

L'apport du TAL aux Humanités numériques

Thierry Poibeau
thierry.poibeau@ens.psl.eu

Master Humanités numériques, PSL, octobre 2025

Boîtes à outils linguistiques

- NLTK: <https://www.nltk.org>
- Gate: <https://gate.ac.uk>
- Spacy : <https://spacy.io/>
- Apprentissage : Weka: <https://www.cs.waikato.ac.nz/ml/index.html>
- Modèles de langage (BERT, GPT3)
 - CamemBERT : <https://camembert-model.fr/>
 - FlauBERT : <https://github.com/getalp/Flaubert>
 - Hugging Face : <https://huggingface.co/>

Qu'est-ce que le TAL peut apporter aux humanités numériques?

- L'informatique peut aider à gérer :
 - de grosses collections de données
 - l'analyse de ces collections
- Combiner l'effort conjoint de groupes de chercheurs plus efficacement (cf. tâches d'annotations)
- Poser des questions qui ne pourraient pas être traitées sans ordinateur
- Cf. Franco Moretti (distant reading)

Morphosyntaxe

- Aider à annoter des corpus au niveau morphosyntaxique
- Aider à analyser la complexité des langues sur le plan morphosyntaxique, et faciliter les études comparatives dans ce domaine
- Aider à identifier des langues marquées par un ensemble de propriétés particulières
- Vérifier l'adéquation d'une description linguistique en la « projetant » sur un corpus test

Analyseur morphosyntaxique

Cette	PRO:DEM	ce
situation	NOM	situation
met	VER:pres	mettre
en	PRP	en
évidence	NOM	évidence
le	DET:ART	le
caractère	NOM	caractère
incontournable	ADJ	incontournable
de	PRP	de
la	DET:ART	le
place	NOM	
du	PRP:det	
travail	NOM	
dans	PRP	
un	DET:ART	
pays	NOM	
industriel	ADJ	
.	SENT	

- Nombreux outils, pour de nombreuses langues
- Application relativement simple à développer pour une nouvelle langue

SEM - Segmenteur-Étiqueteur Markovien

Un étiqueteur du français basé sur les CRF, licence GNU.

Part-Of-Speech

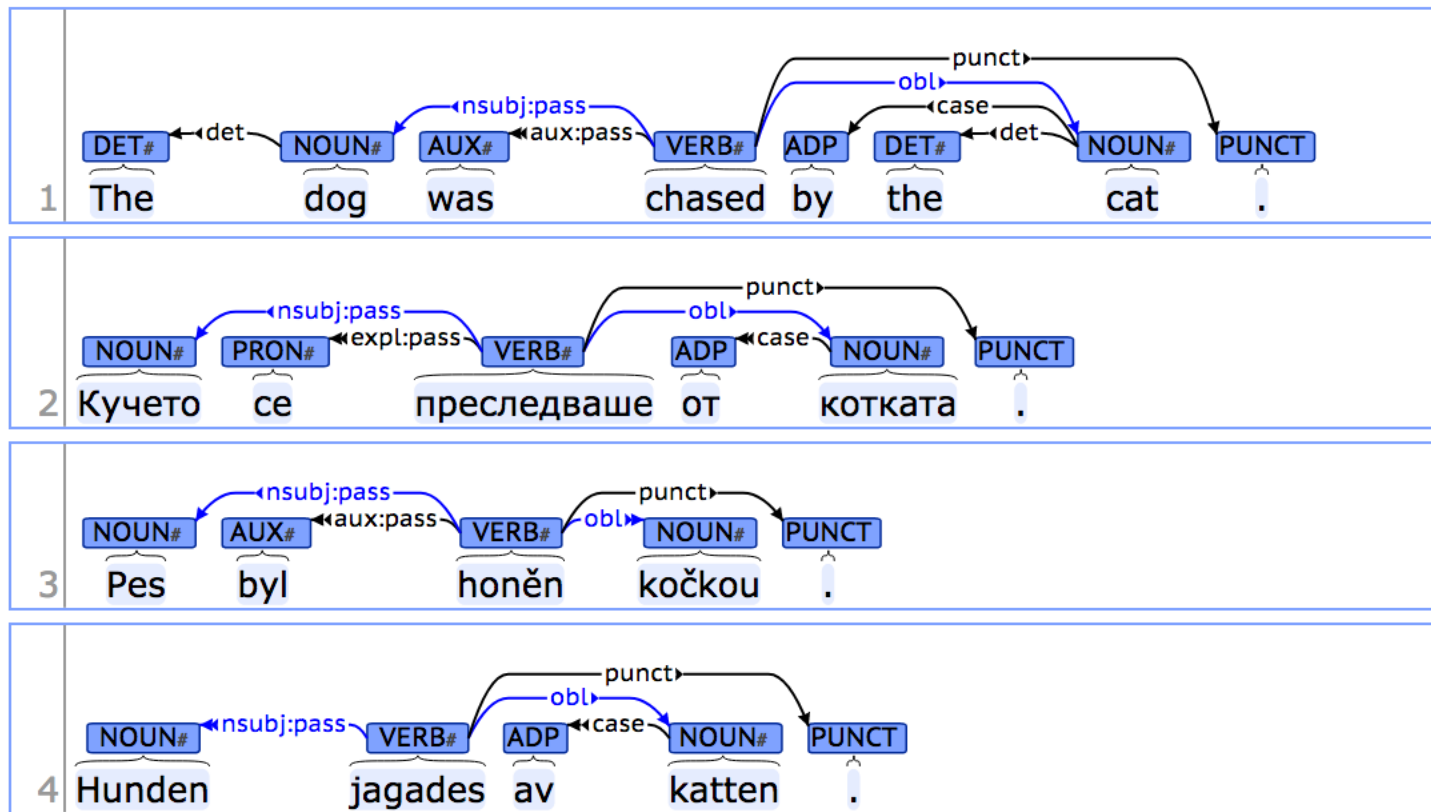
Named Entity

Emmanuel et Brigitte Macron rentrent à Paris ce mardi 21 août. C'est la fin des deux semaines de vacances pour le couple dans le Var, au fort de Brégançon.


















































































Analyse grammaticale

- Disponibilité de corpus annotés pour de nombreuses langues et d'analyseurs correspondants à ces langues
- Universal Dependencies + CoNLL shared task
 - Une base de données de corpus annotés syntaxiquement
 - Un jeu d'étiquettes pour les relations syntaxiques partagé / unifié (+ particularités par familles de langues)
 - Des analyseurs correspondants

Exemple : phrases avec verbe au passif



Données disponibles

▶		Afrikaans	1	49K		IE, Germanic				
▶		Amharic	1	10K	    	Afro-Asiatic, Semitic				
▶		Ancient Greek	2	417K	  	IE, Greek				
▶		Arabic	3	Latin treebanks						
▶		Armenian	1							
▶		Bambara	1	▶	PROIEL	199K	(L)(F)	 	   	★★★★★
▶		Basque	1	▶	ITTB	291K	(L)(F)		   	★★★★★
▶		Basque	1	▶	Perseus	29K	(L)(F)	  	   	★★★★★
▶		Belarusian	1	See here for comparative statistics of Latin treebanks						
▶		Breton	1	10K	     	IE, Celtic				
▶		Bulgarian	1	156K	   	IE, Slavic				
▶		Buryat	1	French treebanks						
▶		Cantonese	1							
				▶	ParTUT	28K	(L)(F)	  	   	★★★★★
				▶	GSD	402K	(L)(F)	   	   	★★★★★
				▶	Sequoia	70K	(L)(F)	   		★★★★★
				▶	PUD	24K	(F)	 	  	★★★☆☆
				▶	FTB	573K	(L)(F)			★☆☆☆☆
				▶	Spoken	34K	(L)(F)		  	☆☆☆☆☆

See here for comparative statistics of French treebanks.

Analyse syntaxique multilingue (« tâche partagée »)

CoNLL 2018 Shared Task

Home
Data | Baseline Models
Evaluation | TIRA | Results
EPE 2018 | EPE Results
Paper Submission Guidelines
Organization | Registration | Timeline

CONTACT

udst-orgs@googlegroups.com

NEWS

July 2: **MAIN RESULTS**
June 25: Eval script 1.2
June 21: Test phase extended to July 1
June 17: Baseline results
June 1: Test phase started
May 2: Baseline models
April 15: **TRAINING DATA**
March 6: Trial data
Jan 26: **Registration open**

Multilingual Parsing from Raw Text to Universal Dependencies

A [CoNLL](#) 2018 shared task.

The proposed task is a follow-up of the [CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies](#). We first summarize aspects that will be new in 2018; then we provide a more detailed description of the shared task for readers who are not familiar with the 2017 task.

There will be three main evaluation metrics. None of them is more important than the others and we will not combine them into a single ranking. Participants who want to decrease task complexity may concentrate on improvements in just one metric; however, all participating systems will be evaluated with all three metrics, and participants are strongly encouraged to output all relevant annotation (syntax + morphology), even if they just copy values predicted by the baseline model.

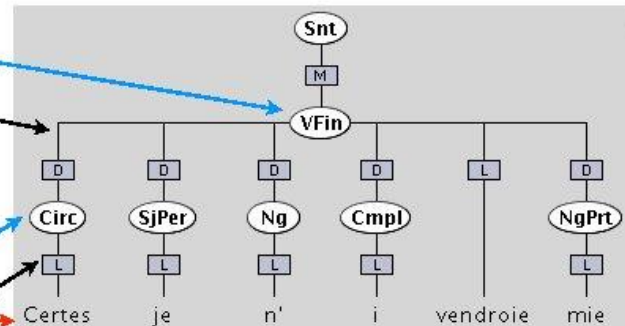
The three metrics are described in more detail [here](#). All three include word segmentation and labeled dependency relations. One of them is identical to the 2017 main metric so that results can be compared. The other two metrics focus on content words and include morphological features and lemmatization, respectively.

Instead of surprise languages, there will be a category of low-resource languages that have little or no training data. The names of the languages, as well as whatever sample data may be available, will not be kept as surprise.

There will be new languages that were not part of the 2017 evaluation (Afrikaans and Serbian already satisfy the requirements; others may be available when the training data is released).

Corpus annoté d'ancien français

```
<?xml id='beroul_pb:1_b:19_1263221020.72'>
  <graph root='_1263221020.72'>
    <terminals>
      <t word='Certes' id='w26_00095' pos='ADV' lemma='certes' />
      <t word='je' id='w26_00097' pos='PRO_pers' lemma='je' />
      <t word='n' id='w26_00098' pos='PRO clit' lemma='ne' />
      <t word='i' id='w26_00099' pos='PRO clit' lemma='y' />
      <t word='vendroie' id='w26_00100' pos='VER' lemma='venir' />
      <t word='mie' id='w26_00101' pos='ADV' lemma='mie' />
    </terminals>
    <nonterminals>
      <nt id='_452409.15' cat='Ng'>
        <edge label='L' idref='w26_00098' />
      </nt>
      <nt id='_221023.93' cat='VFin'>
        <edge label='D' idref='_452410.38' />
        <edge label='D' idref='_452418.9' />
        <edge label='D' idref='_452406.05' />
        <edge label='D' idref='_452407.65' />
        <edge label='D' idref='_452409.15' />
        <edge label='I' idref='w26_00101' />
      </nt>
      <nt id='_452410.38' cat='NgPrt'>
        <edge label='I' idref='w26_00101' />
      </nt>
      <nt id='_452418.9' cat='Cmpl'>
        <edge label='L' idref='w26_00099' />
      </nt>
      <nt id='_452406.05' cat='Circ'>
        <edge label='L' idref='w26_00095' />
      </nt>
      <nl id='_221020.72' cat='Snt'>
        <edge label='M' idref='_221023.93' />
      </nl>
      <nl id='_452407.65' cat='SjPer'>
        <edge label='L' idref='w26_00097' />
      </nl>
    </nonterminals>
  </graph>
</?xml>
```



Projet ANR SRCMF *Syntactic Reference Corpus of Medieval French* (2009-2012), de Sophie Prévost

Sémantique et pragmatique

- Questions de sémantique lexicale
- Désambiguïsation sémantique (définir le sens des mots en contexte)
- Portée de divers éléments linguistiques (négation, conjonction), quantifieurs, modaux, etc.
- Analyse de la référence et de la coréférence
- Analyse de la métonymie, des métaphores et autres figures rhétoriques
- Comparaison de l'usage des honorifiques dans différentes langues
- etc.

Analyse lexicale

- Exemple : SketchEngine (version gratuite disponible pour les universitaires)
 - <https://www.sketchengine.eu/>
 - Conçu à partir des travaux d'Adam Kilgariff
 - Beaucoup utilisé en lexicographie (Collins)
 - Permet d'avoir une vision « analytique » de l'usage d'un mot
 - Fréquence, concordances, etc.
 - Modificateurs et compléments, classés par construction syntaxique (par ex. Pour un V, liste des N apparaissant en position sujet)
 - Comparaison de mots (*strong* vs *powerful*), de langues (usage de deux mots de sens similaires dans deux langues différentes, etc.)
 - Importantes couches de TAL « transparentes (constitution des corpus, analyse, affichage)

SketchEngine

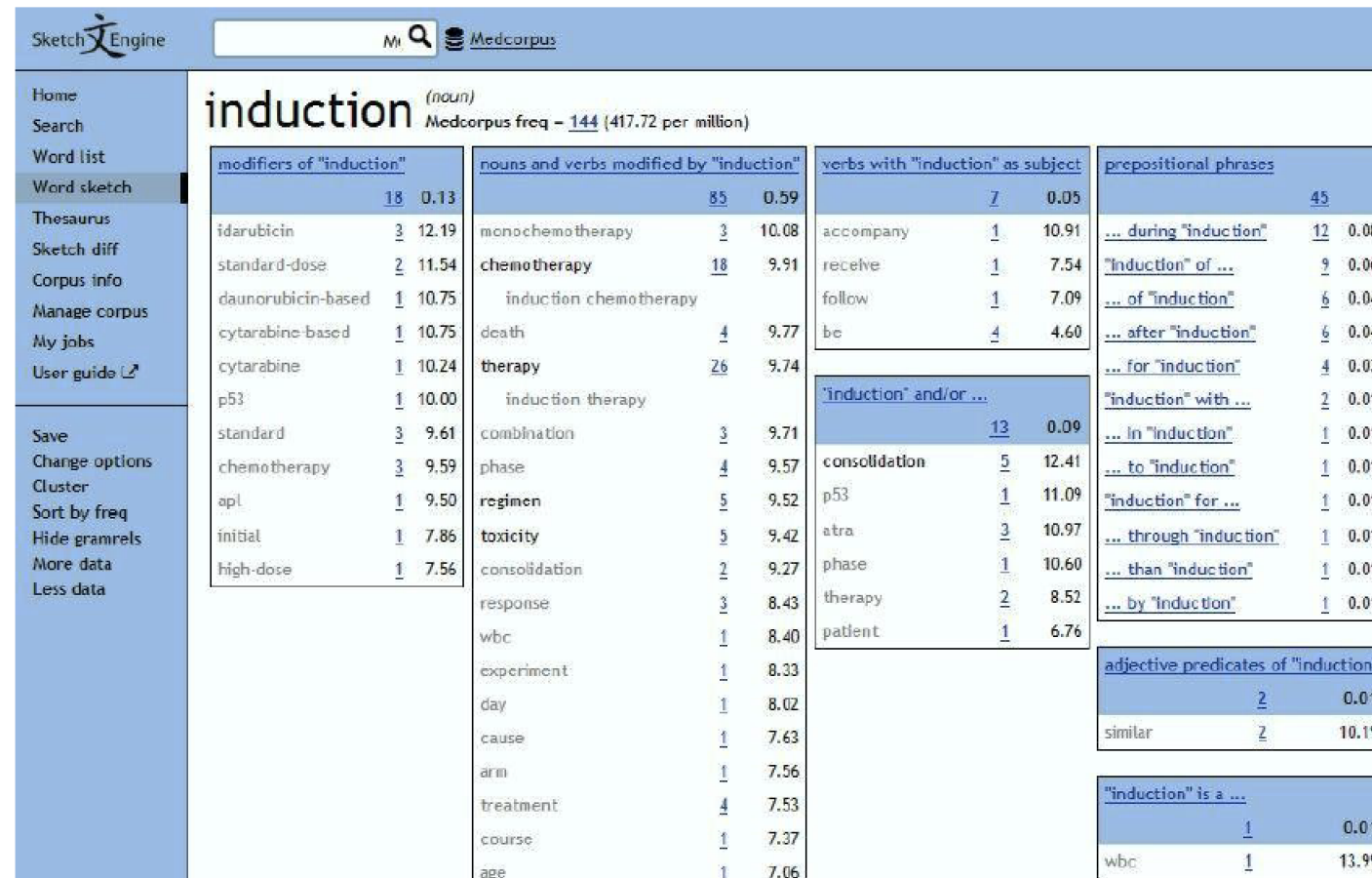
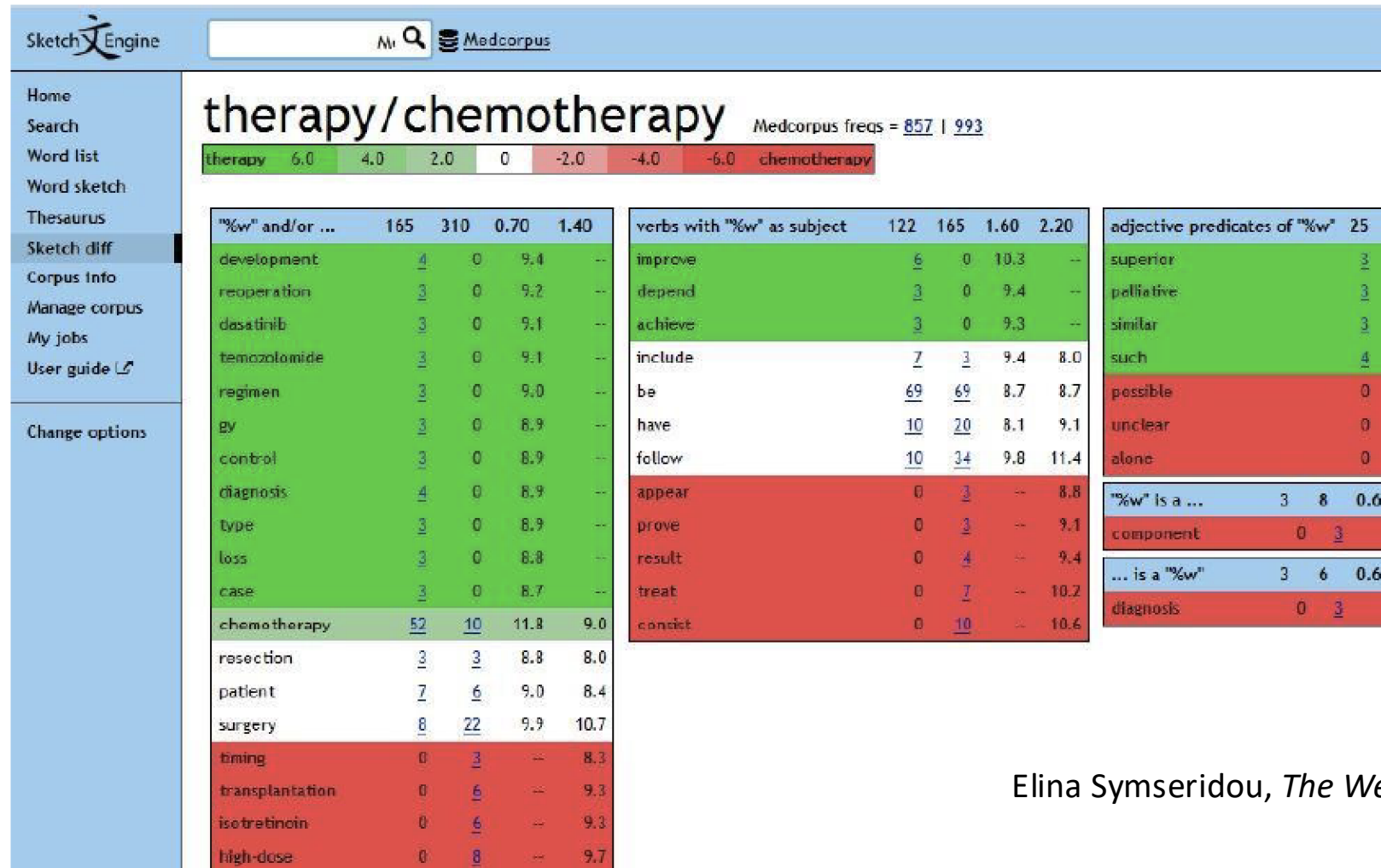


Figure 3. Word Sketch results for the term induction

SketchEngine



Elina Symseridou, *The Web as a Corpus*

Figure 4. Sketch Difference for the comparative study of the terms therapy and chemotherapy

SketchEngine

SketchEngine [Medcorpus](#)

Home
Search
Word list
Word sketch
Thesaurus
Sketch diff
Corpus info
Manage corpus
My jobs
User guide ↗

Save
Change options
Cluster
Sort by freq
Hide grammrels
More data
Less data

leukemia (noun) Medcorpus freq = 294 (852.85 per million)
Click on collocates to access reciprocal bilingual search

λευχαιμία OALcorpus freq = 598 (993.61 p)

modifiers of "leukemia"	445	1.51
acute	<u>164</u>	12.94
acute lymphoblastic leukemia		
lymphoblastic	<u>66</u>	11.98
of acute lymphoblastic leukemia		
promyelocytic	<u>53</u>	11.75
newly diagnosed acute promyelocytic leukemia		
myeloid	<u>20</u>	10.41
acute myeloid leukemia		
childhood	<u>25</u>	10.28
childhood acute lymphoblastic leukemia		
chronic	<u>8</u>	9.05
lymphocytic	<u>6</u>	8.76
myeloid	<u>5</u>	8.50
acute	<u>5</u>	8.49
myelomonocytic	<u>4</u>	8.19
refractory	<u>4</u>	8.12
plasma	<u>4</u>	8.06
monocytic	<u>3</u>	7.78

words before "leukemia"	1,420	2.37
οξεία	<u>124</u>	10.73
οξεία λεμφοβλαστική λευχαιμία		
λεμφοβλαστική	<u>67</u>	10.44
οξεία λεμφοβλαστική λευχαιμία		
Οξεία	<u>78</u>	10.24
μυελογενής	<u>55</u>	10.20
μυελογενής λευχαιμία (
Χρόνια	<u>40</u>	9.57
Χρόνια μυελογενής λευχαιμία		
λεμφοκυτταρική	<u>34</u>	9.56
χρόνια λεμφοκυτταρική λευχαιμία		
χρόνια	<u>47</u>	9.52
χρόνια μυελογενής λευχαιμία		
μυελογενή	<u>24</u>	9.08
οξεία μυελογενή λευχαιμία		
προμυελοκυτταρική	<u>14</u>	8.30
οξεία προμυελοκυτταρική λευχαιμία		

words just before "leukemia"	530	0.89
λεμφοβλαστική	<u>71</u>	11.91
οξεία λεμφοβλαστική λευχαιμία		
μυελογενής	<u>60</u>	11.70
Χρόνια μυελογενής λευχαιμία		
λεμφοκυτταρική	<u>36</u>	11.02
χρόνια λεμφοκυτταρική λευχαιμία		
μυελογενή	<u>27</u>	10.63
οξεία μυελογενή λευχαιμία		
οξεία	<u>32</u>	10.24
οξεία λευχαιμία		
προμυελοκυτταρική	<u>14</u>	9.72
οξεία προμυελοκυτταρική λευχαιμία		
μυελομονοκυτταρική	<u>14</u>	9.72
μυελομονοκυτταρική λευχαιμία		
μυελοβλαστική	<u>11</u>	9.38

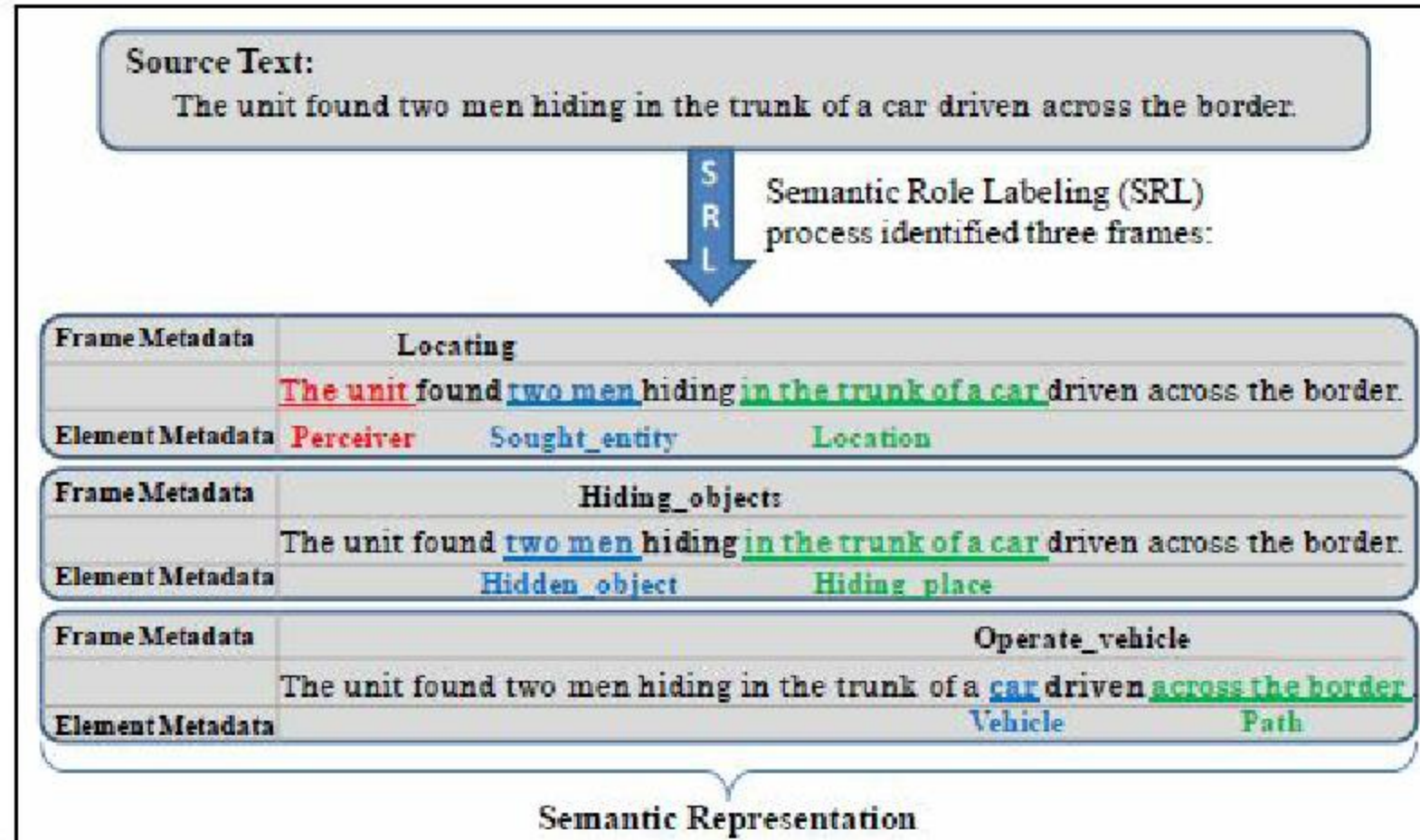
Elena Symseridou, *The Web as a Corpus*

Figure 6. Bilingual Word Sketch presentation of the term leukemia

Lexiques structurés : Framenet

- *Frame semantics* (Fillmore, 1976)
 - Frame: structure conceptuelle (situation prototypique)
 - Frame element : participant ayant un rôle dans la situation
- *Barkeley Framenet Project*
 - Base de données encodant un lexique Framenet pour l'anglais
 - 615 frames, 8000 unités lexicales
- Projet à la fois linguistique et informatique
- Cf. Aussi Wordnet, Ontonotes, etc. DicoValence pour le français, etc.

FrameNet



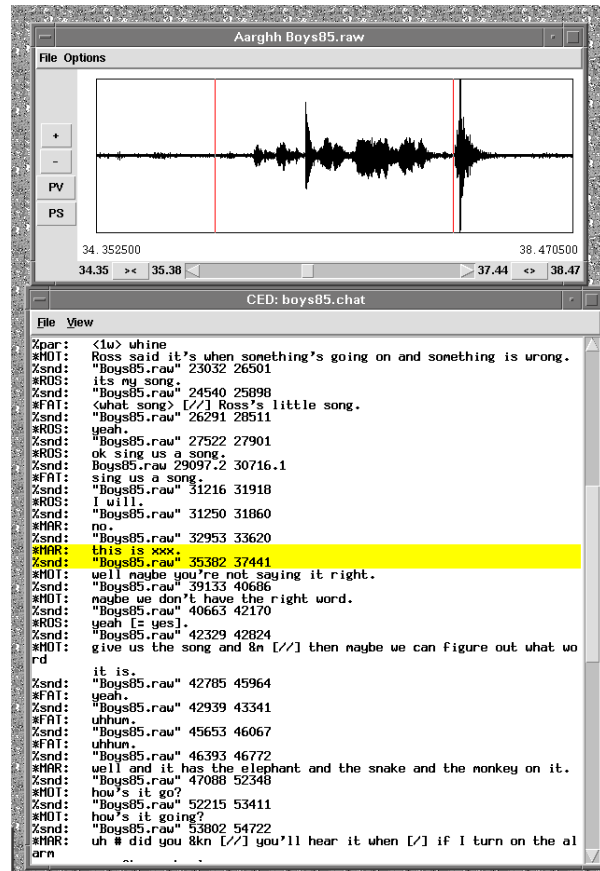
Au-delà de la phrase...

- Anaphore
 - Relations discursives (RST *Rhetorical Structure Theory*)
 - etc.
-
- Importance des recherches sur corpus
 - Equipes mixtes linguistes / TAListes
 - Analyse fondées sur des théories linguistiques variés
 - Formalisation / implémentation
 - Tests sur des corpus importants

Acquisition du langage

- Un des domaines de la linguistique les plus avancés pour l'utilisation des ordinateurs en linguistique
- Base de données *Childes* (CHILd Language Data Exchange System, McWhinney, <https://childes.talkbank.org/>, depuis 1984)
- Base de corpus : *CHILDES*.
- Format de transcription : CHAT.
- Logiciel de transcription et utilisation des corpus : CLAN + commandes.
- Plusieurs millions de mots au total, 32 langues, accès libre
- 3000+ publications associées depuis le début

Corpus multimédia



File name indicating the age of the child

Dependent tiers:
%can = cantonese
%xgl1 = sign language gloss 1

Main tier combining the two dependent tiers

Video

Time

Begin time and end time of the utterance

Clan - [YC020809_utt.cha]

```
73 *MOT: 哥哥? •48800_49780•
74 *BON: 哥哥 me1? •48880_49800•
75 *MOT: 哥哥有 wo3. •49900_51190•
76 *CHI: <le2-IX_obj [= picture]> [% sim]! •51000_52560•
77 %can: le2. •51000_52560•
78 %xgl1: IX_obj [= picture]! •51451_52346•
79 *MOT: 呢個俾哥哥. •52360_53930•
80 *BON: 呢個 aa3 呢個妹妹 lei4 gaa3. •54000_56000•
81 *CHI: IX_obj [= picture]. •55476_56721•
82 %xgl1: IX_obj [= picture]. •55476_56721•
83 *BON: 你睇吓有條辮 gaa3 wo3. •56000_57800•
84 *BON: 紮住個頭髮. •57800_59950•
85 *CHI: <IX_obj-xxx [= picture]> [% sim]. •57841_59531•
86 %xgl1: IX_obj [= picture] •57841_59531•
87 %can: xxx. •57850_59103•
88 *BON: 紮咗 wo3. •61110_62500•
89 *MOT: 呢啲魚呢啲魚蛋. •61600_64000•
90 *CHI: xxx. •62800_64500•
91 %can: xxx. •62800_64500•
92 *MOT: 係 aa3 紮辮 aa4 係 aa3 紮辮. •64400_66500•
93 *BON: 係 wo1 紮咗辮! •64500_68000•
94 *MOT: 魚蛋 aa3 有魚蛋. •66900_69200•
95 *BON: waa3@i 做乜挖鼻哥! •68000_69700•
96 *MOT: 魚蛋. •69200_70410•
97 *MOT: 好痕 aa3 佢. •71000_72900•
98 *CHI: IX_obj [= tissue]. •71114_74049•
99 %xgl1: IX_obj [= tissue]. •71114_74049•
100 *MOT: 擰紙巾 aa3. •73000_74300•
101 *BON: 要唔要? •74300_75400•
102 *CHI: <IX_obj-bookbook@f [= tissue]> [% sim] •75009_77314•
103 %xgl1: IX_obj [= tissue] •75009_77314•
104 %can: bookbook@f. •75263_76300•
105 *MOT: bookbook@f 有 aa3 呢度有. •76200_77700•
106 *MOT: waa3@i 有睇睇 read@s:en [x 3]. •77700_79850•
041109[E][CHAT] 74
```

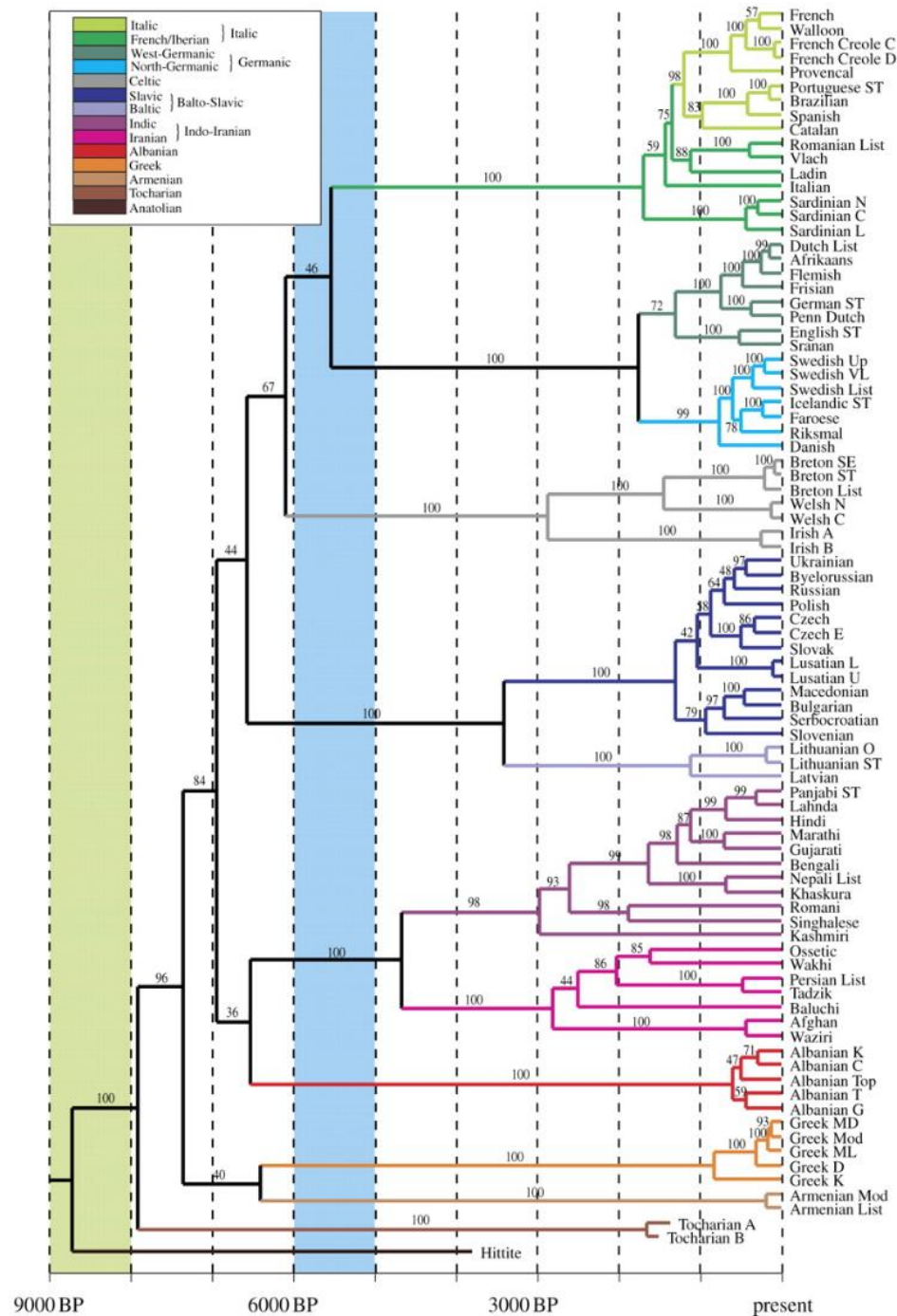
Etude de l'évolution des langues

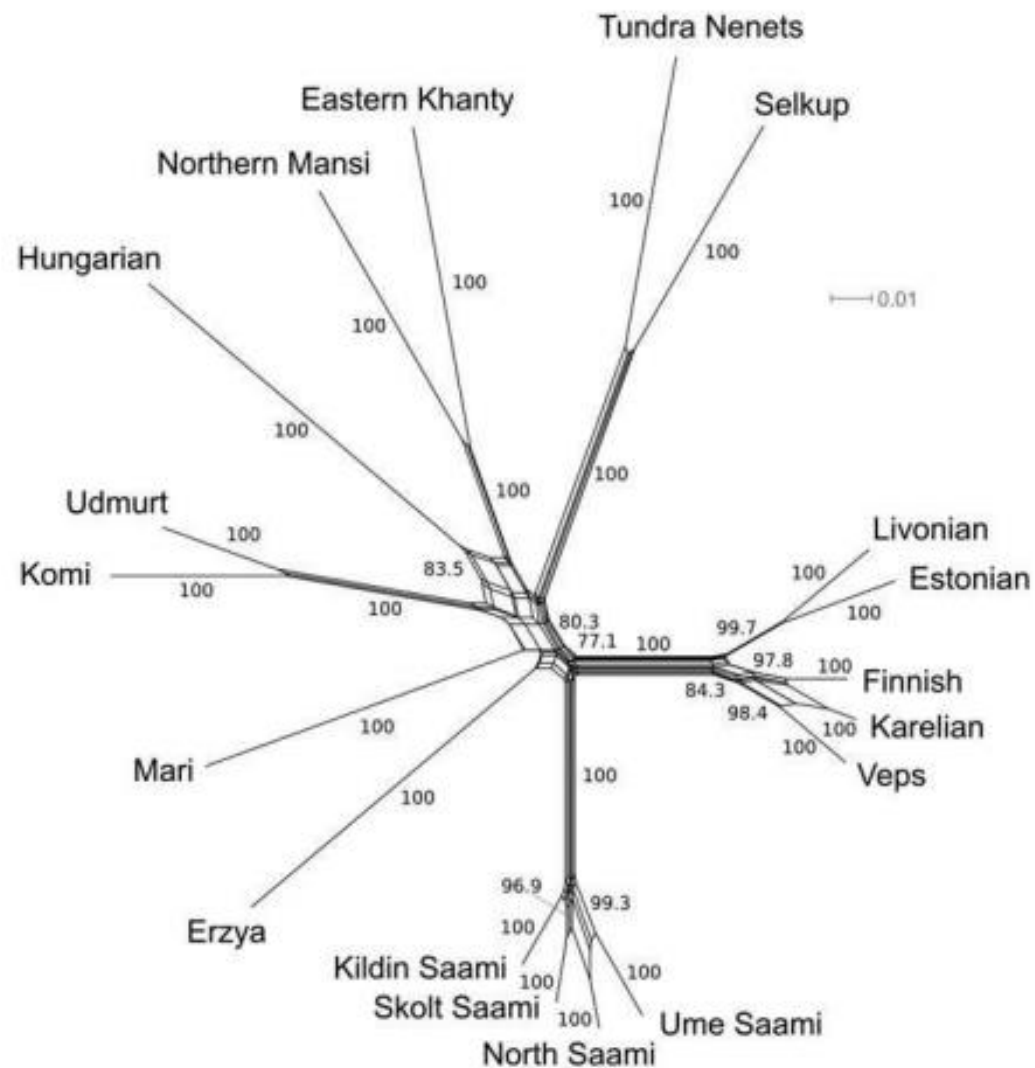
- Reconstruction des liens entre langues à partir de propriétés linguistiques (listes de Swadesh, autres données linguistiques)
- Recherches extrêmement populaires
- Validité scientifique ?

Cf. Nakhleh et al. (2005):

- A model of language change that allows for both genetic (common source) and borrowing (contact) relationships between varieties
- The model can estimate, on the basis of 294 lexical, phonological and morphological characters for 24 languages, the extent to which early Indo-European languages developed in isolation from each other

A dated phylogenetic tree of 87 Indo-European languages, Nakhleh et al. (2005)





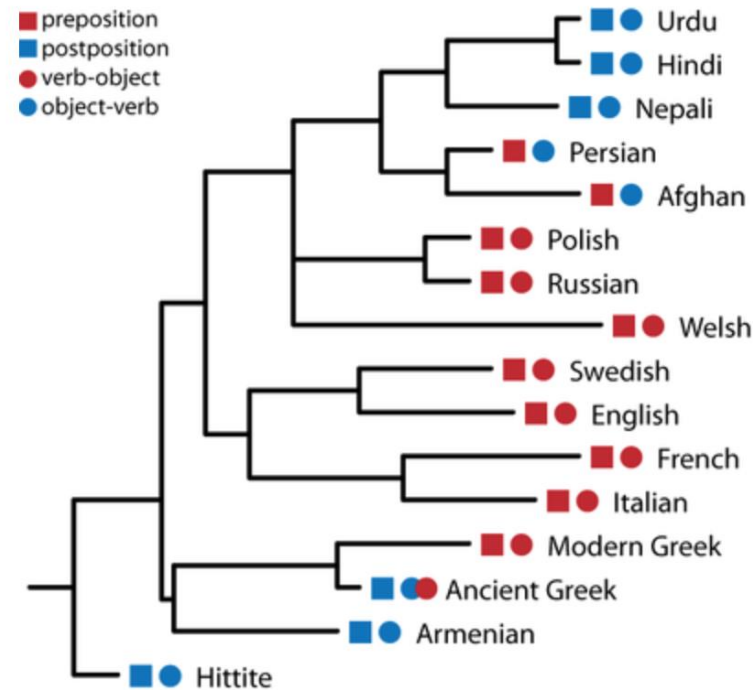
Liens entre langues finno-ougriennes, groupe Bedlan

Application à la typologie

Etude de la variation linguistique entre langues, sur une base linguistique.

Exemple à partir de WALS

- Permettre d'analyser la variation typologique
- Fournir des résultats lisibles (dendrogramme)
- Généralisation possible à un grand nombre de traits



Correlations between preposition/postposition and verb-initial/verb-final final word order features in part of the Indo-European family of languages.

Le renouveau du Deep Learning

Analyse distributionnelle

- You shall know a word by the **company it keeps** (**Firth**, J. R. 1957:11)
- Contexte : permet de prédire le sens d'un mot (mots qui partagent le même contexte ont un sens proche)
- Sémantique, mais aussi syntaxe
- Hyper performant si assez de données, assez de puissance de calcul
- Word2Vec (2013), BERT (2018), GPT3

Modèles de langage

- En [traitement automatique des langues](#), un **modèle de langage** est un [modèle statistique](#) qui modélise la distribution de séquences de mots, plus généralement de séquences de symboles discrets (lettres, phonèmes, mots), dans une [langue naturelle](#). Un modèle de langage peut par exemple prédire le mot suivant une séquence de mots¹.
- [BERT](#) et [GPT-3](#) sont des modèles de langage.

Wikipedia

Applications

- GPT (Mistral, Claude) : génération de texte (prédire le mot suivant)
- BERT : modèle « à trou » (prédire le mot le plus probable dans une séquence) *Pierre était _ de Sophie. Il la suivait partout.*
- Mais en fait, BERT est un modèle « à tout faire » : morphosyntaxe, syntaxe, entités nommées... question-réponse... (GPT3 aussi, en partie, pour résumé et QR)
- Mémorisation + généralisation à partir de milliards de contextes (cf. nombre de paramètres)

En bref

- Modèles extrêmement complexes, impossibles à réentraîner complètement, mais très facilement utilisables
 - Cf. Hugging Face, Spacy
- Adaptation au domaine
 - Fine tuning (juste la dernière couche du réseau de neurones adaptée au domaine)
- Problème de biais, données d'entraînement, coût énergétique
- Langues peu dotées (quid au-delà des 20-100 langues actuellement traitées ?)

Résumé

- TAL : domaine essentiellement informatique aujourd'hui, mais des liens sont possibles (nécessaires ?) avec les Humanités numériques
 - Cf. Transparents précédents
 - Cf. Aussi psycholinguistique, sociolinguistique, etc.
- Nombreuses applications, dans de nombreux domaines
 - Basiques : gérer des corpus, des données volumineuses
 - Moins basiques : annotation semi-automatique / assistée de gros corpus
 - Complexes : modélisations, données hétérogènes