“Les principes de la traduction automatique statistique”

Auteur: François Yvon (LIMSI, CNRS)

Traduction humaine / Bio-traduction

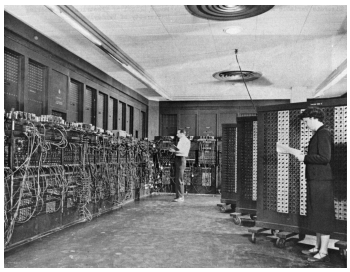
- Haute qualité
- Coûteuse
  - Expertise rare
  - Formation longue
  - Délais

Traduction humaine outillée (CAT)

- Réutilise au mieux les traductions passées
  - Recopie des « matchs » exacts
  - Adaptation des « matchs » partiels
  - Traduction humaine en cas de besoin
- Qualité des mémoires de traduction (TM)
- Idéal: une TM par type de document
- TM évolutives, collaboratives, etc

La traduction « automatique » (TA)

- Instantanée
- Disponible 24h/24
- Traite du texte tout venant
- Coût dérisoire
  (voire gratuit)
- Qualité imprévisible & incontrôlable
  (du médiocre, du grotesque, du très bon)



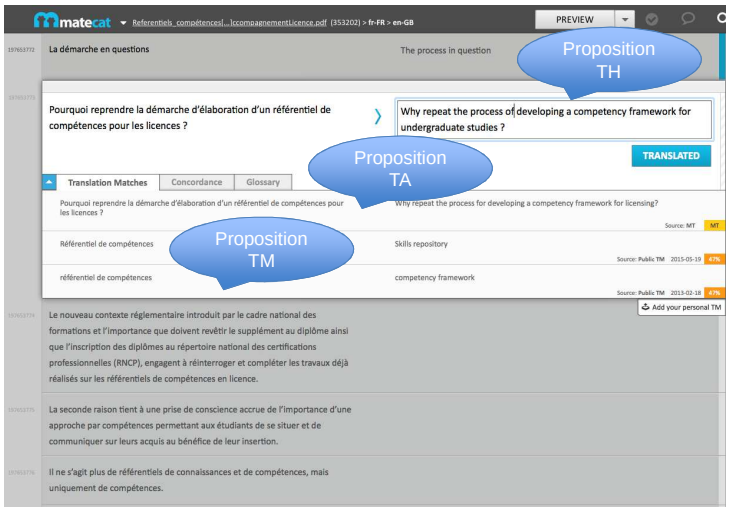
Le meilleur des deux mondes ?

- Tâches répétitives, non solvables: TA
  (faible valeur ajoutée)

La traduction statistique amplifie l'utilisation des TM:
- décompose les segments
- multiplie les « matchs »

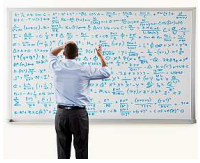
- Tâches créatives : HT, CAT
  (haute valeur ajoutée)

Une hybridation possible

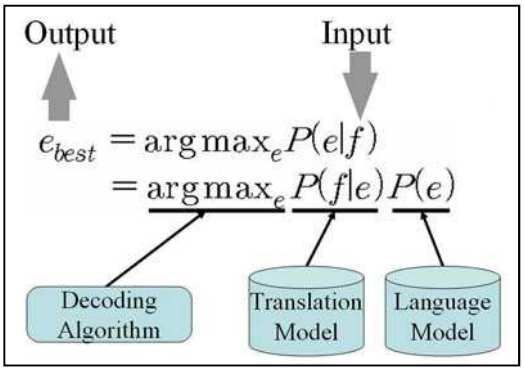


Un long chemin vers la TA statistique

- TA 1960-1990
  - Formulation de règles, systèmes experts
  - « *Logiciellisation* » du problème (sous-tâches, modules, tests)
- Constat: La TA est « *AI – complete* »
- TA Statistique 1990-
  - Construit des modèles (**approximations**)
  - A partir de **données**
  - Solution imparfaite:
    - Intégration dans les WF de traduction
    - Pré / post-édition



Principes de la TA statistique



« La meilleure traduction (e_{best}) est celle qui est la plus probable parmi toutes les traductions (e) possibles de la phrase d'entrée (f). »

La qualité de e dépend de deux termes:
- P(f | e) *: f et e sont bien appariés*
- P(e) *: e est une phrase bien formée*

En TAS ces termes sont appris à partir de données

Deux sources de connaissance

Les données pour la TAS:

- Traductions humaines (bitextes)
- Textes en langue cible

- De grande qualité
- En grande qualité

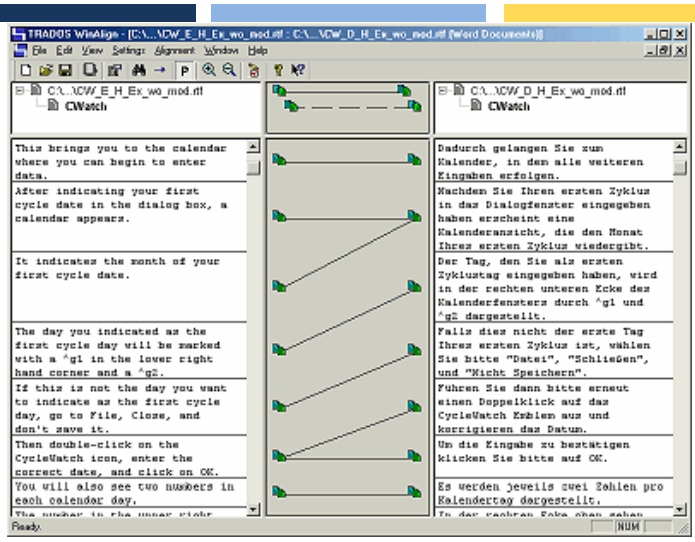
GERMAN	ENGLISH	FRENCH
Einleitung	Introduction	Introduction
I. Von dem Unterschiede der reinen und empirischen Erkenntnis	*I. Of the difference between Pure and Empirical Knowledge*	*I. De la différence de la connaissance pure et de la connaissance empirique.*
Daß alle unsere Erkenntnis mit der Erfahrung anfange, daran ist gar kein Zweifel; denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verstandestätigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenstände zu verarbeiten, die Erfahrung heißt? Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.	That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare to connect, or to separate these, and so to convert the raw material of our sensuous impressions into a knowledge of objects, which is called experience? In respect of time, therefore, no knowledge of ours is antecedent to experience, but begins with it.	Que toute notre connaissance commence avec l'expérience, cela ne soulève aucun doute. En effet, par quoi notre pouvoir de connaître pourrait-il être éveillé et mis en action, si ce n'est par des objets qui frappent nos sens et qui, d'une part, produisent par eux-mêmes des représentations et, d'autre part, mettent en mouvement notre faculté intellectuelle, afin qu'elle compare, lie ou sépare ces représentations, et travaille ainsi la matière brute des impressions sensibles pour en tirer une connaissance des objets, celle qu'on nomme l'expérience? Ainsi, chronologiquement, aucune connaissance ne précède en nous l'expérience et c'est avec elle que toutes commencent.

Exploitation des bitextes

- Identification des phrases « parallèles » par alignement
- Identification d'équivalences lexicales: alignement de mots + probabilités de traduction
- Régularités syntaxiques de surface: modèles de langue

GERMAN	ENGLISH	FRENCH
Einleitung	Introduction	Introduction
I. Von dem Unterschiede der reinen und empirischen Erkenntnis	*I. Of the difference between Pure and Empirical Knowledge*	*I. De la différence de la connaissance pure et de la connaissance empirique.*
Daß alle unsere Erkenntnis mit der Erfahrung anfange, daran ist gar kein Zweifel; denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verstandestätigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenstände zu verarbeiten, die Erfahrung heißt? Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.	That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare to connect, or to separate these, and so to convert the raw material of our sensuous impressions into a knowledge of objects, which is called experience? In respect of time, therefore, no knowledge of ours is antecedent to experience, but begins with it.	Que toute notre connaissance commence avec l'expérience, cela ne soulève aucun doute. En effet, par quoi notre pouvoir de connaître pourrait-il être éveillé et mis en action, si ce n'est par des objets qui frappent nos sens et qui, d'une part, produisent par eux-mêmes des représentations et, d'autre part, mettent en mouvement notre faculté intellectuelle, afin qu'elle compare, lie ou sépare ces représentations, et travaille ainsi la matière brute des impressions sensibles pour en tirer une connaissance des objets, celle qu'on nomme l'expérience? Ainsi, chronologiquement, aucune connaissance ne précède en nous l'expérience et c'est avec elle que toutes commencent.

Alignement des phrases



Alignement des mots

			Sm.	Lg.
		CLASSIC SOUPS		
清燉雞	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯	58.	Chicken Rice Soup	1.85	3.25
雞麵	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃茄蛋	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞	62.	Regular Wonton Soup	1.10	2.10
酸辣	63.	Hot & Sour Soup	1.10	2.10
蛋	64.	Egg Drop Soup	1.10	2.10
雲 蛋	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜	66.	Tofu Vegetable Soup	NA	3.50
雞玉米	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮	69.	Seafood Soup	NA	3.50

Alignement des mots

			CLASSIC SOUPS	Sm.	Lg.
清 燉 雞 湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)		1.50	2.75
雞 飯 湯	58.	Chicken Rice Soup		1.85	3.25
雞 麵 湯	59.	Chicken Noodle Soup		1.85	3.25
廣 東 雲 吞	60.	Cantonese Wonton Soup		1.50	2.75
番 茄 蛋 湯	61.	Tomato Clear Egg Drop Soup		1.65	2.95
雲 吞 湯	62.	Regular Wonton Soup		1.10	2.10
酸 辣 湯	63.	Hot & Sour Soup		1.10	2.10
蛋 花 湯	64.	Egg Drop Soup		1.10	2.10
雲 吞 湯	65.	Egg Drop Wonton Mix		1.10	2.10
豆 腐 菜 湯	66.	Tofu Vegetable Soup		NA	3.50
雞 玉 米 湯	67.	Chicken Corn Cream Soup		NA	3.50
蟹 肉 玉 米 湯	68.	Crab Meat Corn Cream Soup		NA	3.50
海 鮮 湯	69.	Seafood Soup		NA	3.50

Alignement des mots

- Principes sous-jacents
  - Cooccurrences lexicales bilingues
  - Contraintes de localité (proximité source => proximité cible)
  - Contraintes de fertilité (chaque mot cible ne traduit que peu de mots source – et réciproquement)
- Ne connaît que les mots du corpus
- Problème difficile pour l'humain, mal posé pour la machine
- Résultat souvent imparfait

Traduction statistique

I love the boy.
J'aime le garçon.
I love the dog.
J'aime le chien.
They love the dog.
Ils aiment le chien.
They talk to the girl.
Ils parlent à la fille.
They talk to the dog.
Ils parlent au chien.
I talk to the mother.
Je parle à la mère.

Aligned Data

Traduction statistique

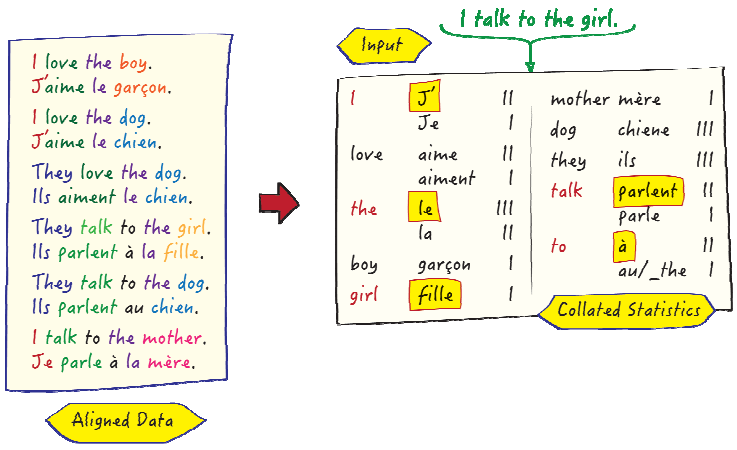
I love the boy.
J'aime le garçon.
I love the dog.
J'aime le chien.
They love the dog.
Ils aiment le chien.
They talk to the girl.
Ils parlent à la fille.
They talk to the dog.
Ils parlent au chien.
I talk to the mother.
Je parle à la mère.

Aligned Data

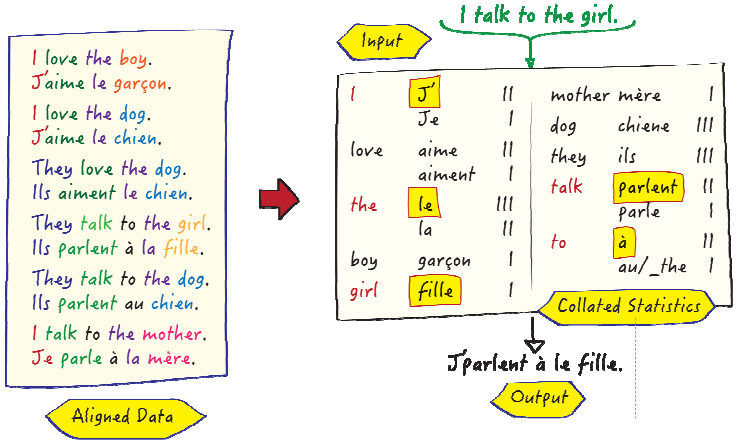
I	J'	II	mother	mère	I
	Je	I	dog	chiene	III
love	aime	II	they	ils	III
	aiment	I	talk	parlent	II
the	le	III		parle	I
	la	II	to	à	II
boy	garçon	I		au/_the	I
girl	fille	I			

Collated Statistics

Traduction statistique



Traduction statistique



Traduction statistique

I love the boy.
J'aime le garçon.
I love the dog.
J'aime le chien.
They love the dog.
Ils aiment le chien.
They talk to the girl.
Ils parlent à la fille.
They talk to the dog.
Ils parlent au chien.
I talk to the mother.
Je parle à la mère.

Aligned Data

I	talk	to	the	girl
J'	parlent		au	le fille
2/3	2/3	1/3	2/5	1/1
Je	parle	à	la	fille
1/3	1/3	2/3	2/5	1/1

Comment choisir la meilleure?

Traduction statistique

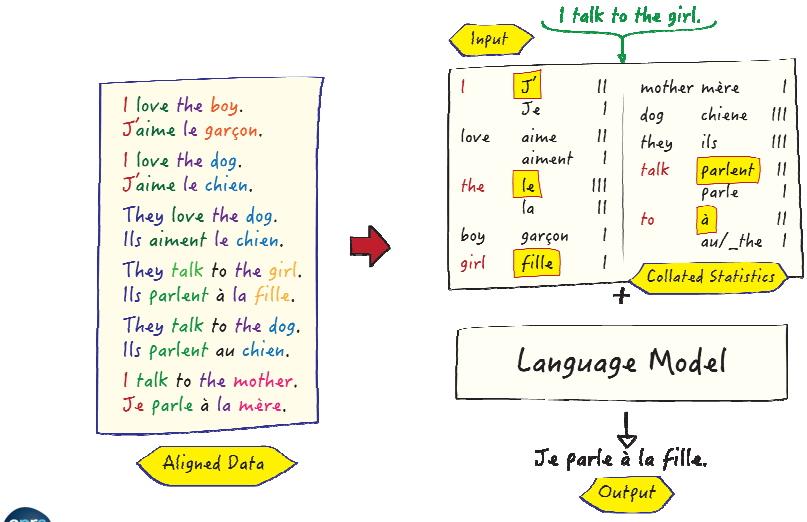
I love the boy.
J'aime le garçon.
I love the dog.
J'aime le chien.
They love the dog.
Ils aiment le chien.
They talk to the girl.
Ils parlent à la fille.
They talk to the dog.
Ils parlent au chien.
I talk to the mother.
Je parle à la mère.

Aligned Data

Le modèle de langue

- Contraintes de surface en cible
- Suites de mots possibles + leurs probabilités
- Appris sur des corpus monolingues
  - *Je parle* : OK
  - *J' parlent* : KO
  - *la fille* : OK
  - *le fille* : KO
- *Je parle à la fille >> J' parlent à le fille*

Traduction statistique

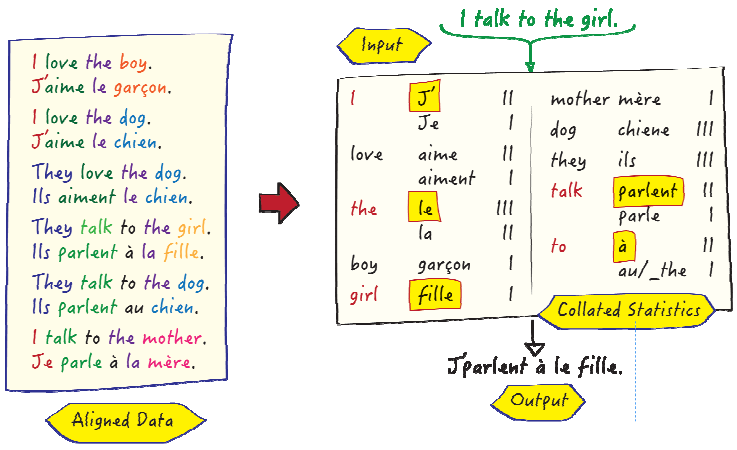


Traduire : des mots aux segments

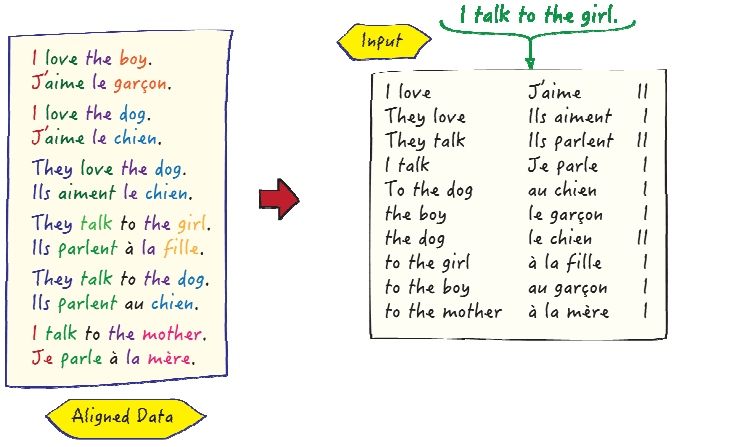
- Traduire des mots isolés est risqué
  - perte du contexte source;
  - erreurs d'accord (*le fille* …),
  - idiomes, mots-composés, termes, etc.
- La « fluidité » repose sur le modèle de langue
- Bien meilleur
  - Préférer des segments longs
  - En dernier ressort : mots isolés

the girl : la fille
to the girl : à la fille
I talk : Je parle

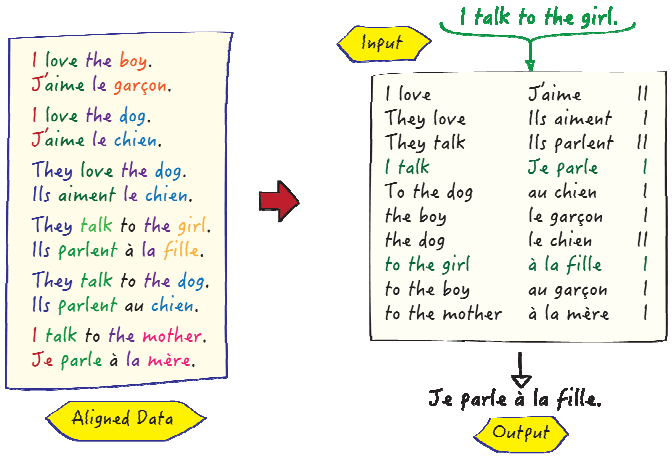
TAS« à base de mots »



TAS « à base de segments »



TAS « à base de segments »



TAS « à base de segments »

- Traductions visiblement meilleures
- Technologie mature & efficace: Google, Microsoft, Baidu, Logiciels professionnels (CAT, TA)

- Implémentation publique: Moses
- Système le plus utilisé
- Développé par des projets européens (2007-)
- Utilisé dans MT@EC (DGT)

MOSES CORE

Dépasser l'état de l'art

- Meilleures analyses linguistiques
  - Syntaxe en langue cible
  - Prise en compte de la morphologie
  - Traduction « sémantique »
  - Modélisation du discours
- Meilleurs modèles statistiques
  - Multiplication des modèles
  - Réseaux neuronaux profonds
- Meilleure utilisation des données
  - Pivot, transfert entre langues
  - Sélection des corpus
- Progression régulière … mais trop lente
- Amélioration garantie: plus de données, de meilleure qualité !

La place des données en TAS

- La TAS est exclusivement à base de données
- Connaissance extraite de grands corpus
  - Correspondances bilingues (données parallèles)
  - Régularités de surface (textes cibles)
  - Peut inclure: dictionnaire, terminologies bilingues, etc.
- Limitations majeure:
  - Qualité de la TA dépend de la « qualité » des données
  - Pas de données : pas de TA

Qualité des bitextes

Plus de résultats

But always remember: garbage in, garbage out! No conversion will make a bad picture good.	Mais souvenez -vous, mauvaises données à l'entrée, mauvais résultats à la sortie. Aucune conversion ne rendra belle une image de mauvaise qualité.
Like they say, garbage in, garbage out.	C'est un marché en pleine expansion.
"Garbage in, garbage out."	"À données inexactes, résultats erronés".
If you put in garbage, you get garbage out.	Les résultats obtenus à partir de données fausses sont eux aussi faux.
If you put in garbage, you get garbage out.	Les résultats obtenus à partir de données fausses sont eux aussi faux.
However, the data provided to these bodies is not always of the highest quality and "if you punch in garbage you get out garbage".	Toutefois, les données communiquées à ces organismes ne sont pas toujours de la plus haute qualité, "ce qui est préjudiciable à la valeur du produit final".
I want this garbage out of here.	Je veux que ces détritus soient sortis d'ici.

Des segments vraiment parallèles

Qualité des bitextes

Je voudrais encore soulever brièvement deux problèmes concernant la déclaration.	I should like to conclude by dealing briefly with two problems relating to the declaration.
Nous publions ici la déclaration finale.	We reproduce here the Declaration adopted at that meeting.
Je condamne également la déclaration du ministre polonais.	I also condemn the statement made by the Polish Minister.
Mme Sriswasdi appuie la déclaration du Brésil.	Ms. Sriswasdi said that she supported the statement by Brazil.
Assigner un numéro à la déclaration de conformité est optionnel.	It is optional to assign a number to the declaration of conformity.
Une copie de la déclaration d'immatriculation.	A copy of the declaration of registration, if applicable.

Des exemples adaptés, représentant les « bons » sens

Qualité des bitextes

the tax return

L'évaluation finale par les autorités peut prendre jusqu'à trois ans à compter de la présentation de la déclaration d'impôts.	The final assessment by the authorities can take up to three years following the submission of the tax return.
L'évaluation finale par les autorités peut prendre jusqu'à trois ans à compter de la présentation de la déclaration d'impôts.	The final assessment by the authorities can take up to three years after submitting the tax return.
L'année fiscale court du 1er avril au 31 mars et la déclaration d'impôts doit être présentée avant le 30 novembre de l'année suivante.	The tax year runs from 1 April to 31 March and the tax return must be submitted by 30 November of the following year.
L'avantage procuré par ce régime devrait être calculé sur la base du montant déduit une deuxième fois du bénéfice net tel qu'enregistré dans la colonne spéciale de la déclaration d'impôts réservée aux activités éligibles.	The benefit of this scheme should be calculated on the basis of the amount deducted a second time from the net profit as recorded on the tax return in the special column for promoted activities.
Il ne figure pas sur la déclaration d'impôts du père.	He's not listed as a dependant on his father's tax returns.

Des exemples adaptés, fournissant des longs « matchs »

Les données pour CEF.AT

- Pour délivrer des services multilingues de qualité pour tous les citoyens et les administrations publiques…
- CEF.AT demande des données de qualité:
  - DGT, Agences européennes, ONU, etc.
  - Administration centrale, agences et établissements publics, ONG, etc.
- En grande quantité

European Language Resource Coordination
Connecting Europe Facility

Merci de votre attention