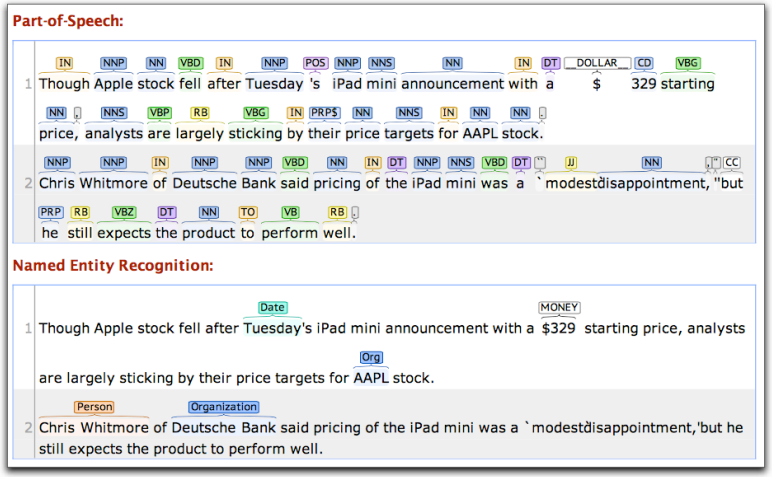
http://nlp.stanford.edu:8080/corenlp/process





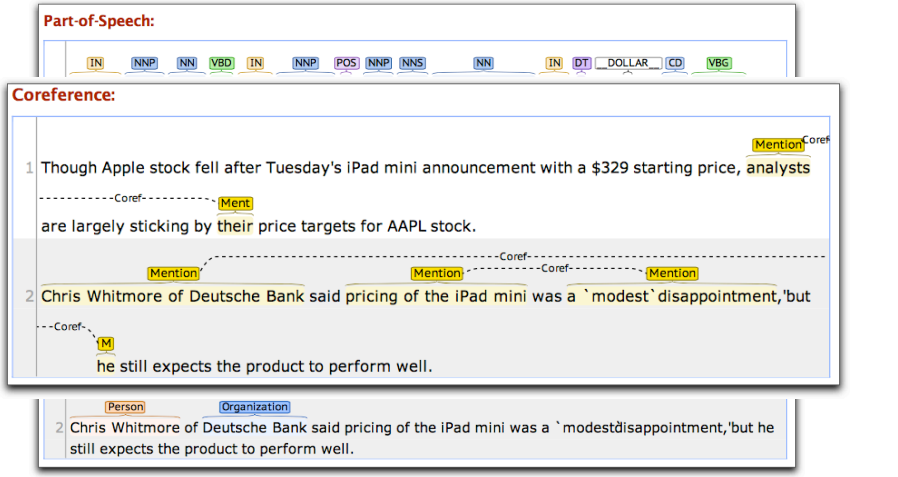
Stanford CoreNLP (3)

Though Apple stock fell after Tuesday's iPad mini announcement with a $329 starting price, analysts are largely sticking by their price targets for AAPL stock. Chris Whitmore of Deutsche Bank said pricing of the iPad mini was a "modest disappointment," but he still expects the product to perform well.

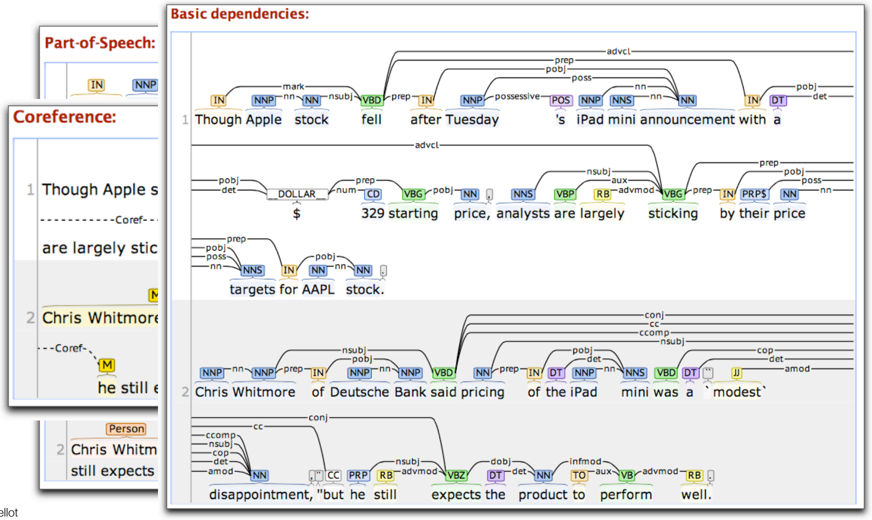


Stanford CoreNLP (3)

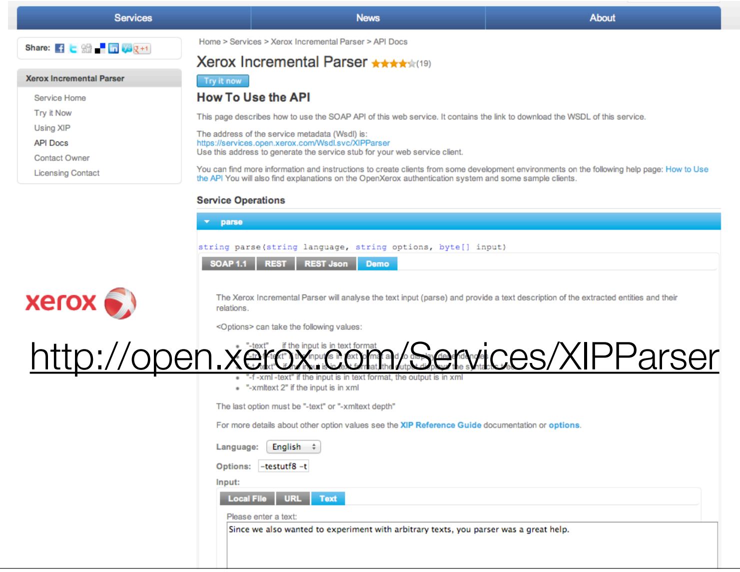
Though Apple stock fell after Tuesday's iPad mini announcement with a $329 starting price, analysts are largely sticking by their price targets for AAPL stock. Chris Whitmore of Deutsche Bank said pricing of the iPad mini was a "modest disappointment," but he still expects the product to perform well.



Stanford CoreNLP (3)



Analyseur Xerox XIP



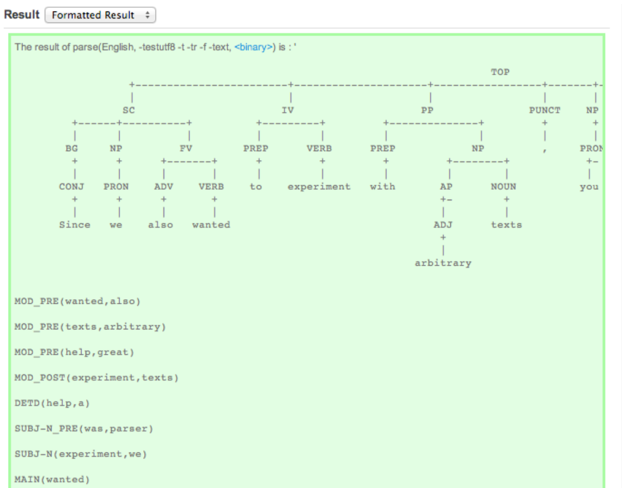
xerox

http://open.xerox.com/Services/XIPParser

Analyseur Xerox XIP (2)

- *Since we also wanted to experiment with arbitrary texts, you parser was a great help.*

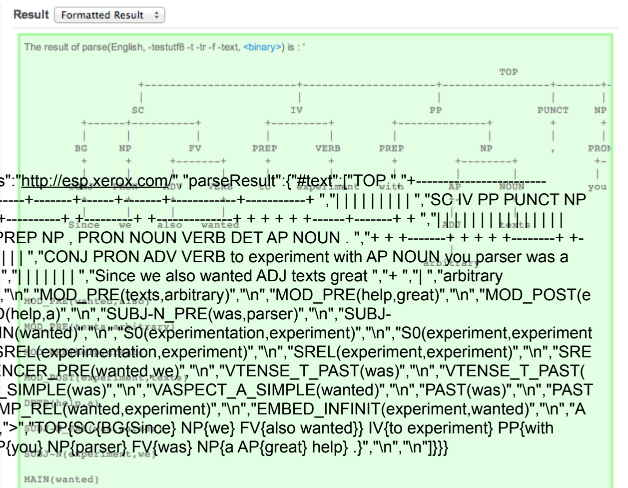
xerox



Analyseur Xerox XIP (2)

- *Since we also wanted to experiment with arbitrary texts, you parser was a great help.*

xerox

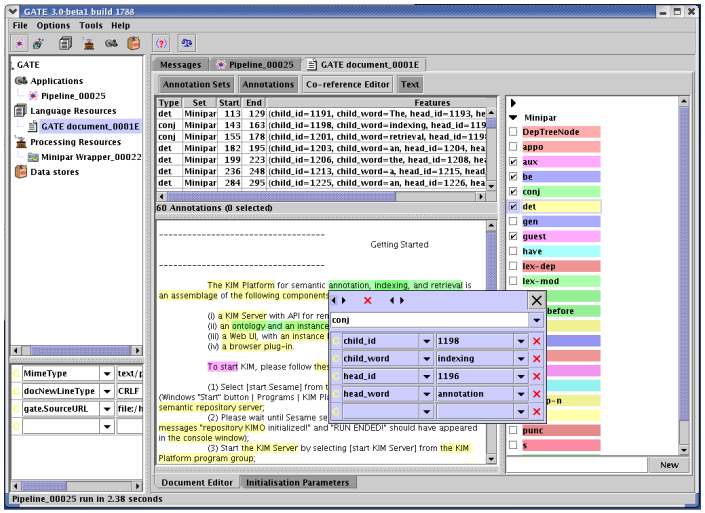


{"parseResponse":{"@xmlns":"http://esp.xerox.com/","parseResult":{"#text":["TOP ","+--------------------+--------------------+----------------+-------+------+-------+----------+----------+ ","| | | | | | | | | ","SC IV PP PUNCT NP NP FV NP PUNCT ","+------+----------+ +---------+ +-------------+ + + + + + +------+-------+ + ","| | | | | | | | | | | | | | | ","BG NP FV PREP VERB PREP NP , PRON NOUN VERB DET AP NOUN . ","+ + +-------+ + + + +--------+ +- + +- +- + + ","| | | | | | | | | | | | | | | ","CONJ PRON ADV VERB to experiment with AP NOUN you parser was a NADJ help ","+ + + + +- + + ","| | | | | | | ","Since we also wanted ADJ texts great ","+ ","| ","arbitrary ","MOD_PRE(wanted,also)","\n","MOD_PRE(texts,arbitrary)","\n","MOD_PRE(help,great)","\n","MOD_POST(experiment,texts)","\n","DETD(help,a)","\n","SUBJ-N_PRE(was,parser)","\n","SUBJ-N(experiment,we)","\n","MAIN(wanted)","\n","S0(experimentation,experiment)","\n","S0(experiment,experiment)","\n","S0(help,help)","\n","SREL(experimentation,experiment)","\n","SREL(experiment,experiment)","\n","SREL(help,help)","\n","EXPERIENCER_PRE(wanted,we)","\n","VTENSE_T_PAST(was)","\n","VTENSE_T_PAST(wanted)","\n","VASPECT_A_SIMPLE(was)","\n","VASPECT_A_SIMPLE(wanted)","\n","PAST(was)","\n","PAST(wanted)","\n","ORDER_TEMP_REL(wanted,experiment)","\n","EMBED_INFINIT(experiment,wanted)","\n","ATTRIB(parser,help)","\n","0",">","TOP{SC{BG{Since} NP{we} FV{also wanted}} IV{to experiment} PP{with NP{AP{arbitrary} texts}} , NP{you} NP{parser} FV{was} NP{a AP{great} help} .}","\n","\n"]}}}

MiniPar Parser

http://gate.ac.uk/sale/tao/splitch17.html

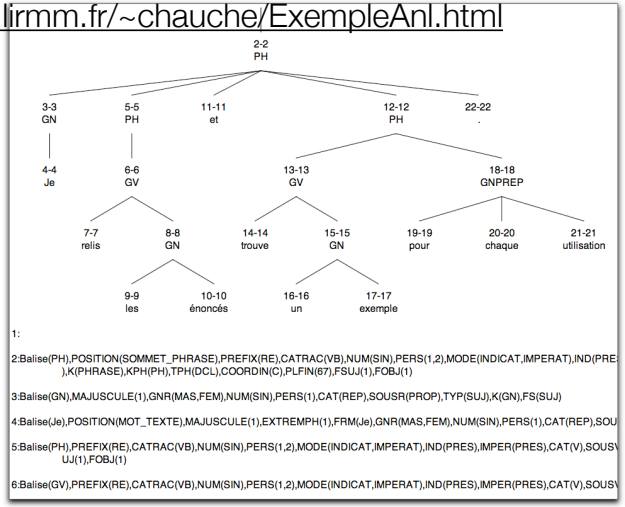
- Partie du discours (POS), lemme, tête, dépendances



LIRMM - Sygmart

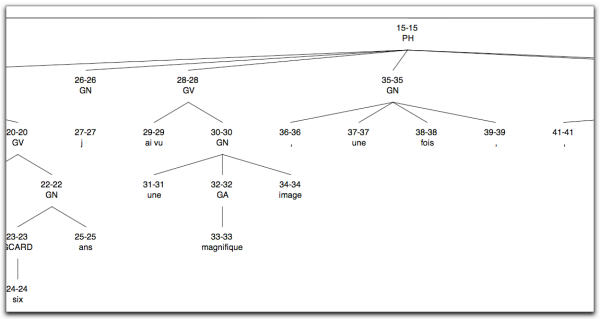
- *Je relis les énoncés et trouve un exemple pour chaque utilisation.*

http://www.lirmm.fr/~chauche/ExempleAnl.html



LIRMM : Sygmart

- Analyse complète du Petit Prince (http://www.sygtext.fr/Le_petit_prince.svg)



Ressources linguistiques

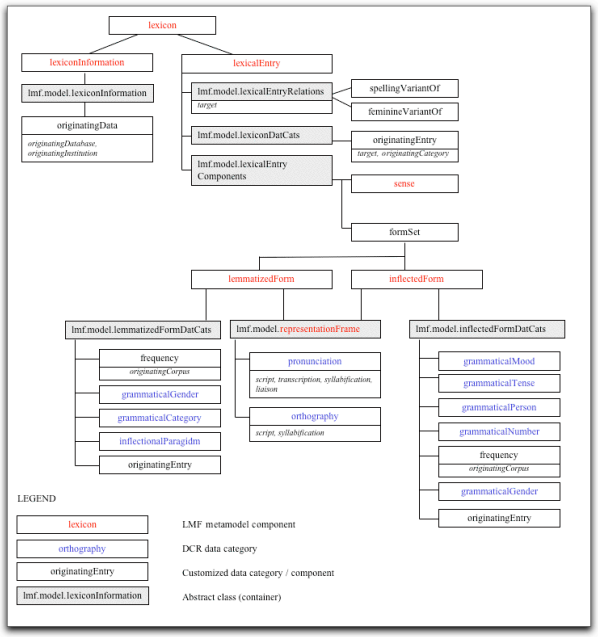
http://www.lexique.org

Dictionnaire de formes fléchies : Morphalou

Dictionnaire de formes fléchies : Morphalou

http://www.cnrtl.fr/lexiques/prolex/

Lexique multilingue de noms propres : Prolex



Lexiques liés à la subjectivité, aux opinions...

- Anglais :
  - The MPQA (Multi-Perspective Question Answering) Subjectivity Lexicon
  - SentiWordNet (valeurs >0 ou <0 aux synsets de WordNet)

```
# POS   ID      PosScore        NegScore        SynsetTerms     Gloss
a       00001740        0.125   0       able#1  (usually followed by `to') having the necessary means or [...]
a       00002098        0       0.75    unable#1        (usually followed by `to') not having the necessary means or [...]
a       00002312        0       0       dorsal#2 abaxial#1      facing away from the axis of an organ or organism; [...]
a       00002527        0       0       ventral#2 adaxial#1     nearest to or facing toward the axis of an organ or organism; [
a       00002730        0       0       acroscopic#1    facing or on the side toward the apex
a       00002843        0       0       basiscopic#1    facing or on the side toward the base
a       00002956        0       0       abducting#1 abducent#1  especially of muscles; [...]
a       00003131        0       0       adductive#1 adducting#1 adducent#1      especially of muscles; [...]
a       00003356        0       0       nascent#1       being born or beginning; [...]
a       00003553        0       0       emerging#2 emergent#2   coming into existence; [...]
```

- Espagnol : http://lit.csci.unt.edu/~rada/downloads/SpanishSentimentLexicons.tar.gz

WordNet

http://wordnet.princeton.edu/

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117097	81426	145104
Verb	11488	13650	24890
Adjective	22141	18877	31302
Adverb	4601	3644	5720
Totals	155327	117597	207016

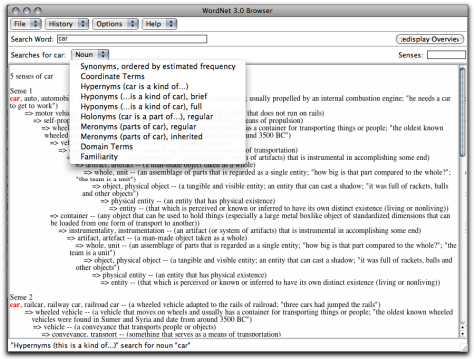


Table 1. Semantic Relations in WordNet

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry

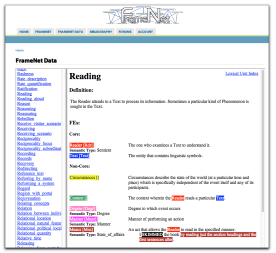
Note: ***N*** *= Nouns* ***Aj*** *= Adjectives* ***V*** *= Verbs* ***Av*** *= Adverbs*

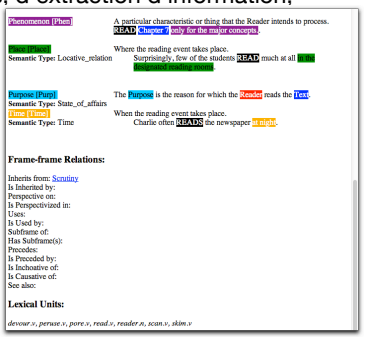
WordNet: A Lexical Database for English
George A. Miller
COMMUNICATIONS OF THE ACM November 1995/Vol. 38, No. 11

WOLF : Wordnet Libre du Français http://alpage.inria.fr/~sagot/wolf.html

FrameNet

- Hypothèse : les mots peuvent être compris selon la manière (structure conceptuelle) dont ils sont employés
- >> 1000 structures sémantiques conceptuelles décrivant des relations entre *participants*
- >> 170 000 phrases étiquetées --> *semantic role labeling*
  - Application aux tâches de questions-réponses, d'extraction d'information, de détection d'inférence (*textual entailment*)





VerbNet

- http://verbs.colorado.edu/~mpalmer/projects/verbnet.html
- Verbes organisés en *classes* (cohérence syntaxique et sémantique) : rôles thématiques, restrictions sur ses arguments, *frames*

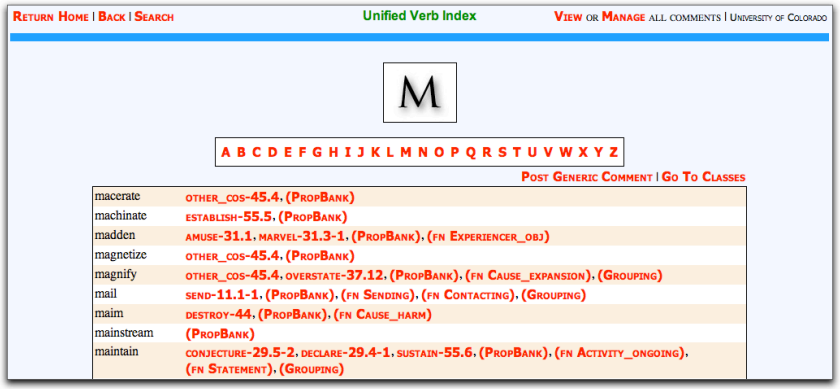
Table 2: Thematic roles and example classes that use them

Actor:	used for some communication classes (e.g., Chitchat-37.6, Marry-36.2, Meet-36.2) when both arguments can be considered symmetrical (pseudo-agents).
Agent:	generally a human or an animate subject. Used mostly as a volitional agent, but also used in VerbNet for internally controlled subjects such as forces and machines.
Asset:	used for the Sum of Money Alternation, present in classes such as Build-26.1, Get-13.5.1, and Obtain-13.5.2 with 'currency' as a selectional restriction.
Attribute:	attribute of Patient/Theme refers to a quality of something that is being changed, as in (The price)att of oil soared. At the moment, we have only one class using this role Calibratable cos-45.6 to capture the Possessor Subject Possessor-Attribute Factoring Alternation. The selectional restriction 'scalar' (defined as a quantity, such as mass, length, time, or temperature, which is completely specified by a number on an appropriate scale) ensures the nature of Attribute.
Beneficiary:	the entity that benefits from some action. Used by such classes asBuild-26.1, Get-13.5.1, Performance-26.7, Preparing-26.3, and Steal-10.5. Generally introduced by the preposition 'for', or double object variant in the benefactive alternation.
Cause:	used mostly by classes involving Psychological Verbs and Verbs Involving the Body.
Location, Destination, Source:	used for spatial locations.
Destination:	end point of the motion, or direction towards which the motion is directed. Used with a 'to' prepositional phrase by classes of change of location, such as Banish-10.2, and Verbs of Sending and Carrying. Also used as location direct objects in classes where the concept of destination is implicit (and location could not be Source), such as Butter-9.9, or Image impression-25.1.
Source:	start point of the motion. Usually introduced by a source prepositional phrase (mostly headed by 'from' or 'out of'). It is also used as a direct object in such classes as Clear-10.3, Leave-51.2, and Wipe instr-10.4.2.
Location:	underspecified destination, source, or place, in general introduced by a locative or path prepositional phrase.
Experiencer:	used for a participant that is aware or experiencing something. In VerbNet it is used by classes involving Psychological Verbs, Verbs of Perception, Touch, and Verbs Involving the Body.
	used only in the Calibratable-45.6 class, to specify the range or degree of change, as in The price of oil

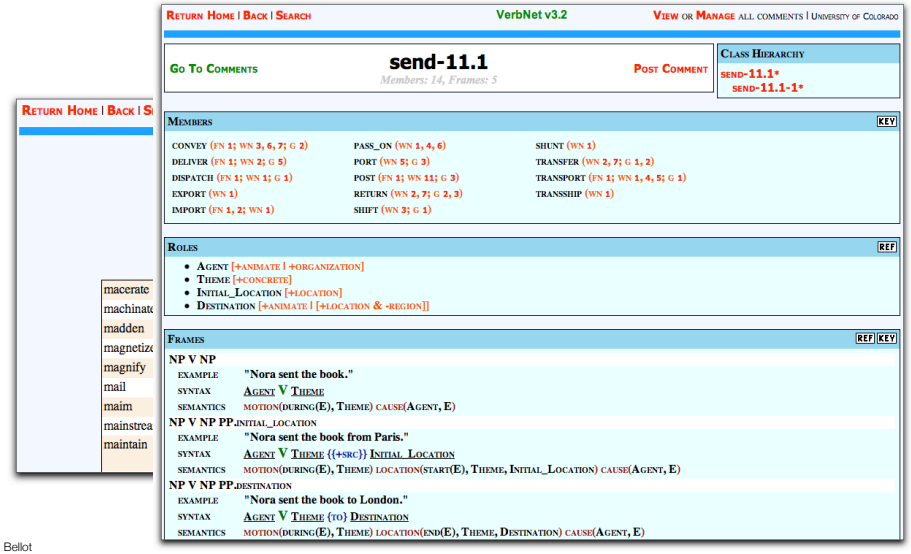
The index has...
6272 total verbs represented
6033 total VerbNet links
270 total VerbNet main classes
202 total VerbNet subclasses
5649 total PropBank links
4186 total FrameNet links
2199 total Grouping links

Verbnet (2)

http://verbs.colorado.edu/verb-index/index/M.php

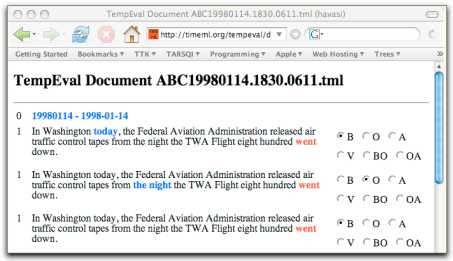


Verbnet (2)



Corpus avec annotations TimeML

- TimeBank : 183 articles journalistiques annotés
- TempEval 1 : relations temporelles entre *événements* principaux *(before, after, overlap)*
- TempEval 2
  — multilingue (Chinese, English, French, Italian, Korean and Spanish)



Entités nommées, bases de connaissances

Conférences MUC

- MUC (Message Understanding)
  - 7 catégories d'entités

  people, organization, location, time, date, money, percentage
  - peu réaliste pour des tâches plus complexes (ex. questions-réponses)

http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

Table 2: Maximum Results Reported in MUC-3 through MUC-7 by Task

Evaluation/ Tasks	Named Entity	Coreference	Template Element	Template Relation	Scenario Template	Multilingual
MUC-3					R < 50% P < 70%	
MUC-4					F < 56%	
MUC-5					EJV F < 53% EME F < 50%	JJV F < 64% JME F < 57%
MUC-6	F < 97%	R < 63% P < 72%	F < 80%		F < 57%	
MUC-7	F < 94%	F < 62%	F < 87%	F < 76%	F < 51%	
Multilingual						
MET-1	C F < 85% J F < 93% S F < 94%					
MET-2	C F < 91% J F < 87%					

Legend:

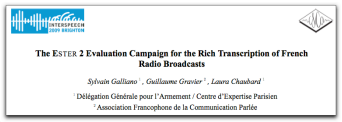
R = Recall P = Precision F = F-Measure with Recall and Precision Weighted Equally

E = English C = Chinese J = Japanese S = Spanish

JV = Joint Venture ME = Microelectronics

Evaluations ESTER

- Transcriptions audio (300 h. transcrites, 1600 h. non transcrites)
- 7 catégories (*persons, locations, organizations, human products, amounts, time and functions*) et 38 sous-catégories



The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts

Sylvain Galliano [1], Guillaume Gravier [2], Laura Chaubard [1]

[1] Délégation Générale pour l'Armement / Centre d'Expertise Parisien

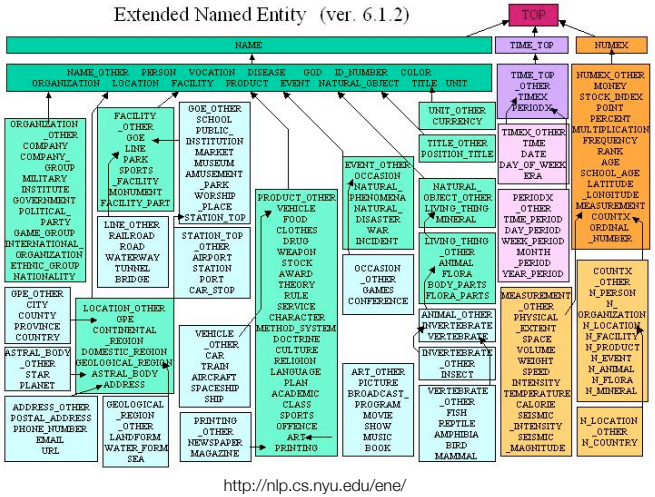
[2] Association Francophone de la Communication Parlée

Table 6: NE task overall performance for each participating site (Slot Error Rate, Precision, Recall and F-measure)

%WER	ref. transcript			LIMSI transcript 12.11			LIA transcript 17.83			IRISA transcript 26.09		
Sites	%S	%P	%R	%S	%P	%R	%S	%P	%R	%S	%P	%R
LIA	23.9	86.46	71.85	**43.4**	79.52	59.45	**51.6**	76.51	55.02	**56.8**	72.26	49.02
LIMSI	30.9	81.15	70.94	45.3	75.13	62.33	55.5	70.50	57.52	61.2	66.13	50.67
LINA	37.1	80.75	55.48	54.0	71.98	44.01	60.4	68.76	40.84	65.2	63.66	35.66
LI Tours	33.7	79.39	65.82	50.7	71.36	54.16	80.8	56.59	46.46	82.9	51.28	42.38
LSIS	35.0	82.65	73.07	55.3	70.23	58.39	86.5	70.36	28.66	88.6	67.03	25.22
Synapse	9.9	93.02	89.37	44.9	76.39	67.16	60.7	70.26	59.21	66.2	65.95	52.71
Xerox	**9.8**	93.61	91.50	44.6	58.91	70.06	—	—	—	—	—	—

La hiérarchie de Sekine

- Plus de 200 types



http://nlp.cs.nyu.edu/ene/

La hiérarchie étendue de Sekine

ENE			Examples
Name_Other			Barbaro, Bubbles, Max, Maggie
Person			Bush, Michael Jackson, Elizabeth II, LeBron Raymone James
God			Zeus, Indra, Danu, Ra
Organization	Organization_Other		the Capone Family, Department of Computer Science, CS Dept., general affairs department
	International_Organization		UN, League of Nations, Pacific Island Forum, SEATO
	Show_Organization		The Cleveland Orchestra, The Beatles, the Bolshoi Ballet troupe, Sex Pistols
	Family		The House of Hamilton, Clan Henderson, Tokugawa clan, Koga family
	Ethnic_Group	Ethnic_Group_Other	White people, Jew, Slavic peoples, Mongoloid race, Japanese Diaspora
		Nationality	Japanese, Israeli, American, American people
	Sports_Organization	Sports_Organization_Other	the Breen Gym, UCLA Bruins, Ma family army, Shinagawa Jogging Club
		Pro_Sports_Organization	New York Yankees, Seattle, NYY, Manchester United
		Sports_League	NFL, National Basketball Association, Atlantic Coast Conference, National League West
	Corporation	Corporation_Other	Association for Computational Linguistics, National Rifle Association, NHK, BBC
		Company	Toyota, SONY, CNN, Microsoft
		Company_Group	Tata Group, JR, the Big Three, Big Four auditors
	Political_Organization	Political_Organization_Other	Palestine Liberation Organization, Clinton Regime, Tokugawa shogunate, Ayyubid dynasty
		Government	National Security Council, Ministry of Finance, the United States Senate, USTR
		Political_Party	Democratic Party, Bharatiya Janata Party, Conservative Party, LDP
		Cabinet	Thatcher's Cabinet, Major's Cabinet, Tanaka's Cabinet, Koizumi's Cabinet
		Military	Self-Defense Forces, US Air Force, Royal Navy, UN forces
Location	Location_Other		Times Square, Ground Zero, Three Views of Japan, Garden of Eden
	Spa		Hakone Spa, Fukuchi Spa, Hakuba Spa, Yunoyama Spa
	GPE	GPE_Other	Taiwan, Hong Kong, Puerto Rico, French Polynesia, Macau
		City	New York City, Brooklyn, Sydney, Rio de Janeiro
		County	West Chester County, Madison County, Orange County, Shima District
		Province	Osaka Prefecture, NY, Kansas, Nova Scotia, Nagorno-Karabakh
		Country	the United States, Japan, UK, Vatican City
	Region	Region_Other	
		Continental_Region	North America, Asia, the Caribbean area, NIES
		Domestic_Region	New England, East Coast, the South, Upper New York
	Geological_Region	Geological_Region_Other	Grand Canyon, Altamira Cave, Great Barrier Reef, Ayers Rock
		Mountain	Mount Everest, K2, Mt. Fuji, Alps
		Island	Florida Keys, Key West, Gilbert Islands, Iriomote
		River	Mississippi River, Hudson River, Yangtze River, Danube
		Lake	Lake Michigan, Lake Baikal, Dead Sea, Great Lakes
		Sea	Pacific Ocean, Sea of Japan, Sunda Strait, English Channel
		Bay	Bay of Bengal, Delaware Bay, Persian Gulf, Gulf of Guinea
	Astral_Body	Astral_Body_Other	Andromeda Galaxy, Solar System, Halley's Comet, Callisto
		Star	Antares, Sirius, North Star, Barnard's Star
		Planet	Sun, Earth, Moon, Icarus
		Constellation	Taurus, Cassiopeia, Argo Navis, Lepus
	Address	Address_Other	
		Postal_Address	715 Broadway, 7th floor, New York, NY 10003 USA, 715 Broadway, 10003
		Phone_Number	(212) 123-4567, 911, ext445
		Email	sekine@cs.nyu.edu
		URL	http://nlp.cs.nyu.edu/sekine/index-jp.html
	Facility_Other		Empire State Building, Cooper Dam, Fulmar oilfield, Eiffel Tower
	Facility_Part		8th floor, room #1204, second basement, Runway 13R-31
	Archaeological_Place	Archaeological_Place_Other	Archaeological Ruins at Moenjodaro, Cahokia Mounds State Historic Site, Angkor, Masada
		Tumulus	Daisen-Kofun, Zhaoling, Ringlemere barrow, Great Pyramid of Giza

Evaluation QUAERO

http://www.quaero.org/

- Plus d'un million de mots annotés
- Entités et composants

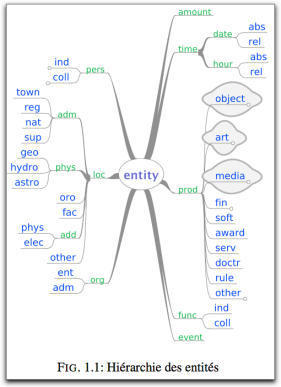


FIG. 1.1: Hiérarchie des entités

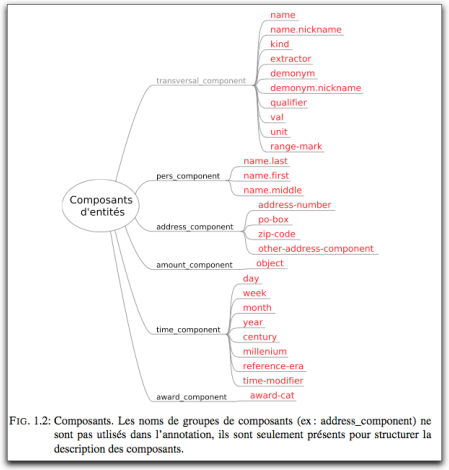


FIG. 1.2: Composants. Les noms de groupes de composants (ex : address_component) ne sont pas utlisés dans l'annotation, ils sont seulement présents pour structurer la description des composants.

Evaluation QUAERO (2)

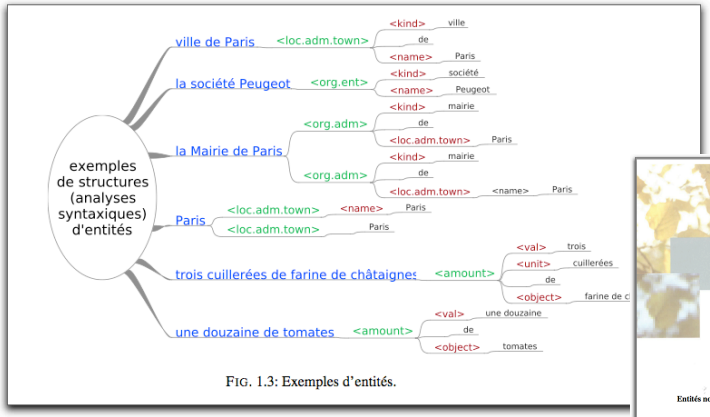


FIG. 1.3: Exemples d'entités.

Rosset, Grouin, Zweigenbaum (2011)
http://quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf

Stanford NLP

http://nlp.stanford.edu/ner/index.shtml

- Approche statistique (Markov maximum entropie, CRF)

  Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

- Evaluation dans le cadre de CoNLL *Conference on Computational Natural Language Learning* http://www.cnts.ua.ac.be/conll2003

  types reconnus : persons, organizations, locations, times and quantities

Corpus		# Word Tokens		# Entities		# Features		Exact Match Score (conlleval)			
Name	Language	Train	Test	Types	Instances	Φ(X)	λ/f(X,Y)	Prec	Rec	F_1	Classifier
CoNLL 2002	Dutch news testa (devset)	218737	37761	4	2616	838524	4192620	78.99%	77.33%	78.15%	pure CMM
CoNLL 2002	Dutch news testb	218737	68994	4	3941	838559	4192795	80.48%	78.96%	79.71%	pure CMM
CoNLL 2002	Spanish news testa (devset)	273037	52923	4	4352	776511	3882555	78.01%	76.19%	77.09%	pure CMM
CoNLL 2002	Spanish news testb	273037	51533	4	3559	776444	3882220	81.24%	81.03%	81.14%	pure CMM
CoNLL 2003	English news testa (devset)	219553	51578	4	5942	738378	3691890	91.37%	91.22%	91.29%	pure CMM
CoNLL 2003	English news testa (devset)	219554	51578	4	5942			92.15%	92.39%	92.27%	postprocessed CMM
CoNLL 2003	English news testb	219553	46666	4	5648	738378	3691890	85.65%	85.41%	85.53%	pure CMM
CoNLL 2003	English news testb	219554	46666	4	5648			86.12%	86.49%	86.31%	postprocessed CMM
CoNLL 2003	German news testa (devset)	220189	51645	4	4833	1079044	5395220	77.12%	61.37%	68.35%	pure CMM
CoNLL 2003	German news testa (devset)	220189	51645	4	4833			75.36%	60.36%	67.03%	postprocessed CMM
CoNLL 2003	German news testb	220189	52098	4	3673	1079037	5395185	79.23%	63.65%	70.59%	pure CMM
CoNLL 2003	German news testb	220189	52098	4	3673			80.38%	65.04%	71.90%	postprocessed CMM
CoNLL 2003	English news testa (devset)	219553	51578	4	5942	616918	11532202	91.64%	90.93%	91.28%	CRF (closed task)
CoNLL 2003	English news testa (devset)	219553	51578	4	5942	633786	12285708	93.28%	92.71%	92.99%	CRF (with distsim)
CoNLL 2003	English news testb	219553	46666	4	5648	633786	12285708	88.21%	87.68%	87.94%	CRF (with distsim)

Stanford NER (2)

Feature	NER	TF
Current Word	✓	✓
Previous Word	✓	✓
Next Word	✓	✓
Current Word Character n-gram	all	length ≤ 6
Current POS Tag	✓	
Surrounding POS Tag Sequence	✓	
Current Word Shape	✓	✓
Surrounding Word Shape Sequence	✓	✓
Presence of Word in Left Window	size 4	size 9
Presence of Word in Right Window	size 4	size 9

Table 2: Features used by the CRF for the two tasks: named entity recognition (NER) and template filling (TF).

Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling

Jenny Rose Finkel, Trond Grenager, and Christopher Manning
Computer Science Department Stanford University

http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

CoNLL					
Approach	LOC	ORG	MISC	PER	ALL
B&M LT-RMN	–	–	–	–	80.09
B&M GLT-RMN	–	–	–	–	82.30
Local+Viterbi	88.16	80.83	78.51	90.36	85.51
NonLoc+Gibbs	88.51	81.72	80.43	92.29	86.86

Table 5: F_1 scores of the local CRF and non-local models on the CoNLL 2003 named entity recognition dataset. We also provide the results from Bunescu and Mooney (2004) for comparison.

```
U.N.        NNP  I-NP  I-ORG
official    NN   I-NP  O
Ekeus       NNP  I-NP  I-PER
heads       VBZ  I-VP  O
for         IN   I-PP  O
Baghdad     NNP  I-NP  I-LOC
.           .    O     O
```


Détection des entités nommées : un système symbolique open source à la disposition de la communauté

Nathalie Friburger, Denis Maurel

A l'origine

2000-2001 création de Cassys
- Système permettant de gérer une cascade de transducteurs (aujourd'hui intégré à Unitex)
- NER limitée aux personnes, lieux, organisations

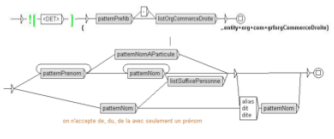
2007-2010 participation de CasEN à Ester2, EPAC, Variling
- Extension aux entités : montants, événements, fonctions, productions humaines

Refonte de CasEN
- Architecture de la cascade
- Nommage des graphes
- Ecriture des graphes
- Outils de deboggage

Pourquoi réécrire CasEN ?
- De nombreuses modifications apportées par des personnes différentes
- Maintenance difficile
- Nouvelles fonctionnalités d'Unitex : morphologie, contextes à exclure ou pas
- Modification du fonctionnement de CasSys / intégration à Unitex

Différents types de transducteurs

Fonction des transducteurs identifiable par le préfixe de leur nom
- **pers, loc, org, event, fonc, prod, amount, time** : reconnaissent des entités nommées correspondant au type indiqué (peuvent rassemble plusieurs transducteurs)
  ex : orgCommerceEtranger reconnait des organisations commerciales ayant un nom étranger
  ex : amount reconnait différentes mesures (monnaie, température, longueur, etc.).
- **tool** : transducteur outil.
  Ex : toolChercheSigleAvecPoints recherche les sigles contenant des points (*C.G.T.* au lieu de *CGT* aujourd'hui).
- **list** : transducteur de listes
- **pattern** : transducteurs masques, décrivent des motifs utilisant des codes Unitex pour sélectionner des séquences de mots (contraintes morphologiques, syntaxiques, lexicales etc.)
- **tag** : transducteur étiquette, permettent d'insérer des étiquettes dans des entités imbriquées



ettes « extérieures » , placées sur les ENs, générale {xxxx,.y1+y2+…+y3}

y1, y2, …, yn peuvent être recherchés ansducteurs unitex par des masques plus écifiques.

<y2+y3>, <y2–y3>

{{Gerhard,.N+Prénom}{Aigner,.N+nom},.entity+pers+hum+grfpersPrenomNom}, le {secrétaire général,.entity+fonc+pol} de l'{UEFA,.entity+org+grforgMetier}, a dans ses placards des projets de {Ligue européenne,.entity+org+grforgDivers}.

Ex : <pers+hum> → {{Gerhard,.N+Prénom}{Aigner,.N+nom},.entity+pers+hum+grfpersPrenomNom}
<fonc> de <org> → { {secrétaire général ,.entity+fonc+pol} de l' {UEFA ,.entity+org+grforgMetier},.

Friburger N., Maurel D. (2004), Finite-state transducer cascade to extract named entities in texts, *Theoretical Computer Science*, vol. 313, 94-104.

Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I., Nouvel D. (2011), Cascades autour de la reconnaissance des entités nommées, *TAL* 52-1.

CasEN http://tln.li.univ-tours.fr/Tln_CasEN.html

personne (*pers*)	humain réel ou fictif (*pers.hum*) animal réel ou fictif (*pers.anim*)	civilité (*pers.hum.title*) titre professionnel (*pers+hum+fonc*) adjectif ethnique (*pers+hum+ethnic*) gentilés et adjectifs toponymiques (*pers+hum+gent*) nationalité (*pers+hum+nat*) dynastie (*pers+hum+dyn*)
fonction (*fonc*)	politique (*fonc.pol*) militaire (*fonc.mil*) administrative (*fonc.admi*) religieuse (*fonc.rel*) aristocratique (*fonc.ari*)	
organisation (*org*)	politique (*org.pol*) éducative (*org.edu*) commerciale (*org.com*) média & divertissement (*org.div*)	
lieu (*loc*)	géographique naturel (*loc.geo*) axe de circulation (*loc.line*) construction humaine (*loc.fac*)	
	région administrative (*loc.admi*)	ville (*loc.admi.ville*)
	adresse (*loc.addr*)	adresse postale (*loc.addr.post*) téléphone et fax (*loc.addr.tel*) adresse électronique (*loc.addr.elec*)
production humaine (*prod*)	produit (*prod.obj*) récompense (*prod.award*) œuvre artistique (*prod.art*) production documentaire (*prod.doc*)	
date et heure (*time*)	date (*time.date*)	date absolue (*time.date.abs*) date relative (*time.date.rel*)
	heure (*time.hour*) adverbe de date (*time.advdate*)	
montant (*amount*)	valeur physique (*amount.phy*)	durée (*amount.phy.dur*) température (*amount. phy.temp*) longueur (*amount. phy.len*) surface et aire (*amount.phy.area*) volume (*amount.phy.vol*) poids (*amount. phy.wei*) vitesse (*amount. phy.spd*) autre (*amount.phy.other*)
	valeur monétaire (*amount.cur*) octets (*amount.computer*)	
évènement (*event*)	histoire (*event+hist*) célébration (*event+cel*) fête (*event+feast*) manifestation (*event+manif*) météorologie (*event+meteo*)	

CasEN http://tln.li.univ-tours.fr/Tln_CasEN.html

personne (*pers*)	humain réel ou fictif (*pers.hum*)	civilité (*pers.hum.title*)

Donnons quelques exemples de reconnaissance:

« Au pire de la crise, {à l'automne dernier,.entity+time+date+rel+grftimeDateRelative}, nous avons détenu jusqu'à 20 % de liquidités dans notre portefeuille », indique {{{ Denis,.N+Prénom} { Remacle,.N+nom},.entity+pers+hum}, {gérant d'{Amplitude Pacifique,.entity+org+com} ,.entity+job} ,.entity+pers+hum+grfpersPrenomNom}, une sicav de {La Poste,.entity+org+com+grforgDico}.

« C'est à nos clients de décider s'ils souhaitent ou non consacrer une partie de leur patrimoine à l'{Asie,.entity+loc+admi+grflocPays} » , souligne {{Pierre,.N+Prénom} {Ciret,.N+nom},.entity+pers+hum+grfpersPrenomNom}, de la {Compagnie financière {{Edmond,.N+Prénom} {de Rothschild,.N+nom} ,.entity+pers+hum} ,.entity+org+com+grforgCommerceGauche}.

Ils ne peuvent pas, en revanche, faire l'impasse sur la {Bourse de {Hongkong , .entity+loc+admi},.entity+org+com+grforgCommerceGauche}, car cette place représente près de la moitié de la capitalisation boursière de la région.{S} Pour sa part, {{ Pierre-Alexis,.N+Prénom} { Dumont,.N+nom},.entity+pers+hum+grfpersPrenomNom}, de {State Street Banque,.entity+org+com+grforgCommerceDroite}, s'est réfugié sur le marché australien, relativement épargné par la tourmente.

{Théâtre Gérard-Philipe,.entity+org+div+grforgDivertissementSorties}, {59, {boulevard Jules-Guesde,.entity+loc+line}, 93000 {Saint-Denis,.entity+loc+ville},.entity+loc+addr+post+grflocAddr}.

Selon une étude de l'{Autorité de régulation des télécommunications,.entity+org+grforgDivers} {{ART,.entity+org+ grforgOrgSuiviDeParentheses}}, le taux d'équipement devrait dépasser les 50 % {en 2002,.entity+time+date+abs+grftimeAnneeSiecle}.

Une évaluation a été réalisée sur un extrait du journal Le Monde daté du 1er janvier 1999, 7 articles, soit 7 070 mots pour 83.2 ko. En voici les résultats :

- Sur 582 entités nommées présentes, 432 ont été reconnues, avec, en plus, 11 fausses reconnaissances, soit :
  - Précision : 97,52%
  - Rappel : 74,23%
- Sur les entités reconnues, 390 types étaient corrects et 3 types comportaient une erreur due à la métonymie (par exemple *France* comme *toponyme* au lieu d'*organisation*).
  - Sans compter les erreurs dues à la métonymie :
    - Précision : 88,71%
    - Rappel : 67,53%
  - En comptant les erreurs dues à la métonymie :
    - Précision : 88,04%
    - Rappel : 67,01%
- Sur les entités reconnues et bien typées, 384 sous-types étaient corrects, soit :
  - Précision : 85,52%
  - Rappel : 65,98%
- Enfin, à nouveau sur les entités reconnues, 393 étaient correctement balisées, 20 avaient un début défectueux et une fin correcte, 16 un début correct et une fin défectueuse, 3 un début et une fin défectueux, soit :
  - Précision : 90,97%
  - Rappel : 67,53%

Entités et propriétés

- *Infoboxes* de Wikipédia
- Ontologies YAGO, SUMO...

Table 5: Size of other ontologies

Ontology	# Entities	# Facts
SUMO [114]	20,000	60,000
Ponzetto et al. [118]	n/a	110,000
WordNet [66]	117,659	821,492
Cyc [106]	300,000	3,000,000
TextRunner [13]	n/a	7,800,000
YAGO	1,700,000	15,000,000
DBpedia [9]	1,950,000	103,000,000

Yago: A Large Ontology from Wikipedia and WordNet

Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum

MPI–I–2007–5-003 December 2007

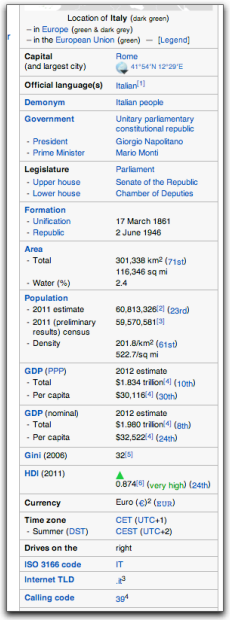
Table 4: Largest relations in YAGO

Relation	# Facts	Relation	# Facts
hasUTCOffset	12724	hasWonPrize	13645
livesIn	15185	writtenInYear	16441
originatesFrom	16876	directed	18633
hasPredecessor	19154	actedIn	22249
hasDuration	23652	bornInLocation	24400
hasImdb	24659	hasArea	26781
hasProductionLanguage	27840	produced	30519
hasPopulation	30731	isOfGenre	33898
hasSuccessor	46658	establishedOnDate	69529
hasWebsite	79779	created	83627
locatedIn	125738	diedOnDate	168037
subClassOf	211979	bornOnDate	350613
givenNameOf	464816	familyNameOf	466969
inLanguage	2389627	isCalled	2984362
type	3957223	means	4014819

Table 6: Simple queries on YAGO

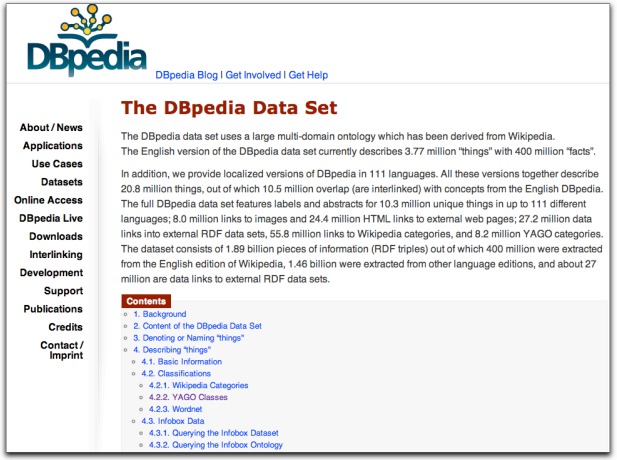
Query	Result
Who was Einstein's doctoral advisor? `Einstein hasDoctoralAdvisor $x`	`$x=Alfred Kleiner`

Who is named after a place in Africa?
`$place locatedin Africa`
`$name means $place`
`$name familynameof $who`
	`$who=Gabriel Sudan`
	and 22 more



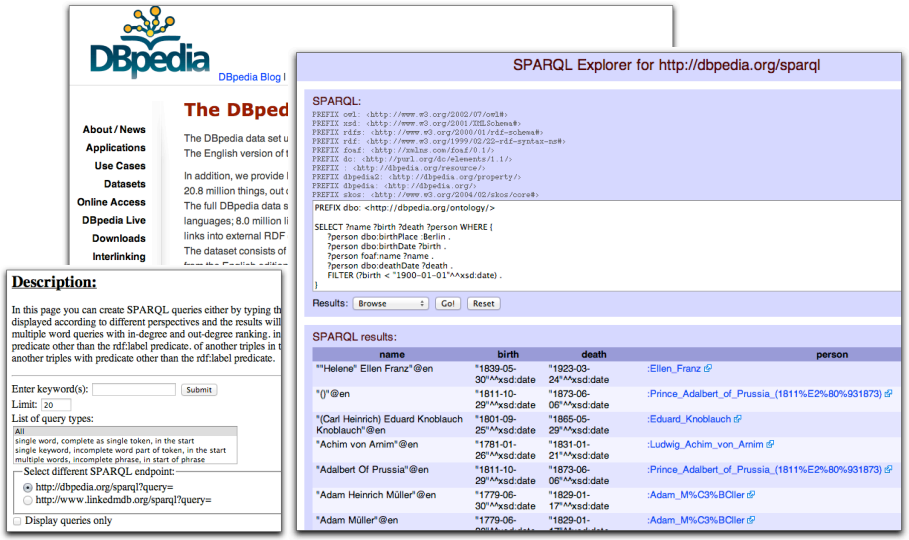
DBpedia

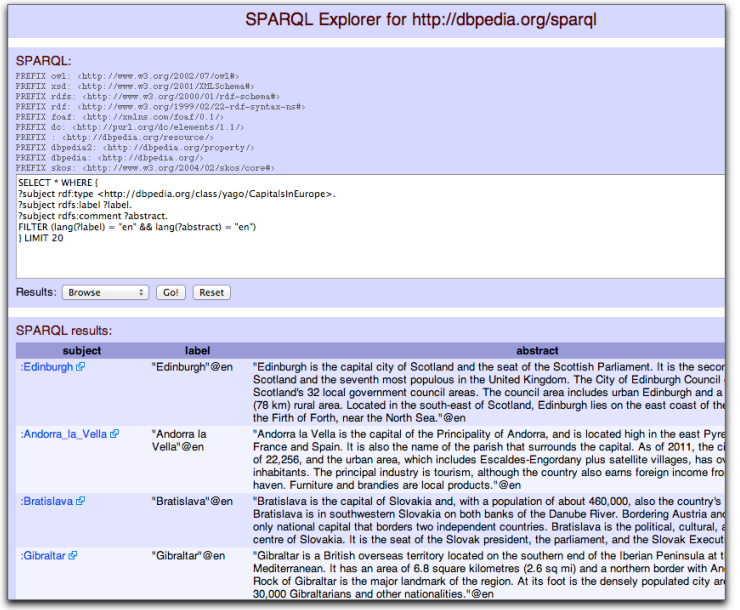
http://dbpedia.aksw.org/dbpedia_demo/dbpedia/tutorials/ranked_keyword_search/demo.php



DBpedia

http://dbpedia.aksw.org/dbpedia_demo/dbpedia/tutorials/ranked_keyword_search/demo.php





SPARQL:

```
PREFIX owl: <http://www
PREFIX xsd: <http://www
PREFIX rdfs: <http://ww
PREFIX rdf: <http://www
PREFIX foaf: <http://xm
PREFIX dc: <http://purl
PREFIX : <http://dbpedi
PREFIX dbpedia2: <http:
PREFIX dbpedia: <http:/
PREFIX skos: <http://ww
SELECT * WHERE {
?subject rdf:type <http:/
?subject rdfs:label ?label.
?subject rdfs:comment ?c
FILTER (lang(?label) = "er
} LIMIT 20
```

Results: Browse

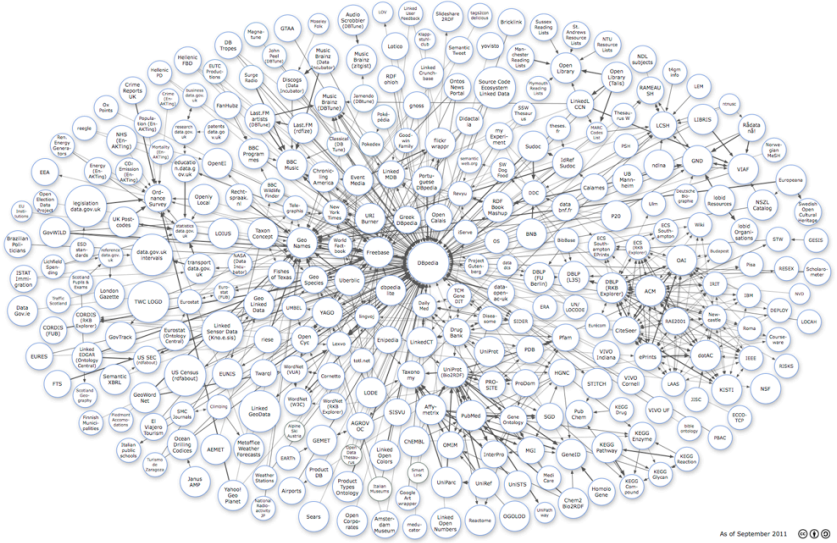
SPARQL results:

subject
:Edinburgh
:Andorra_la_Vella
:Bratislava
:Gibraltar

dbpedia2:partType	"Districts"@en
dbpedia2:highestElevation	514
dbpedia2:codeType	:Vehicle_registration_plates_of_Slovakia
dbpedia2:map1Locator	"Bratislava Region"@en
dbpedia2:map1Size	128
dbpedia2:imageCaption	"Bratislava Montage"@en
dbpedia2:symbolType	"Coat of arms"@en
dbpedia2:utcOffset	"+1"@en
dbpedia2:mapCaption	"Location in Slovakia"@en
dbpedia2:singleLine	"yes"@en
dbpedia2:population	462603
dbpedia2:postalCode	8
dbpedia2:janPrecipitationMm	42
dbpedia2:febPrecipitationMm	37
dbpedia2:junPrecipitationMm	61
dbpedia2:augPrecipitationMm	52
dbpedia2:latNs	"N"@en
dbpedia2:longD	17
dbpedia2:longM	6
dbpedia2:longS	35
dbpedia2:mayor	:Milan_Ft%C3%A1%C4%8Dnik
dbpedia2:populationDate	2012
dbpedia2:populationDensity	"auto"@en
dbpedia2:establishedType	"First mentioned"@en
dbpedia2:areaUrban	853
dbpedia2:areaMetro	2053
dbpedia2:countryFlag	"true"@en
dbpedia2:elevation	134
dbpedia2:statistics	"[http://www.statistics.sk/mosmis/eng/prvav2.jsp?txtUroven=410190&lst MOS/MIS]"@en
dbpedia2:government	"City council"@en
dbpedia2:river	:Little_Danube
dbpedia2:river	:Danube
dbpedia2:river	:Morava_(river)

LINKINGOPENDATA
W3C SWEO Community Project

http://linkeddata.org



As of September 2011

http://richard.cyganiak.de/2007/10/lod/

Extraction d'informations

Exemple

Annotation de références bibliographiques avec des automates à états finis

Approches symboliques avec Unitex

Exemple : Annotation de documents OCRisés (UNESCO)
M. Softi, P. Bellot et al., 2011

26866-26914 Royaume-Uni-United Kingdom

26866 **Boss**, Medard : A psychiatrist discovers India. – Henry A. Frey. – London : Wolff, 1965. 192, 31/6. *Dt: Indienfahrt eines Psychiaters*

26867 **Boutrais**, Cyprien M. : Christ the vision. – London : Burns & Oates. viii, 85, 6/–. *Fr: Vivante clarté*

26868 **Buys**, J. : Christ, our light. – Peter McConville. – London : Chapman. 191, 7/6. *Fr: Jésus Christ, lumière du monde*

26869 **Carrier**, Hervé : The sociology of religious belonging. – Arthur J. Arrieri – London : Darton, Longman & Todd, 1965. 335, 30/–. *Fr: Psycho-sociologie de l'appartenance religieuse*

26870 **Congar**, Yves : Dialogue between Christians. – P. Loretz. – London : G. Chapman. vii, 472, 50/–. *Fr: Chrétiens en dialogue*

26871 — Jesus Christ. – Luke O'Neill. – London : Chapman. 223, 25/–. *Fr: Jésus Christ*

26872 — Tradition and traditions. – Michael Naseby & Thomas Rainborough. – London : Burns & Oates. xx, 536, 90/–. *Fr: La tradition et les traditions*

26873 **Cristiani**, Léon : Why we believe. – Mark Pontifex. – London : Burns & Oates. 123, 41/–. *Fr: Nos raisons de croire*

26874 **Daniel-Rops**, Henry [J.C.H. Petiot] : A fight for God : 1870-1939. – John Warrington. – London : Dent. x, 452, 48/–. *Fr: L'Église des révolutions : un combat pour Dieu*

26875 **Desbuquois**, Gustave : Horn. – Kathleen Pond. – Tenbury Wells : Fowler Wright Books, 1965. 202. *Fr*

26876 **Dournes**, Jacques : God in Vietnam. – Rosemary Sheed. – London : G. Chapman. 203. 30/–. *Fr: Dieu aime les païens*

26877 **Dreze**, A : Living in Christ, liturgy and sacraments. – F.M. Gale & Jennifer Nicholson. – London : G.Chapman. 209, 8/6. *Fr: Jésus Christ notre vie*

26878 **Ebeling**, Gerhard : The Lord's prayer in today's world. – James W. Leitch. – London : S.C.M. Press. 139, 7/6. *Dt: Vom Gebet: Predigten über das Unser-Vater*

26879 — The nature of faith. – Ronald G. Smith. – London : Collins. 191, 7/6. *Dt: Das Wesen des Christlichen Glaubens*

26880 — Theology and proclamation. – John Reches. – London : Collins. 186, 28/–. *Dt: Theologie und Verkündigung*

26881 Ecumenical Dialogue in Europe. – Fletcher Fleet. – London : Lutterworth. 83, 12/6. *Dt: Dialogue œcuménique, les rencontres œcuméniques des Dombes*

26882 **Emery**, Pierre Y. : The communion of saints. – D.J. & M. Watson. – London : Faith P. xiii, 256. *Fr: L'unité des croyants au ciel et sur la terre*

26883 **Floristan**, Casiano : The parish : eucharistic community. – John F. Byrne. – London : Sheed & Ward, 1965. 240, 11/6. *Esp: La parroquia, comunidad eucarística*

26884 **Geiselmann**, Josef R. : The meaning of tradition. – W.J. O'Hara. – London : Burns & Oates. 123, 15/–. *Dt: Die Heilige Schrift und die Tradition*

26885 **Häring**, Bernhard : The new covenant. – R.A. Wilson. – London : Burns & Oates, 1965. xi, 281, 42/–. *Dt: Gabe und Aufrag der Gnadenzeit*

26886 **Hérin**, Charles V. : Spirituality of love. – David Martin. – London : Herder, 1965. 243, 40/–. *Fr: Spiritualité de l'amour*

26887 **Jaeger**, Lorenz *Erzbischof von Paderborn* : A stand on Ecumenism : The Council's Decree. – Hilda Graef. – London : G. Chapman, 1965. xiii, 242, 30/–. *Dt: Das Konzilsdekret über den Ökumenismus*

26888 **Jeanne d'Arc**, *Sœur* : Witness and consecration. – Martin Murphy. – London : G. Chapman. 312, 30/–. *Fr: Le religieuse dans l'Église et dans le monde actuel*

26889 **Jungmann**, Josef A. : The Eucharistic prayer. – Robert L. Barley. – London : Burns & Oates. vii, 55, 6/6. *Dt: Das Eucharistische Hochgebet*

26890 — Liturgical renewal in retrospect and prospect. – Clifford Howell. – London : Burns & Oates, 1965. 45, 4/6. *Dt: Liturgische Erneuerung*

26891 — The liturgy of the word. – H.E. Winstone. – London : Burns & Oates. 82, 8/6. *Dt: Wortgottesdienst im Lichte von Theologie und Geschichte*

26892 — The place of Christ in liturgical prayer. – A. Peeler. – London : G. Chapman, 1965. xx, 300, 42/–. *Dt: Die Stellung Christi im liturgischen Gebet*

26893 **Kitamori**, Kazoh : Theology of the pain of God. – London : SCM Press. 183, 22/6. *Jap: Kami no Itami no Shingaku*

26894 **Kraus**, Hans-J. : Worship in Israel. – Geoffrey Buswell. – Oxford : Blackwell. xi, 246, 37/6. *Dt: Gottesdienst in Israel*

26895 **Küng**, Hans : The church and freedom. – Cecily Hastings. – London : Sheed & Ward, 1965. 42, 6/–. *Dt: Kirche in Freiheit*

26896 **Lemaître**, Solange : Hinduism. – J.F. Brown. – London : Burns & Oates. 128, 4/–. *Fr: Hindouisme ou Sanâtana Dharma*

26897 **Léonard**, Émile G. : A history of protestantism. (1:) The Reformation. – Joyce M.H. Reid. – London : Nelson, 1965. xiv, 461, 90/–. *Fr: Histoire générale du protestantisme (1:) La Réformation*

26898 **Lescrauwaet**, Josephus : The Bible on Christian unity. – N.D. Smith. – London : Sheed & Ward, 1965. 186, 10/6. *Ned: De Bijbel over de christelijke eenheid*

26899 **Lévy-Bruhl**, Lucien : The " soul " of the primitive. – Lilian A. Clare. – London : Allen & Unwin, 1965. 351, 35/–. *Fr: L'âme primitive*

26900 **Linnemann**, Eta : Parables of Jesus. – John Sturdy. – London : S.P.C.K. xiv, 216, 35/–. *Dt: Gleichnisse Jesu, Einführung und Auslegung*

26901 Liturgy in development. – H.J.J. Vaughan. – London : Sheed & Ward, 1965. ix, 187, 12/6. *Ned: Eredienst in ontwikkeling*

26902 **Martin**, Bernard : Healing for you. – A.A. Jones. – London : Lutterworth, 1965. 194, 25/–. *Fr: Vous-tu guérir?*

26903 **Monden**, Louis : Sin, liberty and law. – Joseph Donceel. – London : Chapman. x, 181, 30/–. *Ned: Vrijheid genaden*

26904 **Müller**, Alois : Obedience in the church. – Hilda Graef. – London : Burns & Oates. 190, 15/–. *Dt: Das Problem von Befehl und Gehorsam im Leben der Kirche*

26905 **Orvieto**, Laura : Stories of Greece and the barbarians. – Barbara Whelpton. – London : Burke. 191, 12/6. *It: Storie della storia del mondo greche e barbare*

26906 **Overduin**, Jan : Adventures of a deserter. – Harry Van Dyke. – Exeter : Paternoster P. 152, 5/–. *Ned: Het overstaan van een deserteur*

26907 **Pihl**, Herman G. & **Beckman**, Karin : Follow me. Stories from the Bible. – Edward Carney. – London : Black. 95, ill. 8/6. *Swe*

26908 **Pol**, Willem van de : World Protestantism. – T. Zuydwijk. – London : Sheed & Ward, 1965. xii, 346, 15/–. *Ned: Het Wereldprotestantisme*

26909 **Rad**, Gerhard von : Deuteronomy : a commentary. – Dorothea Barton. – London : SCM Press. 211, 35/–. *Dt: Das fünfte Buch Mose: Deuteronomium*

26910 **Rahner**, Karl : Everyday things. – M.H. Heelan. – London : Sheed & Ward, 1965. 41, 6/–. *Dt: Alltägliche Dinge*

26911 **Ratzinger**, Joseph : Christian brotherhood. – W.A. Glen-Doepel. – London : Sheed & Ward. ix, 94, 9/–. *Dt: Die christliche Brüderlichkeit*

26912 **Régamey**, Pie R. : Non-violence and the Christian conscience. – Thomas Merton. – London : Darton, Longmann & Todd. 272, 25/–. *Fr: Non-violence et conscience chrétienne*

26913 Religious orders in the modern world. – Walter Mitchell. – London : G. Chapman, 1965. viii, 172, 25/–. *Fr: Les religieux aujourd'hui et demain*

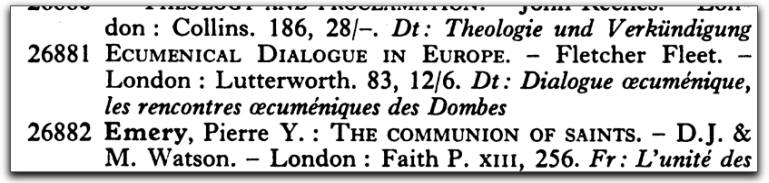
26914 Roman Catholic Church. *2d Vatican Ecumenical Council.* Declaration on religious liberty. – Thomas Anthill. – London : Catholic Truth Society. 19, 1/–. *Lat: De libertate religiosa*

EARIA-2012

Approches symboliques avec Unitex

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

Exemple : Annotation de documents OCRisés (UNESCO)
M. Softi, P. Bellot et al., 2011



don : Collins. 186, 28/–. *Dt: Theologie und Verkündigung*

26881 Ecumenical Dialogue in Europe. – Fletcher Fleet. – London : Lutterworth. 83, 12/6. *Dt: Dialogue œcuménique, les rencontres œcuméniques des Dombes*

26882 **Emery**, Pierre Y. : The communion of saints. – D.J. & M. Watson. – London : Faith P. xiii, 256. *Fr: L'unité des*

EARIA-2012

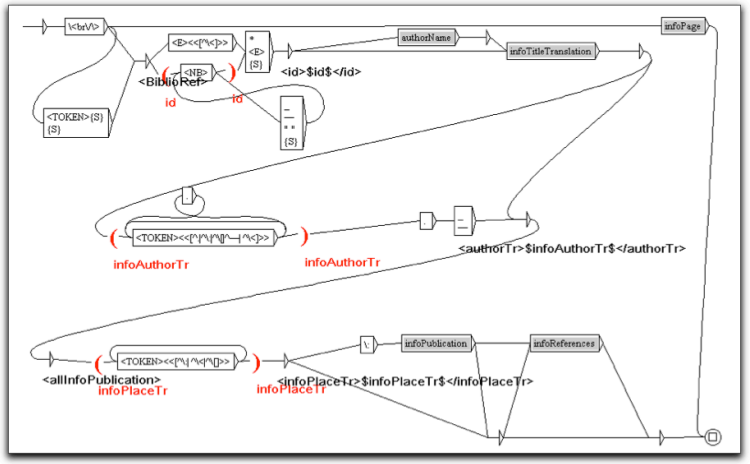
Etiquetage sorties OCR (2)

```
<!--XML OUTPUT-->OCR Output
<pageInfo><br/>159-171 INDEX TRANSLATIONUM</pageInfo><br/>
<country>ESPAGNESPAIN</country><br/>

<italic>Traductions annoncées dans la « Bibliografía española e hispano-americana »de juillet, août et septembre 1935.
Translations listed in the « Bibliografía española e hispano-americana »fort the months of July, August and September 1935.</italic>
<category>Philosophie, Religion.Philosophy, Religion.</category><br/>

<br/>159. AGUSTỸN (San), Kempis Agustiniano.Máximas de..... sobre la Vida cristiana (Recopiladasde sus Obras por el P. Antonio Tonna-Barthet). — P.Francisco Mier. — Barcelona, Ed. Litúrgica Española,1935, 8«, 635.

<br/>160. BAETEMAN, J., Breves Respuestas a lasObjeciones contra la Religión. — Juan Nada. — Barcelona, Ed. Poliglota, 1935, 8o, 48, Ptas. 0.80.
<italic>Francis, Français, French.</italic>

<br/>161. BERDIAEFF, Nicolás, El Cristianismo yla Lucha de Clases. Dignidad del Cristianismo eindignidad de los Cristianos. — María Cardona. —Madrid, Espasa Cal pe, 1935, 8», 208, Ptas. 5.

<br/>162. DELMAS, F. A., y BOLL, M., La Persona+shy;lidad humana. Su Análisis. — J. Albiñana Mompó. — ■Madrid, Aguilar, 1935, 8o, 270, Ptas. 6 (Biblioteca deldeas y Estudios contemporáneos).

<italic>Francés, Français, French [La Personnalité humaine :Son Analyse].</italic>

<br/>163. DESCARTES, Reglas para la Direcciónde Espíritu. — Manuel Mindán. - — Madrid, Revistade Occidente, 1935, 8°, 179, Ptas. 6 (Textos filosóficos).
```

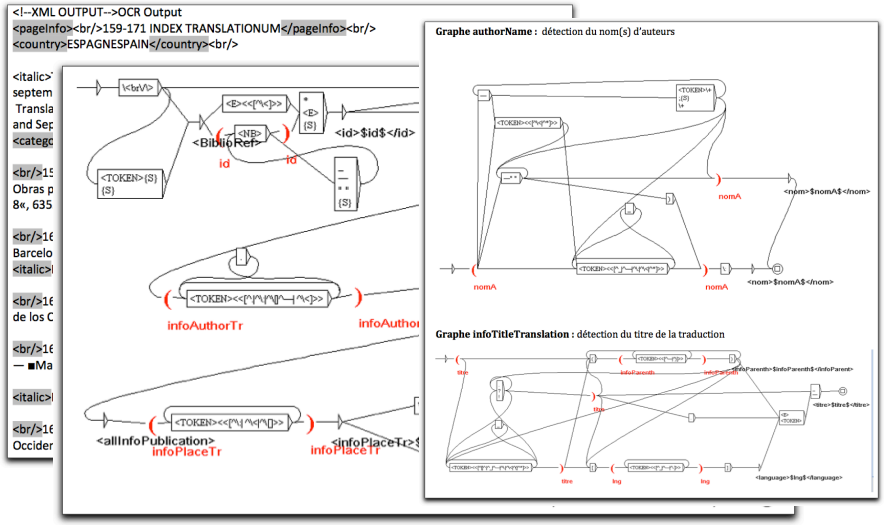

EARIA-2012

Etiquetage sorties OCR (2)

<!--XML OUTPUT-->OCR Output
<pageInfo>
159-171 INDEX TRANSLATIONUM</pageInfo>
<country>ESPAGNESPAIN</country>



Etiquetage sorties OCR (2)



Etiquetage sorties OCR (3)

```
<!__XML OUTPUT-_>
{S}OCR Output
<pageInfo>421_464 Allemagne-Germany</pageInfo><br/>
{S}421 — <italic>N.T. : </italic>
<BiblioRef>
        <id></id>
        <titre>Die Bibel.{S} Die Heilige Schrift d.{S} Neuen Bundes.</titre>
        <allInfoPublication>
                <infoPlaceTr>Freiburg i.{S} Br.{S}</infoPlaceTr>
                <infoPublisher>Herder</infoPublisher>
                <publicationInfo>1968 (4.{S} Aufl.) vni, 276, 48,2.90.</publicationInfo>
        </allInfoPublication>
        <references>Gr</references>
</BiblioRef>
<BiblioRef>
        <id>422</id>
        <nom>— </nom>
        <language>Niederdt</language>
        <titre>Dat Nie Testament in unse Moderspraak.</titre>
        <authorTr>Johannes Jessen</authorTr>
        <allInfoPublication>
                <infoPlaceTr>Göttingen</infoPlaceTr>
                <infoPublisher>Vandenhoeck</infoPublisher>
                <publicationInfo>+ Ruprecht,1968 (6.{S} Aufl.) 494, 10.80.</publicationInfo>
        </allInfoPublication>
        <references>Gr</references>
```


Exemple

Annotation de références bibliographiques par apprentissage automatique

Approches probabilistes pour l'annotation

- **Modèles de Markov Cachés (*Hidden Markov Models* - HMM)**
  - — modèles génératifs (apprentissage des paramètres qui maximisent les probabilités «observations + étiquettes»)
  - — hypothèses d'indépendance (les observations sont indépendantes)
  - — tient compte des états futurs (algo. Forward-Backward) (même si l'état suivant dépend seulement de l'état actuel)
- **Modèles de Markov à Maximum d'Entropie (*MaxEnt Markov Models - MEMM*)**
  - — modèles discriminants (apprentissage des paramètres qui maximisent les probabilités des étiquettes sachant les observations)
  - — ne pose pas l'hypothèse d'indépendance des observations (mots)
  - — ne tient pas compte des états «futurs»
- **Champs Conditionnels Aléatoires (*Conditional Random Fields* - CRF)**
  - — modèles discriminants tenant compte de *features* très diverses
  - — ne pose pas l'hypothèse d'indépendance
  - — tient compte des états «futurs»

Conditional Random Fields (CRF)

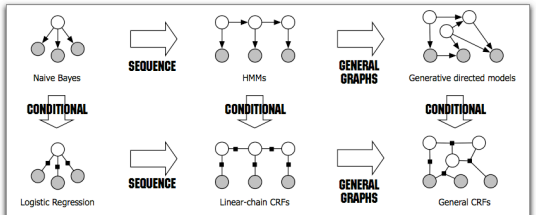


Figure 1.2 Diagram of the relationship between naive Bayes, logistic regression, HMMs, linear-chain CRFs, generative models, and general CRFs.

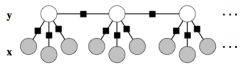
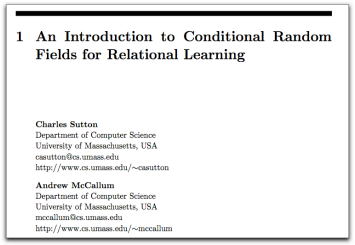


Figure 1.3 Graphical model of an HMM-like linear-chain CRF.

Figure 1.4 Graphical model of a linear-chain CRF in which the transition score depends on the current observation.



1 An Introduction to Conditional Random Fields for Relational Learning

Charles Sutton
Department of Computer Science
University of Massachusetts, USA
casutton@cs.umass.edu
http://www.cs.umass.edu/~casutton

Andrew McCallum
Department of Computer Science
University of Massachusetts, USA
mccallum@cs.umass.edu
http://www.cs.umass.edu/~mccallum

http://people.cs.umass.edu/~mccallum/papers/crf-tutorial.pdf

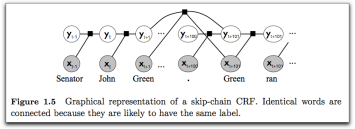


Figure 1.5 Graphical representation of a skip-chain CRF. Identical words are connected because they are likely to have the same label.

EARIA-2012

Conditional Random Fields (CRF)

Conditional Random Fields - Lafferty et al. (2001)

Un modèle discriminant pour l'étiquetage de données séquentielles
$\mathbf{x} = x_1 \ldots x_T$: une séquence d'entrée
$\mathbf{y} = y_1 \ldots y_T$: une séquence d'état
$D = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$: un ensemble de N exemples

La probabilité conditionnelle d'un CRF

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\}, \quad (1)$$

$\theta = \{\theta_k\} \in R^K$: un vecteur de paramètre
$\{f_k(y_t, y_{t-1}, \mathbf{x}_t)\}_{k=1}^{K}$: un ensemble des fonctions de charactéristiques
$Z(\mathbf{x})$: une fonction de normalisation

Figure 1.4 Graphical model of a linear-chain CRF in which the transition score depends on the current observation.

Figure 1.5 Graphical representation of a skip-chain CRF. Identical words are connected because they are likely to have the same label.

EARIA-2012

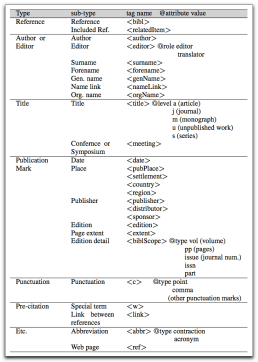
BILBO - Analyse automatique de documents

http://bilbo.hypotheses.org/

Aix*Marseille université openedition REVUES.ORG CALENDA HYPOTHESES.ORG

Google research cnrs

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE



Type	sub-type	tag name	@attribute value
Reference	Reference	<bibl>	
	Included Ref.	<relatedItem>	
Author or Editor	Author	<author>	
	Editor	<editor>	@role editor translator
	Surname	<surname>	
	Forename	<forename>	
	Gen. name	<genName>	
	Name link	<nameLink>	
	Org. name	<orgName>	
Title	Title	<title>	@level a (article) j (journal) m (monograph) s (unpublished work) s (series)
	Conference or Symposium	<meeting>	
Publication	Date	<date>	
Mark	Place	<pubPlace> <settlement> <country> <region>	
	Publisher	<publisher> <distributor> <sponsor> <edition>	
	Edition		
	Page extent	<extent>	
	Edition detail	<biblScope>	@type vol (volume) pp (pages) issue (journal num.) isen part
Punctuation	Punctuation	<c>	@type point comma (other punctuation marks)
Pre-citation	Special term	<w>	
	Link between references	<link>	
Etc.	Abbreviation	<abbr>	@type contraction acronym
	Web page	<ref>	

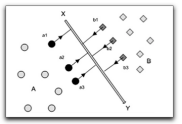
	Examples
Non bibl note	26. La nature euro-centrée du projet était encore plus apparente dans la version originale du texte, qui, comme nous l'avons déjà mentionné, était appelé " Europe élargie ".
Multi bibl note	27. " Une Europe sûre dans un monde meilleur. Stratégie européenne de sécurité ", Bruxelles, 12 décembre 2003. Pour un commentaire critique, voir Toje A., " The 2003 European security strategy:A critical appraisal ", *European Foreign Affairs* Review, vol. 10, n°1, 2005, pp. 117-134.
Part. info.	31. Voir par exemple la communication de la Commission relative au " Renforcement de la politique européenne de voisinage ", COM (2006) 726 final.

EARIA-2012

BILBO - Données d'apprentissage

- Token, features, tags -> Données d'apprentissage
- Environnement MALLET http://mallet.cs.umass.edu/
- Combinaison de SVM (segmentation du texte et classification) et de CRF (annotation)

```
Deux titres retiennent également l'attention : il s'agit des périodiques
médicaux, dont Lassus et Sabatier conservent des collections complètes,
en particulier les
<bibl>
 <hi rend="italic">
   <title level="m">Medical essays and observations</title>
 </hi>
 <c type="guillemot_left">«</c>
 by
 <publisher>a Society in Edinburgh</publisher>
 <c type="guillemot_right">»</c>
</bibl>
<bibl>


Research in karst hydrogeology and geomorphology, which was carried out
in the 1970s, 1980s and 1990s
<c type="parenthesis_left">(</c>
<bibl>
 <author>
   <surname>Delannoy</surname>
 </author>
 <c type="comma">,</c>
 <date>1997</date>
 <link target="Delannoy1997"/>
</bibl>
<c type="parenthesis_right">)</c>
has shown that karst gives excellent examples of thermodynamic systems
whose structure, function and evolution are determined by surrounding
factors, particularly the geomorphological (gravitational energy) and
```



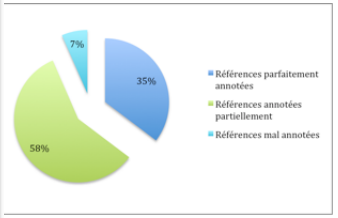
```
BOSERUP ALLCAP surname
E. ALLCAP INITIAL forename
, c
1965 ALLNUMBERS date
, c
The ITALIC FIRSTCAP title
conditions ITALIC ALLSMALL title
of ITALIC ALLSMALL title
Agricultural ITALIC FIRSTCAP title
Growth ITALIC FIRSTCAP title
: ITALIC c
The ITALIC FIRSTCAP title
Economics ITALIC FIRSTCAP title
of ITALIC ALLSMALL title
Agrarian ITALIC FIRSTCAP title
Change ITALIC FIRSTCAP title
under ITALIC ALLSMALL title
Population ITALIC FIRSTCAP title
Pressure ITALIC FIRSTCAP title
, c
Aldine FIRSTCAP publisher
, c
Chicago FIRSTCAP place
, c
218 ALLNUMBERS extent
p ALLSMALL abbr
. c
```

BILBO - Evaluation détaillée

Table 2: Evaluation on UMich Library level 1 data including all files (upper table) and excluding 'ark' and 'mbot' (lower table).

Tool	Precision			Recall			F-measure		
	total	author	title	total	author	title	total	author	title
BILBO	**0.68**	**0.86**	**0.84**	**0.69**	*0.86*	*0.73*	**0.68**	**0.86**	**0.78**
ParsCit	0.57	*0.79*	0.73	0.49	0.63	0.52	0.53	0.70	0.63
Biblio	0.57	0.72	0.62	0.55	0.57	**0.83**	0.56	0.64	0.71
Freecite	*0.65*	0.76	*0.76*	*0.62*	**0.94**	0.68	*0.63*	*0.83*	*0.72*
Grobid	-	-	-	-	-	-	-	-	-

Tool	Precision			Recall			F-measure		
	total	author	title	total	author	title	total	author	title
BILBO	*0.63*	**0.79**	**0.85**	**0.64**	*0.86*	*0.70*	**0.64**	**0.82**	**0.77**
ParsCit	0.51	0.69	0.71	0.41	0.53	0.51	0.45	0.60	0.59
Biblio	0.57	*0.72*	0.60	0.59	0.63	**0.88**	0.58	0.67	0.71
Freecite	0.59	0.71	0.75	0.57	**0.89**	0.68	0.58	*0.78*	*0.71*
Grobid	**0.64**	0.67	*0.79*	*0.62*	0.80	0.63	*0.63*	0.73	0.70



Y.M. Kim, P. Bellot et al. (ACM DocEng 2012, LREC 2012)

Systèmes de questions-réponses

Les moteurs de questions-réponses

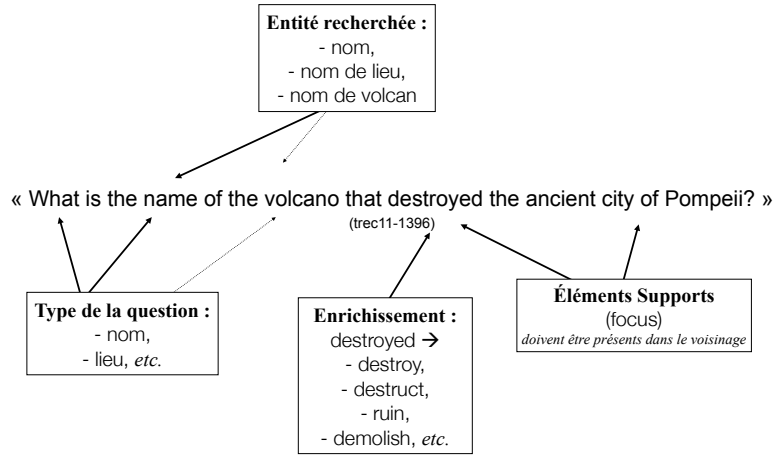
- Questions factuelles (*quelle est la capitale ....*)
- Questions booléennes (*est-ce que .... ?*)
- Questions "liste" (*quelles sont .... ?*)
- Les questions complexes : "comment ?", "pourquoi ?" *etc.*

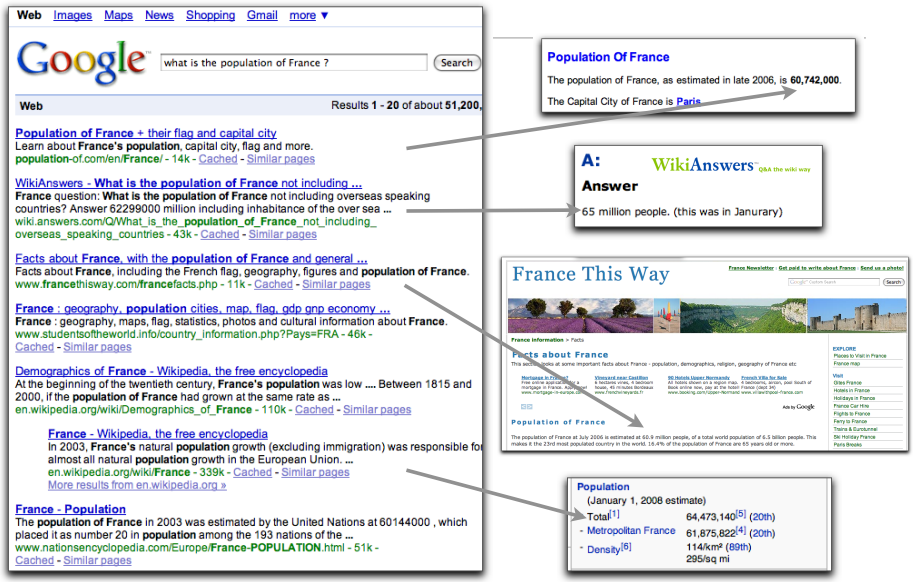
Référence	Libellé de la question
GF222	Qui est le maire de Bastia ?
GF30	Combien de personnes souffrent d'acné en Suisse ?
GF266	Quelle est la monnaie nationale en *Hongrie* ?
GF219	A combien de kilomètres de Paris se trouve la gare de Tours ?
GF178	Comment s'appelle le président Tchétchène ?
GF6	Quel âge a l'abbé Pierre ?
GF232	Combien y a t il de chômeurs en Europe ?
GF245	Qui est le frère de la princesse Leia ?
GF17	Combien y a t il d'habitants en *Lettonie* ?
GF29	Quel grade occupe *Juan Carlos Rolon* dans la marine ?
GF273	Quand est mort *Kurt Cobain* ?
GF105	Quel est le nom du roi du *Maroc* ?
GF298	Quelle est la capitale de Terre Neuve ?
GF147	En quelle année *Hitler* est arrivé au pouvoir ?
GF176	Qui est le président d'*Aérospatiale* ?
GF99	Combien de personnes sont mortes dans des accidents de la route en 1997 ?
GF132	Où se situe *San Cristobal* de Las Casas ?
GF206	Quand a été votée la loi *Evin* ?
GF84	En combien de langues a été publié le Petit Prince ?
GF78	Quel journal publie chaque année le top 50 des personnalités ?

Questions évaluation EQUER 2005 (Grau et al.)

Analyses complexes

- **Objectif** : extraire de la question le + d'informations → localiser la réponse





Classification des questions

- Objectifs :
  - déterminer le type de la question (factuelle, procédurale, définitoire...)
  - déterminer le type de réponse attendue (nom propre, nom de lieu, quantité numérique...)
- Méthodes :
  - à base de règles et de patrons syntaxiques
  - à base d'apprentissage automatique (approche bayésienne, SVM...)
  - méthodes mixtes
  - exploitation des POS, de bigrammes, des lemmes, de ressources...
- Performances :
  - env. 90% F-Mesure sur quelques dizaines de classes

Recherche documentaire dans QR

- Pour la recherche de documents : peu importe le modèle (booléen, vectoriel, probabiliste...)
- Pour la recherche de phrases : c'est différent ! [Radev, 2002]
  - **Combinaison de modèles de langage meilleure que Okapi**
    (ordonnancement des phrases trouvées par Google, Alltheweb, Northernlight)
  - **Pour 132 questions (sur 200), la réponse est dans les 20 premières phrases**

◊ TREC 2000

542 questions ont une réponse juste dans les 50 premiers documents (576 dans les 200 premiers)

◊ TREC 2001 (500 questions) : redondance

- 37 questions ont leur réponse dans au moins 25 documents de la collection complète
- 138 questions ont leur réponse dans au moins 10 documents de la collection complète

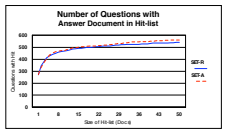


Figure 1. Comparison of number of questions with a "correct" document in the hit list against hit-list depth, for our search engine (SET-R) and AT&T's (SET-A).

Prager *et al.*,

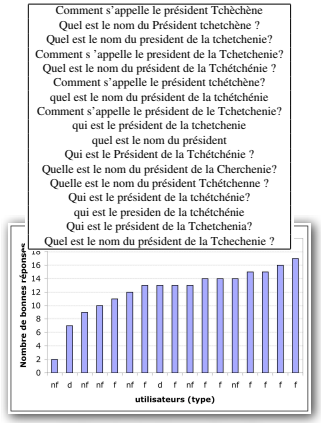
Guru / SMART

En moyenne : document contenant la bonne réponse en position 2 (MRR = 0,49)

Robustesse et suggestions de réécritures

- Evaluation sur des questions "réelles" du système SQuaLIA (Gillard et al.)
- 20 questions X 19 utilisateurs

(9 Français langue maternelle, 6 étrangers vivant en France, 2 dyslexiques)



SITBON, BELLOT, BLACHE, (LREC, 2008)

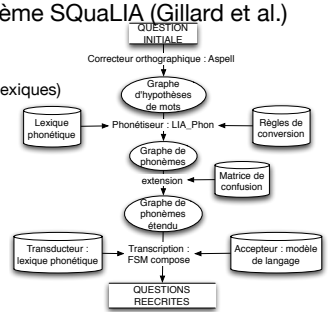


FIG. 4.10: *La procédure de correction orthographique conduisant à la génération d'hypothèses de questions utilise un correcteur orthographique standard, un phonétiseur (LIA_Phon) et un système de transcription automatique (figure reprise de Sitbon et al. (2008c)).*

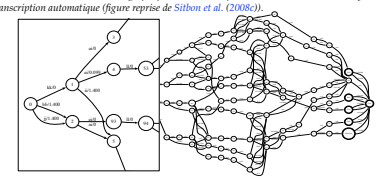


FIG. 4.12: *Exemple d'hypothèses phonétiques (phonème/poids, numéro de saxad) produites par LIA-Phon à partir des hypothèses graphémiques du correcteur Aspell et de la question : kel aje a la Bepierre (Quel âge a l'abbé Pierre ?) — figure reprise de Sitbon et al. (2008c).*

Quelques thèses récentes...

- « *Réordonnancement d'hypothèses dans un système de questions-réponses*», Guillaume BERNARD, LIMSI Paris-Sud, 2011
- « *Extraction d'information générique à partir de textes fondée sur une analyse linguistique profonde* », Ludovic JEAN-LOUIS, CEA LVIC, Paris-Sud, 2011
- « *Validation de réponses dans un système de questions réponses* », Arnaud GRAPPY, LIMSI Paris-Sud, 2011
- « *Désignations nominales des événements — Etude et extraction automatique dans les textes* », Béatrice ARNULPHY, LIMSI Paris-Sud, 2012
- « *Extraction automatique de segments textuels, détection de rôles, de sujets et de polarités* », Rémi LAVALLEY, LIA Avignon, 2012
- ...

Validation de réponses (A. Grappy, 2011)

Critère	Système
	Termes communs
Mots communs	[Kozareva et al. 2006 ; Herrera et al. 2006] [Rodrigo et al. 2006 ; M.A.Garcia-Cumbreras et al. 2007] [Téllez-Valero et al. 2007 ; Téllez-Valero et al. 2008] [Ferrández et al. 2008 ; J.Castillo 2008] [Ferrandez, et al. 2007]
Entités nommées	[Rodrigo et al. 2006 ; Herrera et al. 2006] [Rodriguo, et al. 2007 ; Ferrández et al. 2008] [Kozareva et al. 2006 ; Téllez-Valero et al. 2008]
Similarité de termes grâce à WordNet	[M.A.Garcia-Cumbreras et al. 2007] [J.Castillo 2008 ; Ferrández et al. 2008]
Distinctions morphosyntaxiques	[Téllez-Valero et al. 2007 ; Ferrández et al. 2008] [Téllez-Valero et al. 2008]
Mesure cosine	[Ferrández et al. 2008 ; J.Castillo 2008]
TF*IDF	[Ferrández et al. 2008 ; J.Castillo 2008]
Programmation dynamique	[Ferrández et al. 2008]
Distance de Jaro	[Ferrández et al. 2008]
	Proximité des termes
Plus longue chaîne commune	[Kozareva et al. 2006 ; Bosma & Callison-Bursh 2006] [M.A.Garcia-Cumbreras et al. 2007] [Téllez-Valero et al. 2007 ; Ferrandez et al. 2007]
N-grammes	[Kozareva et al. 2006 ; Herrera et al. 2006] [M.A.Garcia-Cumbreras et al. 2007 ; Ferrandez et al. 2007] [J.Castillo 2008 ; Rodrigo et al. 2006]
Distance de Levenshtein	[Ferrandez et al. 2007 ; Ferrández et al. 2008] [J.Castillo 2008]
Mesure rouge	[Ferrandez et al. 2007]
	Spécifiques validation de réponses
Vérification du type de la réponse	[Téllez-Valero et al. 2007 ; Téllez-Valero et al. 2008]
Contraintes sur la réponse	[Téllez-Valero et al. 2007 ; Ferrández et al. 2008] [Téllez-Valero et al. 2008]
Redondance de la réponse	[Téllez-Valero et al. 2008]

TAB. 1.1 – Critères de validation de réponses

Validation de réponses (A. Grappy, 2011)

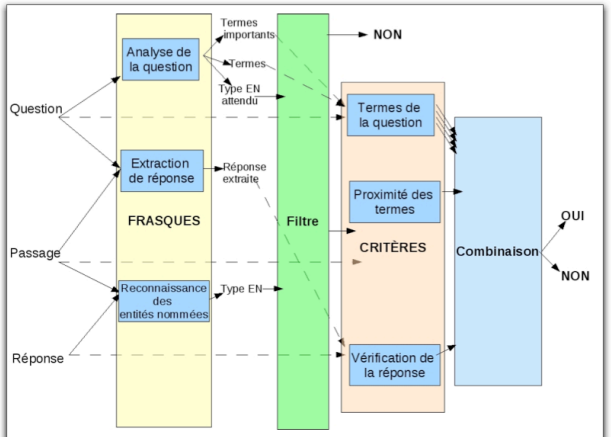


FIG. 3.3 – Mécanisme de validation de réponses : le système AVAL

TAB. 1.1 – Critères de validation de réponses

Validation de réponses (A. Grappy, 2011)

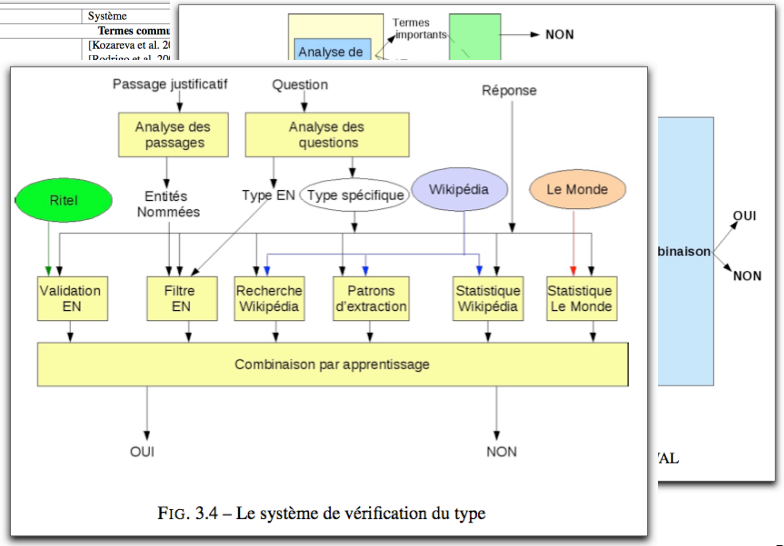


FIG. 3.4 – Le système de vérification du type

Liens entre entités - (L. Jean-Louis, 2011)

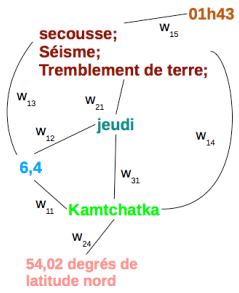
Texte	Graphe d'entités
{PRINCIPAL}Un **tremblement de terre** de magnitude **6,4** sur l'échelle ouverte de Richter s'est produit **jeudi** au large de la péninsule du **Kamtchatka**, dans l'Extrême-Orient russe, ont rapporté les Services de surveillance sismique américains (USGS). {PRINCIPAL}La **secousse** a eu lieu à **01 H 43 GMT**. Son hypocentre a été localisé à environ **54,02 degrés de latidude nord**, entre la pointe sud du **Kamtchatka** et le nord de l'**archipel des Kouriles**, a précisé l'USGS. {CONTEXTE}La région du **Kamtchatka** et des **Kouriles**, située sur la "ceinture de feu" du Pacifique, est régulièrement secouée par de très violents **séismes**. {SECONDAIRE}Quelques heures plus tôt, un autre puissant **tremblement de terre** de magnitude **6,8** sur l'échelle de Richter s'était produit à environ 1.800 km plus au sud, dans le **nord du Japon**, faisant au moins une centaine de blessés.	

FIG. 4.3 – Exemple de graphe d'entités nommées au niveau du document

Liens entre entités - (L. Jean-Louis, 2011)

{PRINCIPAL}Un **tremble**
l'échelle ouverte de
péninsule du **Kamt**
rapporté les Servic
(USGS).
{PRINCIPAL}La **secous**
hypocentre a été loc
nord, entre la poi
l'**archipel des Kouri**

{CONTEXTE}La région d
la "ceinture de feu"
par de très violents **s**

{SECONDAIRE}Quelques
tremblement de te
Richter s'était produi
nord du Japon, fais

FIG. 4.3 – Exe

Description des *features*	FEAT-BASE	FEAT-LEX	FEAT-NOLEX
Types des entités E1 et E2	✓	✓	✓
POS des entités E1 et E2	✓	✓	✓
Mots des entités E1 et E2	✓		
Bigrammes de mots de E1 et E2	✓	✓	✓
Mots entre E1 et E2	✓	✓	
Bigrammes de mots entre E1 et E2	✓	✓	
POS des mots entre E1 et E2	✓	✓	✓
Nb mots entre E1 et E2	✓	✓	✓
Bigrammes de POS de mots entre E1 et E2		✓	✓
Nb relations syntaxiques entre E1 et E2		✓	✓
Chemin syntaxique entre E1 et E2		✓	✓
Position relative et POS[1]		✓	✓
Nb d'entités entre E1 et E2		✓	✓
Nb mentions d'événements entre E1 et E2		✓	✓
POS des deux mots avant/après E1		✓	✓
POS des deux mots avant/après E2		✓	✓

TAB. 1 – *Features* pour la classification de relations

Liens entre entités - (L. Jean-Louis, 2011)

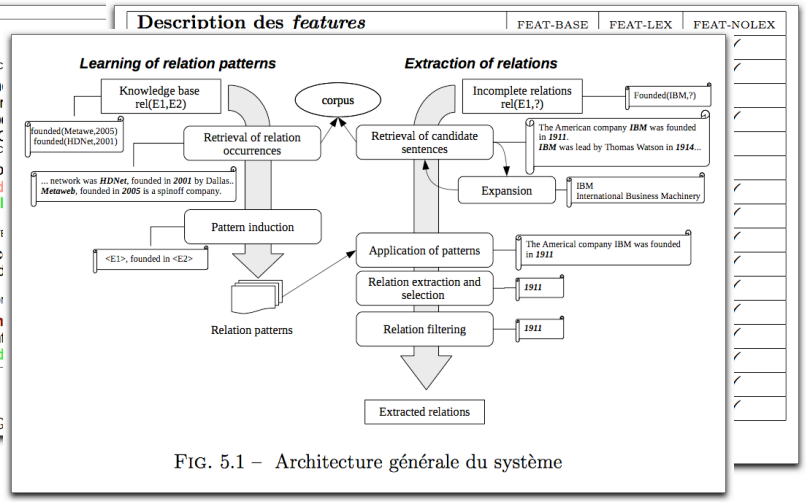


FIG. 5.1 – Architecture générale du système

Plateformes logicielles

Infrastructure for Human Language Technology

GATE

http://gate.ac.uk

The University Of Sheffield.

Free

Open source, licensed under LGPL allowing unrestricted commercial use, hosted on SourceForge.

100% Java

Runs on **any platform** supporting Java 5 or later. Developed and tested daily on Linux, Windows, and Mac OS X.

Mature and actively supported

In development **since 1996**; now at version 5.0; around 20 active developers.

Comprehensive

Support for manual annotation, performance evaluation, information extraction, [semi-]automatic semantic annotation, and many other tasks.
Over **50 plugins** included with the standard distribution, containing over 70 resource types. Many others available from independent sources.

Standards-based

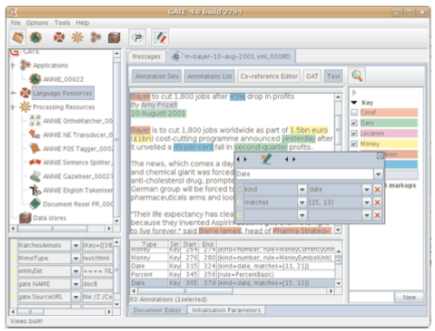
Reference implementation in **ISO** TC37/SC4 LIRICS project; supports XCES, ACE, TREC etc. formats; founder member of **OASIS**/UIMA committee.

Efficient

Optimisations included with the latest version provide a 20 to 40% speed and memory usage improvement. Highly efficient finite state text processing engine; many plugins with linear execution time.

Popular

Assessed as "outstanding" and "internationally leading" by an anonymous EPSRC peer review.
Used at thousands of sites: companies, universities and research laboratories, all over the world. ~35,000 downloads/year.
Rolling funding for more than 15 staff at the University of Sheffield.



Data Management

Pluggable input filters with out of the box support for XML, HTML, PDF, MS Word, email, plain text, etc.

Common in-memory data model built around stand-off annotation, documents and corpora.

Persistent storage layer with support for XML, Oracle, PostgreSQL, or Java serialisation. I/O interoperation with many other systems.

Standard Algorithms

Ready made implementations for many typical NLP tasks such as tokenisation, POS tagging, sentence splitting, named entity recognition, co-reference resolution, machine learning, etc.

User Interface

Comprehensive tool set for data editing and visualisation, rapid application development, manual annotation, ontology management.

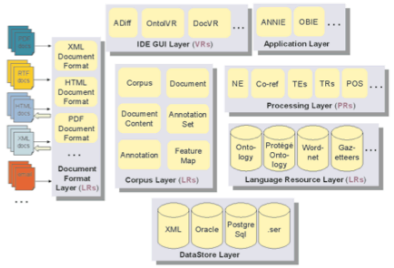
Integration

Leveraging the power of other projects such as:

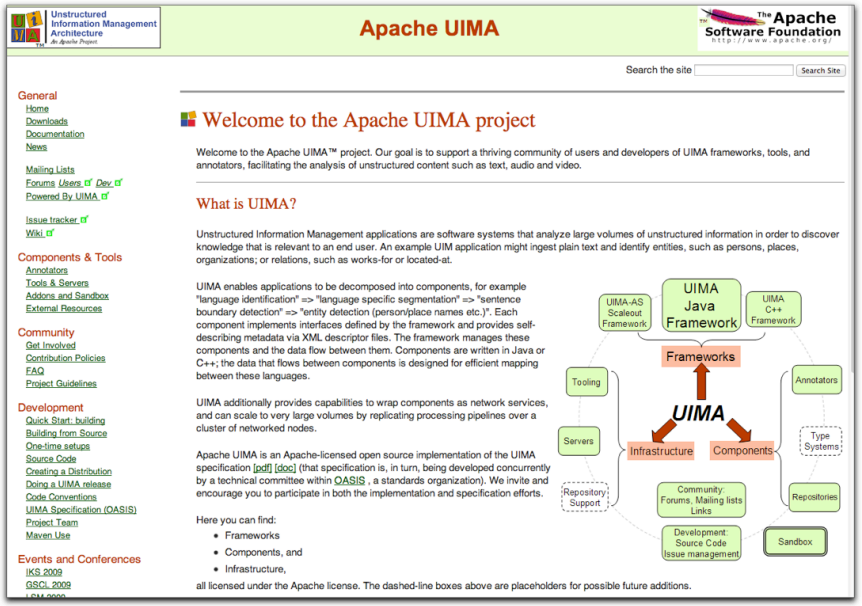
- **Information Retrieval:** Lucene (Nutch, Solr), Google and Yahoo search APIs, MG4J;
- **Machine Learning:** Weka, MaxEnt, SVMLight, etc.;
- **Ontology Support:** Sesame and OWLIM;
- **Parsing:** RASP, Minipar, and SUPPLE;
- **Other:** UIMA, Wordnet, Snowball, etc.

Community and Support

Friendly and active community of developers and users offers efficient help. Commercial support available from Ontotext and Matrixware.



http://uima.apache.org



http://uima.apache.org

Conclusions...

- La RI sur des textes (et audio) fait partie du TAL
- Le TAL utilise des approches symboliques / numériques, des méthodes d'apprentissage automatique, des ressources et bases de connaissances
- Tâches complexes à effectuer et à évaluer (y compris pour un humain)
  — au cœur de nombreuses (toutes ?) campagnes d'évaluation en RI
- Traitements souvent lourds : question de la *pertinence* de les utiliser
- Les améliorer avant de les utiliser en RI ? (complexité, robustesse...)
- Apprendre à mieux les intégrer au sein des modèles de RI ?
  — quelle information ? à quel moment ?
  — architecture informatique (intégration)
  — combinaison de scores, confiance...

Quelques livres...