Traduction automatique et occitan

Gérard Ligozat

CEROC, 11 mai 2011

Plan de l'exposé

1. Introduction Historique : 50 ans de traduction automatique
2. L'approche statistique
3. Application à l'occitan
4. Conclusion

Introduction

Introduction

Quelques termes :

- traitement automatique des langues (TAL, TALN)
  - messagerie, réservation de places, analyse de contenu, résumé automatique, système de question/réponse,
  - etc.
- reconnaissance de la parole
  - oral $\rightarrow$ écrit
- traduction automatique
  - écrit $\rightarrow$ écrit
  - oral $\rightarrow$ oral etc.

Introduction

- Idées reçues :
  - traduction automatique : mécanique, mot-à-mot
  - traduction humaine : créativité, intelligence
- La TA comme outil expérimental :
  - étude des résultats de la TA $\rightarrow$ hypothèses
  - test automatisé des hypothèses
  - mise en évidence de phénomènes qui ne se réduisent pas aux hypothèses faites

Introduction

Deux grandes tendances en TAL :

- Méthodes symboliques :
  - à bases de règles < intuition / introspection du chercheur
  - l'avantage de la machine est sa bêtise : si on lui donne une règle, elle l'utilise
  - bon diagnostic des règles
  - le nombre de cas traités est arbitrairement grand
- Méthodes statistiques :
  - à partir de données (*data-based*)
  - la machine peut apprendre à partir d'exemples
  - corpus d'apprentissage / corpus de travail

Introduction

Le plus grand ennemi du chercheur en TAL : l'ambigüité

- grammaticale : *light* N, V, A ?
- syntaxique : *donner à manger à la fenêtre / aux moineaux*
- sémantique : *voler*
- référentielle : *Elle tire la chevillette et le loup la mange*

Historique de la TA

Les débuts de la TA : le décryptage

- Seconde guerre mondiale en Grande-Bretagne : utilisation d'ordinateurs pour casser le code Enigma
- La TA vue comme un problème de décryptage Warren Weaver (1947-1949) :
  When I look at an article in Russian, I say : "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."
- Principe du décryptage : méthode du canal bruité

Modèle du canal bruité

Traduction du russe à l'anglais selon le modèle de Weaver :
la traduction est vue comme rétablissement du message initial avant brouillage

Les années 1950 : l'expérience de Georgetown

- Grandes promesses, crédits importants
- Expérience de Georgetown (1954)
  - 49 phrases russes traduites en anglais
  - 250 mots, 6 règles
  - prématuré / connaissance syntaxiques ?
  - Chomsky : Syntactic structures, 1957 ; Aspects . . . , 1959
- Financement pour des motifs stratégiques et militaires (guerre froide)
- 1957 : Spoutnik "Un Pearl Harbor technologique"

1957 : le Spoutnik

A postcard of 1958 commemorating satellite launches

- Lancement des deux premiers satellites artificiels
- Traduction du russe en anglais
- Veille scientifique

Les années 1960 : le rapport ALPAC

- Optimisme entretenu par l'expérience de Georgetown ; espoir : de 3 à 5 ans
- Demeure la question de la désambigüisation sémantique
  - avocat $\rightarrow$ *avocado* / *lawyer*
  - *the spirit is strong, but the flesh* is weak
- Le rapport ALPAC (1966) (Automatic Language Processing Committee)

Conclusions du rapport ALPAC

- La post-édition d'une sortie de TA n'est pas moins coûteuse / pas plus rapide que la traduction humaine
- Les EU dépensent moins de 20 M de dollars pour la traduction
- Seule une petite partie des publications scientifiques russes mérite d'être traduite
- On dispose de suffisamment de traducteurs humains
- Favoriser la recherche linguistique théorique et améliorer les méthodes de traduction humaine
- Abandon de nombreux projets de recherche dans le domaine aux EU
- En France (Bernard Vauquois, Grenoble), au Canada, poursuite de travaux

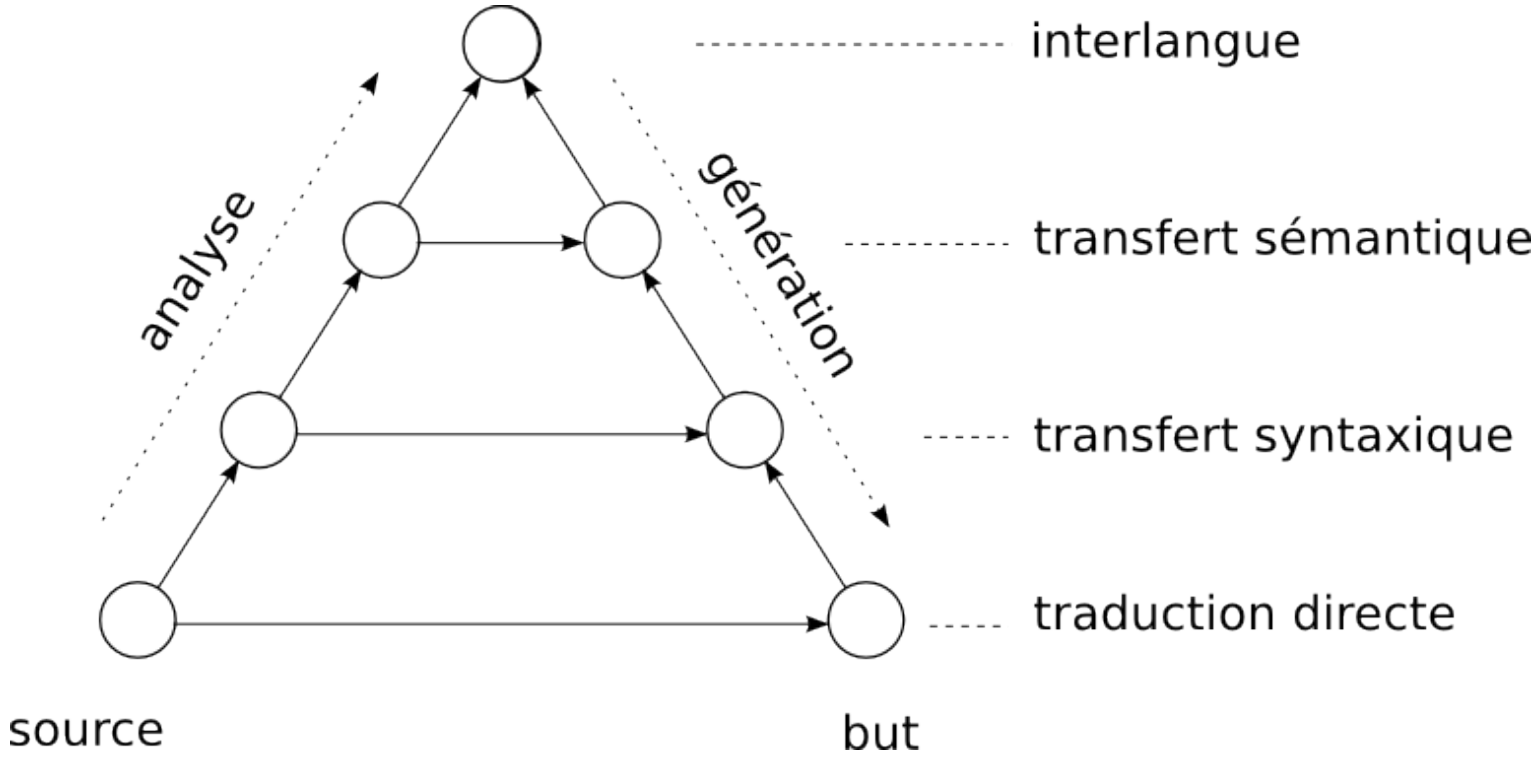
Les années 1970

- En dépit du rapport ALPAC, développement de systèmes utilisant l'approche symbolique
- Système METEO pour traduire les bulletins météo, Montréal, à partir de 1976
- Fondation de **SYSTRAN** en 1968
  - russe → anglais utilisé par US Air Force depuis 1970
  - acheté par la Communauté Européenne en 1976 français → anglais

Deux approches du TAL : symbolique / statistique

- Approche symbolique, souvent à base de règles
- Utilise des connaissances linguistiques (morphologie, syntaxe) et/ou sémantiques
- exemple typique : systèmes utilisant grammaires à la Chomsky : LFG, tree grammars, HPSG, unification / patrons (patterns)
- Avant les années 1990, l'approche statistique est utilisée exclusivement pour la reconnaissance de la parole

Le triangle de la traduction



Les années 1980

- 1980 : Logos, METAL Pan American Health Association, Washington, espagnol $\rightarrow$ anglais (depuis 1970)
- Au Japon, 1980-1990, nombreux projets japonais $\rightleftarrows$ anglais
- Projet Eurotra, 1984-1988
  - interlangue, pour les 9 langues de la CE à l'époque

Le tournant des années 1990

- De manière générale en TAL, développement d'une nouvelle approche statistique inspirée par les succès des méthodes statistiques pour la compréhension de la parole
- Projets et méthodes IBM

Méthodes statistiques en TALN

- A la fin des années 1980, IBM Research : les modèles statistiques sont très efficaces pour la reconnaissance de la parole (speech recognition)
- Projet Candide : traduction comme problème d'optimisation statistique
- 1990-2000 prématuré :
  - l'approche syntaxe & interlangue continue de dominer
  - Chercheurs d'IBM $\rightarrow$ Wall street

Méthodes statistiques en TALN

- Depuis 2000 :
  - 1998, un atelier (*workshop*) à la John Hopkins U. : des chercheurs reprogramment les méthodes d'IBM
  - Des outils statistiques deviennent accessibles aux chercheurs
  - DARPA finance des projets de traduction statistique (TIDES, GALE)
  - 11 septembre 2001 : intérêt nouveau pour la traduction arabe $\rightarrow$ anglais
- Technique : puissance de calcul et capacités de stockage accrues

La TA aujourd'hui

- Milieu universitaire : projets, ateliers
- Un problème central : l'évaluation
- Textes parallèles / textes comparables
- Commerce : IBM, Google Méthodes hybrides : intégration de méthodes statistiques dans Systran
- Internet : 50 M de pages par jour sont traduites (Google, Yahoo, Microsoft)
- Campagnes internationales d'évaluation chaque année : NIST

Applications de la TA

- Assimilation : traduire pour comprendre (veille)
- Diffusion : traduire pour publier
- Communication, échanges (email, chat, etc.)
- Accès aux bases de données
- Chaque fonction demande des vitesses et qualités particulières

La TA : rêve et réalité

- Le rêve : la FAHQMT (traduction entièrement automatisée de grande qualité) concevable pour :
  - Domaines délimités (météo, informations horaires train, auto, prise de rendez-vous)
  - Langues contraintes (par exemple documentation compagnies internationales)
- La réalité :
  - *gisting* (donner l'idée, le principal) : services de Systran pour Yahoo, Google, 500 M de mots par jour
  - Intégration parole / écrit (téléphone), intégration reconnaissance de la parole / TA statistique relativement aisée : méthodes communes
  - TA sur téléphones portables intelligents (smartphones) :
    - Sous-langages (icônes, menus, horaires de train) avec reconnaissance optique de caractères (OCR)

La TA : rêve et réalité

- Post-édition : la TA comme aide pour le traducteur humain
- Systèmes de *mémoires de traduction* pour les traducteurs professionnels
- Langues romanes
  - Systèmes libres (*open source*)
  - Alicante, Donostia : espagnol, catalan, galicien
  - Équipe Transducens, Internostrum (espagnol/catalan), Traductor Universia (espagnol/portugais) (traducteurs à états finis, modèles de Markov cachés (HMM), recherche de morceaux (*chunking*) par des techniques finies)

Approches statistiques

Approches statistiques

- Principe : apprentissage par la machine :
  - on entraîne le modèle sur un grand corpus pour fixer les paramètres
  - on le teste
- Pour la TA, utilisation de grands corpus bilingues

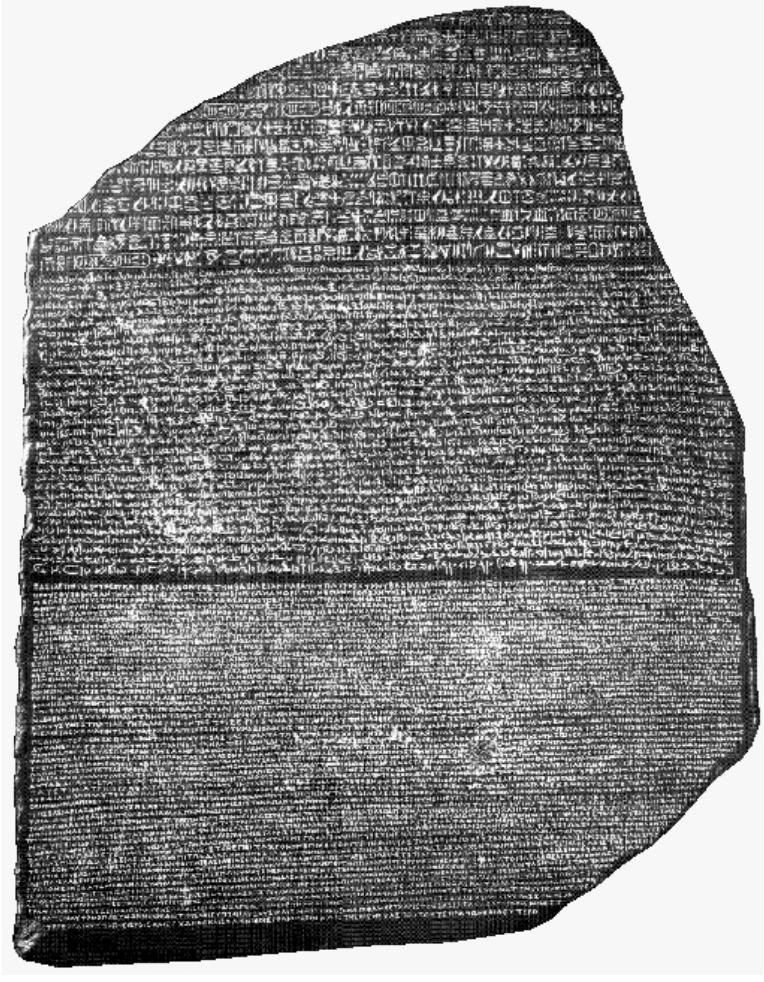
Approches statistiques

- Grands corpus bilingues
  - pour le couple anglais /français, Hansards canadiens : délibérations du Parlement
  - Hansards de Hong Kong, anglais et chinois
  - Corpus fournis par les Nations Unies

Approches statistiques

- Pour la traduction automatique, l'approche statistique utilise des textes parallèles (TP)
- Le traitement automatique de TP est centré sur l'opération d'**alignement**
  - de phrases
  - de mots
  - d'expressions
- Idée : un fragment du texte en L1 correspond à un fragment du texte en L2

Un texte parallèle : la pierre de Rosette



- égyptien en deux graphies :
  - hiéroglyphique
  - démotique
- grec

Un texte parallèle occitan / français : Roland Pécout, Portulan

De ton ajocador auses de bruchs, de rises. Davalas. Mostrar que siás aquí. Reintegrar ton tu. Cubrir teis uelhs d'un agach espinchable. Una còla de pichons, bralhas mascaradas, morets. Te fan la saludada. Respondes a cadun, que cadun es cadun. Un pichonet te pren la man e compta lei dets, puei s'escacalassa. Lei grands volon de cigarretas.

Perché, tu entends des bruits, des rires. Tu descends. Montrer que tu es là, te réintégrer. Te vêtir de regard regardable. Une bande de gamins, aux vêtements déchirés, te salue. Tu réponds à chacun, car chacun est chacun. Un petit te prend la main et compte les doigts, puis éclate de rire. Les grands veulent des cigarettes. (trad. de M.-J. Verny)

Alignement de phrases

- Un extrait de R. Pécout, traduit par M.-J. Verny :

De ton ajocador auses de bruchs, de rises.	Perché, tu entends des bruits, des rires.
Davalas	Tu descends
Mostrar que siás aquí.	Montrer que tu es là, te réintégrer.
Reintegrar ton tu.	
Cubrir teis uelhs d'un agach espinchable.	Te vêtir de regard regardable.

- Ici une correspondance (2:1) : deux phrases de l'occitan correspondent à une seule phrase du français.

Alignement de phrases

- En général, *m* phrases dans L1 peuvent correspondre à *n* phrases dans L2 : on parle de (m : n)
- La plupart des systèmes partent de l'hypothèse que (1:1) est majoritaire
- Omissions : (1 : 0) et ajouts (0 : 1)
- On considère rarement le cas où *m* et *n* sont supérieurs à 2 (fusions complexes)

Alignement de mots

	perché	tu	entends	des	bruits	des	rires
rises							
de							
bruchs							
de							
auses							
ajocador							
ton							
de							

perché (1,2,3) tu (4) entends (4) des (5) bruits (6) des (7) rires (8)

Utilisation du canal bruité

- Exemple : traduire de l'oc au fr
- On considère que l'oc O provient d'une source F déformée par le canal, il s'agit de trouver la meilleure approximation F1 de F.
- Maximiser la probabilité que l'occitan O provienne du français F.

Utilisation du canal bruité

Pour calculer cette probabilité, on a besoin de déterminer :

- la fréquence d'un mot donné F en français ;
- la fréquence de chacune des traductions en occitan de ce mot.

Application à l'occitan

TP occitans : un corpus considérable

- Constatation : une grande partie de la littérature moderne en occitan est en fait constituée de textes parallèles.
- XIXème siècle : beaucoup de textes bilingues (en particulier, Félibrige) exemple : Mirèio, écrit en occitan, et traduit en français par Mistral
- Depuis 1945, textes souvent publiés en occitan seulement, mais traduits en français ultérieurement

Exemple 1 : Mirèio, Cant segound

Cantas, cantas, magnanarello !
Que la culido es cantarello !
Galant son li magnan e
s'endormon di tres ;
Li amourié soun plen de fiho
Que lou bèu tèms escarrabiho,
Coume un vòu de blóundis abiho
Que raubon sa melico i roumanin
dóu gres.

Chantez, chantez,
magnanarelles !
- Car la cueillette aime les chants.
Beaux sont les vers à soie, et ils
s'endorment de leur troisième
somme ;
Les mûriers sont pleins de jeunes
filles
Que le beau temps rend alertes et
gaies,
Telles qu'un essaim de blondes
abeilles
Qui dérobent leur miel au
romarin des champs pierreux.

Exemple 2 : Fabre d'Olivet, Le troubadour

Chaûmès pas-mai ;
hoû !vendemiaires,
hoppo ! dépés ! léû al trabal ;
vous espéroun lous bouteilhaires,
acò's proù chourat, avén sal
prenéz banastas é desquetas,
levadoùs, semals e tinetas.
L'un, qu'acampe ; l'autre
qu'estuja ;
resto plus-rès dins lou tinel,
que la rapuga tout'essuja ;
lou trouilhaire espéro un faissel.

Ne tardez pas d'avantage ; holà ! vendangeurs, debout ! allons, vite, à l'ouvrage ; les pauvres grappilleurs sont là qui vous attendent ; allons, c'est assez prendre du repos ; le travail vous rappelle. Et tôt, prenez ces larges marines d'osier, ces petites corbeilles, ces cornues **propres à conserver le jus de raisin**.

Que l'un cueille le fruit de la vigne, tandis que l'autre l'entassera dans les cornues ; il ne reste bientôt plus rien, dans la cuve à fouler, que la grappe dépouillée, **et le marc aride** ; le fouleur attend une nouvelle cuvée.

Alignement de textes occitans

- Projet MULTEXT-Cataloc (1997), alignement (phrastique) de deux textes :
  - *Lo vesitaire dis Estèllo*, de Rémi Blancon ;
  - *Mirèio*.
- Interrogations sur l'opportunité d'alignement entre langues des statuts sociolinguistiques différents
- Intérêt de l'étude systématique des décalages entre texte occitan et texte français

La Ligosada : une farce en vivaro-alpin (Gap)

- Auteur : l'abbé Borel, curé de Romette, pseudonyme Rob d'Ettemor ;
- Publiée (partiellement) dans le Bulletin de la Société d'Etudes des Hautes-Alpes à partir de 1889, puis sous forme de livre.
- Examen d'un passage
  - six vers occitans, dix français ;
  - registre pseudo-noble en français, ajouts *les sombres voiles*, mais *ça presse* $\rightarrow$ burlesque ;
  - connivence avec le lecteur qu'on amuse.

La Ligosada

Ou ciel, lou lendemà, lusié
mai d'ena estièra,
Quand Lisa se levèc (car èra
matinièra).
Se beté vite à criar : –
Léva-të, moun ami ;
Gros feniant ! es qu'as pas
encara prou durmi ?
Siés struina couma un vèu,
sensa te far de bila ;
Anën, despacha-të, chau
partir par la vila.

La Ligosada

Ou ciel, lou lendemà, lusié mai d'ena estièra,
Quand Lisa se levèc (car èra matinièra).
Se beté vite à criar : –
Léva-të, moun ami ;
Gros feniant ! es qu'as pas encara prou durmi ?
Siés struina couma un vèu, sensa te far de bila ;
Anën, despacha-të, chau partir par la vila.

Au ciel, le lendemain, sous un ciel plein d'étoiles,
De la nuit s'étendaient encor les sombres voiles,
Quand Lise se leva, (matinière elle était),
Et vite à plein gosier à Jean elle criait :
Lève-toi, fainéant, chasse donc la paresse ;
C'est assez de sommeil, lève-toi, car ça presse.
N'as-tu pas honte, Jean, d'être là comme un veau,
Couché de tout ton long ? allons, allons, mon beau,
Tu ne te fais pas bien, je le vois, de la bile ;
Hâte-toi, car il faut t'en aller à la ville.

La Ligosada : alignement par phrase

- Utilisation d'un algorithme classique d'alignement par phrase, celui de Gale et Church ;
  - parmi les paramètres de l'algorithme, choix de marqueurs principaux (., ?,!) et secondaires (;, :, ,)
  - très peu de cognats utilisables.
- Résultats inattendus :
  - l'alignement n'est pas majoritairement de type (1:1) ;
  - de nombreux cas n'entrent pas dans une correspondance (m:n) stricte : Chevauchements, effacements partiels, insertions.

La Ligosada : alignement par mots

- Programmes qui prennent en entrée un texte aligné par phrase, et fournissent un alignement par mots.
- Test de deux types de programme :
  - Giza++, qui implémente des algorithmes IBM, basé sur les mots fréquents ;
  - Anymalign (Adrien Lardilleux), qui utilise des hapax que l'on fait apparaître par échantillonnage (choix de phrase au hasard dans le corpus)
- Travail en cours

Conclusion

Conclusion : utilisation des corpus parallèles

- mise en relief des différences entre textes sources et traductions
- mise en relief des différences entre textes provenant de locuteurs natifs et non-natifs ;
- tâches comparatives augmentant notre connaissance des caractéristiques typologiques et culturelles propres à chaque langue, ainsi que des caractéristiques universelles ;
- nombreuses applications pratiques : lexicographie, enseignement des langues, traduction.

Conclusion : utilisation des corpus parallèles

- Résolution de problèmes de traduction, par constitution de mémoires de traduction ;
- Création et maintenance de textes parallèles ;
- Vérification de traductions, levée d'ambigüité ;
- Correction d'erreurs ;
- Mise en évidence d'idiotismes syntaxiques et lexicaux ;
- Détection des suites préfabriquées.

Conclusion : utilisation des corpus parallèles

- Question de stylistique, phraséologie ;
- Marques de la situation diglossique :
  - dans les choix lexicaux, syntaxiques, stylistiques ;
  - discours implicite tenu au lecteur ;
  - connivence et/ou distanciation ;
  - mise en scène du folklorique, de la différence.

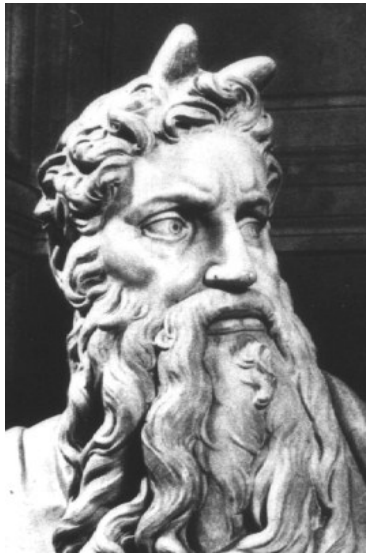
Conclusion : Langues moins répandues

- pour le traitement statistique, nécessité de disposer de corpus de grande taille
- fouille automatique de textes dans ces langues sur Internet
- développement de techniques adaptées aux corpus réduits

Conclusion générale

- Utilisation de méthodes automatiques pour faire apparaître des phénomènes linguistiques de manière :
  - systématique ;
  - quantifiable ;
  - non biaisée.

Traduction humaine . . .



Une erreur de traduction immortalisée par Michel-Ange :
. . . *et ignorabat quod cornuta esset facies sua*
. . .
(Saint Jérôme, Vulgate, Exode 34:29)

Joan Roqueta-Larzac : . . . *sabiá pas, el, que de la pèl de sa cara banejava de lum*