

## Les systèmes de traduction automatique

*Victoria Arranz*

*ELRA/ELDA*

arranz@elda.org

<http://www.elda.org/> or <http://www.elra.info/>

- ELRA/ELDA : mission et activités
- Petit rappel historique
- De la traduction automatique...
- ... de nos jours...
- ... pour quel futur ?

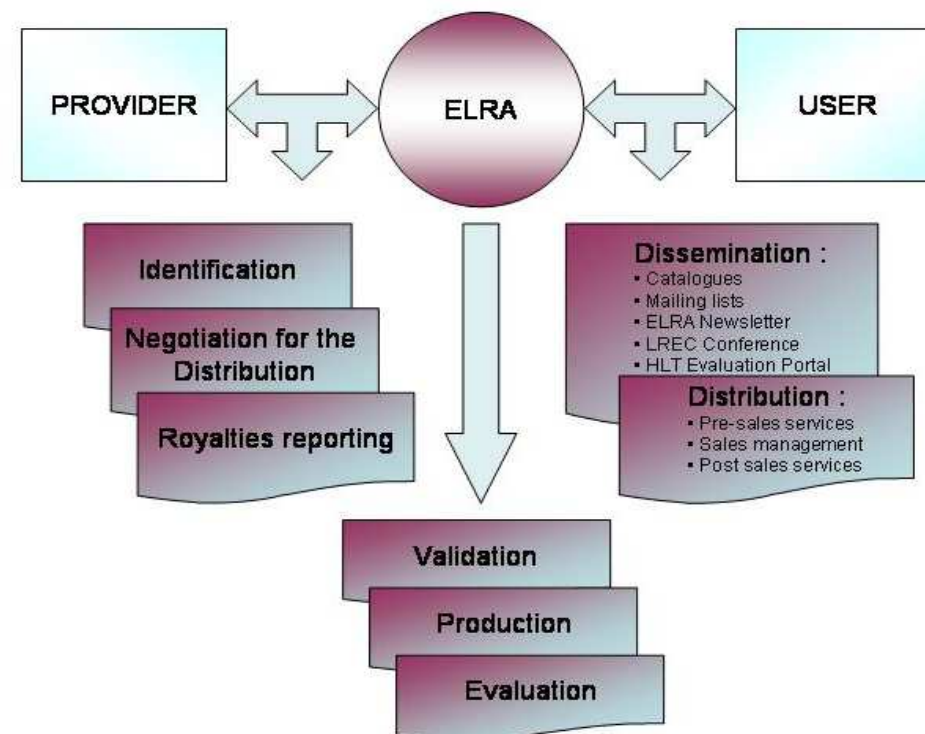
---

# **ELRA/ELDA : mission et activités**

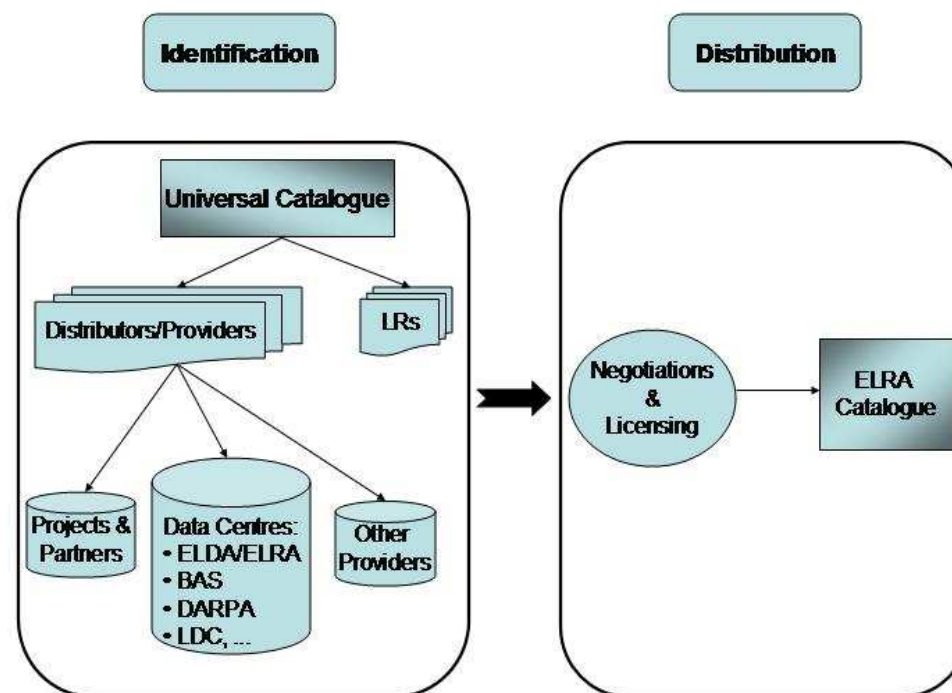
- Création d'ELRA : 1995, financement de la CE (3 ans)  
→ Attirer l'attention de la communauté : besoin urgent de *partager* et *réutiliser* les ressources linguistiques (RLs)

- Corps opérationnel : ELDA

- 4 pôles d'activité



- Recherche de RLs
- Négociations et aspects juridiques
- Catalogue ELRA : <http://catalogue.elra.info/>
- Catalogue Universel : <http://universal.elra.info/>
- Types de RLs : texte, parole, multimodal, évaluation



- Production propre ou coordination de la production
- Projets nationaux, européens et internationaux
- Soutien d'institutions, commerciales ou académiques
- Rôle consultatif :
  - Standards et validation : PCom et VCom
  - Aspects techniques : Format de *codage*, de *stockage*, de *description*...
- Production récentes :
  - bases de données audio (LILA, Orientel...), corpus parallèles (TA de la parole – TC-STAR), corpus multimodaux (annotations de vidéos – CHIL), corpus et packages d'évaluation (TC-STAR, CHIL, campagnes EVALDA...)

- Au premier plan dans le domaine de l'ingénierie linguistique :
  - Evaluation de base : identifier les *créneaux porteurs* pour le développement de nouvelles technologies et le potentiel commercial
  - Evaluation des technologies : évaluer la recherche des *équipes de R&D* et assurer le développement des systèmes en les testant
  - Evaluation d'utilisabilité : tester l'*utilisation finale* sur le terrain
  - Evaluation de l'impact : mesurer les *conséquences* socio-économiques
  - Evaluation de programme : mesurer le *retour sur investissement* aux agences ayant financé des projets
  
- Techniquement, les « campagnes d'évaluation » contribuent à :
  - Produire des données de test ou d'entraînement de qualité et réutilisables
  - Développer des méthodologies, métriques et protocoles d'évaluation
  - Créer des « packages d'évaluation » distribués en tant que LR
  - Dynamiser le domaine de l'ingénierie linguistique

- Projets européens (CHIL, TC-STAR, CLEF...) et programmes Français (Technolange/EVALDA, Passage...)
- Services personnalisés ou par collaboration :
  - Services d'évaluation sur demande
  - RLs personnalisées pour laboratoires et/ou sociétés
  - Evaluation d'outils et technologies TAL :
    - *Traitement du texte* : traduction automatique, alignement multilingue, recherche d'information, annotation syntaxique, extraction terminologique...
    - *Traitement de la parole* : traduction automatique de la parole, identification du locuteur, reconnaissance automatique de la parole, synthèse de la parole...
    - *Interfaces multimodales* : détection multimodale de personnes, identification multimodale des personnes, reconnaissance audiovisuelle de la parole...
- Promotion des activités d'évaluation:
  - Distribution de packages d'évaluation
  - Dissémination sur le *portail d'évaluation*

- Site ELRA : <http://www.elra.info>
- Site ELDA : <http://www.elda.org>
- Portail d'évaluation *HLT* : <http://www.hlt-evaluation.org/>
- Site BLARK : <http://www.blark.org/>
- Conférence internationale sur les ressources linguistiques et l'évaluation : <http://www.lrec-conf.org>
- *Language Resources and Evaluation Journal* (Springer)
- ELRA Newsletter, ELRA Members' News
- Articles de conférences

# Petit rappel historique

- XVII<sup>e</sup> : Descartes et Leibniz
- XX<sup>e</sup> siècle : brevets de George Artsouni (1933) et Peter Smirnov-Troyanskii (1937)
- 1949 : mémo de Warren Weaver → utilisation des ordinateurs pour la traduction, avec objectifs et méthodologies
- 1951 : Bar-Hillel est nommé chercheur en TA à MIT
- 1952 : 1<sup>ere</sup> conférence sur TA
- 1954 : démonstration publique de système de TA par l'Université de Georgetown et IBM :
  - Russe->Anglais
  - 49 phrases
  - Vocabulaire de 250 mots
  - 6 règles grammaticales
  - Résultat : motive la recherche et stimule le financement à grande échelle

- 10 années suivantes :
  - La décennie de l'optimisme : années très actives
  - Approche principale : traduction directe, mais aussi d'autres essais
  - Résultats : grande productivité dans la création de dictionnaires électroniques, techniques d'analyse syntaxique, mais la TA de qualité n'arrive pas → de nombreux problèmes linguistiques ressortent
- 1964 : Le comité ALPAC est établi
- 1966 : Rapport ALPAC, qui conclut sur la TA :
  - Plus lente, moins précise et plus chère que la traduction humaine
  - Il faut investir dans les aides automatiques pour les traducteurs, pas la TA
- Impact imminent aux Etats-Unis : fin des financements gouvernementaux, pourtant Systran et Logos démarrent
- Europe : Canada, France et Allemagne continuent leur recherche

- 1967-1985 :
  - Systran : adopté par la US Air Force (Ru-En, 1970) et la CE (1976)
  - METEO : développé par l'Université de Montréal et mis en place en 1977
  - Autres systèmes : Weidner, GETA-Ariane (Grenoble), SUSY (Saarbrücken), METAL (Austin, Texas), MU (Univ. Kyoto)
  - EUROTRA : projet ambitieux pour les langues de la CE → transfert
- 1985-1995 :
  - Idée de traduction interlangue et systèmes à base de connaissances :
    - Connaissance du monde réel : approche de l'Intelligence Artificielle
    - Connaissance linguistique : DLT (Utretch) et ROSETTA (Phillips, Eindhoven)
  - TA en Chine, Corée, Russie et Europe de l'Est
  - Boom de la TA au Japon : Fujitsu, Toshiba, Matsushita, Mitsubishi...
  - Avancées en techniques statistiques (IBM) et en technologies de la parole pour la traduction oral (Verbmobil, ATR, JANUS)
  - La TA a base de règles continue (Catalyst, Pangloss)
  - Nouveaux Consortia : C-STAR

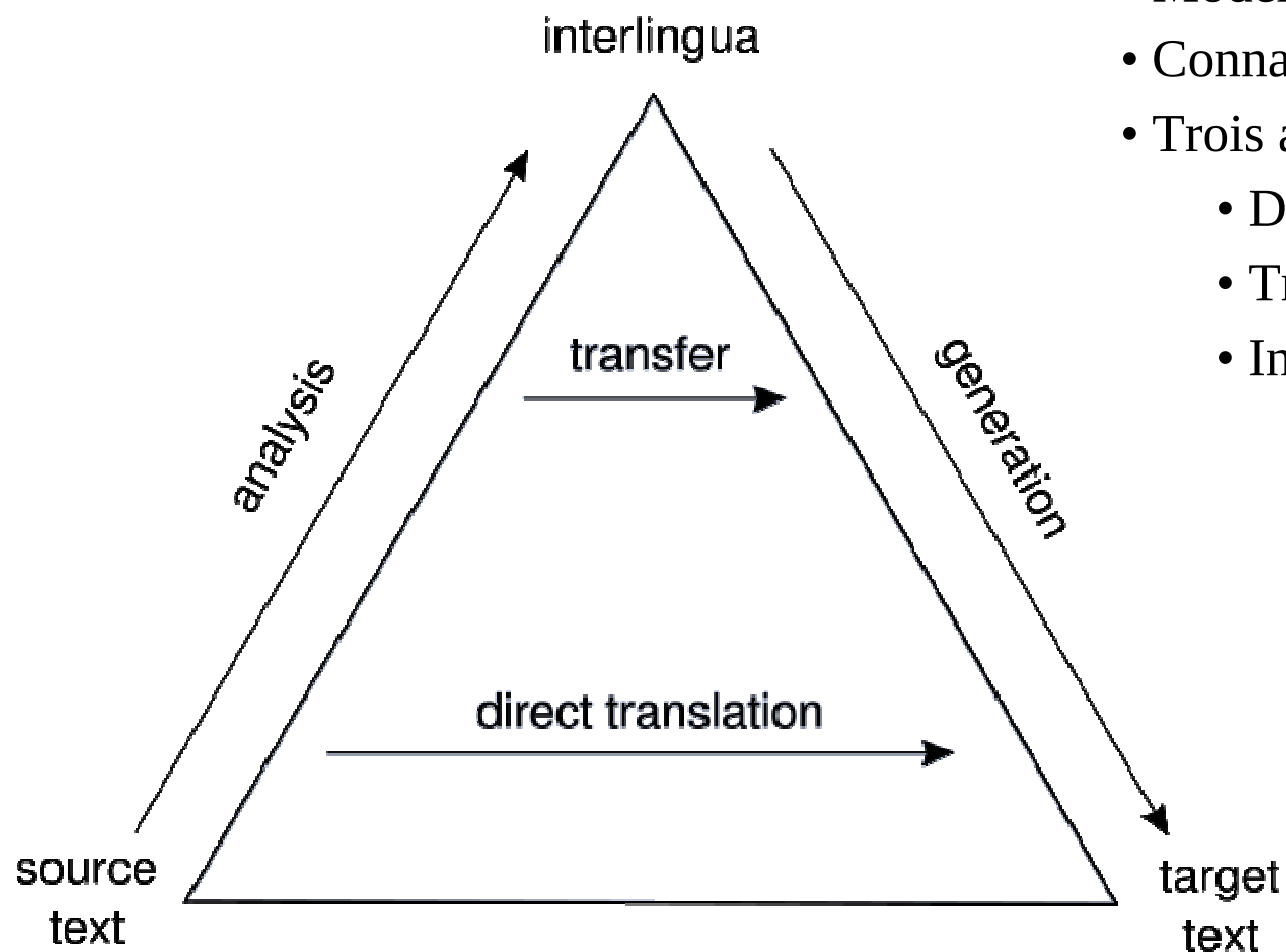
- Une réalité : l'idée initiale d'une TA de haute qualité est remplacée par l'utilisation des outils de traduction et des vocabulaires contrôlés
- Développement et disponibilité des ordinateurs personnels
- Explosion du marché de la traduction
- TA disponible pour PC : Globalink, MicroTac, Systran...
- TA disponible sur Internet : Altavista's Babelfish (technologie Systran) et Google Language Tools (technologie Systran au début et statistique à présent)

# De la traduction automatique...

- Conversion d'un texte en langue source en un texte en langue cible
- Opération complexe :
  - Décodage du sens (texte source) + recodage dans la langue cible
  - Opération cognitive pour le traducteur → challenge pour la machine
  - Plusieurs concepts : « compréhension »... « création »... « fluidité »
- Cette définition provoque plusieurs réactions :
  - Une machine peut-elle générer des textes de bonne qualité ?
  - La qualité peut-elle être assurée si les objectifs sont moins ambitieux ?
  - Une traduction imparfaite est-elle aussi utile ?
- Trois approches principales pour la traduction automatique:
  - A base de règles, à base d'exemples, statistique

- Différents profils utilisateur pour différents besoins :
  - Compréhension
    - Plusieurs langues sources et une langue cible
    - N'importe quel sujet
    - N'importe quel style
    - Résultat peut ne pas être parfait
    - Résultat peut être post-édité
    - Utilisateur : lecteur
  - Dissémination
    - Une langue source et plusieurs langues cibles
    - Un sujet
    - Une entrée contrôlée
    - Traduction bien écrite
    - Résultat sans post-édition
    - Utilisateur : écrivain

- Processus de traduction difficile (pour humains et pour machines)
- Approches naïves : marchent pour certaines applications, mais en général il nous faut plus...
- Quelques problèmes de traduction :
  - Mots individuels ambigus :
    - Morphologie : *avions* (nom / verbe)
    - Polysémie : *rose* (fleur / couleur)
    - Homonymie : *livre* (objet / mesure)
  - Ambiguïté syntaxique, sémantique,
  - Différences lexicales (« trou lexical » *punch* → *donner un coup de poing à*),
  - « trous conceptuels » (esp : *tortilla*),
  - Différences syntaxiques entre langues
  - ...



- Modèle stratifié
- Connaissances linguistiques
- Trois approches :
  - Directe (dictionnaire)
  - Transfert
  - Interlangue

Source: Wikipedia

- Règles écrites manuellement
- Limitations :
  - Problème de couverture pour les informations lexicales et la grammaire
  - Analyse syntaxique : pas toujours une traduction de qualité
  - Coût de développement
- Bons résultats, surtout pour les domaines restreints (FAME, réservation d'hôtel, médical...)
- Renaissance maintenant avec :
  - Disponibilité de grammaires et lexiques de grande couverture
  - Apprentissage de grammaires

- Traductions existantes utilisées comme modèle pour traduire
- Trois composants principaux :
  - Correspondance du texte source avec la base d'exemples
  - Alignement des fragments de traduction dans les exemples cibles
  - Recombinaison des fragments dans le texte cible
- Problèmes :
  - Correspondance basée sur les caractères pouvant ramener des erreurs
  - Alignement de fragments automatique est complexe
  - Recombinaison difficile selon la position du fragment dans les exemples
- Intéressant : utilisation de blocs plus idiomatiques
- Applications : combinaison avec TA à base de règles, mémoire de traduction, traduction statistique

- Processus statistique → un problème de probabilités
- Modèles entraînés sur des corpus alignés parallèles
- 2 modèles linéaires :
  1. Modèle de traduction :
    - Probabilités de trouver 1 mot source avec 1 mot cible
    - Problème 1 : pas uniquement des correspondances 1-1
    - Problème 2 : les mots des corpus parallèles peuvent avoir plusieurs positions différentes dans la phrase
  2. Modèle du langage :
    - Recomposer l'ordre des mots dans la langue cible
    - Probabilités sur les *n-grams* (séquences de mots)
- Plus facilement utilisable, plusieurs produits Open Source « clé en main » disponibles : Moses, Joshua, Apertium...

---

... de nos jours...

- Depuis 1995 :
  - Recherche :
    - Systèmes statistiques (Candide, IBM, nombreuses universités)
    - Systèmes à base d'exemples (M. Nagao, quelques universités)
    - Systèmes à base de règles (Eurolang, KANT, NESPOLE!, JANUS, FAME)
    - **Systèmes hybrides : statistique + règles** (ex. Systran+Lium, ...)
  - Utilisation commerciale de méthodes statistiques, mais limitée :
    - Language Weaver
    - Google
    - Microsoft
  - Situation en évaluation
    - Nombreuses campagnes (NIST, TC-STAR, Euromatrix, IWSLT...)
    - Evolution croissante, multiplication des mesures (MetricsMATR...)

- Situation actuelle :
  - Logiciels de TA allant du grand public aux serveurs d'entreprises
    - *Language Weaver* : <http://www.languageweaver.com/page/home/>
    - *LinguaTec* : <http://www.linguatec.net/>
    - *Morphologic* : <http://www.morphologic.hu/>
    - *Prompt* : <http://www.prompt.com/>
    - *Reverso* : <http://www.softissimo.com/>
    - *Systran* : <http://www.systran.fr/>
    - ...
  - Services gratuits en ligne, même pour des systèmes commerciaux :
    - *SystraNET* : <http://www.systranet.fr/>
    - *Reverso* : <http://www.reverso.net/>
    - *Voila.fr* et *Orange.fr* (technology Prompt) : <http://tr.voila.fr/>
    - *Yahoo's Babelfish* (technologie Systran) : <http://fr.babelfish.yahoo.com/>
  - Plus importante utilisation des logiciels de traduction par des sociétés

- Situation actuelle :
  - La localisation de logiciel ... une affaire en marche :
    - *Amicus TransTec* : <http://www.amicus-transtec.com/> (services de localisation de logiciel pour l'industrie de la traduction, logiciel de traduction, logiciel de gestion du contenu...)
    - *Ta with you* : <http://www.tauyou.com>  
(logiciel comme service : TA adapté aux besoins de l'utilisateur, technologie statistique)
  - Outils d'aide à la traduction et mémoires de traduction : utilisés par environ 80% d'entreprises qui produisent des documentations multilingues (Trados, Déjà Vu, OmegaT, Star Transit, Wordfast...)
- Besoins actuels :
  - Traduction rapide : applications pour Internet, courrier électronique, sites Web, logiciels de discussion instantanée...
  - TA (texte/parole) sur applications portables (téléphone, PDA, iPhone...)
  - Combinaison avec l'image et la vidéo (InterAct...)

---

... pour quel futur ?

- Société de plus en plus multilingue :
  - Union Européenne, gestion de documentation, frontières ouvertes, actions humanitaires,...
- Société de plus en plus multimodale :
  - Outils vocaux, visioconférences...
  - Applications portables
  - Challenge technologique
- Société de plus en plus communicante :
  - Grand échange d'informations via internet
  - Discussion instantanée
- Vers de nouvelles méthodes ?
  - TA par document entier plutôt que phrase à phrase ?

- TA utilisable par une société commerciale ?
- Vrais besoins ? En fonction de l'objectif :
  - Outils d'aide à la traduction ? Mémoires de traduction ?
  - Logiciels de traduction automatique adaptés sur demande ?
  - Besoins en ressources grandissants : corpus parallèles, mais aussi corpus comparables...
- TA intégrée dans d'autres applications ?
  - Ex. aide médicale aux non-parlants
- Evaluation
  - Comment estimer la qualité ?
  - Important pour le chercheur et pour l'acheteur/utilisateur

*Merci pour votre attention*