# Modeling the Complexity of Manual Annotation Tasks: a Grid of Analysis

**Karën Fort**, Adeline Nazarenko, Sophie Rosset

December 13th



# Modeling the Complexity of Manual Annotation Tasks: a Grid of Analysis

**Karën Fort**, Adeline Nazarenko, Sophie Rosset

December 13th


# Manual annotation: notoriously costly

**Penn Treebank** [Marcus et al., 1993]:

- 4.8 million tokens annotated with POS ⇒ learning phase of 1 month, to reach 3,000 words/h
- 3 million tokens annotated in syntax ⇒ learning phase of 2 months, to reach 475 words/h

**Prague Dependency Treebank** [Böhmová et al., 2001]:

- 1.8 million tokens annotated with POS and syntax

⇒ 5 years, 22 persons (max. 17 in parallel), 600,000 dollars

**GENIA** [Kim et al., 2008]:

- 9,372 sentences annotated in microbiology (proteins and gene names)

⇒ 5 part-time annotators, 1 senior coordinator and 1 junior for 1.5 year

# Some solutions...

- tag dictionary [Carmen et al., 2010]:
  - \+ simple to implement
  - \- bias

# Some solutions...

- tag dictionary [Carmen et al., 2010]:
  - \+ simple to implement
  - \- bias
- pre-annotation:
  - \+ gain in time and consistency
    [Marcus et al., 1993, Fort and Sagot, 2010]
  - \- bias

# Some solutions...

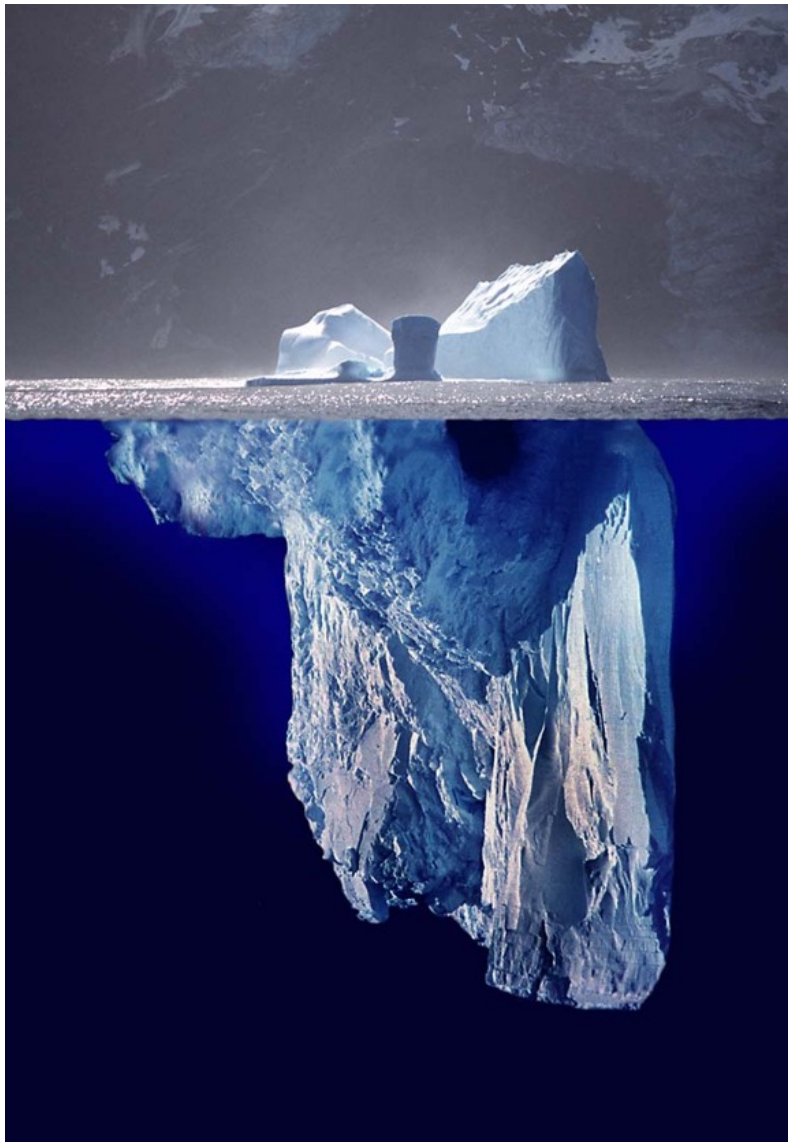
- tag dictionary [Carmen et al., 2010]:
  - \+ simple to implement
  - \- bias
- pre-annotation:
  - \+ gain in time and consistency
    [Marcus et al., 1993, Fort and Sagot, 2010]
  - \- bias
- active learning:
  - \+ gain in time [Cohn et al., 1995, Engelson and Dagan, 1996]
  - \- bias and not so simple to implement

# Some solutions...

- tag dictionary [Carmen et al., 2010]:
  - \+ simple to implement
  - \- bias
- pre-annotation:
  - \+ gain in time and consistency [Marcus et al., 1993, Fort and Sagot, 2010]
  - \- bias
- active learning:
  - \+ gain in time [Cohn et al., 1995, Engelson and Dagan, 1996]
  - \- bias and not so simple to implement
- crowdsourcing:
  - GWAPs: real cost rarely estimated [Chamberlain et al., 2013]
  - microworking (MTurk): quality and ethical issues [Fort et al., 2011]

# Some solutions...

# ... to a problem that is still little known



# Some traces

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]
- Good practices:
  - formats [Ide and Romary, 2006]
  - organization [Bontcheva et al., 2010]
  - evaluation [Krippendorff, 2004]
- Partial methodologies: agile annotation
  [Bonneau-Maynard et al., 2005, Voormann and Gut, 2008]
- Some insights from cognitive science [Tomanek et al., 2010]

## Some traces

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]
- Good practices:
  - formats [Ide and Romary, 2006]
  - organization [Bontcheva et al., 2010]
  - evaluation [Krippendorff, 2004]
- Partial methodologies: agile annotation
  [Bonneau-Maynard et al., 2005, Voormann and Gut, 2008]
- Some insights from cognitive science [Tomanek et al., 2010]

# What is complex in manual annotation?

1 Introduction

2 Analysing the complexity of an annotation campaign

3 What to annotate?

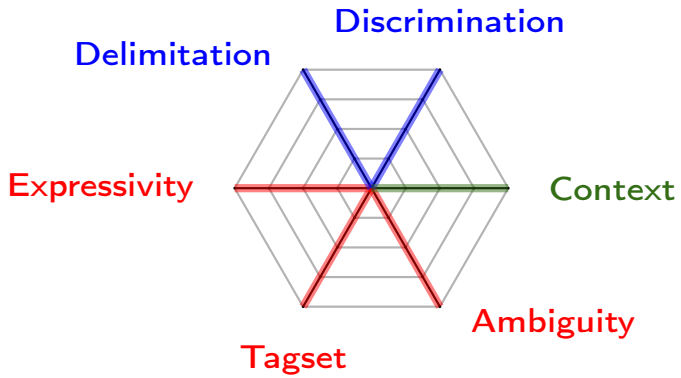
4 How to annotate?

5 Synthesis

6 Conclusion and prospects

# Complexity dimensions

- 5 independent dimensions:
  - 2 related to the localisation of annotations
  - 3 related to the characterisation of annotations
- 1 not independent: the context




- Scale from **0** (null complexity) to **1** (maximal complexity) to allow for the comparison between campaigns
- Independent from the volume to annotate and the number of annotators

# Elementary Annotation Task (EAT)

From a complex task, to several elementary tasks:

**Criteria**

An annotation task may be decomposed into at least two EATs if the used tagset can be decomposed into reduced and independent tagsets.

$\rightarrow$ may correspond to several successive annotation steps or not

# Example: gene renaming

1. Identification of gene names in the source signal:

   *The **yppB** gene complemented the defect of the recG40 strain. **yppB** and **ypbC** and their respective null alleles were termed "**recU**" and "**recU1**" (recU:cat) and "**recS**" and "**recS1**" (recS:cat), respectively.*

2. Identification of gene couples expressing a renaming relation:

   *The **yppB** gene complemented the defect of the recG40 strain. **yppB** and **ypbC** and their respective null alleles were termed **"recU"** and "**recU1**" (recU:cat) and **"recS"** and "**recS1**" (recS:cat), respectively.*

# Discrimination

Parts-of-speech [Marcus et al., 1993], pre-annotated :

*I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.*

Gene renaming[Fort et al., 2012], no pre-annotation:

*The yppB:cat and ypbC:cat null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ B. subtilis strain. The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU:cat) and recS and "recS1" (recS:cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.*

# Discrimination

Parts-of-speech [Marcus et al., 1993], pre-annotated :

> *I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.*

Gene renaming[Fort et al., 2012], no pre-annotation:

> *The yppB:cat and ypbC:cat null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ B. subtilis strain. The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU:cat) and recS and "recS1" (recS:cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.*

$\Rightarrow$ **more difficult** if the units to annotate are scattered, in particular if the segmentation is not obvious.

# Discrimination

The discrimination weight is all the more high as the proportion of what *should* be annotated as compared to what *could* be annotated is low.

## Definition

$$Discrimination(Flow) = 1 - \frac{|Annotations(Flow)|}{\sum_{i=1}^{LevelSeg} |UnitsObtainedBySeg_i(Flow)|}$$

$\Rightarrow$ Need for a reference segmentation

Parts-of-speech[Marcus et al., 1993] :

*I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.*

$$Discrimination_{PTB_{POS}} = 0$$

Gene renaming[Fort et al., 2012] :

*The yppB:cat and ypbC:cat null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ B. subtilis strain. The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU:cat) and recS and "recS1" (recS:cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.*

$$Discrimination_{Identification} = 0,9$$
$$Discrimination_{Renaming} = 0,95$$

# Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  *Madame Chirac* $\rightarrow$ *Monsieur et Madame Chirac*

# Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  *Madame Chirac* $\rightarrow$ *Monsieur et Madame Chirac*
- **decompose** a discriminated unit into several elements:
  *le préfet Érignac* $\rightarrow$ *le préfet Érignac*

# Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  *Madame Chirac* → *Monsieur et Madame Chirac*
- **decompose** a discriminated unit into several elements:
  *le préfet Érignac* → *le préfet Érignac*
- or **group** together several discriminated units into one unique annotation:
  *Sa Majesté*
  *le roi Mohamed VI* → *Sa Majesté le roi Mohamed VI*

# Boundaries delimitation

## Definition

$$Delimitation(Flow) = min\left(\frac{Substitutions + Additions + Deletions}{|Annotations(Flow)|}, 1\right)$$

$$Delimitation_{Identification} = 0$$
$$Delimitation_{Renaming} = 0$$

$$Delimitation_{PTB_{POS}} = 0$$

$$Délimitation_{EN_{TypesSubtypes}} = 1$$
$$Délimitation_{EN_{Components}} = 0,3$$

# Expressiveness of the annotation language

## Definition

The degrees of expressiveness of the annotation language are the following:

- 0.25: type languages
- 0.5: relational languages of arity 2
- 0.75: relational languages of arity higher than 2
- 1: higher-order languages

$$Expressiveness_{Identification} = 0.25$$
$$Expressiveness_{Renaming} = 0.25$$

# Dimension of the tagset

<table>
<tr><th colspan="3">Person</th><th colspan="3">Function</th></tr>
<tr><td colspan="2">pers.ind (individual person)</td><td>pers.coll (group of persons)</td><td colspan="2">func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><th colspan="3">Location</th><th colspan="3">Production</th></tr>
<tr><td>administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td>physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td>facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><td>prod.object (manufactured object)</td><td>prod.serv (transportation route)</td><td>prod.fin (financial products)</td></tr>
<tr><td colspan="3"></td><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td colspan="3"></td><td>prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><th colspan="3">Organization</th><th colspan="3" rowspan="3">Time</th></tr>
<tr><td colspan="2">org.adm (administration)</td><td>org.ent (services)</td></tr>
<tr><th colspan="3">Amount</th></tr>
<tr><td colspan="3">amount (with unit or general object), including duration</td><td colspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td>time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
</table>

Types and sub-types used for structured NE annotation [Grouin et al., 2011]

# Dimension of the tagset

| **Person** | |
|---|---|
| *pers.ind* (individual person) | *pers.coll* (group of persons) |

| **Function** | |
|---|---|
| *func.ind* (individual function) | *func.coll* (collectivity of functions) |

| **Location** | | |
|---|---|---|
| administrative (*loc.adm.town*, *loc.adm.reg*, *loc.adm.nat*, *loc.adm.sup*) | physical (*loc.phys.geo*, *loc.phys.hydro*, *loc.phys.astro*) | facilities (*loc.fac*), oronyms (*loc.oro*), address (*loc.add.phys*, *loc.add.elec*) |

| **Production** | | |
|---|---|---|
| *prod.object* (manufactured object) | *prod.serv* (transportation route) | *prod.fin* (financial products) |
| *prod.doctr* (doctrine) | *prod.rule* (law) | *prod.soft* (software) |
| *prod.art* | *prod.media* | *prod.award* |

| **Organization** | |
|---|---|
| *org.adm* (administration) | *org.ent* (services) |

| **Time** | |
|---|---|
| *time.date.abs* (absolute date), *time.date.rel* (relative date) | *time.hour.abs* (absolute hour), *time.hour.rel* (relative hour) |

| **Amount** |
|---|
| *amount* (with unit or general object), including duration |

Level 1: *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* $\rightarrow$ 7 possibilities (degree of freedom = 6).

# Dimension of the tagset

<table>
<tr><th colspan="4">Person</th><th colspan="3">Function</th></tr>
<tr><td colspan="2">pers.ind (individual person)</td><td colspan="2">pers.coll (group of persons)</td><td colspan="2">func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><th colspan="3">Location</th><th colspan="3">Production</th></tr>
<tr><td>administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td>physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td>facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><td>prod.object (manufac-tured object)</td><td>prod.serv (transporta-tion route)</td><td>prod.fin (financial products)</td></tr>
<tr><td colspan="3"></td><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td colspan="3"></td><td>prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><th colspan="3">Organization</th><th colspan="3">Time</th></tr>
<tr><td colspan="2">org.adm (administra-tion)</td><td>org.ent (services)</td><td colspan="2" rowspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td>time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
<tr><th colspan="3">Amount</th></tr>
<tr><td colspan="3">amount (with unit or general object), includ-ing duration</td><td colspan="3"></td></tr>
</table>

Level 1: *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* $\rightarrow$ 7 possibilities (degree of freedom = 6).
Level 2: *prod.object*, *prod.serv*, *prod.fin*, *prod.soft*, *prod.doctr*, *prod.rule*, *prod.art*, *prod.media*, *prod.award* $\rightarrow$ 9 possibilities (degree of freedom = 8).

# Dimension of the tagset

<table>
<tr><th colspan="3">Person</th><th colspan="3">Function</th></tr>
<tr><td colspan="2">pers.ind (individual person)</td><td>pers.coll (group of persons)</td><td colspan="2">func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><th colspan="3">Location</th><th colspan="3">Production</th></tr>
<tr><td>administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td>physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td>facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><td>prod.object (manufac-tured object)</td><td>prod.serv (transporta-tion route)</td><td>prod.fin (financial products)</td></tr>
<tr><td colspan="3"></td><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td colspan="3"></td><td>prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><th colspan="3">Organization</th><th colspan="3">Time</th></tr>
<tr><td colspan="2">org.adm (administra-tion)</td><td>org.ent (services)</td><td colspan="2" rowspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td rowspan="2">time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
<tr><th colspan="3">Amount</th></tr>
<tr><td colspan="3">amount (with unit or general object), includ-ing duration</td></tr>
</table>

Level 1: *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* $\rightarrow$ 7 possibilities (degree of freedom = 6).
Level 2: *prod.object*, *prod.serv*, *prod.fin*, *prod.soft*, *prod.doctr*, *prod.rule*, *prod.art*, *prod.media*, *prod.award* $\rightarrow$ 9 possibilities (degree of freedom = 8).
Level 3: *loc.adm.town*, *loc.adm.reg*, *loc.adm.nat*, *loc.adm.sup* $\rightarrow$ 4 possibilities (degree of freedom = 3).

# Dimension of the tagset

## Degree of freedom

$$\nu = \nu_1 + \nu_2 + \ldots + \nu_m$$

where $\nu_i$ is the maximal degree of freedom the annotator has when choosing the $i^{th}$ sub-type ($\nu_i = n_i - 1$).

## Dimension of the tagset

$$Dimension(Flow) = min(\frac{\nu}{\tau}, 1)$$

where $\tau$ is the threshold from which we consider the tagset to be very large (experimentally determined).

$$Dimension_{Identification} = 0$$
$$Dimension_{Renaming} = 0.04$$
$$Dimension_{NE_{TypesSubtypes}} = 0.34$$

# Degree of ambiguity: residual ambiguity

Using the traces left by the annotators:

*[...] <EukVirus>3CDproM</EukVirus> can process both structural and nonstructural precursors of the <EukVirus* ***uncertainty-type*** *= "too-generic"><taxon>poliovirus</taxon> polyprotein</EukVirus> [...].*

## Définition

$$AmbiguityRes(Flow) = \frac{|Annotations_{amb}|}{|Annotations|}$$

$$AmbiguityRes_{Identification} = 0.04$$
$$AmbiguityRes_{Renaming} = 0.02$$

# Degree of ambiguity: theoretical ambiguity

Proportion of the units to annotate that corresponds to ambiguous vocables.

## Definition

$$AmbiguityTh(Flow) = \frac{\sum_{voc_i=1}^{|Voc(Flow)|} (Ambig(voc_i) * freq(voc_i, Flow))}{|Units(Flow)|}$$

with

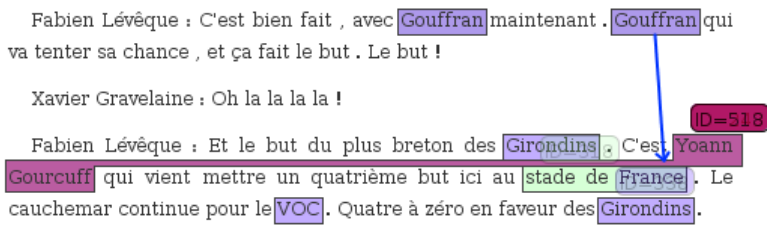
$$Ambig(voc_i) = \begin{cases} 1 & \text{if} \quad |Tags(voc_i)| > 1 \\ 0 & \text{else} \end{cases}$$

$$AmbiguityTh_{Identification} = 0.01$$

$\rightarrow$ Does not apply to renaming relations (2 EATs).

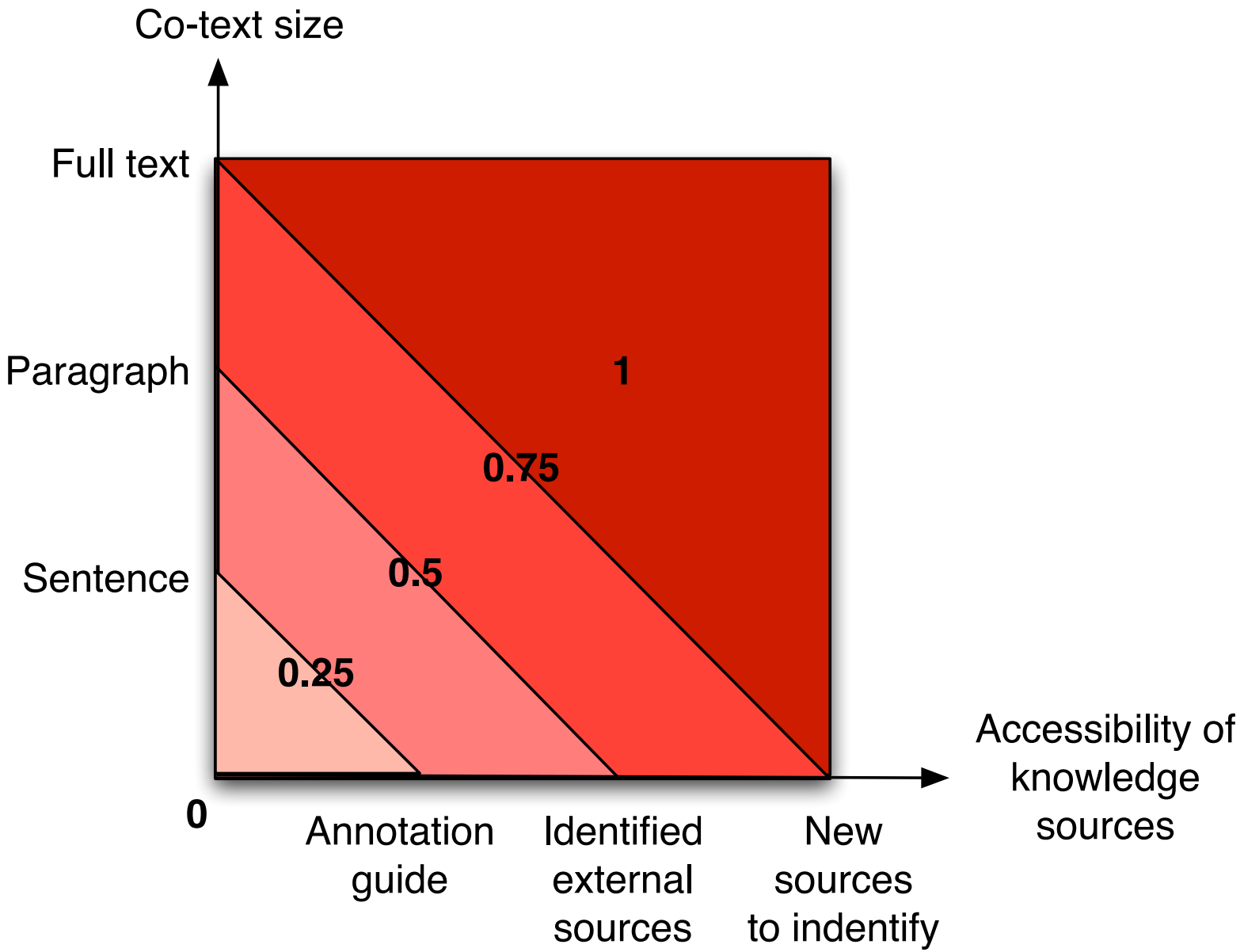
# Context to take into account

- **size of the window** to take into account in the source signal:
  - The sentence:
    *I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.*
  - ... or more:

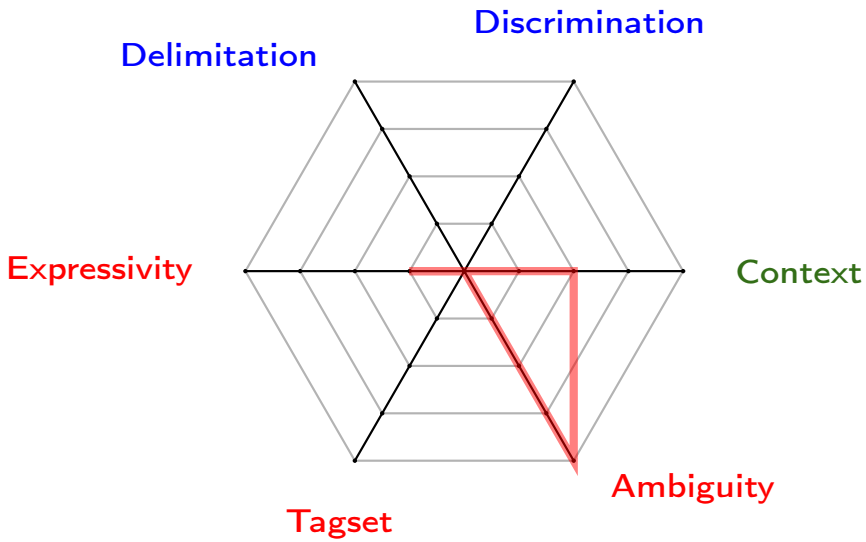



- number of **knowledge elements** to be rallied or degree of accessibility of the knowledge sources that are consulted:
  - annotation guidelines
  - nomenclatures (Swiss-Prot)
  - new sources to be found (Wikipedia, etc.)

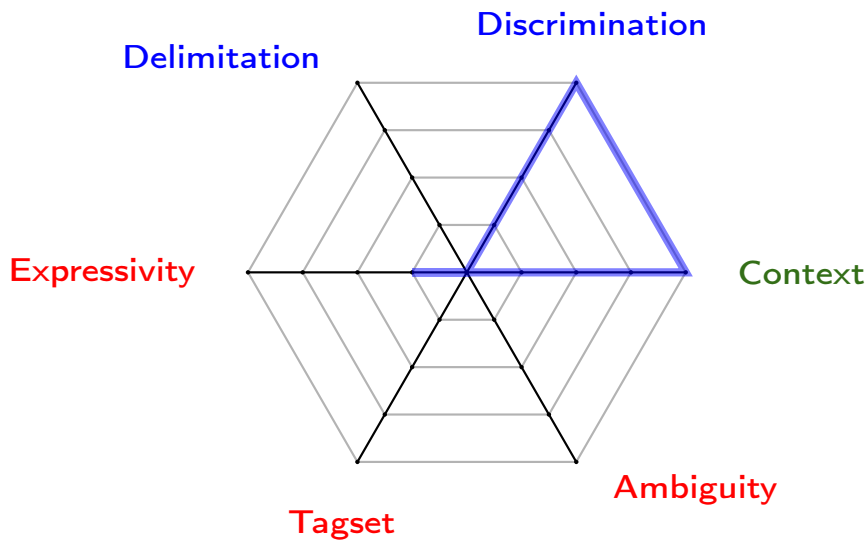
# Weight of the context



# Synthesis of the complexity dimensions

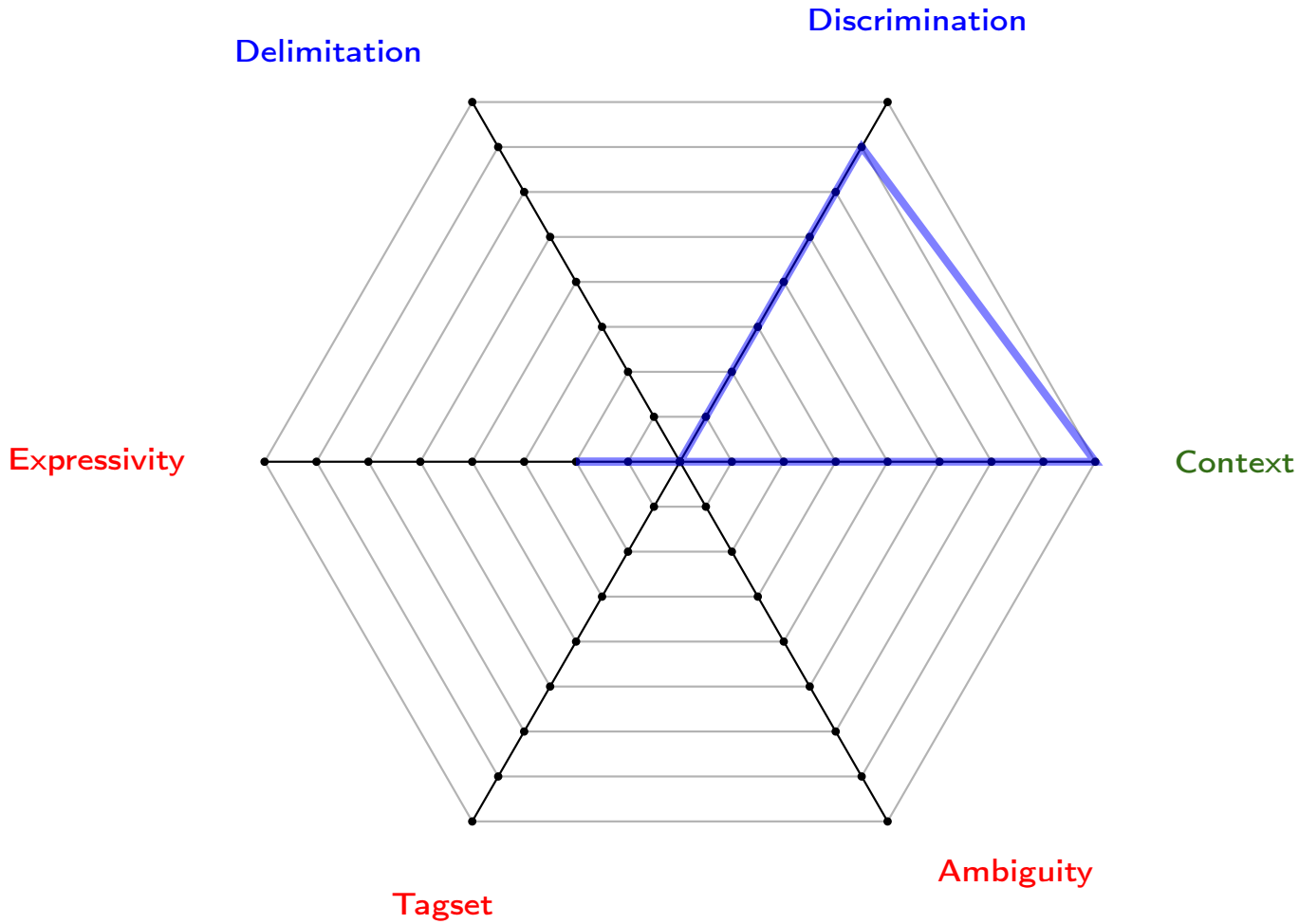



Classification of *it* pronouns as anaphoric or impersonal




Gene names identification

# Synthesis




Gene renaming campaign (2 EATs)

1 Introduction

2 Analysing the complexity of an annotation campaign

3 What to annotate?

4 How to annotate?

5 Synthesis

6 Conclusion and prospects

# Conclusion and prospects

A grid of analysis:

- to use during preparatory work
- to help asking the right questions and finding the appropriate solutions

→ that should be computed more or less automatically

→ that should be integrated as part of annotation tools
[Kaplan et al., 2010, Bontcheva et al., 2010]

Thank you for your attention!


Abeillé, A., Clément, L., and Toussenel, F. (2003).
Building a treebank for French.
In Abeillé, A., editor, *Treebanks*, pages 165 –187. Kluwer, Dordrecht.


Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2001).
The prague dependency treebank: Three-level annotation scenario.
In Abeillé, A., editor, *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.


Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., and Mostefa, D. (2005).
Semantic annotation of the French Media dialog corpus.
In *Proceedings of the InterSpeech*, Lisbonne, Portugal.


Bontcheva, K., Cunningham, H., Roberts, I., and Tablan, V. (2010).
Web-based collaborative corpus annotation: Requirements and a framework implementation.
In Witte, R., Cunningham, H., Patrick, J., Beisswanger, E., Buyko, E., Hahn, U., Verspoor, K., and Coden, A. R., editors, *Proceedings of the*

*workshop on New Challenges for NLP Frameworks (NLPFrameworks 2010)*, La Valette, Malte. ELRA.


Carmen, M., Felt, P., Haertel, R., Lonsdale, D., McClanahan, P., Merkling, O., Ringger, E., and Seppi, K. (2010).
Tag dictionaries accelerate manual annotation.
In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, La Valette, Malte. European Language Resources Association (ELRA).


Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., and Poesio, M. (2013).
Using games to create language resources: Successes and limitations of the approach.
In Gurevych, I. and Kim, J., editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg.


Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1995).
Active learning with statistical models.

In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7, pages 705–712. The MIT Press.

Engelson, S. P. and Dagan, I. (1996).
Minimizing manual annotation cost in supervised training from corpora.
In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 319–326, Morristown, NJ, USA. Association for Computational Linguistics.

Fort, K., Adda, G., and Cohen, K. B. (2011).
Amazon Mechanical Turk: Gold mine or coal mine?
*Computational Linguistics (editorial)*, 37(2):413–420.

Fort, K., François, C., Galibert, O., and Ghribi, M. (2012).
Analyzing the impact of prevalence on the evaluation of a manual annotation campaign.
In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turquie.

7 pages.


Fort, K. and Sagot, B. (2010).
Influence of pre-annotation on POS-tagged corpus development.
In *Proceedings of the Fourth ACL Linguistic Annotation Workshop*, pages 56–63, Uppsala, Suède.


Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011).
Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview.
In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA.
Poster.


Ide, N. and Romary, L. (2006).
Representing linguistic corpora and their annotations.
In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Gène, Italie.


Kaplan, D., Iida, R., and Tokunaga, T. (2010).
Annotation process management revisited.
In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 365 – 366, La Valette, Malte.


Kim, J.-D., Ohta, T., and Tsujii, J. (2008).
Corpus annotation for mining biomedical events from literature.
*BMC Bioinformatics*, 9(1):10.


Krippendorff, K. (2004).
*Content Analysis: An Introduction to Its Methodology,*.
Sage, Thousand Oaks, CA., USA, second edition edition.


Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993).
Building a large annotated corpus of English : The Penn Treebank.
*Computational Linguistics*, 19(2):313–330.


Tomanek, K., Hahn, U., Lohmann, S., and Ziegler, J. (2010).
A cognitive cost model of annotations based on eye-tracking data.

In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, ACL'10, pages 1158–1167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Voormann, H. and Gut, U. (2008).
Agile corpus creation.
*Corpus Linguistics and Linguistic Theory*, 4(2):235–251.