TreeLex++
vers une ressource syntaxico-sémantique

Anna Kupść
avec Antonio Balvet, Pauline Haas, Rafael Marin

December 11, 2018

Présentation générale

- TreeLex
- assignation d'aspect
- analyse d'interactions

Lexique syntaxique: quoi et pourquoi?

- ressource lexicale avec des informations syntaxiques
- structure argumentale peut varier d'une langue à l'autre

FR:	**Sophie** manque **à ses amis**	**SN**, **à+SN**
EN:	**Sophie's friends** miss **her**	**SN**, **SN**
PL:	**Przyjaciele Zosi** tęsknią **za nią**	**SN**, **za+SN_instr**

- apprentissage des langues
- TAL: extraction d'information, QA, analyse, génération, etc.

TreeLex

- une ressource lexicale extraite à partir de French Treebank (FTB)
- extraction automatique guidée par des connaissances linguistiques
- structure argumentale: fonction + nature
- env. 2000 verbes

`http://redac.univ-tlse2.fr/lexiques/treelex.html`

French Treebank

- journal *Le Monde*, (1990-1993)
- annotation automatique: morphologique, constituents, fonctions
- format XML
- validation/vérification manuelle
- env. 560 000 mots (dans la partie catégorisée)

$\Rightarrow$ petit mais riche

`http://ftb.linguist.univ-paris-diderot.fr/`

Exemple

Une quinzaine de militaires libériens ont été transférés à Abidjan.

Texte | XML | PTB | Tiger | CoNLL

```
<SENT argument="ETR" author="MINANGOY ROBERT" date="1990-01-19" nb="1015" textID="4
    <NP fct="SUJ">
        <w cat="D" ee="D-ind-fs" ei="Dfs" lemma="un" mph="fs" subcat="ind">Une</w>
        <w cat="N" ee="N-C-fs" ei="NCfs" lemma="quinzaine" mph="fs" subcat="C">quin
        <PP>
            <w cat="P" ee="P" ei="P" lemma="de">de</w>
            <NP>
                <w cat="N" ee="N-C-mp" ei="NCmp" lemma="militaire" mph="mp" subcat=
                <AP>
                    <w cat="A" ee="A-qual-mp" ei="Amp" lemma="libérien" mph="mp" su
                </AP>
            </NP>
        </PP>
    </NP>
    <VN>
        <w cat="V" ee="V--P3p" ei="VP3p" lemma="avoir" mph="P3p" subcat="">ont</w>
        <w cat="V" ee="V--Kms" ei="VKms" lemma="être" mph="Kms" subcat="">été</w>
        <w cat="V" ee="V--Kmp" ei="VKmp" lemma="transférer" mph="Kmp" subcat="">tra
    </VN>
    <PP fct="P-OBJ">
        <w cat="P" ee="P" ei="P" lemma="à">à</w>
        <NP>
            <w cat="N" ee="N-P-ms" ei="NPms" lemma="Abidjan" mph="ms" subcat="P">Ab
        </NP>
    </PP>
    <w cat="PONCT" ee="PONCT-S" ei="PONCTS" lemma="." subcat="S">.</w>
</SENT>
```

Fonctions et Catégories dans FTB

(MOD ignoré)

fonction	catégorie
SUJ	NP, VPinf, Ssub, COORD
OBJ	NP, AP, VPinf, COORD, Sint, Ssub
DE-OBJ	VPinf, PP, Ssub, COORD
A-OBJ	VPinf, PP, COORD
P-OBJ	PP, AdP, COORD, NP
ATO	Srel, PP, AP, NP, VPpart, COORD, VPinf, Ssub
ATS	NP, PP, AP, AdP, VPinf, Ssub, COORD, VPpart, Sint

Sophie manque **à ses amis**
manquer $\Rightarrow$ SUJ:NP, A-OBJ:NP

Extraction des fonctions

indiquées pour prédicats verbaux:

- VN (noyau verbal): verbe principal, auxiliaires, clitiques, négation
- dépendants de VN:

```
<NP fct="SUJ">L'eclipse lunaire</NP>
<VN>a duré</VN>
<NP fct="OBJ">1h14</NP>
```

- attributs de VN:

```
<VN fct="SUJ">Elle a duré</VN>
<NP fct="OBJ">1h14</NP>
```

- pas de relation explicite entre la tête verbale et ses dépendents
- la tête verbale: le dernier verbe dans VN

Arguments clitiques

- séparer les fonctions accumulées:
  `<VN fct="SUJ/OBJ">Il l'a vue</VN>`
- garder les clitiques spéciaux (sans fonction dans VN)
  `<NP fct="SUJ">Marc</NP>`
  `<VN>s'est évanoui</VN>` ⇒ SUJ:NP,**refl:CL**
  `<VN fct="SUJ">Il y a</VN>`
  `<NP fct="OBJ">une solution</NP>`
  ⇒ SUJ:**il**,OBJ:NP,**obj:y**
- supprimer les doublons:
  `<NP fct="SUJ">Paul</NP>`
  `<VN fct="OBJ/SUJ">en mange-t-il</VN>`
  `<AdP fct="OBJ">beaucoup</AdP>?`
- grouper des réalisations, ex. sujet
  Marc *dort* vs. ***Il*** *dort* ⇒ SUJ:NP

Forme active vs. passive

- argument prépositionnel avec *par* ou *de*
- liste de 62 verbes conjugués avec *être*
- si un verbe conjugué avec *avoir* apparaît avec *être*:
  - ajouter OBJ
  - (supprimer ATS et ajouter ATO)
  - supprimer PP *par* ou *de*
  - `Un autre tiers a pu être acquis par les employés.`

D'autres Modifications

- sujet SN ajouté aux formes impératives et infinitifs
- neutralisations:
  - type de subordonnée (Sint, Ssub)
  - préposition (sauf *de* et *à*)
  - catégorie d'attributs sujet/objet (ATS, ATO)
- ordre des fonctions:

  `SUJ,OBJ,A-OBJ,DE-OBJ,P-OBJ,ATS,ATO,obj,refl`
- réalisations optionnelles (groupée par A.Abeillé):

  `SUJ:NP,OBJ:NP`

  `SUJ:NP`

  $\Rightarrow$ `SUJ:NP,(OBJ:NP)`

Extrait de TreeLex

CSV, point virgule, UTF8

```
gonfler;SUJ:NP, OBJ:NP, DE-OBJ:PP;1
gonfler;SUJ:NP;2
gonfler;SUJ:NP, OBJ:NP;7
entraver;SUJ:NP, OBJ:NP;6
assembler;SUJ:NP, OBJ:NP;3
expliquer;SUJ:NP, OBJ:Ssub, (A-OBJ:PP);22
expliquer;SUJ:NP, OBJ:NP, P-OBJ:PP;1
expliquer;SUJ:NP, (OBJ:NP), (A-OBJ:PP);92
expliquer;SUJ:NP, P-OBJ:PP, refl:CL;5
associer;SUJ:NP, OBJ:NP, (A-OBJ:PP);21
associer;SUJ:NP, OBJ:NP, A-OBJ:PP, P-OBJ:PP;1
associer;SUJ:NP, P-OBJ:PP, refl:CL;3
associer;SUJ:NP, DE-OBJ:PP, refl:CL;1
```

Extrait de TreeLex

CSV, point virgule, UTF8

verb	Frame_realisation	Frame_frequency
gonfler	SUJ:NP, OBJ:NP, DE-OBJ:PP	1
gonfler	SUJ:NP	2
gonfler	SUJ:NP, OBJ:NP	7
entraver	SUJ:NP, OBJ:NP	6
assembler	SUJ:NP, OBJ:NP	3
expliquer	SUJ:NP, OBJ:Ssub, (A-OBJ:PP)	22
expliquer	SUJ:NP, OBJ:NP, P-OBJ:PP	1
expliquer	SUJ:NP, (OBJ:NP), (A-OBJ:PP)	92
expliquer	SUJ:NP, P-OBJ:PP, refl:CL	5
associer	SUJ:NP, OBJ:NP, (A-OBJ:PP)	21
associer	SUJ:NP, OBJ:NP, A-OBJ:PP, P-OBJ:PP	1
associer	SUJ:NP, P-OBJ:PP, refl:CL	3
associer	SUJ:NP, DE-OBJ:PP, refl:CL	1

TreeLex: Résumé

- 1912 verbes (3229 entrées) avec structure argumentale riche
- une entrée: verbe avec **toutes** réalisations du corpus
  - `voler:  SUJ:NP,OBJ:NP,A-OBJ:PP`
  - `voler:  SUJ:NP`
  - `voler:  SUJ:NP,DE-OBJ:PP`
- 465 verbes polylexicales, ex. `courir le risque`, `donner lieu`
- réalisations avec clitiques:
  - verbes pronominaux, ex. `se réjouir`
  - expressions idiomatiques, ex. `s'en sortir`
  - sujet impersonnel, ex. `falloire`
- réalisations (cadres) attestés dans le corpus

Applications de TreeLex?

- une ressource pour apprentissage des langues
- recherche linguistique?
  taille raisonnable mais informations très riches $\Rightarrow$
  généralisations difficiles
- TAL?
  pas de flexion $\Rightarrow$ pas d'applications directes

TreeLex++: Vers une étude linguistique

- étude des propriétés syntaxico-sémantiques
- interactions verbes/adjectifs/noms
- sémantique: propriétés aspectuelles

TreeLex vs. Aspect

- env. 40% de verbes dans TreeLex ont plusieurs cadres ⇒ potentiellement polysémiques (et/ou polyaspectuels)
- différentes réalisations syntaxiques séparées ⇒ ambiguïté artificielle
  - `déplorer:  SUJ:NP,OBJ:Ssub`
  - `déplorer:  SUJ:NP,OBJ:NP`
- pas d'examples dans TreeLex ⇒ difficile de determiner le sens/aspect

Données pour TreeLex++

- un argument optionnel est un argument:
  `SUJ:NP,(OBJ:NP)` $\Rightarrow$ `SUJ:NP,OBJ:NP`
- groupement des différentes réalisations syntaxiques, ex.
  `déplorer:  SUJ:NP,OBJ:NP/Ssub`
- verbes avec **un seul** cadre (dans FTB)
- examples de FTB pour chaque verbe
- verbes polylexicaux exclus

$\Rightarrow$ 1049 verbes $\sim$ 55% TreeLex

Classes Aspectuelles

aspect: propriétés sémantiques généralisées

Vendler (1967)

	dynamique	télique	duratif
ÉTAT	−	−	+
ACTIVITÉ	+	−	+
ACCOMPLISSEMENT	+	+	+
ACHÈVEMENT	+	+	−

Caractéristique aspectuelle: Tests

[+dynamique]	*être en train de Vinf* *Que s'est-il passé hier?*	ACT,ACC,ACH
[+télique]	*V en x temps* *finir de Vinf*	ACC,ACH
[+duratif]	*V (en/pendant) x temps* *commencer de Vinf*	ÉTAT,ACT,ACC

Assignation d'aspect: un exemple

Verbe	example	tests	classe
EXPLOITER	ce marché va attirer de nouveaux opérateurs , venant s'ajouter à ceux qui **exploitent** déjà les chaînes thématiques actuelles	*ils sont en train d'exploiter les chaînes thématiques actuelles* → **[+dynamique]** *ils ont exploité les chaînes thématiques pendant (*en) 10 ans*. → **[+duratif] , [-télique]**	ACT
PRÉTENDRE	Les syndicats s'y opposent puisqu'ils **prétendent** représenter l'ensemble des employés face au patronat	* *Ils s'y opposent puisqu'ils sont en train de prétendre représenter x* → **[-dynamique]**	ETAT

assignation manuelle:

- deux experts
- tests appliqués aux exemples de FTB
- examples **adaptés** aux tests

Données

Verb	Aspect	Telic	DYN	NbArgs	General	GeneralFrame_arg2	GeneralFrame_arg3	GeneralFrame_concat	Freq Rang
abîmer	ACC	yes	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	LOW
abolir	ACH	yes	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	LOW
abonder	ETAT	no	no	I	SUJ	NO	NO	SUJ.NO.NO	LOW
aborder	ACH	yes	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	MID
abréger	ACH	yes	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	LOW
abriter	ETAT	no	no	II	SUJ	OBJ	NO	SUJ.OBJ.NO	LOW
absorber	ACC	yes	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	LOW
absoudre	ACH	yes	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	LOW
abuser	ACH	yes	yes	II	SUJ	DE-OBJ	NO	SUJ.DE-OBJ.NO	LOW
accéder	ACH	yes	yes	II	SUJ	A-OBJ	NO	SUJ.A-OBJ.NO	MID
accentuer	ACT	no	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	MID
accoucher	ACC	yes	yes	II	SUJ	DE-OBJ	NO	SUJ.DE-OBJ.NO	LOW

interaction Aspect vs. arguments

Données

Verb	Aspect	Telic	DYN	NbArgs	General	GeneralFrame_arg2	GeneralFrame_arg3	GeneralFrame_concat	Freq Rang
abîmer	ACC	yes	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	LOW
abolir	ACH	yes	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	LOW
abonder	ETAT	no	no	I	SUJ	NO	NO	SUJ.NO.NO	LOW
aborder	ACH	yes	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	MID
abréger	ACH	yes	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	LOW
abriter	ETAT	no	no	II	SUJ	OBJ	NO	SUJ.OBJ.NO	LOW
absorber	ACC	yes	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	LOW
absoudre	ACH	yes	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	LOW
abuser	ACH	yes	yes	II	SUJ	DE-OBJ	NO	SUJ.DE-OBJ.NO	LOW
accéder	ACH	yes	yes	II	SUJ	A-OBJ	NO	SUJ.A-OBJ.NO	MID
accentuer	ACT	no	yes	II	SUJ	OBJ	NO	SUJ.OBJ.NO	MID
accoucher	ACC	yes	yes	II	SUJ	DE-OBJ	NO	SUJ.DE-OBJ.NO	LOW

interaction Aspect vs. arguments

Aspect: chiffres

aspect	fréquence	%
ACH	532	50.7%
ACC	228	21.7%
ACT	201	19.2%
ETAT	88	8.4%
TOTAL	1049	100%

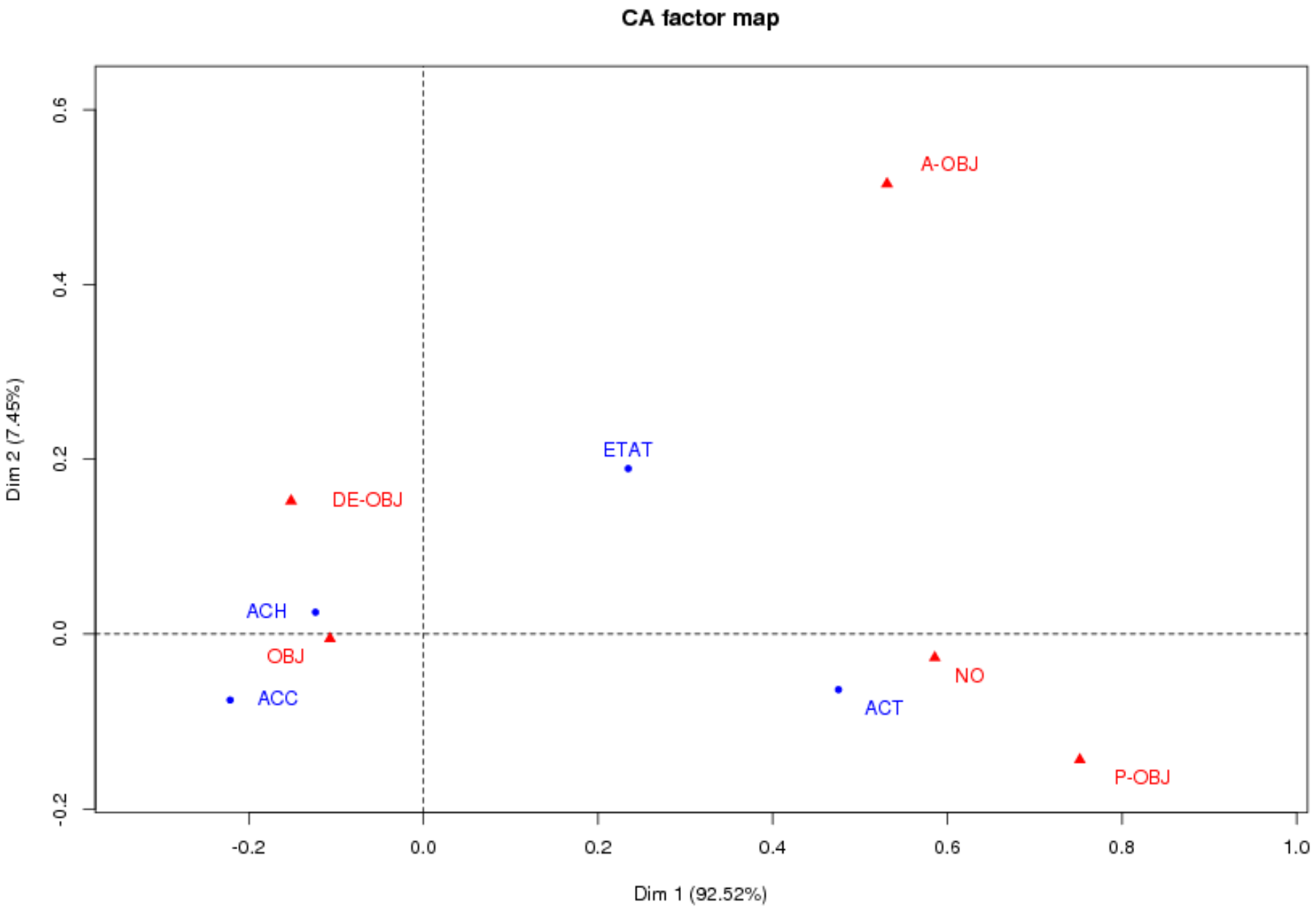
Analyse quantitative

- vérifier des interactions Aspect vs. Structure Argumentale
- peut-on distinguer les 4 aspects?
- hypothèse: il y a une forte association entre aspect et structure argumentale

Analyse de correspondence

- Aspect x Structure Argumentale (ARG1,ARG2,ARG3)
- analyse avec R, bibliothèque FactoMinerR
- ARG1: présent avec tous les verbes $\Rightarrow$ inutile
- ARG2/ARG3 vs. classes aspectuelles

Aspect x ARG2



Interpretation de résultats pour ARG2

- quatre classes indentifiables:
  - ETAT $\neq$ ACT
  - ACH $\sim$ ACC (télique)
- ETAT: pas de realisation ARG2 spécifique
- ACH+ACC: OBJ
- ACH avec DE-OBJ?
- ACT: un seul argument (ARG2 absent: NO) ou P-OBJ?
- A-OBJ: pas d'aspect spécifique ou association négative avec ACH/ACC

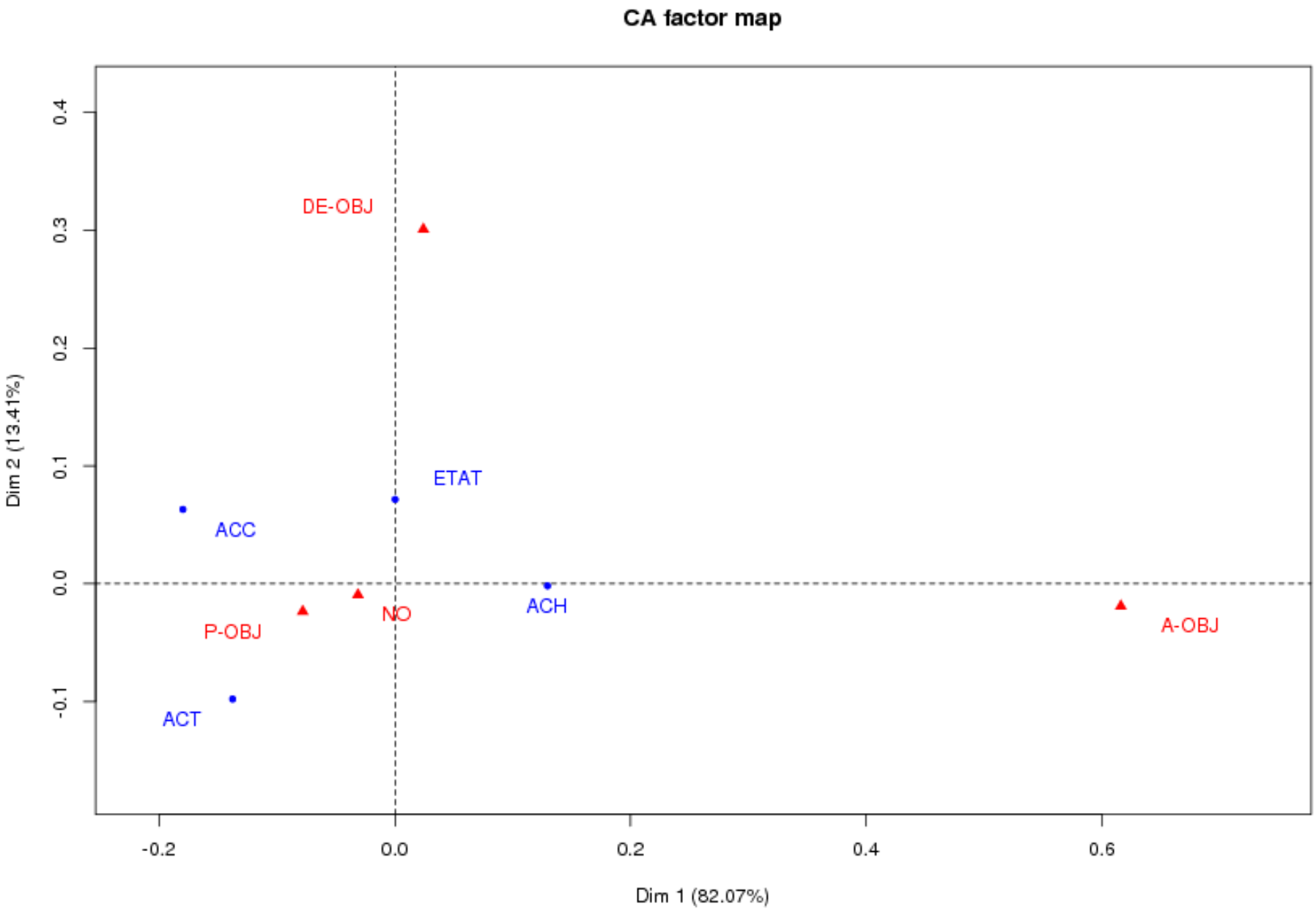
ARG2 sous la loupe

ARG2	A-OBJ	DE-OBJ	NO	OBJ	P-OBJ
ACC	0	4	13	208	3
ACH	8	10	41	464	8
ACT	5	2	46	135	13
ETAT	4	2	14	65	3

ARG2 sous la loupe

ARG2	A-OBJ	DE-OBJ	NO	OBJ	P-OBJ
ACC	0	4	13	208	3
ACH	8	10	41	464	8
ACT	5	2	46	135	13
ETAT	4	2	14	65	3

Aspect x ARG3



Interpretation de résultats pour ARG3

- quatre classes toujours séparées
  - ETAT toujours différent des autres, en particulier de ACT
- ACT et ACC: ARG3 absent (NO) ou P-OBJ?
- ARG3 absent dans d'autres classes
- A-OBJ et DE-OBJ sans classe spécifique

ARG3 sous la loupe

ARG3	A-OBJ	DE-OBJ	NO	P-OBJ
ACC	2	10	208	8
ACH	40	18	458	15
ACT	4	3	188	6
ETAT	4	4	79	1

ARG3 sous la loupe

ARG3	A-OBJ	DE-OBJ	NO	P-OBJ
ACC	2	10	208	8
ACH	40	18	458	15
ACT	4	3	188	6
ETAT	4	4	79	1

Conclusion et discussion

- plus de données nécessaires pour conclusions plus firmes
- biais dans FTB (donc TreeLex):
  - verbes intransitifs (ARG2:NO, ARG3:NO)
  - verbes transitifs directs (ARG2:OBJ)
- mais quatre classes identifiables/séparables ⇒ annotation à plus grande échelle faisable
- verbes téliques avec OBJ et la majorité de ACT intransitifs ⇒ notation FTB/TreeLex vs. syntaxe/sémantique

  ex., `arriver`: inaccusatifs

  Il est arrivé trois enfants. `SUJ:il,OBJ:NP`

  Trois enfants sont arrivés. `SUJ:NP`
- uniquement l'analyse de fonctions ⇒ quid nature?

  ARG: NP,Ssub,Vinf,PP