Turun yliopisto
University of Turku

ASPECTS PLURILINGUES ET ANALYSES QUANTITATIVES DES CMO*

Finnish Internet Parsebank et autres projets

Veronika Laippala
Département de langues et de traduction / Francais, Université de Turku
& Turku Institute for Advanced Studies (TIAS)

*Communications Médiatisées par Ordinateur

Turun yliopisto
University of Turku

KONEEN SÄÄTIÖ

= la fondation Kone

FINNISH INTERNET PARSEBANK

- **Vise à transformer l'Internet finnophone à une ressource linguistique avec les analyses syntaxiques**
- **Financement de la Fondation Kone 2014-2016**
    - Groupe de recherche
    - Département de langues et de traduction + le Département de sciences informatiques

Faculty of Mathematics and Natural Science

Faculty of Humanities

Turun yliopisto
University of Turku

Filip Ginter

Anna Missilä

Jenna Kanerva (Nyblom)

Juhani Luotolahti

Turun yliopisto
University of Turku

TURKU DEPENDENCY TREEBANK (TDT)

- **Collection de textes avec les annotations de syntaxe manuelles**
  - 204 399 mots, 15 126 phrases
    - Wikipédia, Wikinews, articles de la presse, blogues, *Europarl, Jrc-Aquis*, oeuvres de fiction, etc.

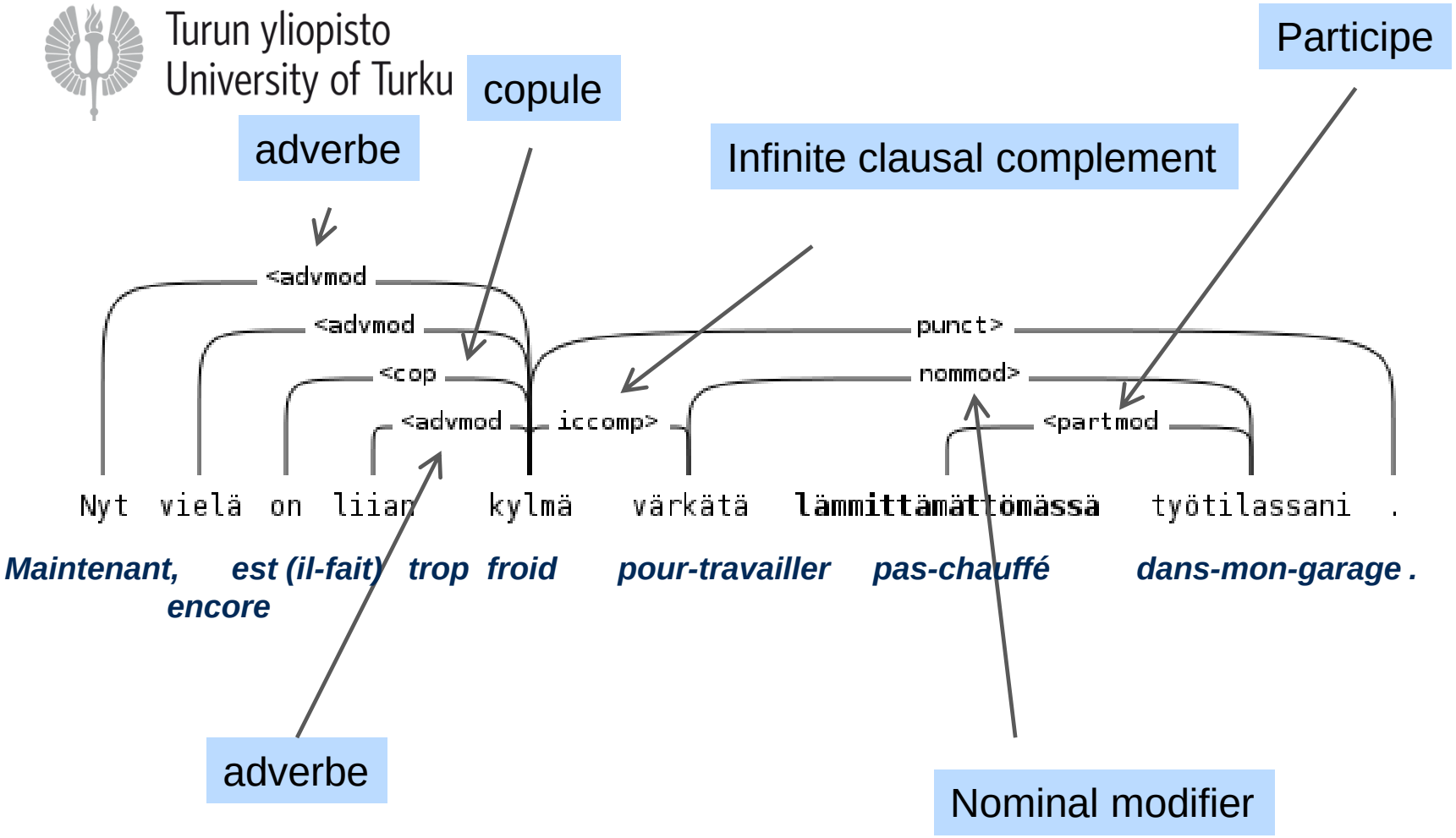
- Haverinen et al.: *Building the essential resources for Finnish: the Turku Dependency Treebank.* Language Resources and Evaluation. 2013

Turun yliopisto
University of Turku

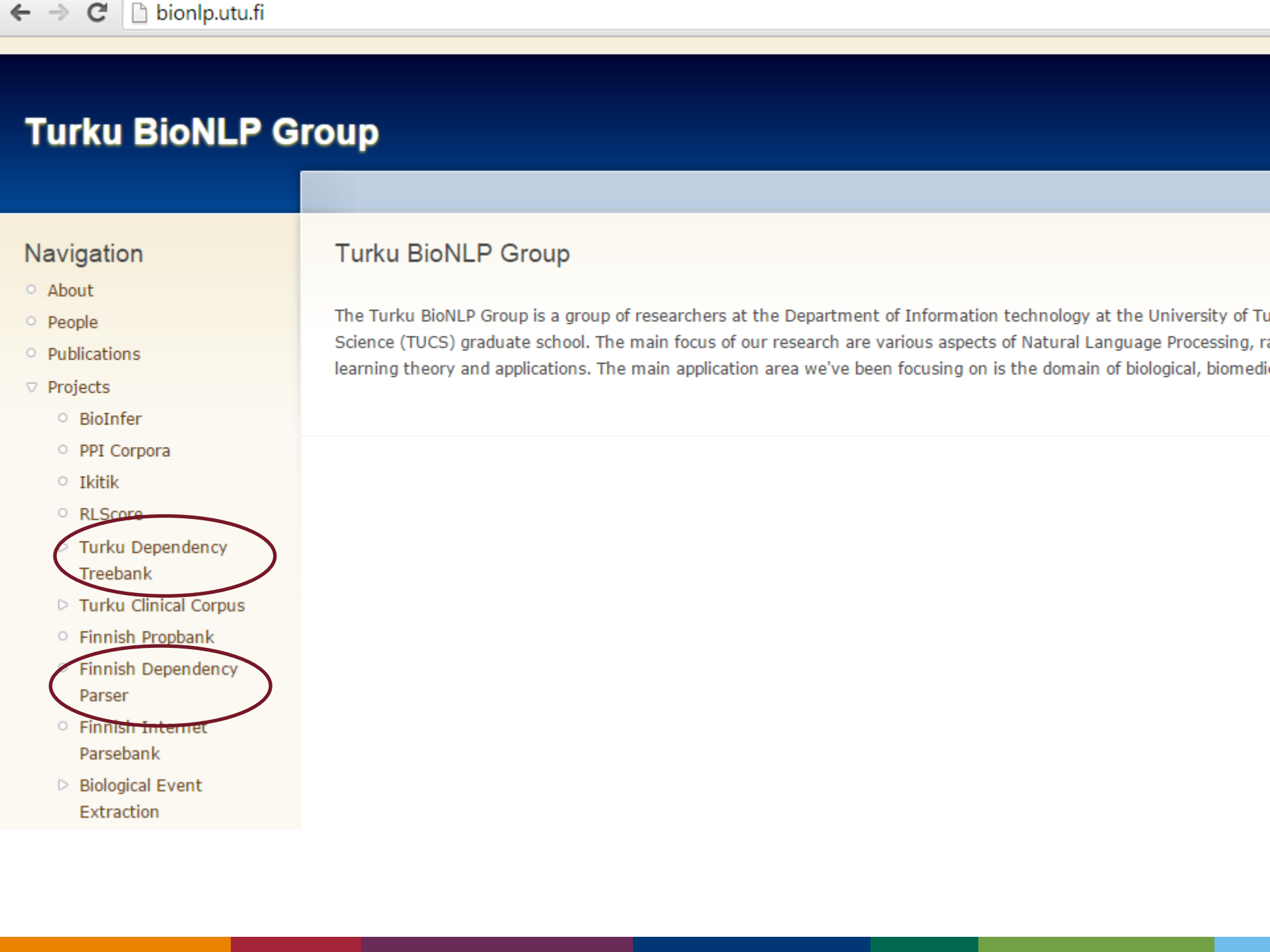
ANNOTATIONS DE SYNTAXE DANS LE TDT

- **Annotation en paires**
  - Inter annotator agreement 91%
- **Annotations suivent le schéma de Stanford (de Marneffe and Manning 2008)**
  - Détaillé, 48 types de dépendances
  - À l'origine pour l'anglais, quelques changements mineurs pour le finnois
  - Utilisé aussi dans les *Google treebanks* (anglais, francais, pourtugais, finnois, allemand, italien, indonesian, japonais, koréen, espagnol, suédois)

Turun yliopisto
University of Turku



Maintenant, il fait encore trop froid pour travailler dans mon garage sans chauffage.


bionlp.utu.fi
Turku BioNLP Group
Navigation
About
People
Publications
Projects
BioInfer
PPI Corpora
Ikitik
RLScore
Turku Dependency Treebank
Turku Clinical Corpus
Finnish Propbank
Finnish Dependency Parser
Finnish Internet Parsebank
Biological Event Extraction
Turku BioNLP Group
The Turku BioNLP Group is a group of researchers at the Department of Information technology at the University of Tu
Science (TUCS) graduate school. The main focus of our research are various aspects of Natural Language Processing, ra
learning theory and applications. The main application area we've been focusing on is the domain of biological, biomedi

Turku BioNLP Group / TDT Browser

Query language - Dependency types - TDT

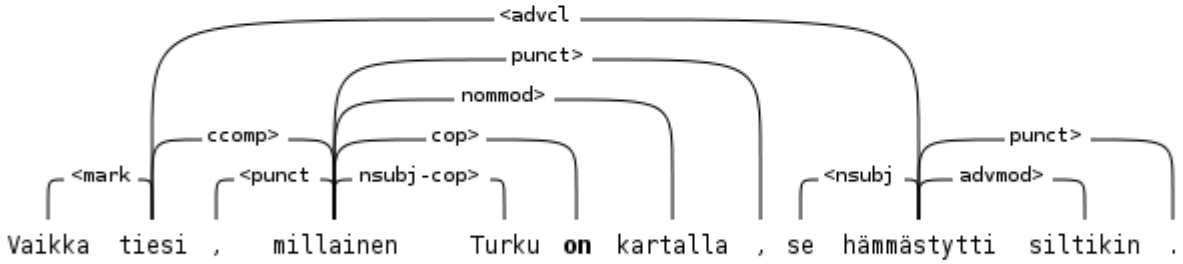
/on/ Search

Example: "_ </nsubj/ _"

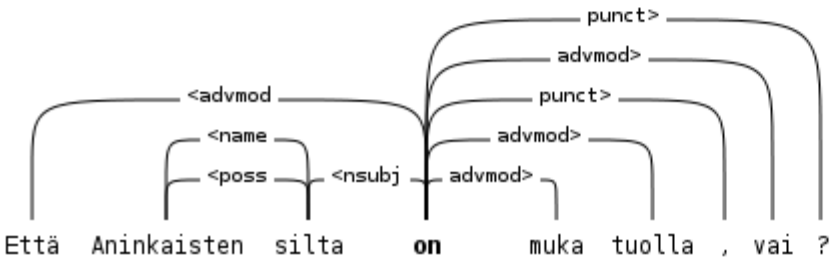
1291 tokens in **1149** sentences matched your query.

Showing results 1-20 (from a total of 1149)

b101-pg1/5



b101-pg1/6



Turun yliopisto
University of Turku

FROM A TREEBANK TO A PARSEBANK*

- L'analyseur syntaxique permet le développement de corpus avec ces analyses
  - Les analyses syntaxiques donnent plus d'information sur le langage utilisé et permettent des recherches plus détaillées
  - **Pourquoi pas faire un tel corpus?**

- Big is beautiful! Pourquoi exclure quelque chose quand il est possible de tout prendre?
  - **Pourquoi pas prendre tout l'Internet finnophone?**

* Une ressource linguistique avec les analyses syntaxiques

Turun yliopisto
University of Turku

TRANSFORMER L'INTERNET À UNE RESSOURCE LINGUISTIQUE

1. PARSEBANK
-Développement technique
-Découpage en mots, phrases, Analyses morphologiques + syntaxiques

2. SOUSCORPUS
-Classement en souscorpus selon genres / registres
-Définir les charactéristiques de ces souscorpus

3. INTERFACE UTILISATEUR
-Développement d'une interface facile à utiliser

Turun yliopisto
University of Turku

1. PARSEBANK

- ***Common crawl* copie tout l'Internet tous les deux ans**
  - → pas besoin de parcourir tout l'Internet
  - → identifier les pages en finnois suffit
  - Aussi un programme qui copie les sites sous .fi
  - En ce moment, 1,5 milliards de mots avec les analyses syntaxiques.

Common Crawl

Turun yliopisto
University of Turku

2. SOUSCORPUS ET LEURS CARACTÉRISTIQUES

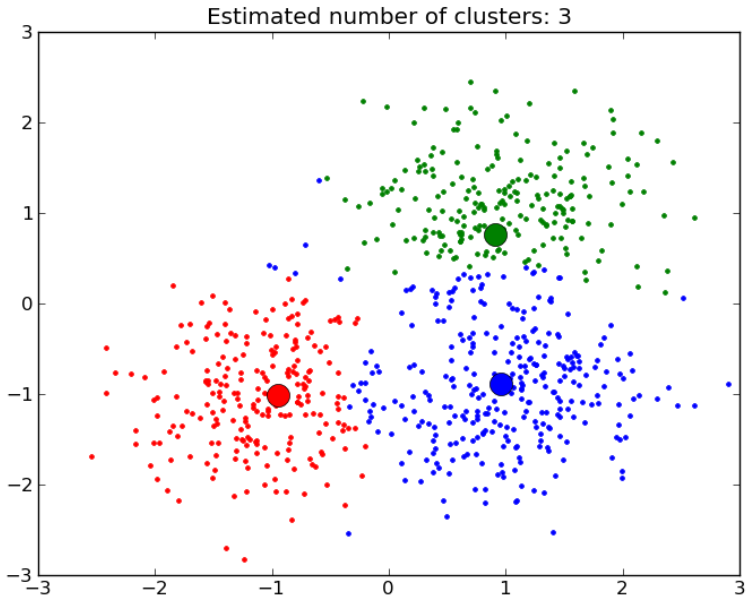
- **Dans le parsebank, il y a tout ce qu'on trouve sur l'Internet…**
  - De chats, de la presse, sites des services sociales, etc.
  - Très hétérogène et pas très facile à utiliser!
- **But de classifier les textes et en définir les caractéristiques**
  - → L'usage plus facile de Parsebank
  - → Nouvelles connaissances sur le langage et la CMO

DÉVELOPPEMENT DES SOUSCOPURS

- **La classification devrait se faire d'une manière *data-driven***
  - Le nombre + thématique des classes ne peuvent pas être décidés à l'avance
  - Il n'est pas possible de lire le parsebank
- **Méthodes possibles pour le classement *le clustering* et *l'apprentissage automatique***
- **Parsebank sera distribué classifié de plusieurs facons**
  - Souscorpus crées à partir de critères différents

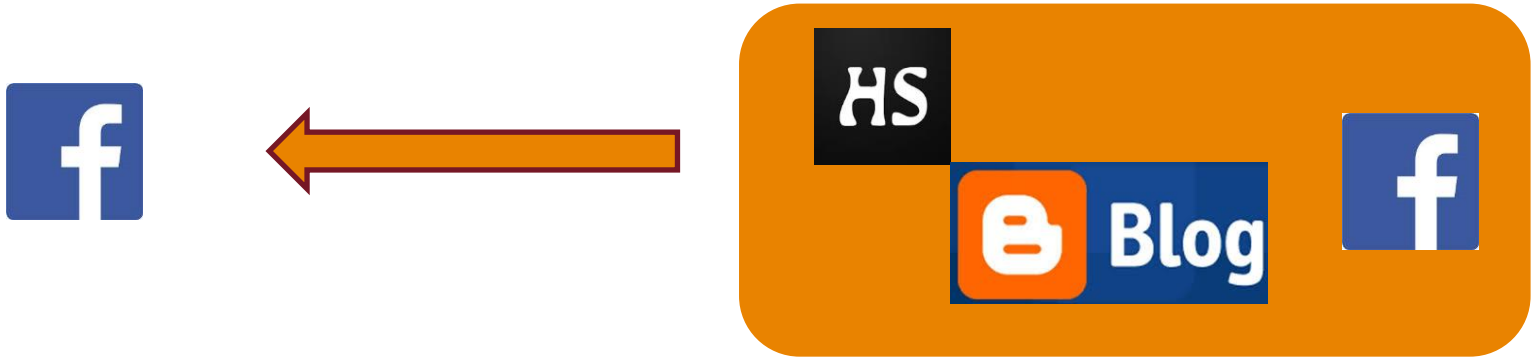
CLUSTERING

- **Une méthode de classification utile quand le matériel est inconnu**
  - → Parsebank!
- **Classes crées selon une mésure de similarité**
  - Similarités lexicales
    - →Classes de textes avec des lexiques similaires
  - Similarités structurales
    - → Structures similaires



CLASSEMENT DES TEXTES AVEC L'APPRENTISSAGE AUTOMATIQUE

- **Extraction des textes similaires que dans un corpus de référence**
  - Par ex. des articles de la presse, des forums de discussion, etc.



- **Classification automatique des textes du parsebank selon les registres/genres**

Turun yliopisto
University of Turku

ETAPE 1: IDENTIFICATION DES TEXTES TRADUITS (PAR LA MACHINE)

- **Déjà fait pour les textes traduits en anglais**
  - Caractéristiques utilisées dans l'apprentissage automatique les n-grammes des classes morphologiques (Aharoni et al. 2014)
  - "Phrase-salad" (Lopez 2008)
- **Nos caractéristiques des n-grammes syntaxiques**
  - Petits sous-arbres des analyses de dépendance
- **Résultats actuels sur les textes traduits par la machine, mais plus tard aussi pour les textes traduits par des humains**

→ **Extraction des textes traduits du parsebank**

→ **l'examen des caractéristiques des textes traduits**

ETAPES 2, 3, ETC

- **Identification d'autres registres**
- **Jusqu-où peut-on aller avec seulement des caractéristiques syntaxiques?**
    - Variation interne: blogues et blogues?
    - aussi pour le francais!

DECRIRE LES SOUSCORPUS

- **Caractéristiques des souscorpus montrées aux utilisateurs**
  - → Usage plus facile de Parsebank
  - → Information sur l'usage de la langue sur l'Internet

- **Méthodes possibles par ex. *factor analysis* (Biber e.g. 1992, 1995) *& analyse des mots-clés (key word analysis)* (Scott & Tribble 2006)**

ANALYSE DES MOTS-CLÉS

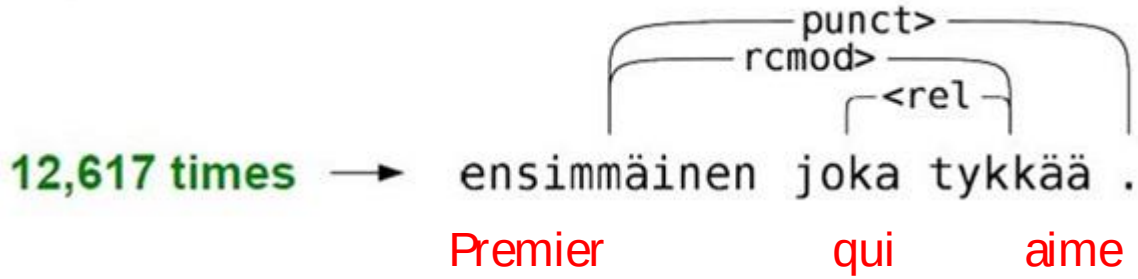
- **Scott & Tribble 2006**
- **Différences statistiquement significatives entre des corpus à l'aide des listes de mots clés**
  - Listes des mots / combinaisons de mots dont la fréquence dans un texte est plus grande qu'attendue

    → "style" ou thématique de texte

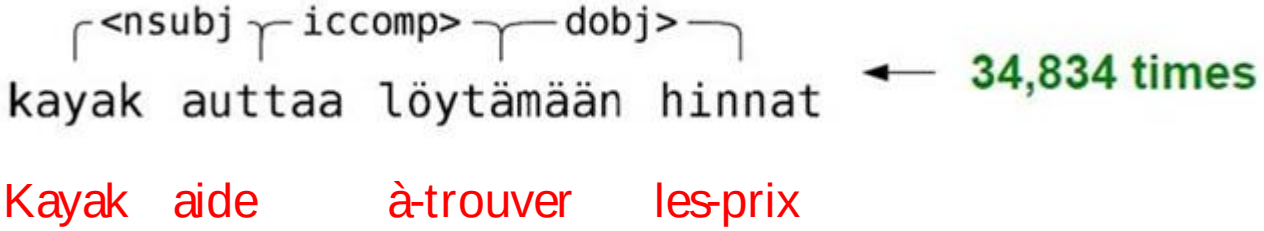
ANALYSE DES MOTS-CLÉS → ANALYSES DES STRUCTURES CLÉS (?)

- **Les analyses syntaxiques permettent d'étendre l'analyse des mots-clés aux structures syntaxiques**
  - Combinaisons (n-grammes) des fonctions syntaxiques
    - N-grammes syntaxiques = petits sous-arbres des analyses de dépendance
  - Variation de la structure du texte

  → Structures syntaxiques caractéristiques de certains registres?


punct>
rcmod>
<rel
12,617 times
ensimmäinen joka tykkää .
Premier qui aime


conj>
punct>
<name
14,859 times
bangkok , new york


<nsubj
iccomp>
dobj>
34,834 times
kayak auttaa löytämään hinnat
Kayak aide à-trouver les-prix

Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/978-1-61499-442-8-184

Syntactic N-gram Collection from a Large-Scale Corpus of Internet Finnish

Jenna KANERVA [a,1], Juhani LUOTOLAHTI [a],
Veronika LAIPPALA [b] and Filip GINTER [a]

[a] *Department of Information Technology, University of Turku, Finland*
[b] *Department of French Studies, University of Turku, Finland*

Abstract. In this paper, we report on the development of a large-scale Finnish Internet parsebank, currently consisting of 1.5 billion tokens in 116 million sentences. The data is fully morphologically and syntactically analyzed and it has been used to extract flat and syntactic n-gram collections, as well as verb-argument and noun-argument n-grams. Additionally, distributional vector space representations of the words are induced using the *word2vec* method. All n-gram collections as well as the vector space models are made available under an open license.

Keywords. Finnish, syntactic parsing, n-grams, syntactic n-grams, large-scale

syntactic-ngram-builder

An open-source tool to generate syntactic n-grams from a syntactically parsed data. The syntactic n-grams follow the same format as used in the Google Ngram Collection (http://googleresearch.blogspot.fi/2013/05/syntactic-ngrams-over-time.html).

Input

At the moment the only supported input format is CONLL-09. The extended n-grams are defined for the Stanford Dependencies (SD) scheme.

Generating n-grams

```
python build_ngrams.py input.conll09 --ngrams --args --out_dir output_directory
```

aelinco-en.blogs.uva.es/2014/07/29/7th-international-conference-on-corpus-linguistics-valladolid-spain-5-7-march-2015-uni

CILC 2015

7th International Conference on Corpus Linguistic

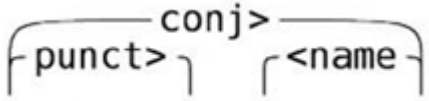
Home News Committees Call for papers Programmes Submission of proposals Proceed

7th International Conference on Corpus Linguistics Valladolid (Spain), 5-7 March 2015 University of Valladolid

29 July, 2014 by aelinco No comments

“Syntactic ngrams as keystructures reflecting typical syntactic patterns in Finnish”

- Wikipédia, de la presse, forums de discussion, littérature
- Des n-grammes syntaxiques sans items lexicaux mais avec les classes grammaticales



N punct N N

Anna Missilä, Filip Ginter, Jenna Kanerva, Outi Paloposki and Veronika Laippala:
Recognising (machine) translated texts in Finnish-language Internet

- **Détection par les n-grammes syntaxiques**
- **Caractéristiques syntaxiques des textes traduits**
  - **Variation selon la langue de départ**
  - **Variation selon le genre / registre**

L’idée principale est d’utiliser les analyses syntaxiques dans la description de la variation des textes

Caractéristiques syntaxiques des registres / genres ?

Turun yliopisto
University of Turku

Veronika Laippala

Blog, comment and discuss! A quantitative study of French Internet texts using automatic morpho-syntactic analysis.

My project aims at a quantitative analysis of the characteristics and distinguishing features of a large collection of different Internet texts in French, using automatic morphological and full dependency syntax analyses, i.e. the detection of word forms and their functions in the sentence as well as the identification of the entire sentence structure.

Turun yliopisto
University of Turku

Veronika Laippala

Blog, comment and discuss! A quantitative study of French Internet texts using automatic morpho-syntactic analysis.

My project aims at a quantitative analysis of the characteristics and distinguishing features of a large collection of different Internet texts in French, using automatic morphological and full dependency syntax analyses, i.e. the detection of word forms and their functions in the sentence as well as the identification of the entire sentence structure.

Turun yliopisto
University of Turku

TURKU CORPUS ON WRITTEN INTERACTION BASED ON COMPUTER-MEDIATED COMMUNICATION ON THE INTERNET

- **Allemand, suédois, finnois, anglais, francais**
- **Corpus comparables**
- **2008**

Turun yliopisto
University of Turku

- **Discussions du journal *Le Monde* (73 610 mots)**
  - La crise financière, la crise en Géorgie, Barack Obama, Sarah Palin
- **Discussions concernant la vie des jeunes adultes et étudiants (72 253 mots)**
  - La recherche de l'emploi, le CPE (contrat première embauche), questions financières
- **10 éditoriaux du journal *Le Monde* suivis de commentaires des lecteurs (3 603 + 22 343 mots)**
  - La crise financière, les élections présidentielles aux États-Unis, différentes questions liées à la politique française
- **15 chats avec des politiciens ou d'autres personnalités publiques (234 540 mots)**
  - www.rue89.com, www.libération.fr, www.20minutes.fr

Journal of Pragmatics

Volume 43, Issue 6, May 2011, Pages 1610–1627

Postcolonial pragmatics

Blogging politics in various ways: A typology of French politicians' blogs

Lotta Lehti

Show more

DOI: 10.1016/j.pragma.2010.11.017

Get rights and content

- 80 blogues de politiciens francais
- 874 billets (266,475 mots)
- septembre 2007
- 3316 commentaires (425,084 mots)

The French Social Media Bank: a Treebank of Noisy User Generated Content

Djamé Seddah, Benoît Sagot, Marie Candito, Virginie Mouilleron, Vanessa Combet

	# sent.	# tokens	avg. Length	std dev.
Doctissimo	771	10834	14.05	10.28
high noisiness subcorpora	36	640	17.78	17.63
other subcorpora	735	10194	13.87	9.74
JeuxVideos.com	199	3058	15.37	14.44
Twitter	216	2465	11.41	7.81
high noisiness subcorpora	93	1126	12.11	8.51
other subcorpora	123	1339	10.89	7.20
Facebook	452	4200	9.29	8.17
high noisiness subcorpora	120	1012	8.43	7.12
other subcorpora	332	3188	9.60	8.49

ORTOLANG

Open Resources and TOols for LANGuage

CoMeRe Repository: Corpora of Computer-Mediated Communication in French

- Polititweets (502 085 sanaa)
- Chat (IRC) (~4 000 000 viestiä)

- POST: xml:id: cmr-get-c008-a167 | when-iso: 2004-02-07T00:30 | who: #cmr-get-c008-p27729 | alias: SmileyEyes | type: chat-event | subtype: conne
  p: SmileyEyes(~54baab66e86ea5@fe22be2970bef9f2) vient d'arriver ! #18-25ans.
- POST: xml:id: cmr-get-c008-a168 | when-iso: 2004-02-07T00:30 | who: #cmr-get-c008-p27724 | alias: peuge | type: chat-message
  p: oui Jenni2004 a va bien
- POST: xml:id: cmr-get-c008-a169 | when-iso: 2004-02-07T00:31 | who: #cmr-get-c008-p27729 | alias: SmileyEyes | type: chat-event | subtype: decor
  p: SmileyEyes(~54baab66e86ea5@fe22be2970bef9f2) s'est déconnecté: Connection reset par peer
- POST: xml:id: cmr-get-c008-a170 | when-iso: 2004-02-07T00:31 | who: #cmr-get-c008-p27724 | alias: peuge | type: chat-message
  p: bonne année

Style in French Politicians' Blogs: Degree of Formality

LOTTA LEHTI AND VERONIKA LAIPPALA
University of Turku
urn:nbn:de:0009-7-37980

Abstract

We describe the degree of formality of language in French politicians' blogs, with specific focus on comparing blog posts and blog comments. The degree of formality is investigated in a corpus of posts and comments in 80 blogs through a cluster of features derived both from traditional French sociolinguistics and from studies of informal computer-mediated communication. The features examined are 1) syntactic (omission rate of the negative particle *ne* and forms of Yes/No-questions), 2) lexical (frequency of colloquialisms and of acronyms and non-standard spelling), and 3) prosodic (frequency of repetitive punctuation and emoticons). The analysis shows that the language used in the French politicians blogs is overall relatively standard. However, the language politicians use in their blog posts is more standard than the language used by commenters – the latter ranges from strictly formal to highly colloquial.

IMPEC
INTERACTIONS MULTIMODALES PAR ECRAN

Edition 2014
Du 2 au 4 juillet 2014
Université Lumière Lyon 2
16-18 quai Claude Bernard
69007 Lyon – France

Actes du colloque IMPEC 2014

Les formes interrogatives totales à travers divers modes de discussion en ligne : quelle connotation de niveau de langue ?

Lehti, Lotta ; Laippala,Veronika

Turun yliopisto
University of Turku

RÉSULTATS

	Commentaires blogs de politicien (n = 979)	Facebook (n= 86)	Forum Le Monde (n=181)	Editorial Le Monde (n= 64)	Forum Étudiants (n= 129)	Chat (n=877)
SV	35.2 %	**62.8 %**	**61.9 %**	37.5 %	**51.9 %**	20.2 %
ESV	1.8 %	4.7 %	5.5 %	4.7 %	7.8 %	4.7 %
VS	**62.9 %**	32.6 %	32.6 %	**57.8 %**	40.3 %	**75.1 %**

VS les plus fréquentes en rouge

SV les plus fréquentes en mauve

Turun yliopisto
University of Turku

RÉPARTITION DANS LES CHATS

	Chat (n=877)	Chat (van Compernolle and Williams 2009) (n= 167)
SV	20.2 %	97.7 %
ESV	4.7 %	0.6 %
VS	**75.1 %**	**1.7 %**

MAIS : ORTHOGRAPHE NON STANDARD

- ***Ne pensez vous pas*** ***que les Parisiens usagers ou travailleurs dans ces structures méritent mieux que ces contrats précaires et les changements incessants de personnel qui en découlent ?*** **(Chat Rue89)**
- ***Ce concours sera t'il conservé ?*** **(20 minutes chat)**

→ **Quelle connotation de niveau de langue ?**

Turun yliopisto
University of Turku

QUELLES MÉTHODES?

- **Analyse *data-driven* pour trouver les traits distinctifs entre les textes**
  - Au lieu des caractéristiques choisies à l'avance
  - Souvent les travaux se concentrent sur quelques caractéristiques prédéfinies
- **Comment se caractérisent les différents genres CMO étudiés?**

- **Interaction entre la thématique et les traits syntaxiques, lexicaux, les traits CMO…**
  - Quelles différences causées par le sujet, quelles par ”le format” (discussion synchrone/asynchrone, blogue, etc.)?
    - Blogue de politicien plutôt formel, mais un autre blogue peut-être pas?
    - Chat sur une thématique sociale plutôt formelle, mais sur une autre thématique peut-être pas?

Turun yliopisto
University of Turku

Caractéristiques lexicales (=sujet / thématique)

Caractéristiques CMO

genre

Caractéristiques structurelles

Variation personnelle selon le scripteur

Turun yliopisto
University of Turku

QUELLES MÉTHODES?

- **Dans la pratique?**
  - *Factor analysis*
  - mots-clés, structures clés
  - Une combinaison de … ?
  - ….?

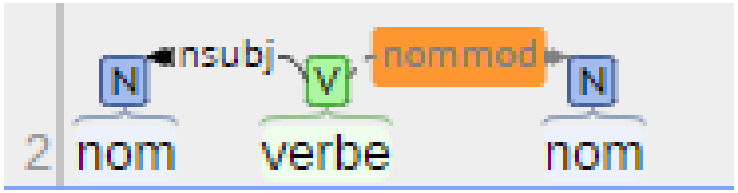
STRUCTURES CLÉS D'UN CORPUS DE PRESSE EN FINNOIS

Classe grammaticale

Type de dépendance

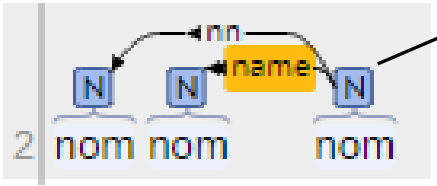
1 n__nsubj_2++v__root_0++n__nommod_2++

Numéro identifiant du gouverneur



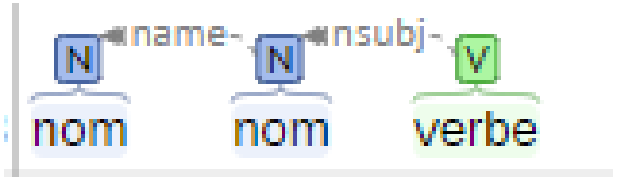
2 n__nn_3++n__name_3++n__nsubj_0++

noun compound modifier

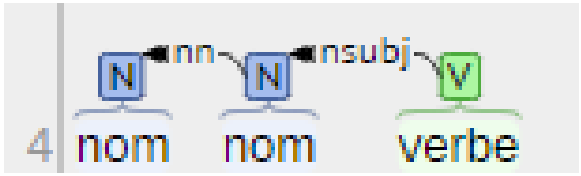


nsubj

3 n__name_2++n__nsubj_3++v__root_0++



4 n__nn_2++n__nsubj_3++v__root_0++



ANALYSE SYNTAXIQUE

- **Besoin d'adapter pour les CMO?**
- **Le schéma de Stanford très détaillé → utile dans l'analyse**
- **Notre pipeline des n-grammes syntaxiques pour le schéma de Stanford**

http://universaldependencies.github.io/docs/

Universal Dependencies

Universal | Basque | Bulgarian | Czech | English | Finnish | French | German | Greek | Hebrew | Hungarian | Irish | Italian | Japanese | Korean | Persian | Spanish | Swedish | ...

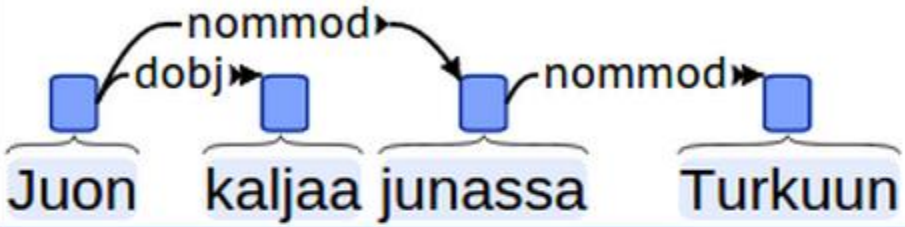
Introduction to Universal Dependencies

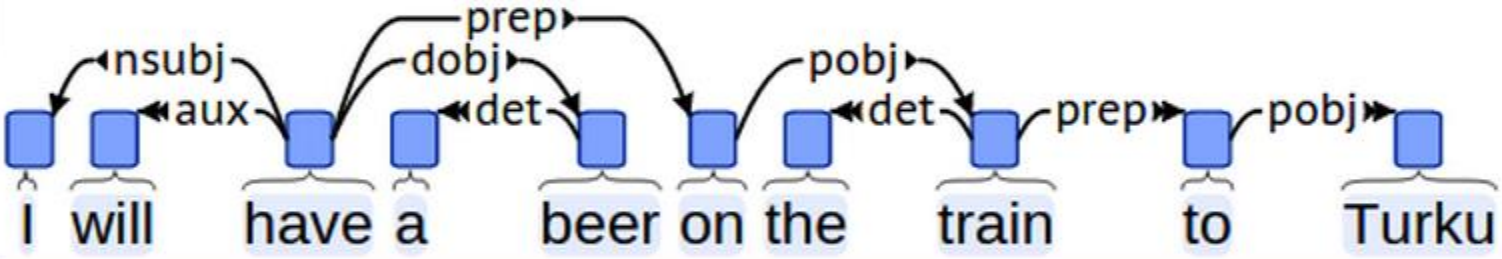
- Tokenization
- Morphology
  - General principles
  - Universal POS tags (single document)
  - Universal features (single document)
  - Language-specific features
- Syntax
  - General principles
  - Specific constructions
  - Universal dependency relations (single document)
  - Language-specific relations

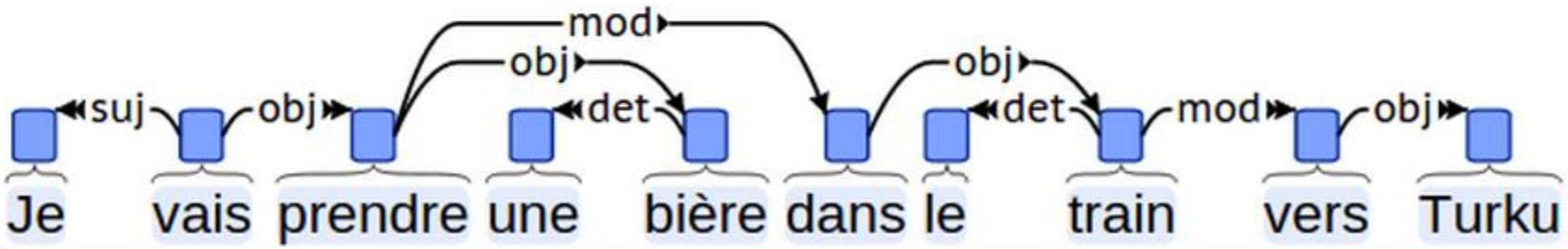
CoNLL-U format

This is the online documentation and example treebank for Universal Dependencies, version 1 (2014-10-01). We intend to treat version 1 as stable for at least the next year, but we may subsequently make further revisions based on experiences using it to treebank a range of languages. If you plan to use the scheme yourself, please get in touch so that we can avoid problems with conflicting versions.

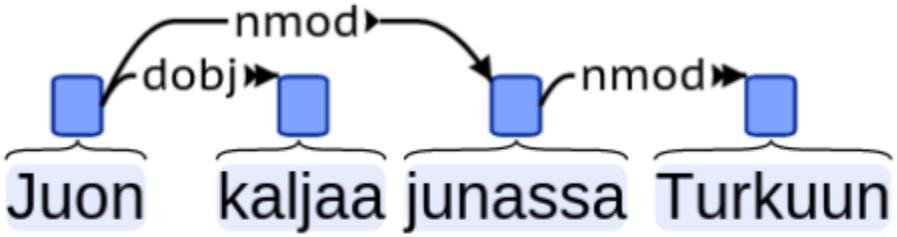
"Every man for himself!"



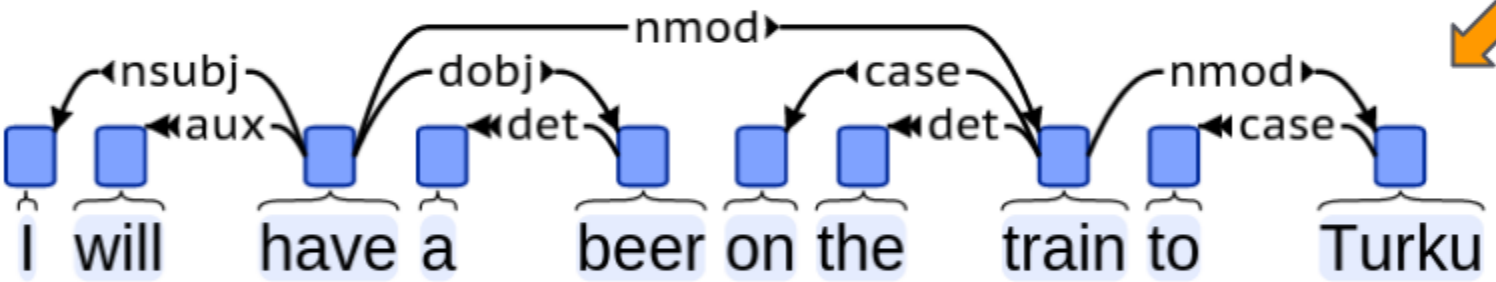


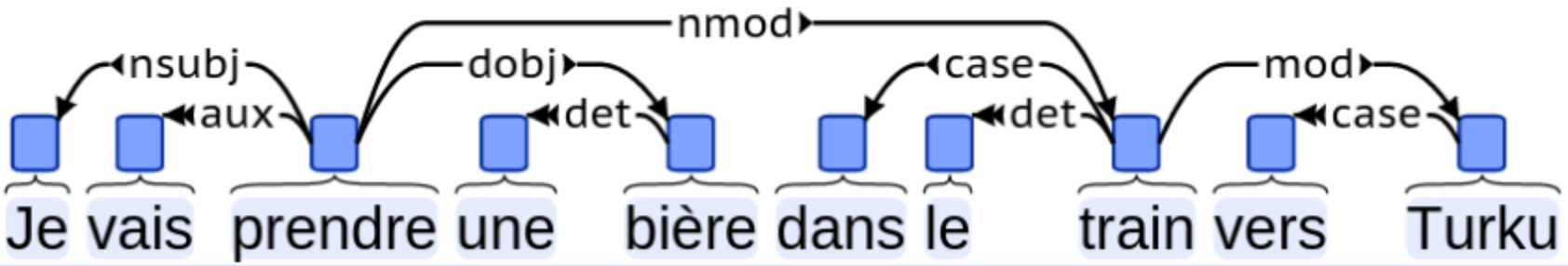


“United we stand!”



Same tree!

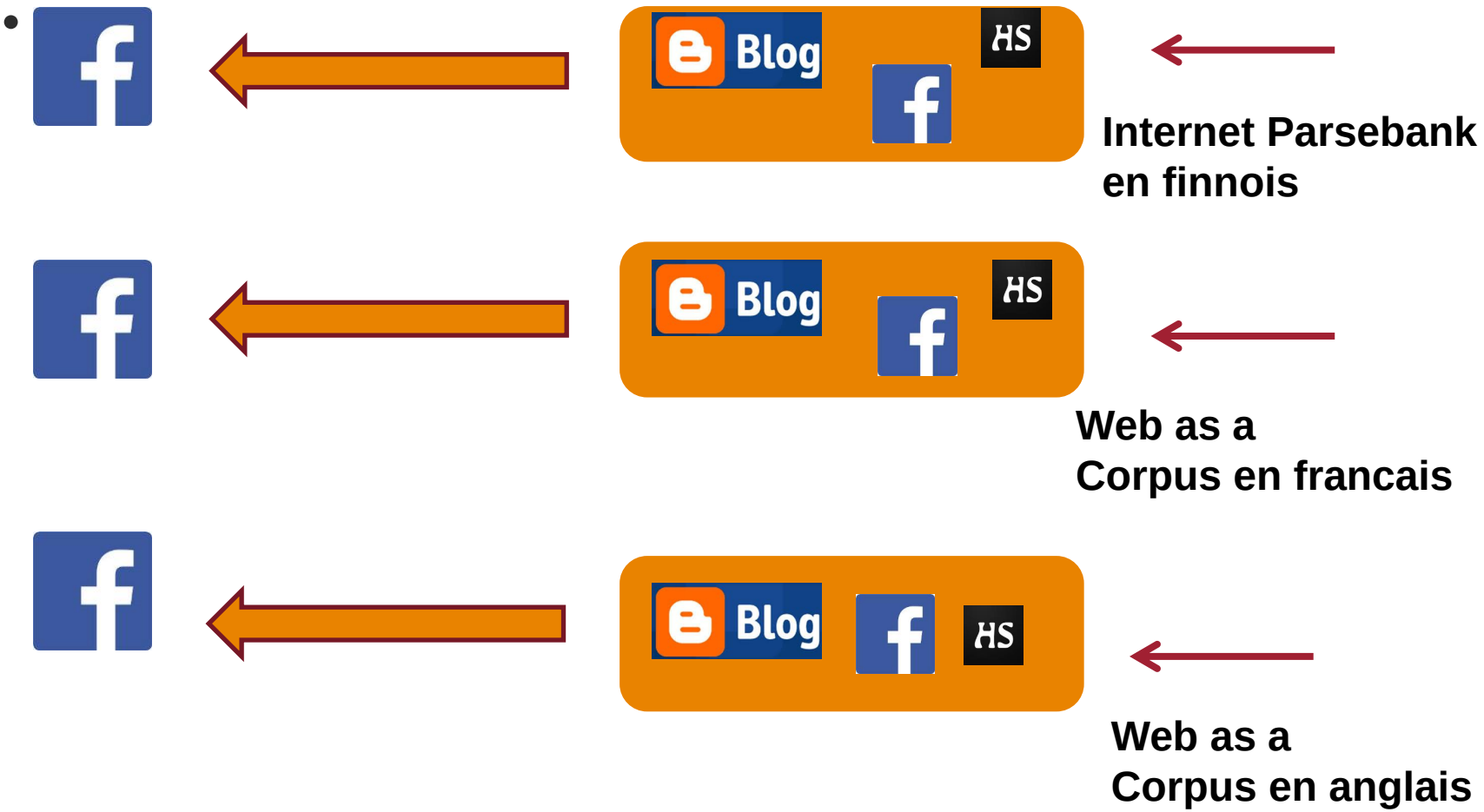




Universal Parser

- Each treebank in UD is of sufficier size and quality to train a parser
- During 2015, we can expect this to happen for a number of languages (and that certainly includes Finnish)
- Approaching a “universal parser” for all major languages
  - Any language IN - unified representation OUT
  - GOAL: same (high) accuracy for all languages

CORPUS PLURILIGUES ET LEUR COMPARAISON!



Turun yliopisto
University of Turku

QUELLES MÉTHODES ?
(ENCORE LA MÊME QUESTION)

- **Pour les n-grammes syntaxiques, le schéma de Stanford ou l'UD nécessaire**
  - Méthode statistique de comparaison entre les corpus *random forests? SVM?*
- Comme promis, factor analysis entre les corpus?
- Mots / structures clés seuls ne suffisent pas?
  - Peuvent quand même être informatifs sur l'interaction entre le genre / trais lexicaux (thématique) / traits syntaxiques

Turun yliopisto
University of Turku

A FAIRE TRÈS BIENTÔT

- **Examen des corpus IRC**
  - Discussions du journal *Le Monde*
  - Discussions concernant la vie des jeunes adultes et étudiants
  - Editoriaux du journal *Le Monde* suivis de commentaires des lecteurs
  - Chats avec des politiciens ou d'autres personnalités publiques
- **D'autre blogues que ceux des politiciens, pour la comparaison?**


vastavalo.fi:242502
vastavalo.fi:242502
Merci!
Questions, commentaires?