L’intervention humaine dans le traitement automatique des langues

Caroline Atallah
Maud Colléter

26 janvier 2010

Plan

- Introduction
- Définition des tâches d'après l'observation de différentes études
  - G. Rehm, M. Santini & al.
  - RST TreeBank
  - TREC SQR 2007
  - Penn TreeBank
- Bilan
  - Simplification des schémas
  - Schéma de synthèse
  - Problèmes posés par l'intervention humaine
- Conclusion
- Bibliographie

Introduction

- Intervention humaine :
  - Lors d'une étude impliquant un traitement automatique, un ou plusieurs locuteurs sont amenés à traiter manuellement un ensemble de données.

- Pourquoi ?
  - Effectuer automatiquement une tâche dont le résultat serait similaire à celui obtenu par un humain

Vocabulaire couramment employé

- Pour les tâches : annotation, tag, catégorisation, segmentation, validation, codage, étiquetage, correction, évaluation, …

- Pour les locuteurs : annotateur, codeur, correcteur, juge, naïf / expert, …

Deux constatations

- Une terminologie floue :
  - pas de définition unique
  - pas de correspondance stricte entre tâche et humain
- Une démarche en partie implicite

Objectif

Décrire ces tâches en explicitant l'ensemble des étapes qu'elles nécessitent

Plan

- Introduction
- Définition des tâches d'après l'observation de différentes études
  - G. Rehm, M. Santini & al.
  - RST TreeBank
  - TREC SQR 2007
  - Penn TreeBank
- Bilan
  - Simplification des schémas
  - Schéma de synthèse
  - Problèmes posés par l'intervention humaine
- Conclusion
- Bibliographie

Définition des tâches d'après l'observation de différentes études

- Observation des travaux présentés lors de l'U.E. TAL
- Recherche d'articles
- Synthèse

G. Rehm, M. Santini & al.

- *Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems* (2008)
  - 9 chercheurs intéressés par les genres du Web
  - Objectif : construction d'un corpus de référence
    - liste harmonisée des genres possibles
    - tester cette liste
    - constituer un corpus représentatif des genres

G. Rehm, M. Santini & al.

- Précédentes études :
  - Naïfs ont pour tâche d'assigner des genres à des pages Web
  - « *Other studies have shown that user-based genre labeling usually exhibits a certain kind of fragmentation and a low, at most moderate, inter-rater agreement.* »
  - Problèmes :
    - Taux d'accord faibles
    - Pas de comparaison possible entre les résultats des différentes études

G. Rehm, M. Santini & al.

- Constatations :
  - Notion peu consistante
  - Tâche trop complexe pour des naïfs
- Même tâche avec des experts
  - 7 des 9 auteurs participent
  - 50 pages Web sélectionnées au hasard
  - Pas de liste de genre prédéfinie
  - Possibilité d'assigner plusieurs genres à une même page

G. Rehm, M. Santini & al.

- Attentes :
  - Voir si le taux d'accord augmente avec des experts
  - Homogénéité des genres assignés (« *labels assigned […] make some kind of sense* »)
  - Moitié des catégories *synonymes* ou très similaires
  - Au minimum rattachement possible à des concepts de base (*basic concept*)

G. Rehm, M. Santini & al.

- Constat d'échec :

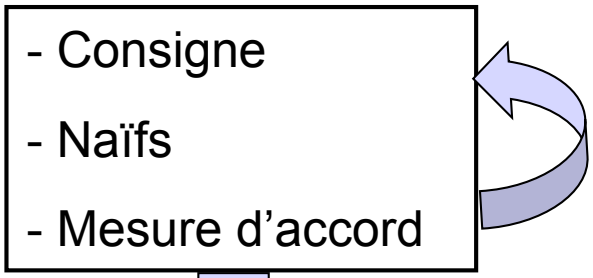
  « *What is needed to arrive at a consistent set of genre labels are annotation guidelines that provide, in a detailed, tranparent, and unambiguous way, a set of ground rules that explain the task of assigning genre labels to web documents.* »

G. Rehm, M. Santini & al.

- Travaux en cours et à venir
  - Discussion pour établir la liste des genres et leurs définitions
  - Exploitation des corpus déjà existants pour tester cette liste
  - Reprise de la discussion jusqu'à l'obtention d'une liste satisfaisante

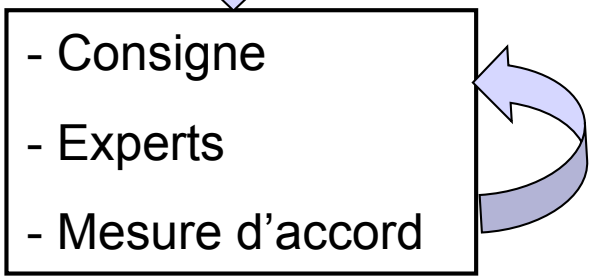
  - Rédaction d'un guide
  - Campagne d'annotation par des utilisateurs du web

G. Rehm, M. Santini & al. : Synthèse



Discussion

Objectif : tester la consistance de la notion



Discussion

Objectif : rendre la notion consistante

RST TreeBank

- *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory.* L. Carlson, D. Marcu, M.E. Okurowski (2001)
  - Un cadre théorique : la RST (Mann & Thompson 1988)
  - Objectif : construire un corpus de référence pouvant servir à l'élaboration de traitements automatiques

RST TreeBank

- Deux tâches :
  - segmentation du texte en unités
  - liaison des unités entre elles à l'aide d'une liste prédéfinie de relations rhétoriques

- Construction d'un guide

  « Because the goal of this effort was to build a high-quality, consistently annotated reference corpus, the task required that we […] specify a rigorous set of annotation guidelines. »

RST TreeBank

- Présentation orale de la RST et de la tâche à accomplir aux participants
- Première exploration :
  - Tâches effectuées sur un corpus restreint
  - Confrontation des résultats obtenus par les différents participants
  - Identification des difficultés
  - Première version du guide

RST TreeBank

- Répétition de la tâche sur un corpus plus large
- Calcul du taux d'accord
- Nouvelle discussion en vue d'une amélioration du guide

→ Ainsi de suite… jusqu'à l'obtention d'un taux d'accord jugé satisfaisant

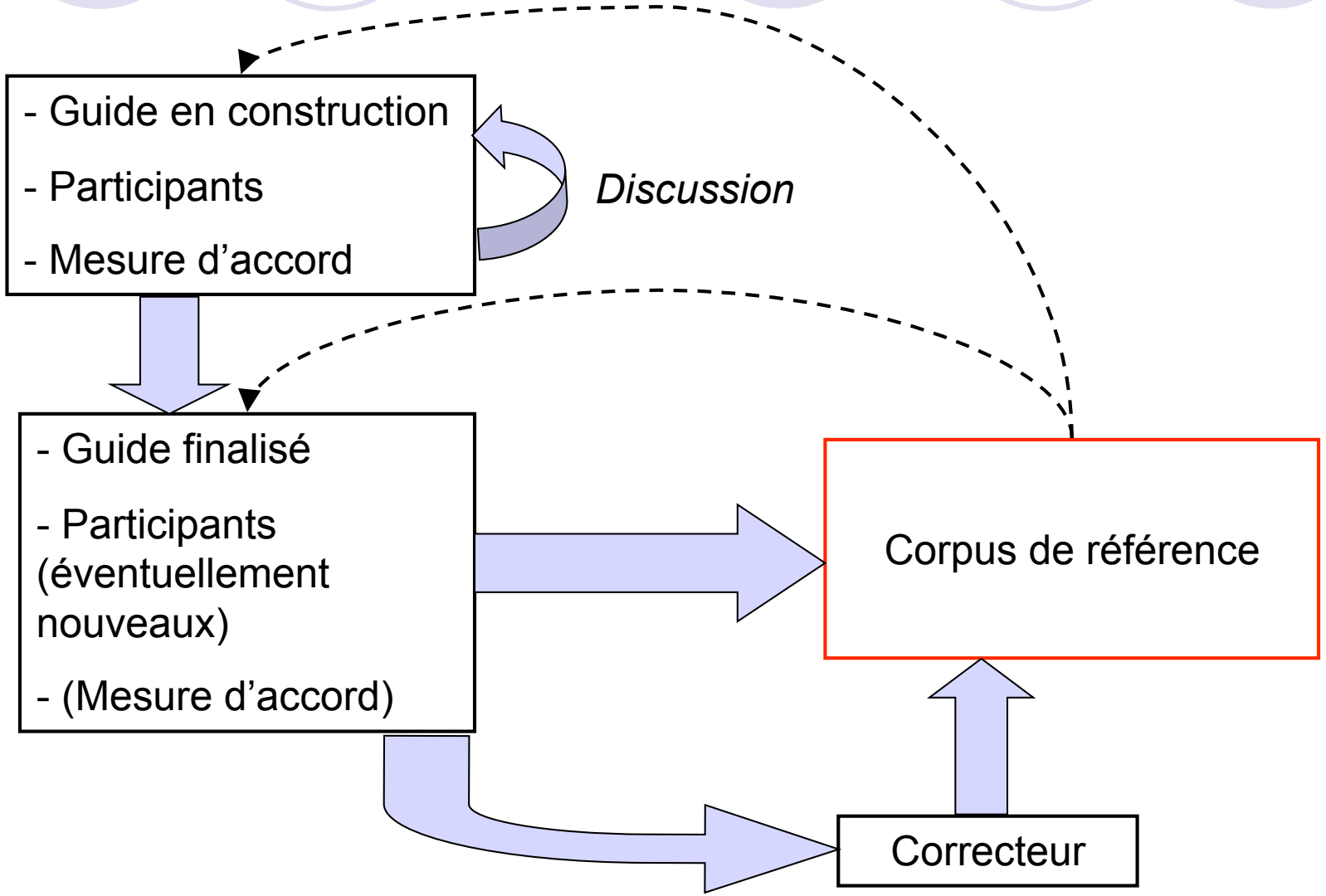
RST TreeBank

- Constatation : tâche de segmentation trop complexe
  - *« At this point, we decided to proceed by pre-segmenting all of the texts on hard-copy, to ensure a higher overall quality to the final corpus. »*
- Conséquences :
  - Segmentation de tous les textes par deux personnes
  - Désaccords résolus par le rédacteur du guide

RST TreeBank

- Un corpus segmenté et un guide finalisé
- Campagne d'annotation :
  - une douzaine de participants
  - une double annotation permettant de juger de la qualité du corpus final
- Un corpus de référence

RST TreeBank : Synthèse



TREC SQR 2007

- *TREC 2007 Question Answering Track Guidelines*
  - Un corpus et des séries de questions susceptibles d'être posées par un utilisateur de système de question-réponse
  - Objectif : permettre à plusieurs candidats de développer leur système en suivant les consignes fournies par NIST
    - Evaluation de tous les systèmes sur les mêmes bases
    - Discussion entre participants et organisateurs

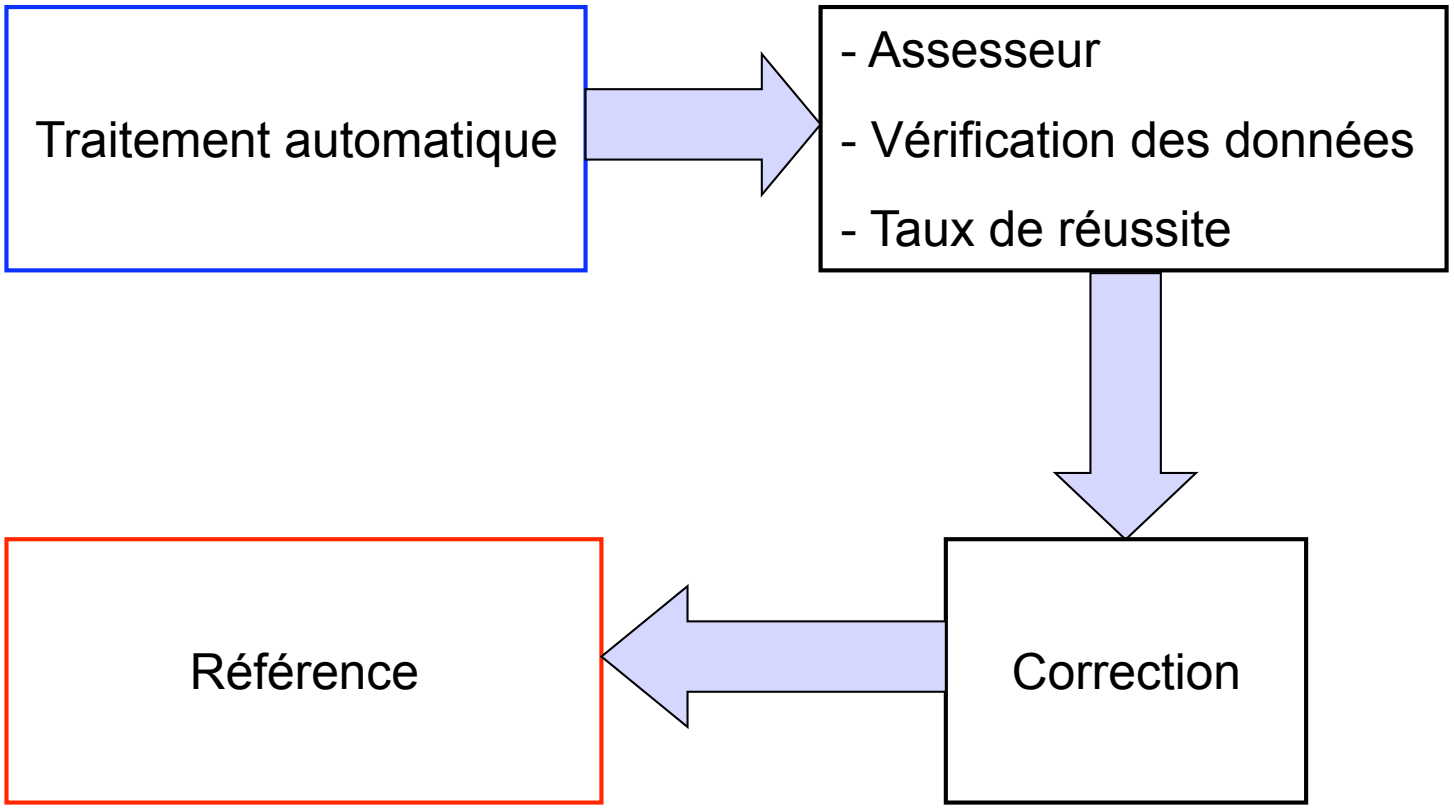
TREC SQR 2007

- Chaque candidat soumet les questions à son système.
- Résultats fournis aux organisateurs
- Un « assesseur »
  - Note la qualité des réponses pour chaque question (5 degrés)
  - Calcule un score (nombre de réponses exactes sur le nombre de questions)
  - Retour personnalisé vers les candidats

TREC SQR 2007

- Mise en commun des données récoltées
  - Obtention des réponses correctes à chaque question
  - Création de la liste des documents pertinents pour chaque question
- Mise en ligne du corpus, des questions et de la liste des documents pertinents pour chaque question
- L'ensemble de ces données peut servir de référence pour évaluer d'autres SQR.

TREC SQR 2007 : Synthèse



Penn TreeBank

- *Building a large annotated corpus of English : the Penn TreeBank*. M. Marcus, B. Santorini, M.A. Marcinkiewicz (1993).
  - Construire un grand corpus annoté de l'anglais-américain
    - annotation des parties du discours
    - annotation de la structure syntaxique
  - Objectif : Améliorer les performances des annotateurs en termes de vitesse, cohérence et exactitude

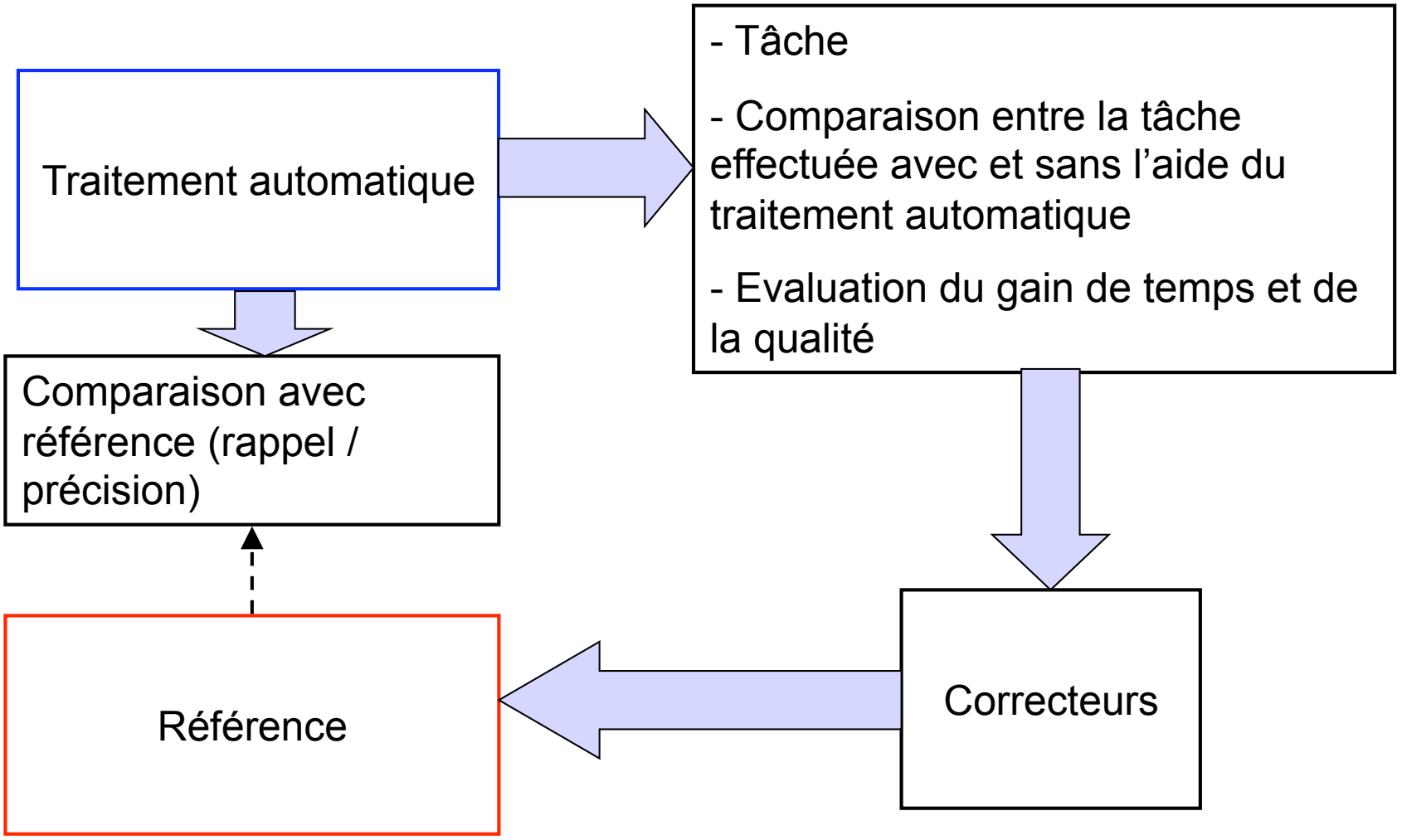
Penn TreeBank

- Deux modes d'annotation des parties du discours :
  - « Tagging » : annotation entièrement manuelle
  - « Correcting » : vérification et correction d'une annotation faite automatiquement
- Constatations
  - Un meilleur accord inter-annotateurs et un gain de temps important pour la tâche de « correcting »
- Conclusion : Le traitement automatique est validé par la tâche *(validation extrinsèque)*.

Penn TreeBank

- Remarques
  - Tagger conçu initialement pour les étiquettes du Brown Corpus
  - Traitement automatique en deux temps :
    - « Tagging »
    - « Mapping »
  - Développement de nouveaux taggers (TreeTagger) se basant sur le Penn TreeBank
    - Penn Treebank utilise ces nouveaux taggers et supprime ainsi les erreurs dues au « mapping »

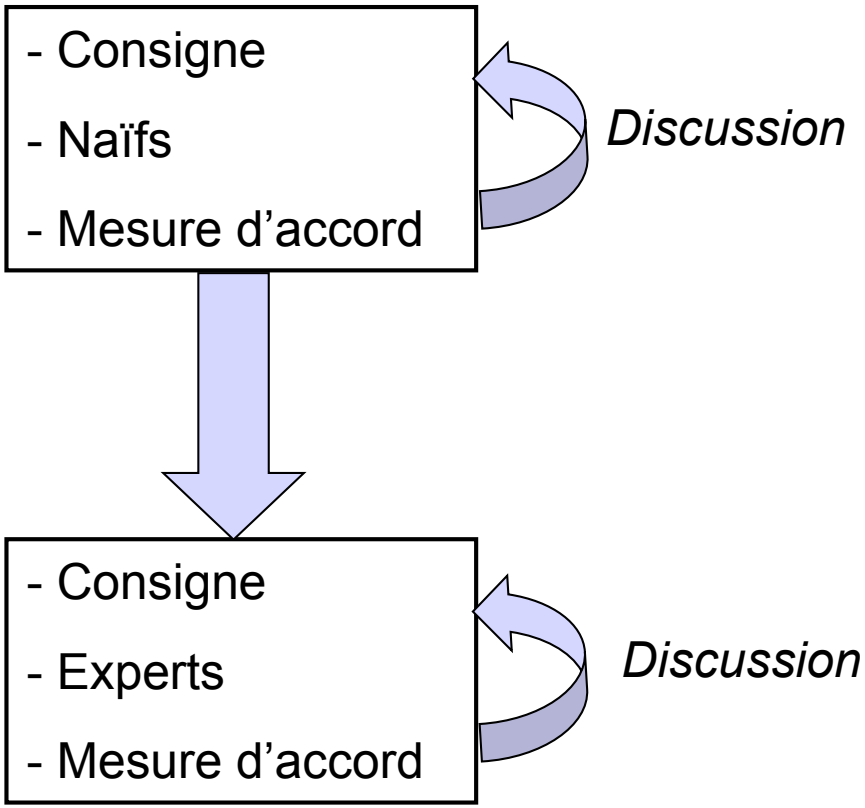
Penn TreeBank : Synthèse



Plan

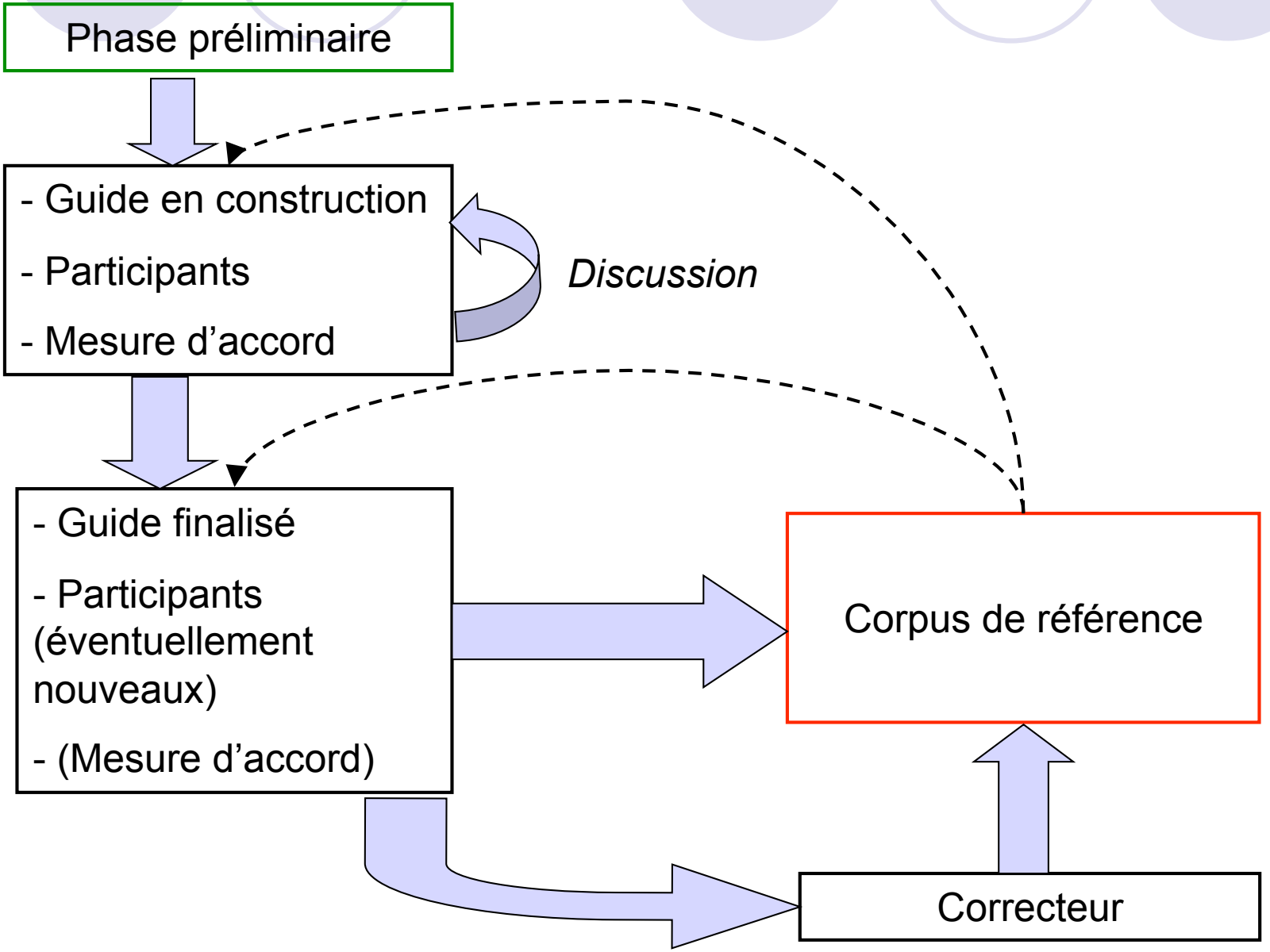
- Introduction
- Définition des tâches d'après l'observation de différentes études
  - G. Rehm, M. Santini & al.
  - RST TreeBank
  - TREC SQR 2007
  - Penn TreeBank
- Bilan
  - Simplification des schémas
  - Schéma de synthèse
  - Problèmes posés par l'intervention humaine
- Conclusion
- Bibliographie

Annotation

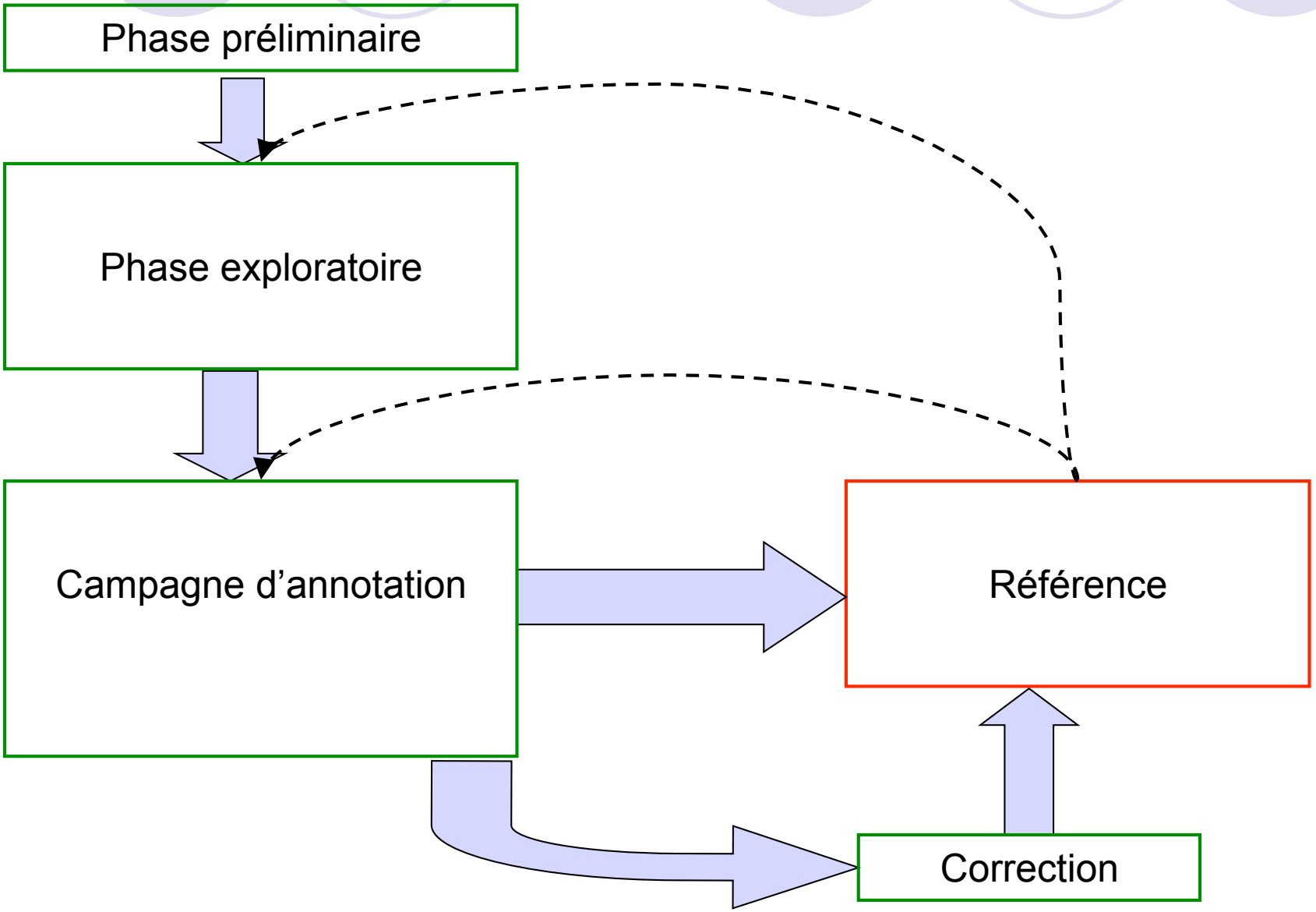


Phase préliminaire

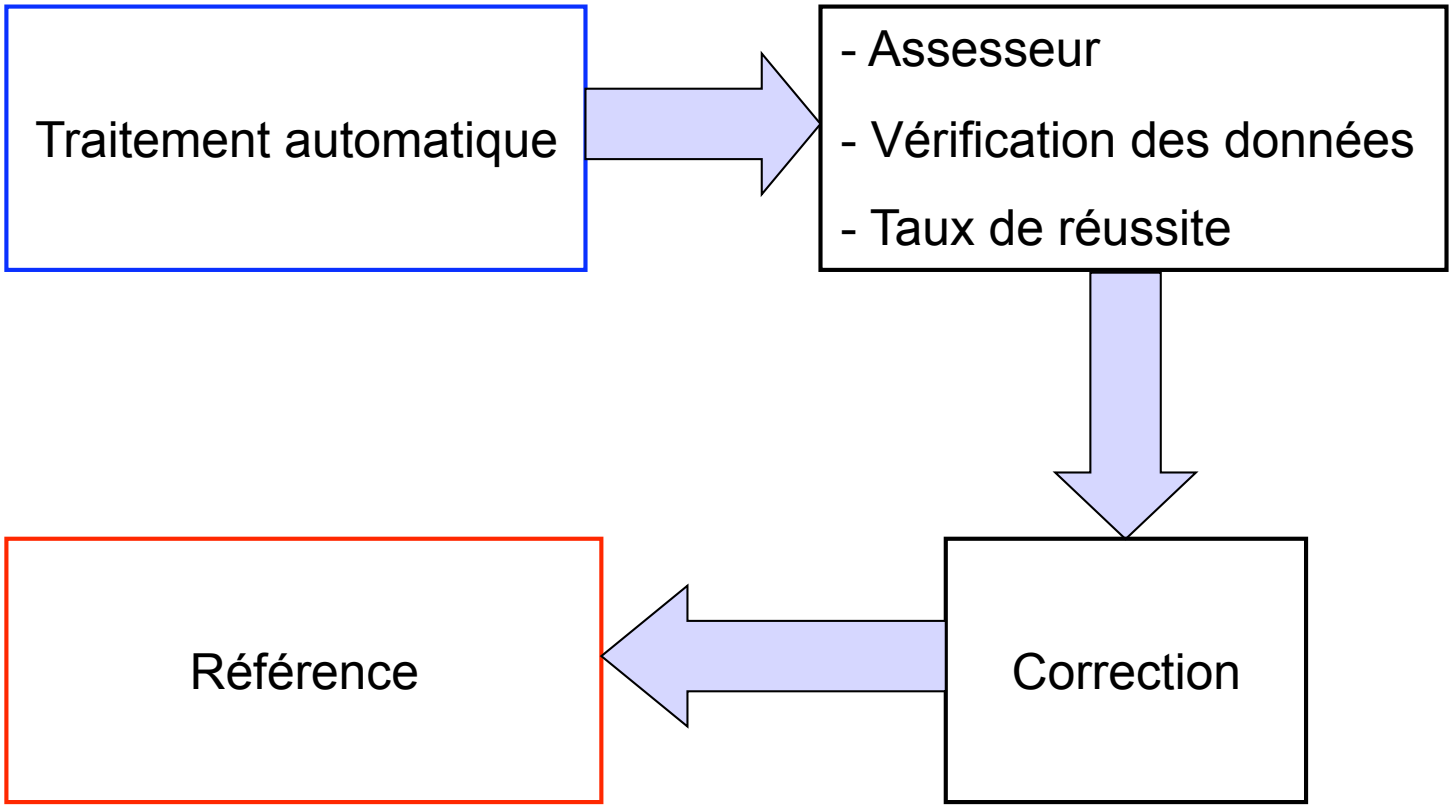
Annotation


Phase préliminaire
- Guide en construction
- Participants
- Mesure d'accord
Discussion
- Guide finalisé
- Participants (éventuellement nouveaux)
- (Mesure d'accord)
Corpus de référence
Correcteur

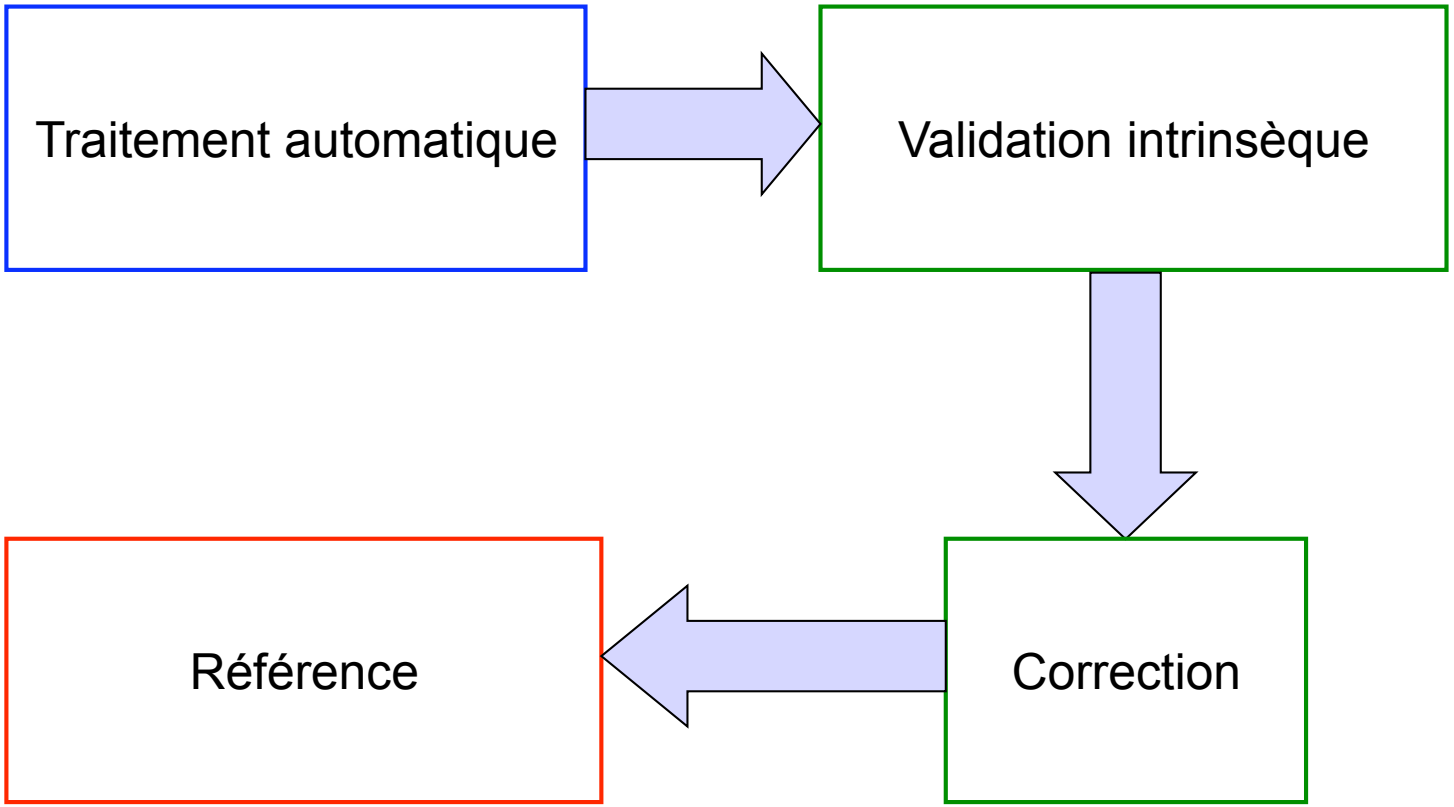
Annotation


Phase préliminaire
Phase exploratoire
Campagne d'annotation
Référence
Correction

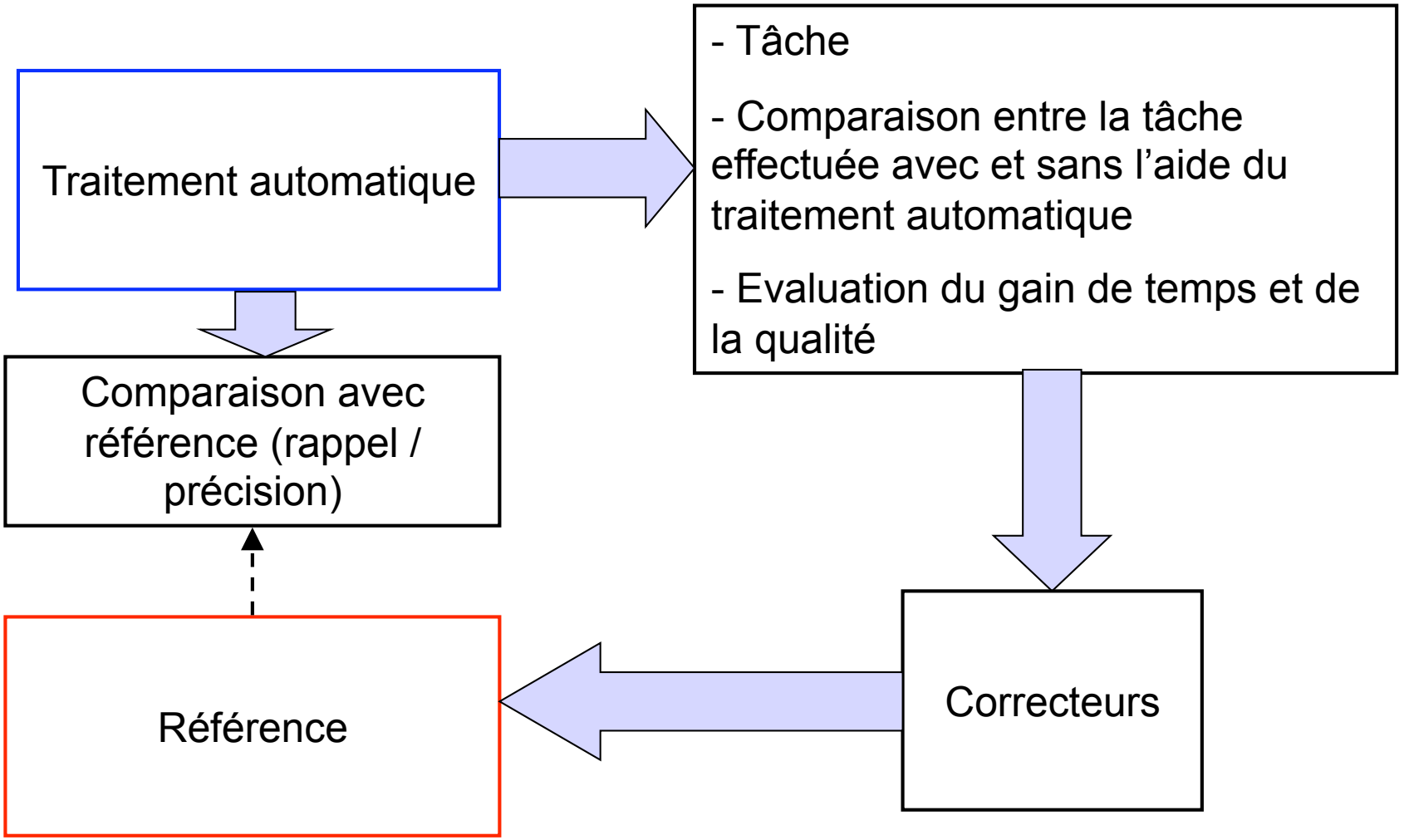
Validation


Traitement automatique
- Assesseur
- Vérification des données
- Taux de réussite
Correction
Référence

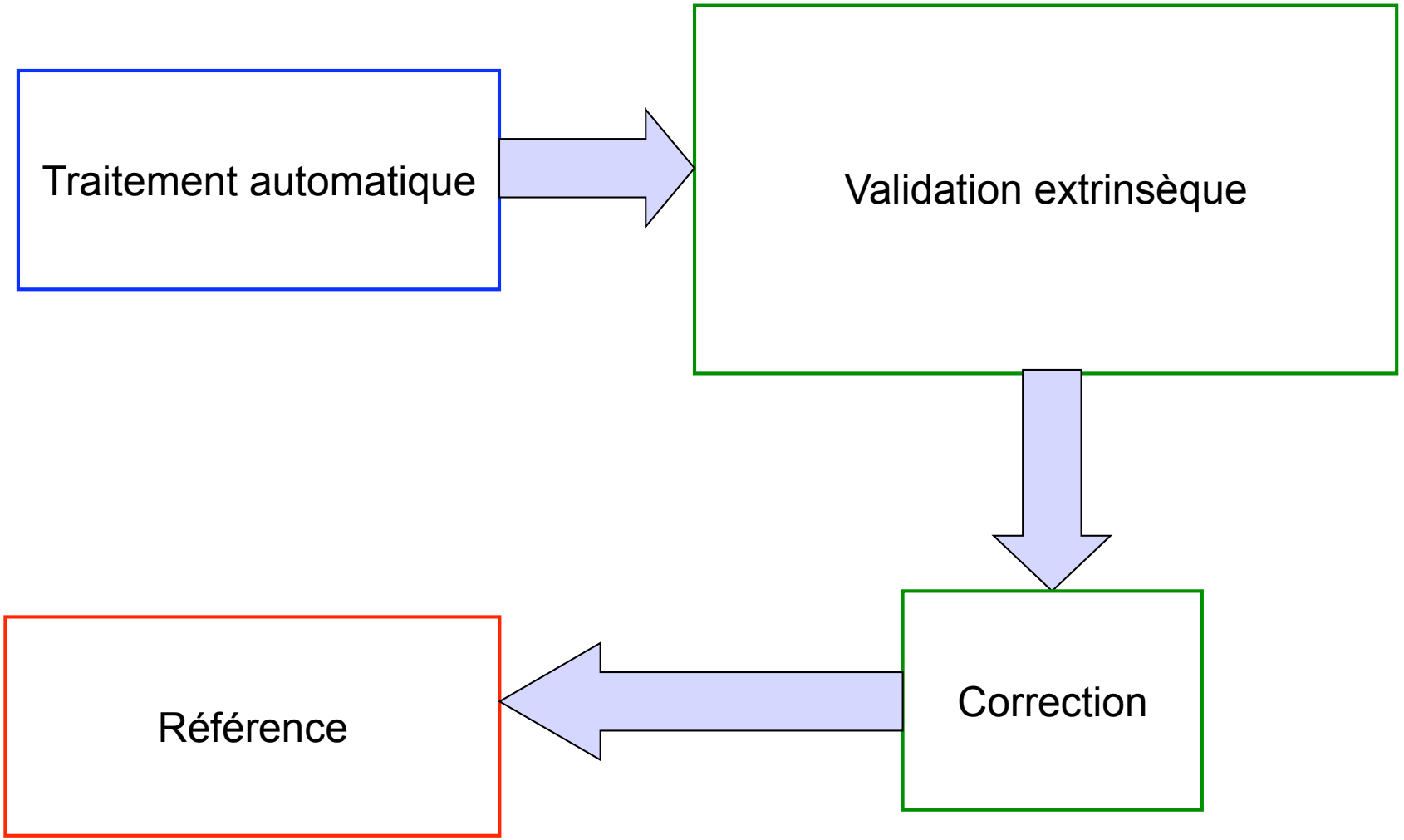
Validation

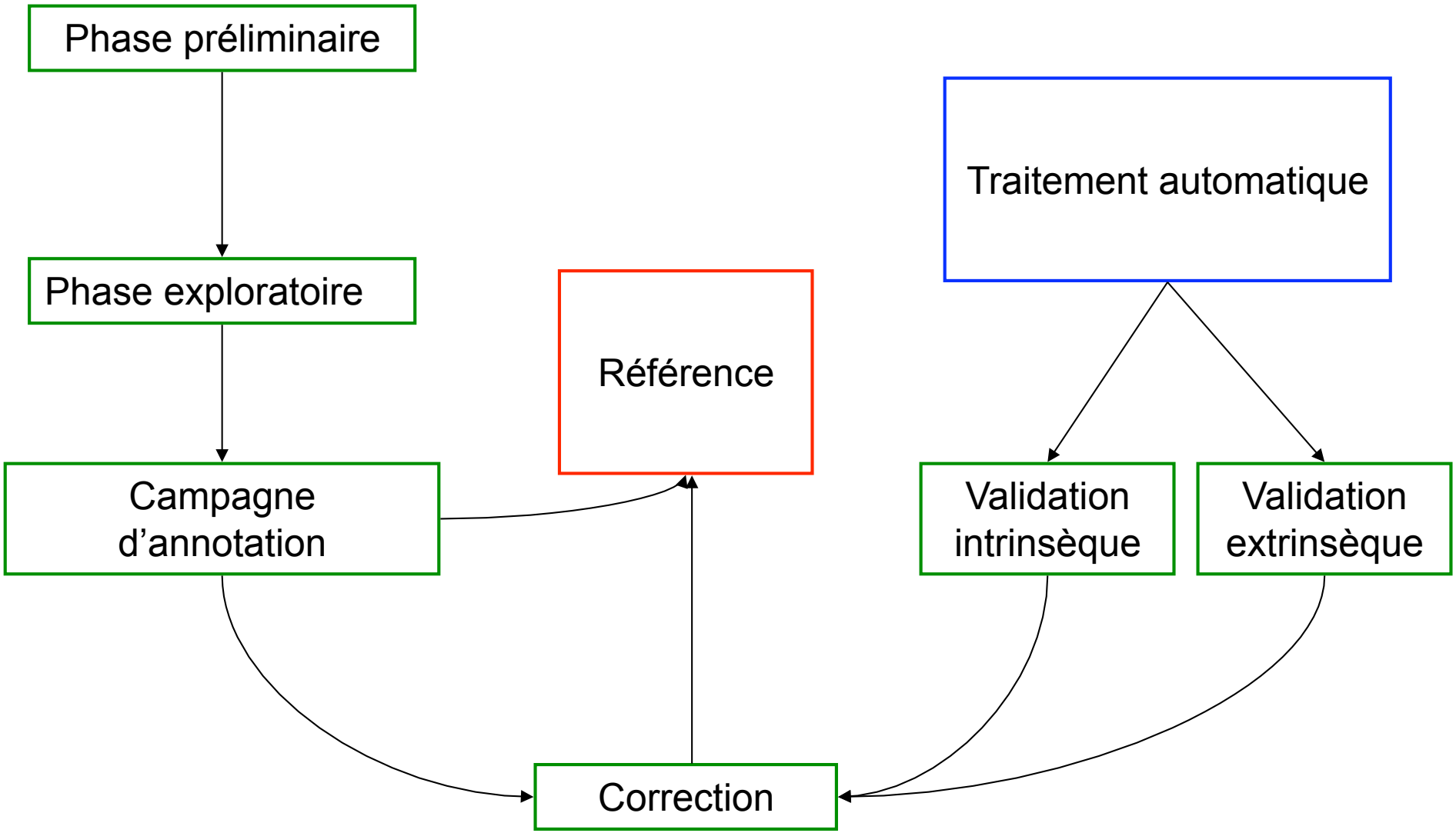

Traitement automatique
Validation intrinsèque
Correction
Référence

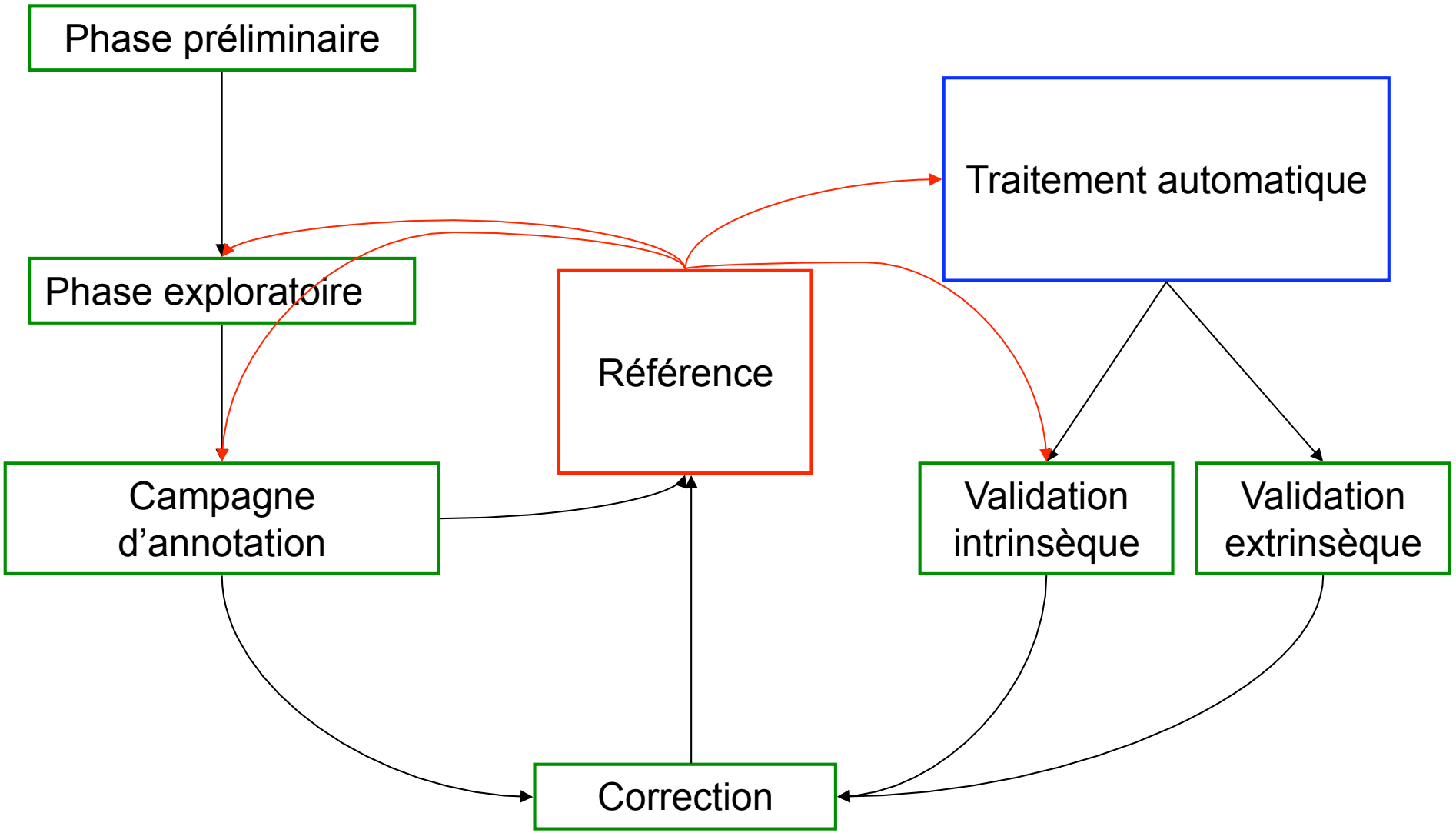
Validation


Traitement automatique
- Tâche
- Comparaison entre la tâche effectuée avec et sans l'aide du traitement automatique
- Evaluation du gain de temps et de la qualité
Comparaison avec référence (rappel / précision)
Référence
Correcteurs

Validation


Traitement automatique
Validation extrinsèque
Correction
Référence


Phase préliminaire
Phase exploratoire
Campagne d'annotation
Référence
Traitement automatique
Validation intrinsèque
Validation extrinsèque
Correction


Phase préliminaire
Phase exploratoire
Campagne d'annotation
Référence
Traitement automatique
Validation intrinsèque
Validation extrinsèque
Correction

Problèmes posés par l'intervention humaine

- Disposer de temps, d'un budget suffisant, d'un certain nombre de participants

- Faire des compromis
  - Sauter des étapes
  - Réduire le temps consacré au dialogue
  - Se contenter de taux d'accords moyens
  - Traiter moins de données
  - Abandonner la double-annotation

Problèmes posés par l'intervention humaine

- Trouver d'autres solutions
  - Exemple du Penn Discourse Treebank
    - Repérage automatique des connecteurs pour permettre une annotation plus rapide
  - Exemple du projet CESTA (campagne Technolangue)
    - Elaborer un traitement automatique permettant d'évaluer une traduction automatique
  - Exemple de Treetagger
    - Exploiter des ressources préexistantes pour évaluer le traitement automatique

Conclusion

- Tentative de mise à plat

- Un schéma qui nous semble cohérent
  - Y retrouvez-vous votre démarche ?...

Bibliographie

Carlson, L., Marcu, D., Okurowski, M.E. (2001). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, Eurospeech 2001, Danemark.

Hamon, O. (2007). Rapport du Projet CESTA : Campagne d'Evaluation des Systèmes de Traduction Automatique. http://www.technolangue.net/IMG/pdf/Rapport_final_CESTA_v1.04.pdf

Marcus, M., Santorini, B., Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English : the Penn TreeBank. *Computational linguistics,* 19(2), 313-330.

Miltsakaki, E., Prasad, R., Joshi, A., Webber, B. (2004). The Penn Discourse TreeBank. *In* LREC 2004, Lisbon, Portugal.

Rehm, G. , Santini, M., Mehler, A. & al. (2008). Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems. *In* LREC 2008, Marrakech, Maroc.

TREC (2007). Question Answering Track Guidelines*. Appel à candidature.* http://trec.nist.gov/data/qa/2007_qadata/qa.07.guidelines.html

Treetagger. Site de l'Université de Stuttgart : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/