Traduction Automatique (TA) et Traduction Assistée par Ordinateur (TAO)

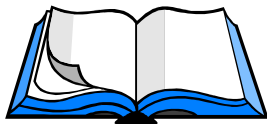
Malek Boualem
France Telecom
R&D/TECH/EASY/LN
16 Février 2006

Biographie de Malek Boualem

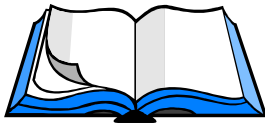
Malek Boualem est ingénieur informaticien et a obtenu, en 1993 à l'université de Nice, le titre de docteur en informatique. Sa thèse, préparée à l'INRIA, avait pour thème le multilinguisme et la traduction automatique multilingue. Avant de rejoindre France Télécom R&D, Malek Boualem a travaillé à l'université de Provence et au laboratoire Parole et Langage du CNRS, où il a obtenu le prix du CNRS/ANVIE 1996 de la valorisation de la recherche scientifique, pour le co-développement d'un éditeur de textes multilingues MtScript. Il a également travaillé en qualité de chercheur au Computing Research Laboratory à l'université du Nouveau Mexique aux USA. Il possède une expérience industrielle, notamment dans le domaine de la localisation multilingue des applications informatiques. Actuellement, Malek Boualem travaille dans l'équipe Traitement Automatique des Langues Naturelles au laboratoire FTR&D/TECH/EASY à Lannion. Il coordonne les activités de recherche en traduction automatique de l'écrit et de l'oral dans le cadre du projet Transat et participe à diverses activités liées au multilinguisme.

Plan

I. Introduction



Texte **Traduction humaine**



Texte

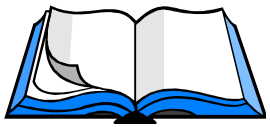
Parole **Traduction humaine**

Parole

➔ **Evolution des besoins**

- Rapidité
- Coût

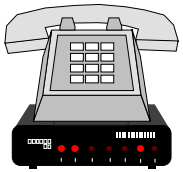
Evolution des besoins ➡ automatisation



Texte

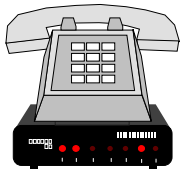
Traduction

Texte

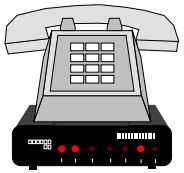


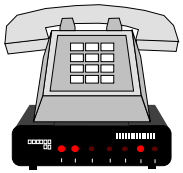
Parole

Traduction



Parole





Parole ➡ Reconnaissance de parole ➡ Texte ➡ Traduction ➡ Texte ➡ Synthèse de parole ➡ Parole

Ce dernier schéma illustre les services du type "Téléphone multilingue"

II. Traitement automatique du langage naturel

La mangge souris lle chat
Orthographe : **NON** ➡ Lexique et morphologie

La mange souris le chat
Orthographe : **OUI**
Syntaxe : **NON** ➡ Grammaire

La souris mange le chat
Orthographe : **OUI**
Syntaxe : **OUI**
Sémantique : **NON** ***sens du mot "souris " (animal ou souris d'ordinateur)*** ➡ Sens

La souris mange le chat
Orthographe : **OUI**
Syntaxe : **OUI**
Sémantique : **OUI**
Pragmatique : **NON** ***la souris ne mange pas le chat*** ➡ Domaine, monde réel

Le phénomène de l'ambiguïté

- **Ambiguïté lexicale :** *la pêche est bonne*
- **Ambiguïté structurale :** *la belle ferme le voile*
- **Ambiguïté référentielle :** *Marc a pris sa voiture*
- **Ambiguïté sémantique :** *le malade doit prendre les médicaments*
- **Ambiguïté idiomatique :** *Mark a explosé de joie*

Autres phénomènes

- **Anaphores** : reprise d'un segment de discours antécédent par un mot qui y renvoie (ex. : ***de l'argent, j'en ai***)
- **Ellipses** : omission de mots dans le discours (ex. : ***j'ai coupé le pain et l'ai mangé***)
- **Etc.**

III. Histoire de la traduction

- **Premières expériences (1946-1954) :** W. Weaver, A.D. Booth (dictionnaires), N. Wiene, A. Turing et J. Bar-Hillel (domaines réduits)
- **Optimisme et persuasion (1954-1966) :** Georgetown et IBM, Mark II (russe ➨ anglais), Spoutnik (1957), Moscou (D. Panov), ALPAC
- **Diversification des stratégies (1966-1980) :** Systran, Logos, Metal, Ariane, modèle pivot, Taum-Meteo.
- **Traduction assistée par ordinateur (1975-1985) :** Traduction interactive, TAO, traduction spécialisée, traducteurs de poche, mémoires de traduction, traduction statistique.
- **Premiers projets collaboratifs (1985-1995) :** Eurotra, etc.
- **Utilisation commerciale et offre en ligne (1990-2000) :** Systran, Reverso, etc.
- **Exploration de mécanismes sémantiques avancés (depuis 2000)**.
- **Traduction de l'oral (depuis 2002).**

Générations des systèmes de traduction

- **1re génération (1954)**
  - Traitements différents pour chaque couple de langues
  - Structures linéaires du langage
  - Traitement syntaxique partiel
- **2e génération (1960)**
  - Trois phases : analyse, transfert et génération
  - Mécanismes de transduction
  - Séparation des données linguistiques des programmes
  - Difficulté de traitement sémantique
- **3e génération (période actuelle)**
  - Objectif : traduction automatique véritable
  - Systèmes à base de connaissances (compréhension du langage)
  - Domaines restreints

IV. Catégories de traduction

1. **MAHT (Machine Aided Human Translation)**
   Traduction humaine + traitement de texte + dictionnaires électroniques + mémoires de traduction

2. **HAMT (Human Aided Machine Translation)**
   Traduction automatique + pré-édition + post-édition

3. **Traduction interactive**
   Ordinateur <-> Opérateur

4. **Traduction automatique**
   Ordinateur seul

5. **FAHQT (Fully Automated High Quality Translation)**
   Véritable traduction automatique de bonne qualité

V. Poste de traduction
(Translator's workbench / TWB)

➔ **Environnement d'outils pour assister le traducteur professionnel**

- Traitement de texte, traitement des documents
- Numérisation de textes
- Ressources monolingues : correcteur orthographique et grammatical, bases terminologiques, dictionnaires, corpus, dictée vocale, etc.
- Ressources bilingues : dictionnaires bilingues, corpus alignés, mémoires de traduction, etc.
- Logiciel de traduction
- Outils de communication

➔ **Usage possible : localisation(*) des applications et des documents**

(*) *Procédé par lequel un produit est adapté aux différentes langues (traduction) et aux différents pays (aspects culturels et pragmatiques) où il sera disponible.*

VI. Techniques de traduction

VI.1. Techniques symboliques ou à base de règles :

- Traduction bilingue ou directe
- Traduction basée sur le transfert
- Traduction basée sur la représentation intermédiaire (pivot)

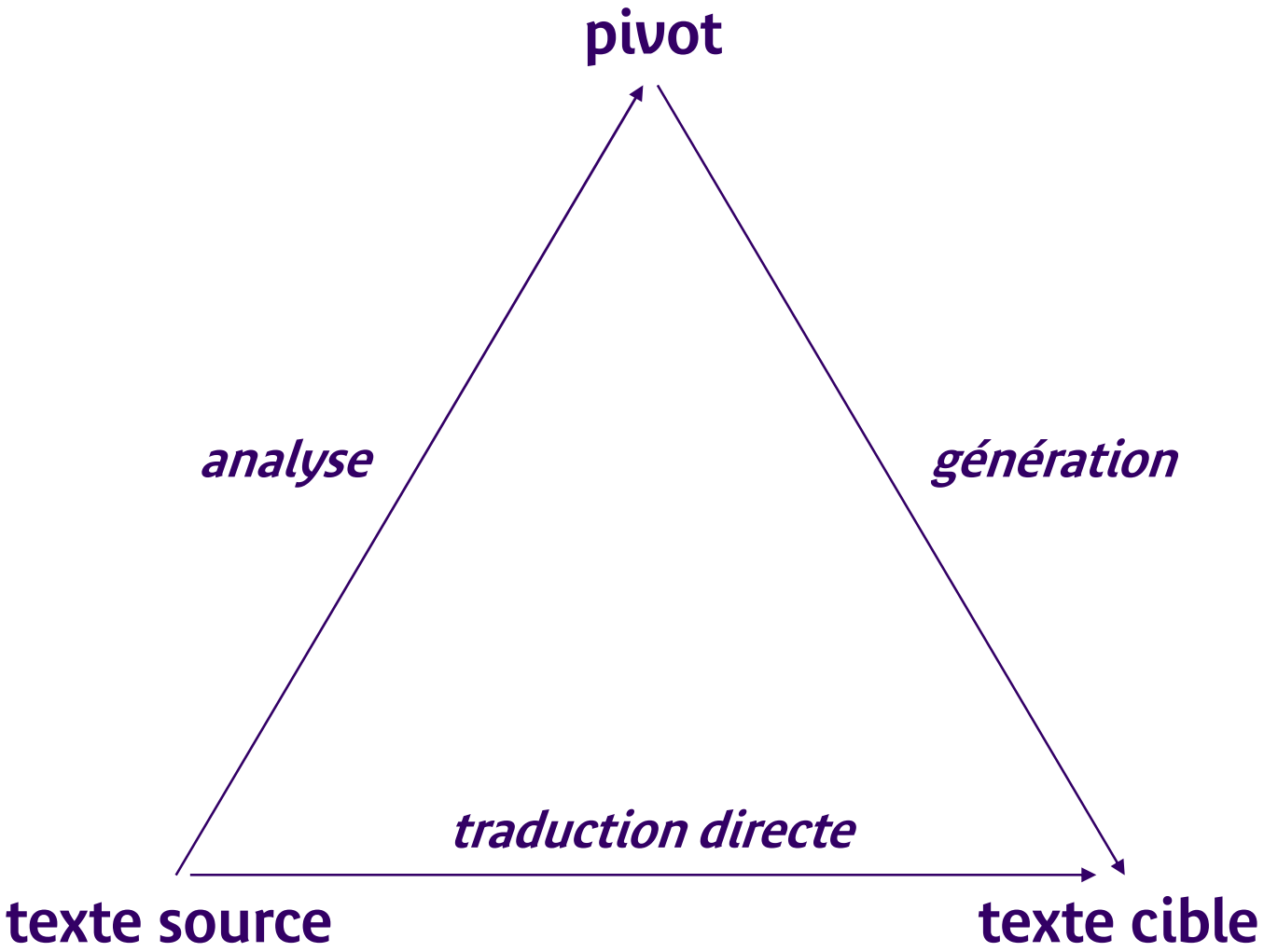
VI.2. Techniques d'apprentissage basées sur les corpus :

- Mémoires de traduction
- Traduction basée sur les statistiques
- Traduction basée sur l'exemple
- Traduction basée sur les connaissances
- …

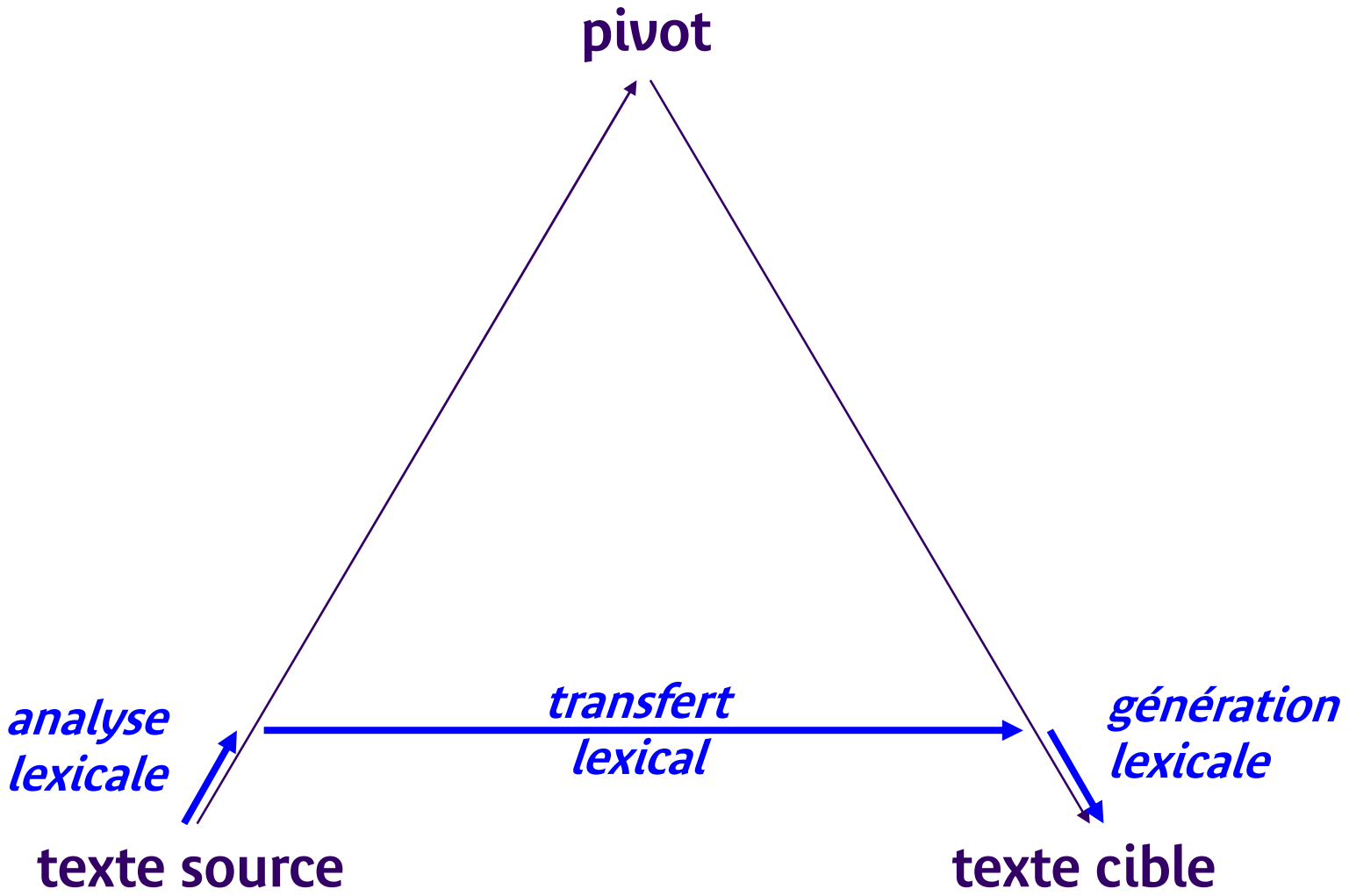
Techniques de traduction

VI.1. Techniques symboliques ou à base de règles

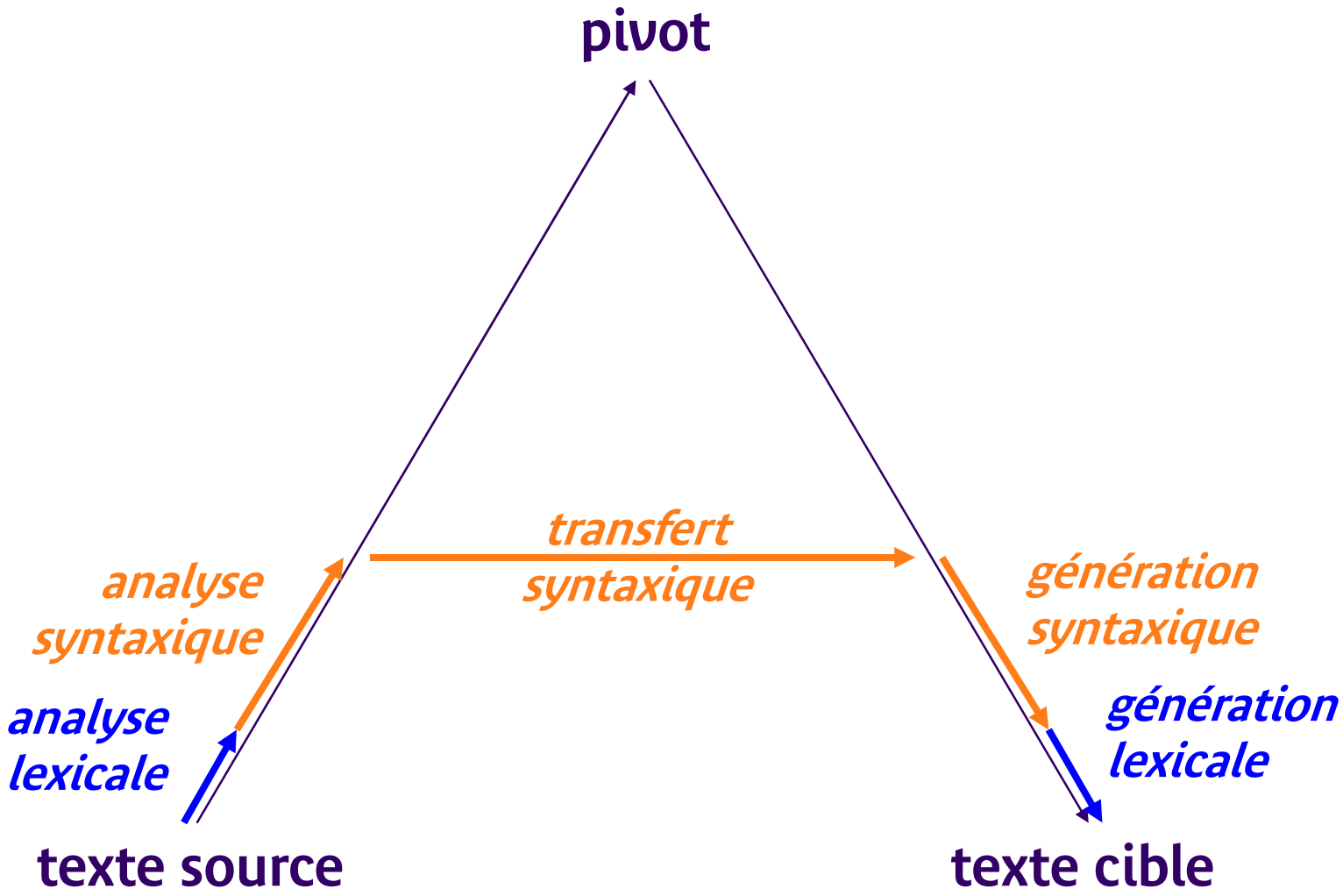
Triangle de Vauquois théorique



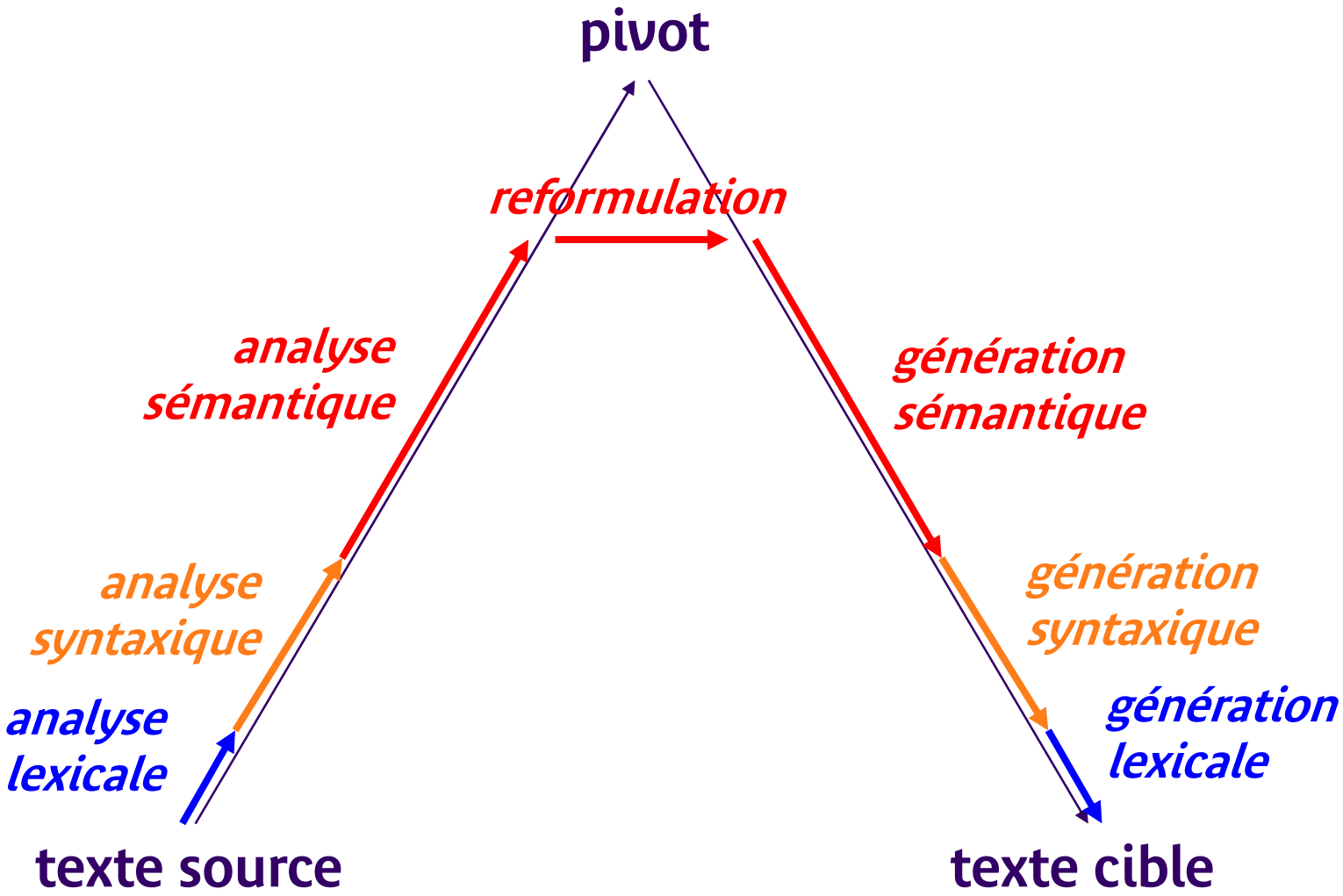
Triangle de Vauquois (1G : Systran)



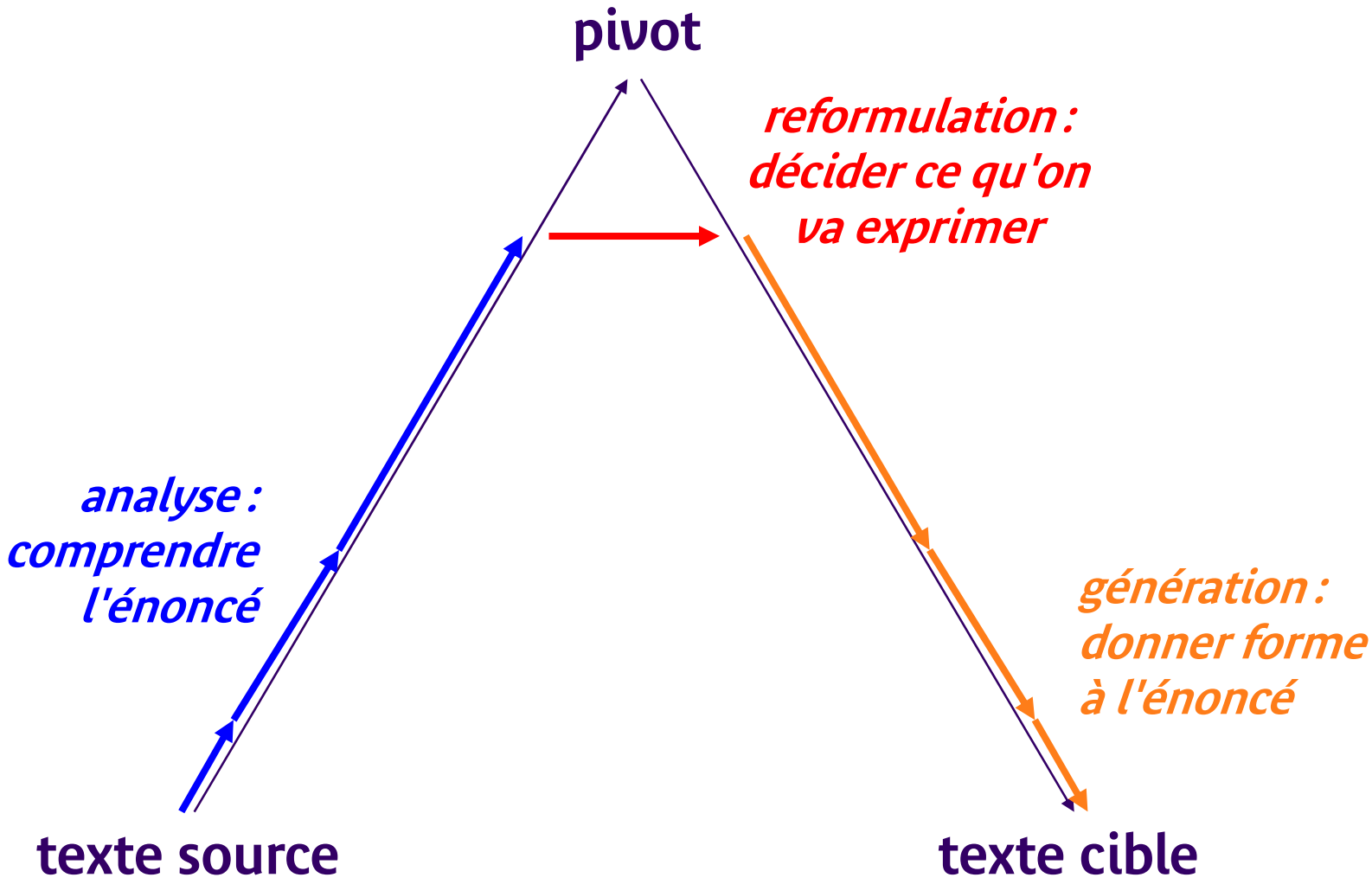
Triangle de Vauquois (2G : Reverso)



Triangle de Vauquois (3G : TiLT)

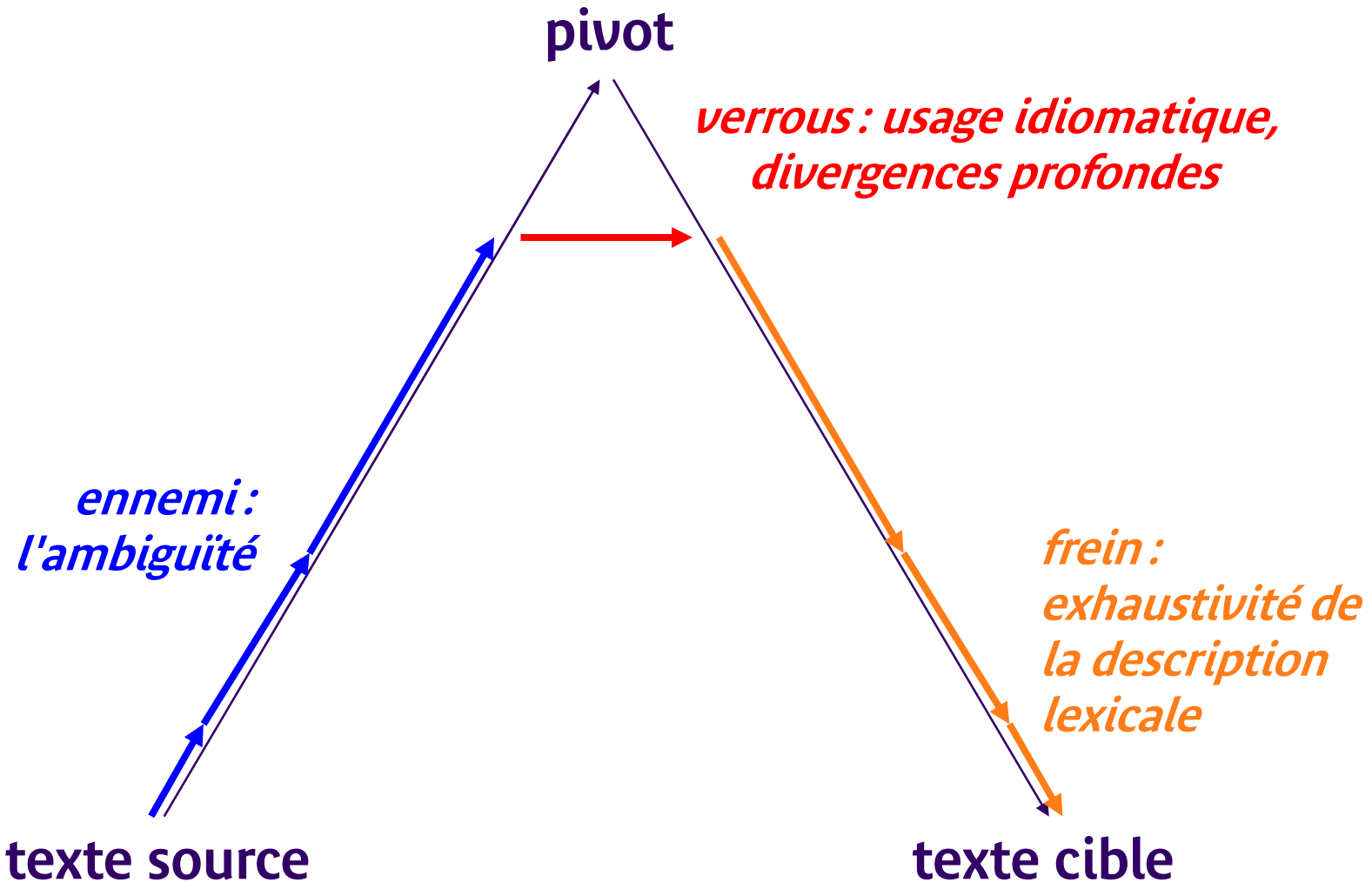


Triangle de Vauquois : grands enjeux



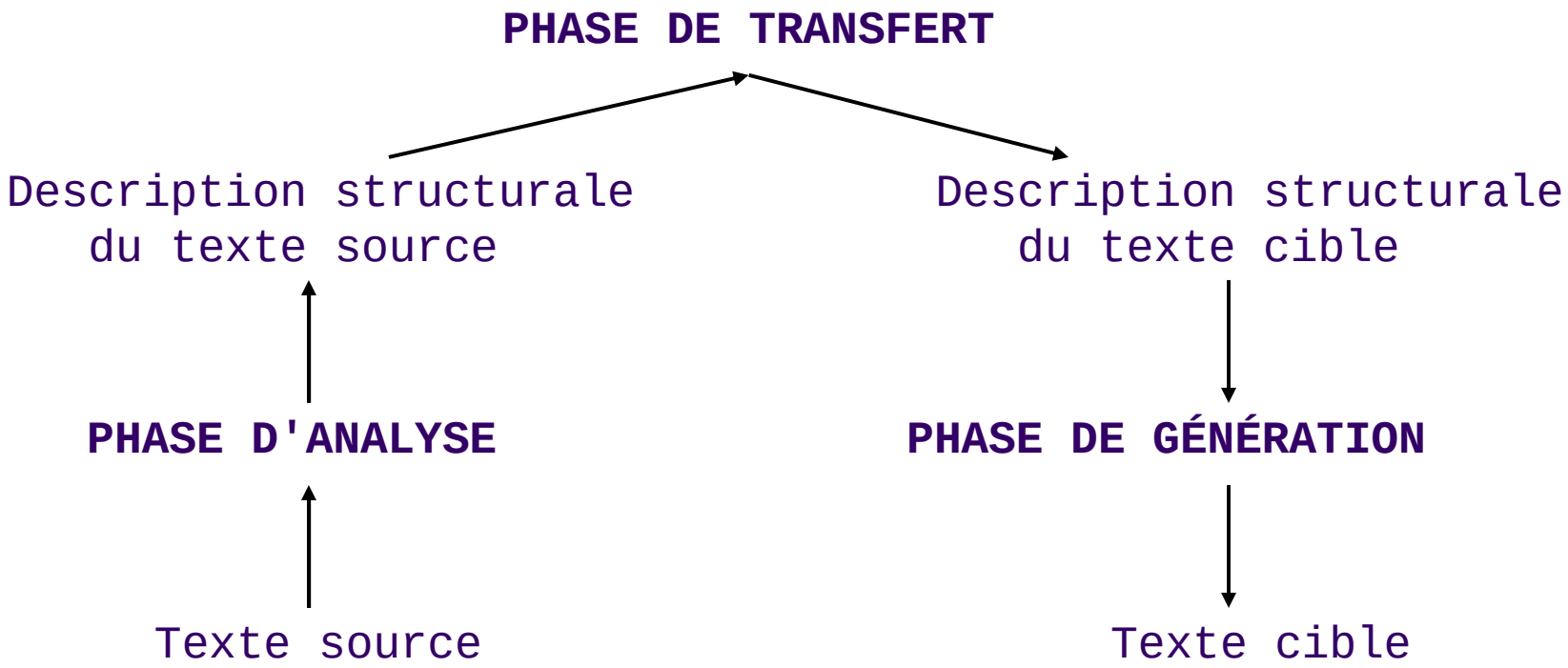
diffusion interne

Triangle de Vauquois : obstacles



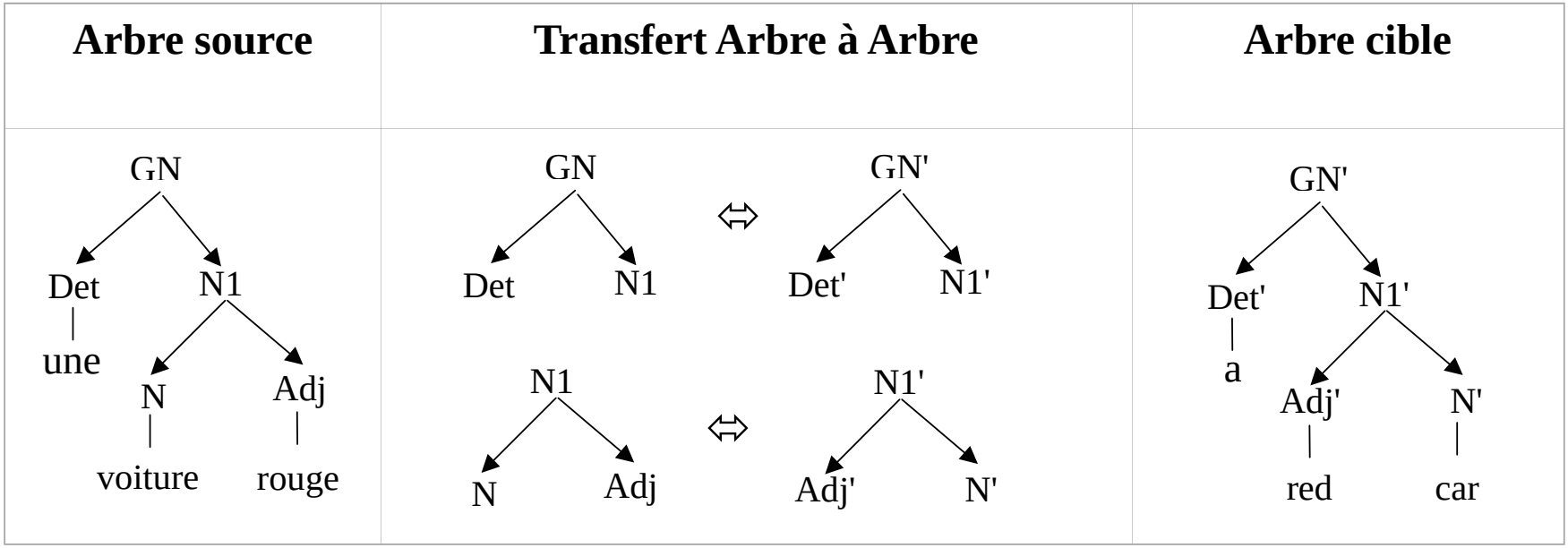
diffusion interne

VI.1.1. Traduction automatique basée sur le transfert



Transfert syntaxique

Vue simplifiée : transfert arbre à arbre



- Algorithme récursif
- Non déterministe
- Top-down

Exemples de systèmes de traduction automatique basés sur le transfert

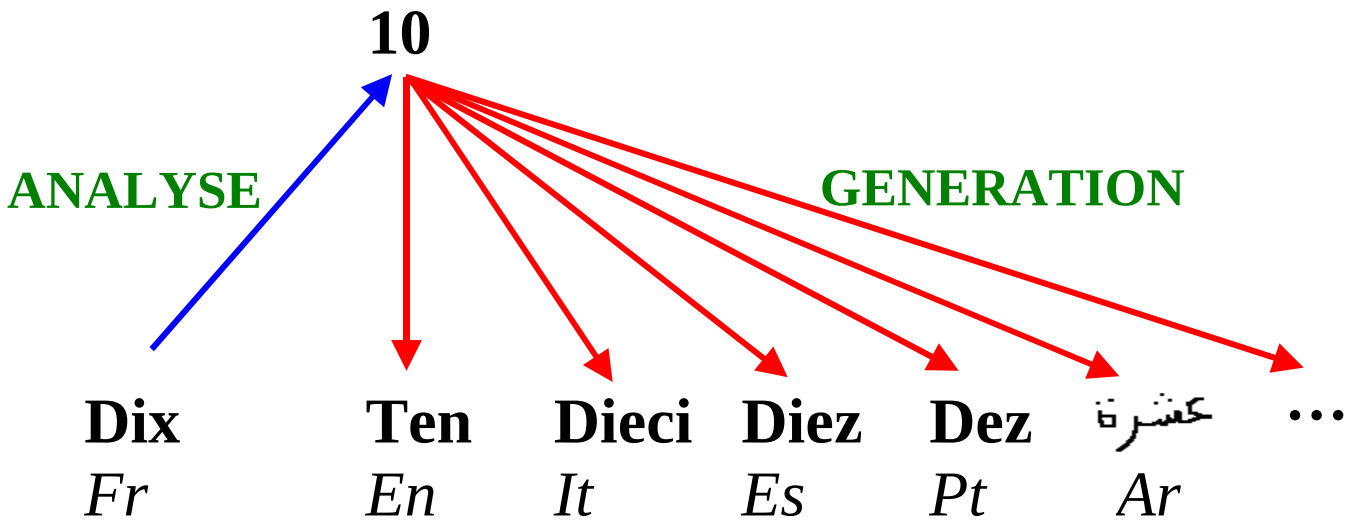
➔SYSTRAN
http://www.systransoft.com/
Version en ligne : http://babelfish.altavista.com/

➔REVERSO
http://www.softissimo.com/
Version en ligne : http://www.reverso.com/

SYSTRAN et REVERSO se partagent la quasi-totalité du marché européen de la traduction automatique à base de règles.

VI.1.2. TA basée sur la représentation intermédiaire (pivot ou interlingua)

➔ Cette technique de traduction qui transite par une représentation abstraite indépendante des langues est idéale pour les traductions multilingues.



Exemples de systèmes basés une représentation pivot

Compte tenu de la complexité de la formalisation d'une structure sémantico-conceptuelle intermédiaire entre les différentes langues, il n'existe encore aucun système opérationnel basé sur la traduction interlingue.

Un certain nombre de travaux de recherche sont en cours, comme :

➔Système ARIANE du GETA (Groupe d'Etude pour la Traduction Automatique)
http://www-clips.imag.fr/geta/

➔Projet UNL (Universal Networking Language)
http://www.undl.org/

➔UNITRAN & IAMTC
http://www.umiacs.umd.edu/labs/CLIP/

➔TILT à France Telecom R&D
http://langnat.elibel.tm.fr/traduction/ (site sécurisé)

Techniques de traduction

VI.2. Techniques d'apprentissage basées sur les corpus

Alignement de textes

Lorsque des textes ont été précédemment traduits, des techniques d'alignement permettent d'extraire de ces textes des bases d'unités de traduction qui pourraient alimenter des mémoires de traduction

A. Alignement au niveau des phrases

B. Alignement au niveau des mots

C. Alignement terminologique

VI.2.1. Mémoire de traduction *(Translation Memory)*

Aide à la traduction humaine (professionnelle) pour éviter les tâches répétitives de traduction de fragments de textes identiques ou similaires

➔Processus de traduction d'un fragment ou UT (Unité de Traduction)

- Recherche de l'UT source dans la mémoire de traduction
- Si une UT identique est identifiée, restitution de sa traduction
- Sinon, déclenchement d'un **"processus de calcul de similarité"**
- Proposition des UT possibles
- Choix d'une UT et mise à jour manuelle de la traduction
- Insertion de la nouvelle UT+ sa traduction dans la mémoire de traduction

Exemples de systèmes basés sur les mémoires de traduction

TRADOS

TRADOS Translation Memory Desktop

TRADOS Translation Memory Server

http://www.trados.com/

IBM

TranslationManager

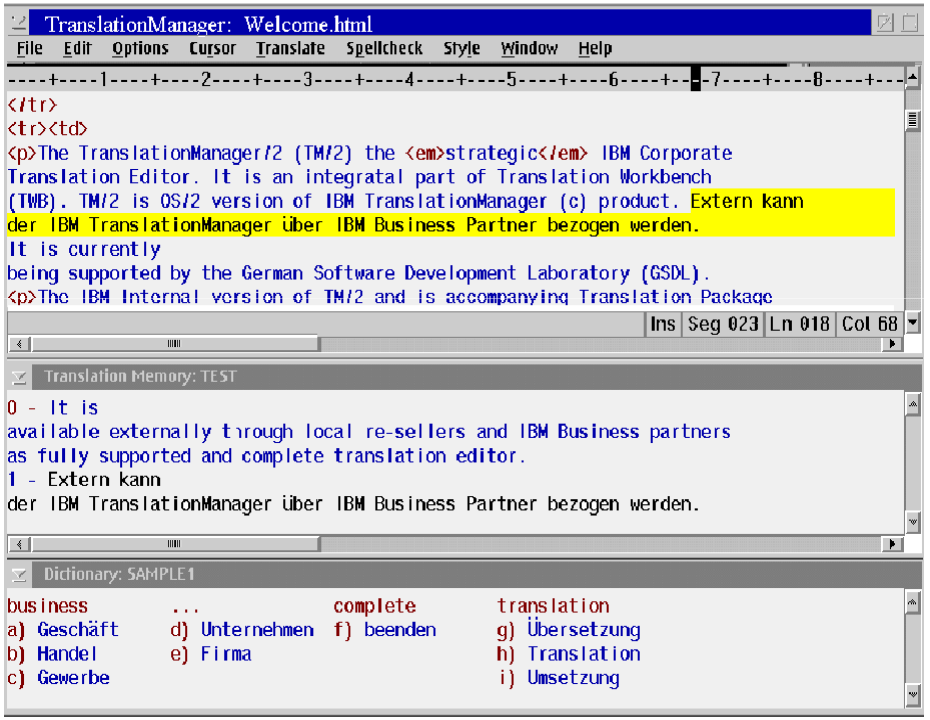
Evolution vers IBM WebSphere

http://www-306.ibm.com/software/

LINGUA & MACHINA

SIMILIS

http://similis.fr/Similis.html



VI.2.2. Traduction basée sur les statistiques

- La traduction basée sur les statistiques utilise de grands volumes de corpus bilingues pour la prédiction des traductions.
- Elle n'utilise pas des connaissances linguistiques, mais elle se base sur les propriétés distributionnelles (mesurables) des mots et des phrases afin d'en déduire les traductions les plus probables
  - Cooccurrence des mots dans les textes sources et cibles
  - Position relative des mots dans les phrases
  - Longueur des phrases
  - etc.
- La traduction basée sur les statistiques est particulièrement appropriée pour la traduction de l'oral.

Exemples de travaux actuels sur la traduction statistique (notamment pour l'oral)

- USA :
  - CMU, IBM, Microsoft, MIT, ISI/USC, AT&T, SRI, …
- Japon :
  - NTT, ATR, NEC, OKI, Nagaoaka University, University of Tokyo
- Allemagne :
  - RWTH, Karlsruhe
- Italie :
  - ITC-IRST
- Espagne :
  - TALP
- Chine :
  - NLPR
- UK :
  - Edinburgh University
- France :
  - CLIPS, LIMSI

VI.2.3. Traduction basée sur l'exemple (EBMT)

- La technique de traduction basée sur l'exemple est une généralisation de la technique basée sur la mémoire de traduction

- Contrairement à la mémoire de traduction où le système reformule les anciennes traductions pour générer de nouvelles traductions, la mémoire d'exemples est utilisée pour en extraire des segments identiques dans les phrases traduites

VI.2.4. Traduction automatique basée sur les connaissances (KBMT)

Cette technique est également considérée comme une technique à représentation intermédiaire du langage

Contrairement au modèle classique de représentation intermédiaire, cette représentation va au-delà de la structure purement linguistique pour couvrir des connaissances du monde réel liées à un domaine d'application particulier

Exemples de travaux sur la traduction basée sur les connaissances

Projet Mikrokosmos :

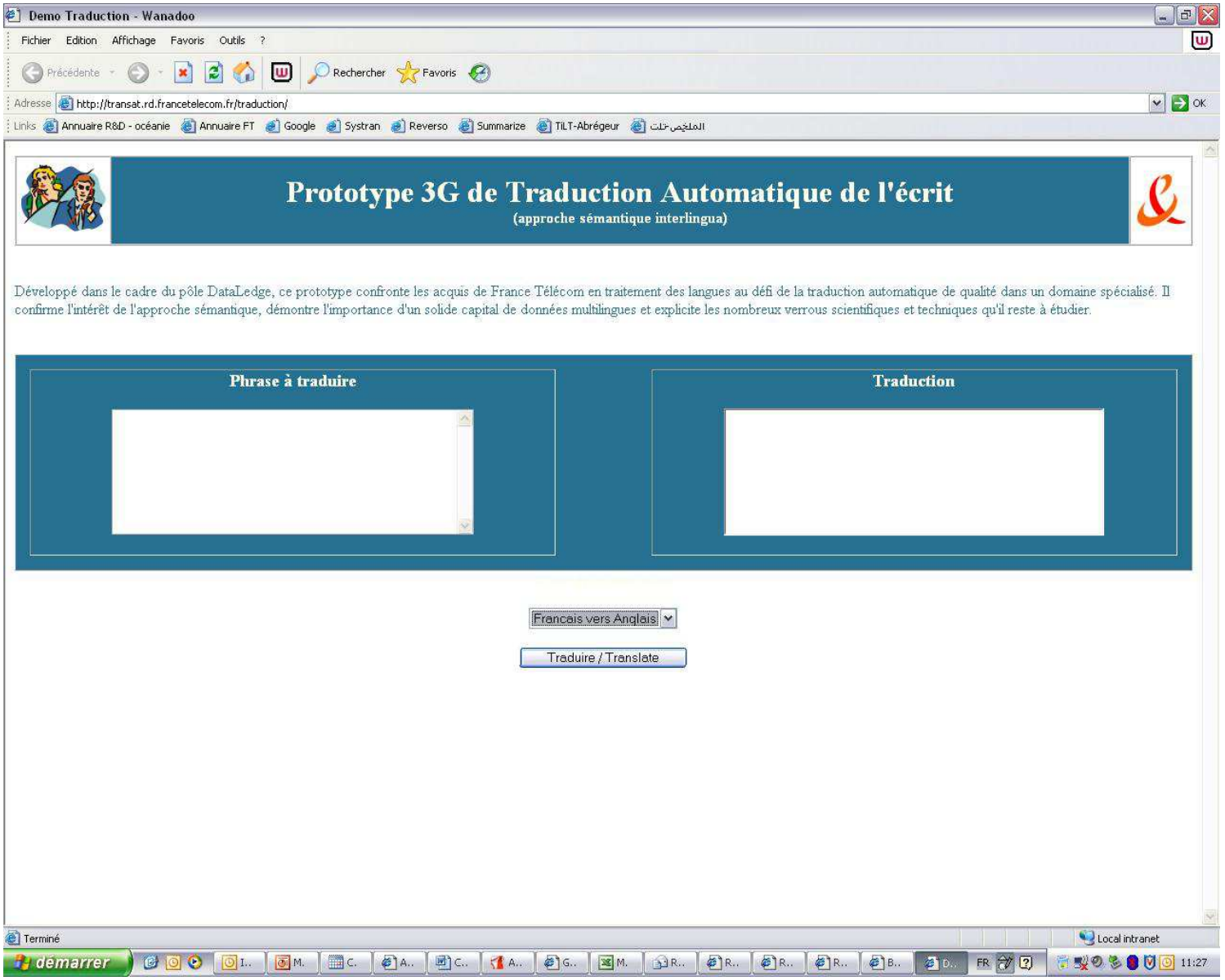
http://crl.nmsu.edu/Research/Projects/mikro/

Travaux à FRANCE TELECOM sur la traduction automatique

Travaux sur la traduction automatique de l'écrit et de l'oral à France Telecom

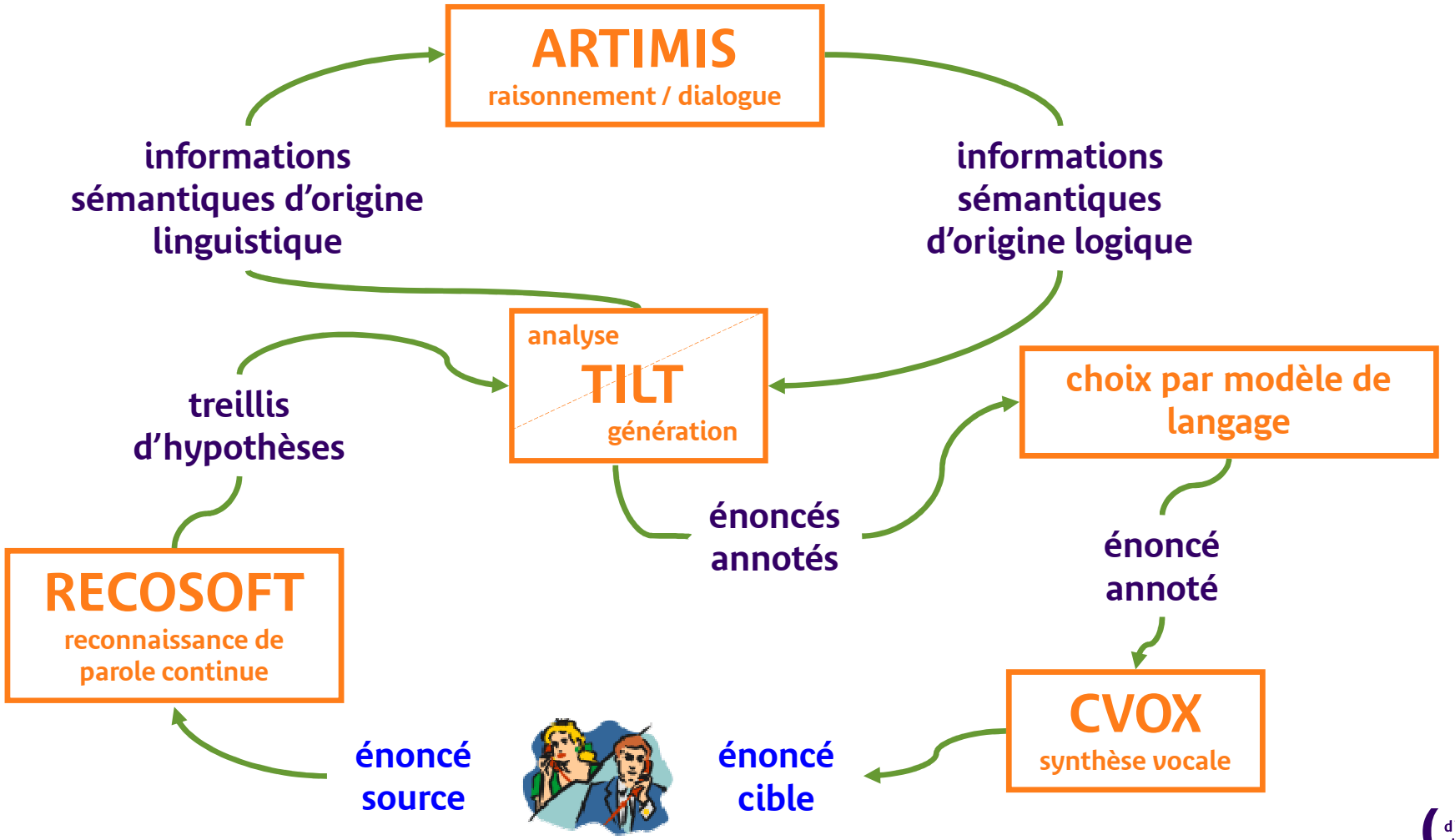
- **Besoins (à France Télécom et en dehors de France Télécom)**
  - Diffusion et Recherche d'informations multilingues.
  - Besoins considérables en traduction de l'oral (services d'urgences, etc.).
  - Réduction des coûts et rapidité.
- **Offre actuelle en systèmes de traduction :**
  - Pas satisfaisante.
- **Background de France Télécom R&D**
  - Technologies innovantes en traitement multilingue du langage écrit et parlé.
- **Travaux sur la traduction automatique à France Télécom R&D**
  - Technologie de traduction 3G intégrant les niveaux sémantique et pragmatique du langage naturel.
  - Mise en œuvre de la traduction automatique à base d'une structure "pivot".
  - Combinaison des approches symboliques et statistiques.
  - Traduction automatique de l'oral.

Travaux sur la TA de l'écrit à France Telecom



Travaux sur la TA de l'oral à France Telecom
Projet TRANSAT - *TRANslation of Speech And Text*

Démonstration **02 96 05 95 61**



diffusion interne