

Le TAL: des mots au contenu

1. Les mots pour le TAL : approche historique
2. Quel contenu pour le TAL ?

Cori M. et J. Léon, « La constitution du TAL, Etude historique des dénominations et des concepts », *TAL*, Vol. 43, n° 3, 2002, pp. 21-55.

1. Les mots pour le TAL : approche historique

1.1 La traduction automatique

1.2 La *Computational Linguistics*

1.3 Le *Natural Language Processing*

1.4 Les mots en français

1.1 La traduction automatique

Texte fondateur : W. Weaver, *Memorandum*, 1949

position « modeste » : traduction littéraire ou parfaite impossible.

- *Machine Translation*
- *Mechanical Translation*

Idée : machine spécifique dédiée à la traduction automatique

1960: rapport de Y. Bar-Hillel

1966: rapport de l'ALPAC (*Automatic Language Processing Advisory Committee*)

Mise à l'index de la TA, suppression des subventions aux Etats-Unis.

Cause : travaux peu fondés scientifiquement, très décevants au regard d'annonces triomphalistes, soupçons d'escroqueries.

Diagnostic : la traduction automatique est impossible.

Conséquences : - faire de la *machine-aided translation* ;

- développer les aspects théoriques (langages formels, analyse syntaxique,...)

1.2 *La Computational Linguistics*

Analyseurs syntaxiques fondés sur des grammaires.

Y. Bar-Hillel, V. Yngve, D.G. Hays (Tesnière)

1962 : *Association for Machine Translation and Computational Linguistics*

1973 : *Association for Computational Linguistics*

1954-1965 : *Mechanical Translation and Computational Linguistics*

1965-1973 : *American Journal of Computational Linguistics*

1973-2004 : *Computational Linguistics*

1.3 Le Natural Language Processing

Apparition de termes plus en relation avec les traitements informatiques qu'avec les études linguistiques

Hays, 1962 : *Automatic Language(-Data) Processing*

années 1970 : *Natural Language Understanding* (lien avec l'IA)

années 1980 : *Natural Language Processing* (terme toujours en vigueur)

années 1990 : *Natural Language Engineering*

Retour vers applications industrielles.

1.4 Les mots en français

Retard sur les Etats-Unis en Traduction automatique.

1959, E. Delavenay, *La machine à traduire*

1959 : création de l'ATALA (Association pour l'étude et le développement de la traduction automatique et de la linguistique appliquée)

Revue : *La Traduction Automatique* (1960-1964)

1965-1991 : *TA Informations*

Impossibilité de traduire *Computational Linguistics*

1959 : Linguistique quantitative : à la fois aspects formels et études statistiques

années 1960 : Linguistique mathématique

années 1970 : Informatique linguistique (et Linguistique informatique)

1985 : cursus de linguistique *et* informatique (Univ. Paris 7).

L'unification sous le terme TAL

Premières occurrences du terme : 1967.

TAL = Traitement automatique des langues

Traitement automatique du langage

TALN : des langues naturelles, du langage naturel

1992 : la revue de l'ATALA devient *TAL*

1993 : création de la licence de Sciences du langage, mention
Traitement automatique des langues

Les nouveaux termes

Industries de la langue (maîtrise mention Industries de la langue)

Ingénierie linguistique

Technologies de la langue

Les développements technologiques rendent accessibles aux particuliers et aux entreprises certaines applications du TAL.

2. Quel contenu pour le TAL ?

2.1 Les tensions internes au TAL

2.2 TAL et linguistique formelle

2.3 Le TAL constitue-t-il une discipline unifiée ?

2.1 Les tensions internes au TAL

Plusieurs clivages dans le TAL :

- applications pratiques vs recherche scientifique
- TAL robuste vs TAL théorique
- linguistique de corpus vs linguistique formelle
- méthodes probabilistes et statistiques vs méthodes symboliques

Les disciplines connexes

Le TAL s'appuie sur plusieurs disciplines internes/externes :

- la linguistique ;
- l'informatique ;
- les mathématiques (algèbre, logique, statistiques et probabilités) ;
- l'Intelligence artificielle, les sciences cognitives.

TAL et disciplines connexes

Confusion des étiquettes, rapports plus ou moins conflictuels entre le TAL et ces divers domaines :

- Les acteurs du TAL proviennent en général de l'un de ces domaines, et ils ont un rattachement académique à l'un de ces domaines
- Les articles de TAL sont publiés soit dans des revues spécialisées, soit dans des revues d'informatique
- Les objectifs et les problématiques diffèrent selon la discipline de référence

Informatique : 1. Construire des logiciels qui répondent à une demande sociale et économique ;

2. Définir des algorithmes généraux, des structures de données formelles ou même des langages de programmation (informatique théorique).

Mathématiques : critères de rigueur importés dans le TAL.

Intelligence artificielle : en concurrence avec le TAL. Le TAL est-il un problème particulier de l'Intelligence artificielle ? Le TAL est vu surtout dans le cadre des problèmes de compréhension.

Sciences cognitives : volonté d'être une discipline unifiante.

2.2 TAL et linguistique formelle

Pourquoi faire du TAL dans la perspective de la linguistique ?

1. Possibilité d'accéder à des grands corpus ;
2. Constitution de modèles formels ;
3. Expérimentation des modèles formels.

Historiquement : les algorithmes d'analyse syntaxique, au cœur de la discipline dans les années 1960.

Des analyseurs aux formalismes syntaxiques

Années 1970 : les ATN, outils mixtes.

Années 1980 : développement des formalismes déclaratifs

- les DCG (*Definite Clause Grammars*, Pereira et Warren) ;
- les FUG (*Functional Unification Grammars*, Kay) ;
- PATR II (*Parse and Translate*, Shieber).

Convergence avec des formalismes linguistiquement motivés:

- les LFG (*Lexical Functional Grammars*, Bresnan) ;
- les GPSG (*Generalized Phrase Structure Grammars*, Gazdar) ;
- les TAG (*Tree Adjoining Grammars*, Joshi).

Présentation unifiée de tous ces formalismes sous la dénomination de *Grammaires d'unification* (Shieber).

Mais, actuellement, étant donné l'importance prise par le TAL robuste, il y a un affaiblissement du lien entre linguistique formelle et TAL.

Développement de nouvelles branches en linguistique, de nouveaux types de travaux :

- constitution d'outils linguistiques (dictionnaires, grammaires) à base de corpus ;
- découverte de faits linguistiques à travers des corpus ;
- modélisation statistique des phénomènes de variation (étude des genres textuels, par exemple).

Regroupés sous le terme de linguistique de corpus.

2.3 Le TAL constitue-t-il une discipline autonome et unifiée ?

Y a-t-il des objets propres au TAL ?

Dans une première époque :

- idée d 'une « machine à traduire » ;
- idée de langages de programmation spécifiques ;
- objets qui sont à mi-chemin du programme et du formalisme : les ATN (*Augmented Transition Networks*, Woods, 1970).

Années 1980 : les systèmes informatiques doivent séparer les programmes (procéduraux) des données (déclaratives).

Motivation : avoir une approche plus rigoureuse.

Conséquence : division du travail entre l'informaticien (qui écrit les programmes) et le linguiste (qui élabore les grammaires et les lexiques).

Il s'ensuit que les acteurs du TAL sont incités à se replier sur une spécialité donnée. D'où la difficulté de l'affirmation du TAL comme discipline scientifique autonome.

Ce sont les ingénieurs qui seuls peuvent se réclamer sans équivoque du domaine.

Conclusion

Il ne faut pas croire que nous sommes dans un domaine dont les frontières sont établies une fois pour toutes, ni même dont l'existence est pérenne.