Consortium Corpus Écrits

GT 9 *Annotation de surface*

Benoît Sagot – Alpage
INRIA & Université Paris–Diderot

Réunion plénière du consortium Corpus Écrits
23 novembre 2012

CORPUS *Infrastructure de Recherche*

ILF
INSTITUT
DE LINGUISTIQUE
FRANÇAISE

Groupe de travail n°9

- **Annotation de surface**
  - Segmentation en unités lexicales
  - Annotation morphosytaxique
  - Analyse morphologique
  - Lemmatisation
  - Annotation en entités nommées

Groupe de travail n°9

- **Objectif : mettre en contact**
  - Les chercheurs qui travaillent sur les concepts, les méthodes et les outils dédiés à l'annotation de surface
  - Les chercheurs qui font usage de telles annotations pour l'exploitation et l'exploration de corpus

En pratique

- Une liste de diffusion
  corpus-ecrits-annotation-surface@groupes.renater.fr
- Un wiki associé
  https://groupes.renater.fr/wiki/corpus-ecrits-annotation-surface/
- 20 membres représentant une quinzaine de laboratoires
  Antonio Balvet, Delphine Bernhard, Basilio Calderone, Karën Fort, Antoine Gautier, Joaquín Giráldez, Natalia Grabar, Cyril Grouin, Linda Hriba, Evelyne Jacquey, Laurence Longo, Claude Martineau, Cristian Martinez, Denis Maurel, Ghassan Mourad, Kamal Naït-Zerrad, Fiammetta Namer, Céline Poudat, Matthias Tauveron
- Contact direct : benoit.sagot@inria.fr

Réunion présentielle
6 juillet 2012

- Prise de contact
- Tour d'horizon des problématiques liées à l'annotation de surface
  - segmentation en unités lexicales et étiquetage morphosyntaxique (B. Sagot)
  - analyse morphologique (D. Bernhard)
  - entités nommées (D. Maurel)
  - annotation manuelle de corpus (K. Fort)

Segmentation en unités lexicales et étiquetage morphosyntaxique

- Distinction entre *token* (unité typographique) et *forme* (unité lexicale, syntaxiquement atomique)
- Cas de divergence entre tokens et formes (composés, amalgames, entités nommées)
- Étiquetage morphosyntaxique
- Lexiques morphologiques

Segmentation en unités lexicales et étiquetage morphosyntaxique

- Quelques lexiques morphologiques libres pour le français
  - DELA (U. Marne-la-Vallée)
  - MULTEXT
  - Morphalou (Atilf)
  - Le*fff* (Alpage)

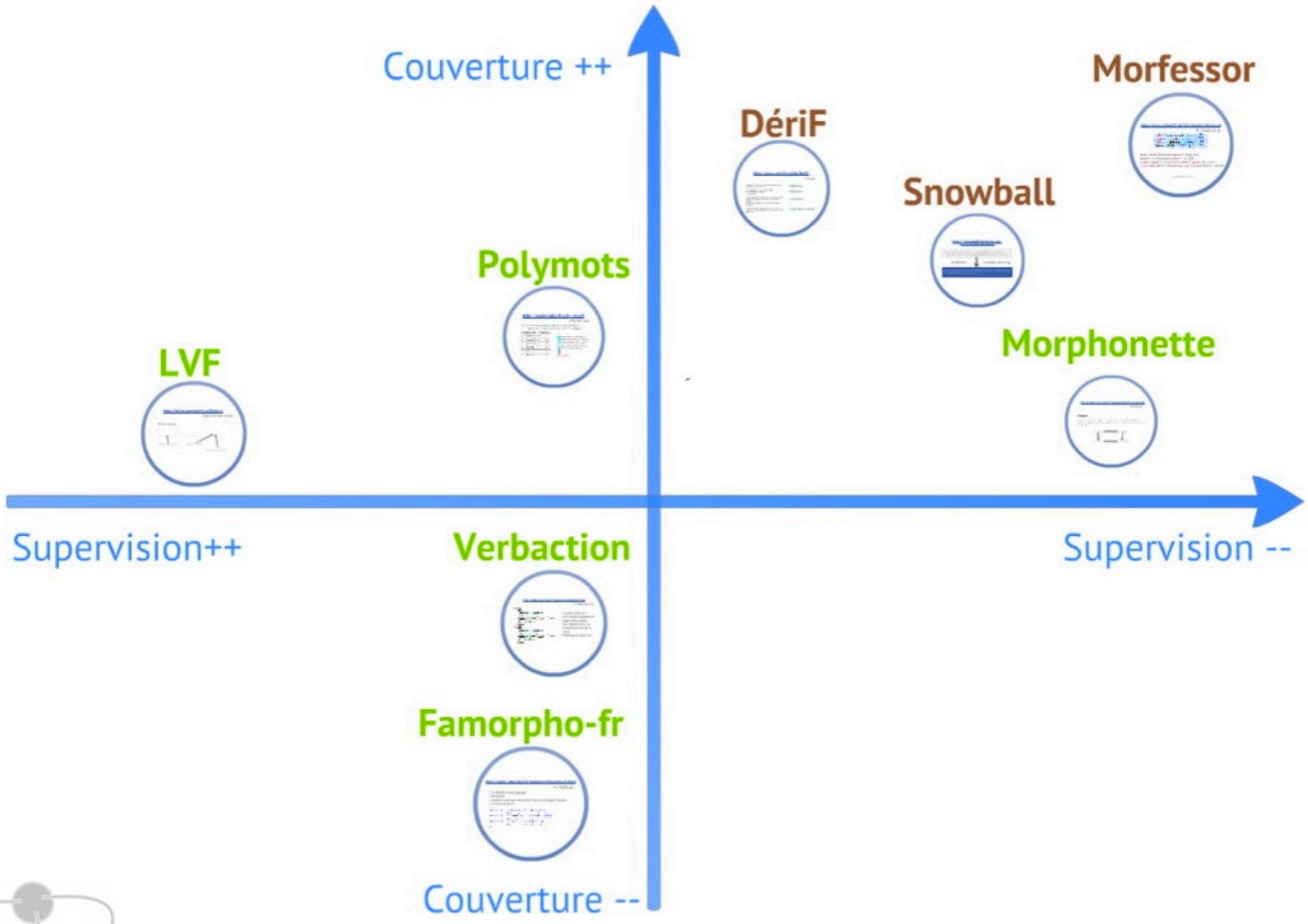
Segmentation en unités lexicales et étiquetage morphosyntaxique

- Quelques outils libres
  - segmentation et autres pré-traitements : UNITEXT (U. Marne-la-Vallée), SxPipe (Alpage)…
  - étiqueteurs : TreeTagger (*moindre qualité*), Morfette (Saarbrücken), MElt (Alpage)…
- Importance du packaging, de la facilité d'installation et d'utilisation

Analyse morphologique

- Étudier la création lexicale
  - analyse linguistique
  - traitement automatique des « mots inconnus »
  - *lexique dynamique*

Analyse morphologique



Entités nommées

- Campagnes d'évaluation et guides associés (MUC, en France ESTER, ESTER2, ETAPE)
- Typologie de base
  - Personnes, lieux, organisations
    Dates, heures
    Pourcentages, valeurs monétaires
- Typologies plus riches, plus complexes aussi, pour les humains comme pour les systèmes automatiques
  - Annotations enchâssées, quantités d'objets (et non plus seulement d'unités de mesure), quelques événements…

Annotation manuelle de corpus

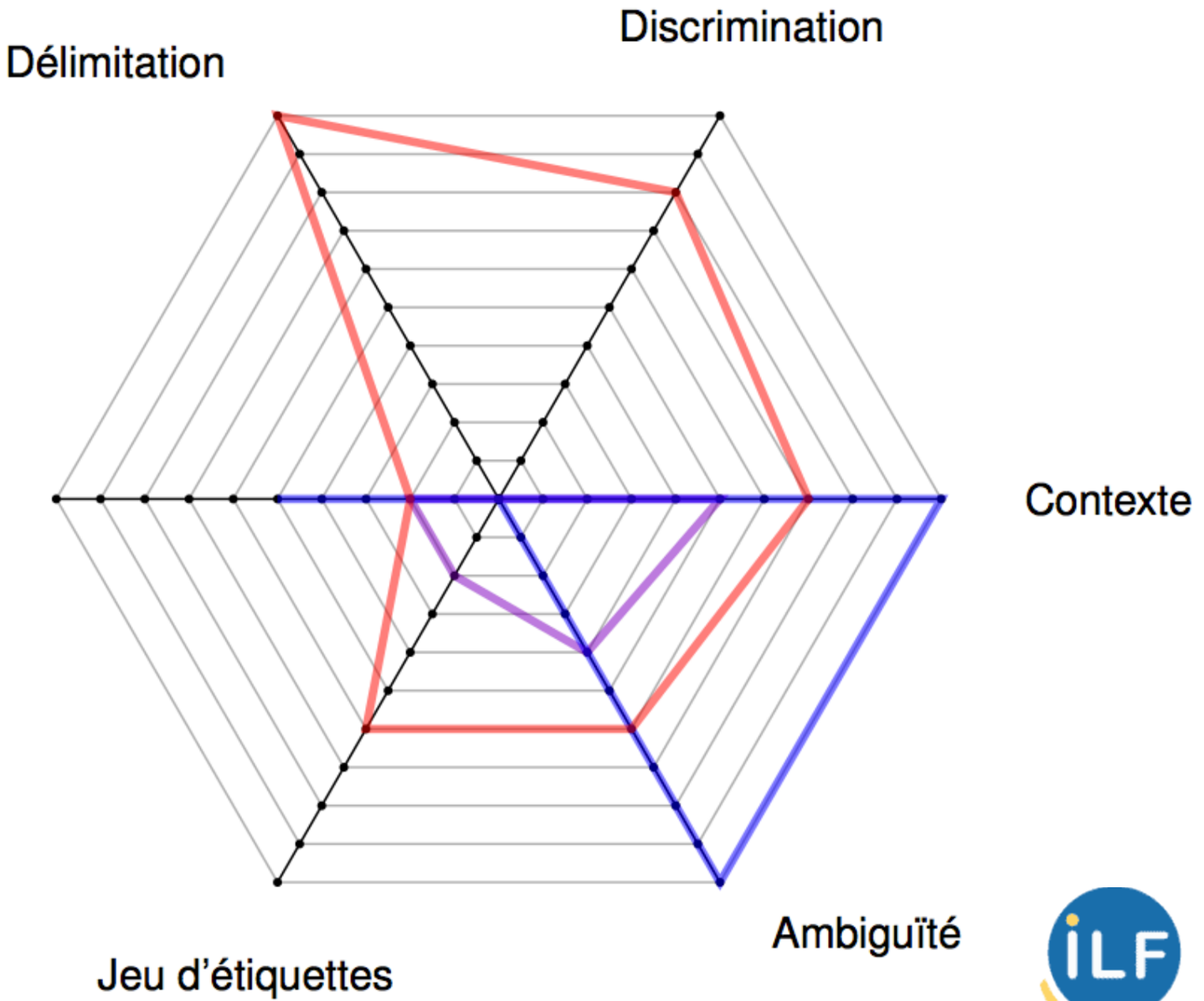
- De nombreux outils dépendent aujourd'hui, pour leur développement comme pour leur évaluation, de corpus annotés manuellement
- L'annotation manuelle est coûteuse
- Elle pose de nombreuses questions

Annotation manuelle de corpus

Penn TreeBank

Renommage de gènes

Entités nommées étendues



CORPUS *Infrastructure de Recherche*

Consortium Corpus Écrits

INSTITUT DE LINGUISTIQUE FRANÇAISE

Annotation manuelle de corpus

- Ces dimensions de complexité sont importantes à identifier en amont
- Impact sur les caractéristiques des outils de pré-annotation à utiliser (ou non) pour accélérer l'annotation manuelle
  - Favoriser le rappel sur la précision?
  - Utiliser des ressources externes

Projets pour 2013

- Mieux comprendre les **besoins des utilisateurs** de telles ressources
  - En particulier, la définition de certaines notions peut être guidée par les utilisateurs (p.ex., quelle unité élémentaire? quelles entités nommées?…)
- **Faciliter l'utilisation** d'outils développés dans un contexte TAL
  - information, formations, travail sur certains outils (p.ex., MElt)

Projets pour 2013

- Recenser les **corpus annotés** qui rentrent dans le cadre des activités du groupe de travail
- Identifier les **ressources lexicales**, leurs limites, leur adéquation ou inadéquation par rapport à certains types d'annotations ou certaines utilisations aval
    - Aspects liés à la néologie et à l'analyse morphologique
- Identifier les besoins et les outils/ressources concernant des **langues autres que le français**

Projets pour 2013

- **Interagir avec les autres groupes de travail concernés**
  - Nouveaux moyens de communication (GT7)
  - Annotations de plus haut niveau (GT8)
  - Exploration de corpus (GT10)

Merci