

Seminar Internaţional „Instrumente pentru asistarea traducerii"
Bucureşti (România) — 28-29 februarie 2008

Séminaire international « Les outils d'aide à la traduction »
Bucarest (Roumanie) — 28-29 février 2008

Avant-propos
Program / Programme
Comunicări / Communications
Sponsorizare / Partenariat
Expozanţi / Exposants

Organizat de către / Organisation et coordination

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Sub înaltul patronaj / Sous le haut pratonage

Sub patronajul Comisarului european pentru multilingvism, Leonard Orban
/
Sous le patronage du commissaire européen pour le **multilinguisme, Leonard Orban**

Comunicări / Communications

Instrumentele de gestiune terminologică / Les outils de gestion terminologique

Conferinţă specială / Intervention spéciale

" Terminologie et travail en réseaux de langues "
Loïc Depecker, professeur, directeur de recherches Université de Paris Sorbonne

Principii de elaborare a unei baze de date terminologice multilingve în domeniul economic
Maria-Antoaneta Lorentz, conf. dr., ASE, Bucureşti

Un instrument flexibil şi eficient — gestionarul de terminologie
Dorina Chiş, Universitatea „Tibiscus" din Timişoara, Facultatea de Limbi Moderne Aplicate

CARTAGO : une terminologie large langues de l'enseignement électronique à distance, dans un contexte de co-élaboration multilingue de documents normatifs
HUDRISIER Henri, Maître de Conférences, Lab. Paragraphe, Université de Paris 8 –MSH Paris Nord, Liaison A de l'ISO/AUF / BEN HENDA Mokhtar, Maître de Conférences, Université Tunis et Bordeaux 3, Liaison A de l'ISO/AUF - Lab. Paragraphe, MSH Aquitaine

Metode de analiză lingvistică a textelor în limba română pentru extragerea terminologic
Luciana Peev, CS I, SC Software ITC, Cluj / Felicia Şerban, CS I, Institutul de Lingvistică şi Istorie Literară „Sextil Puşcariu"

eDTLR - Dicționarul tezaur al limbii române în format electronic
Marius Răschip, Facultatea de Informatică, Universitatea „Al.I. Cuza" Iaşi / Dan Cristea, Institutul de Informatică Teoretică, filiala Iaşi a Academiei Române / Corina Forăscu, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română, Bucureşti

Instrumentele de traducere automată sau asistată / Les outils de traduction automatique ou assistée

Conferinţă specială / Intervention spéciale

" Les technologies de la langue en soutien au multilinguisme "

Joseph Mariani, directeur, Institut pour le multilinguisme et le multimédia dans l'information

- *Sisteme de management al terminologiei şi traducerii asistate de calculator*
  Veronica Ştefan, conf. univ. dr., Universitatea Valahia din Târgovişte / Dumitru Ştefan, mat. drd., director general ATS Târgovişte

- *L'emploi des bases de données dans la traduction des brevets*
  Franco Bertaccini, profesor docent de terminologie şi limbaje sectoriale, director al Laboratorului de Cercetare Terminologică, Forlì / Claudia Lecci, Laboratorul de Cercetare Terminologică, Forlì

- *SDL TRADOS — Soluţia completă pentru traducătorii profesionişti*
  Cristian Cerghedean, coordonator Proiecte, altalingua EES / Lucian Dumitrescu, director asociat, altalingua EES

- *Traducerea asistată de computer a corespondenţei de afaceri engleze-franceze-române*
  Victoria Maxim, Universitatea Tehnică a Moldovei, Catedra Informatică Aplicată

- *TransitXV*
  Cristina Zaharia, Senior Project Manager, Star Information Services & Tools

- *La traduction assistée par ordinateur est-elle annonciatrice d'une mutation du travail du traducteur ?* [résumé]
  Marie-Josée de Saint Robert, JIAMCATT

- *Un Survol des Logiciels Trados, SDLX, Déjà Vu et WordFast de la perspective du traducteur freelance*
  Adina Popa, traductrice freelance

- *Enjeux et perspectives de la traduction automatique dans le domaine juridique*
  Daniela Dincă, Universitatea din Craiova

- Rezoluţia anaforei pentru traducere
  Mihai Alex Moruz, Facultatea de Informatică, Universitatea „Al. I. Cuza", Iaşi

Instrumentele de căutare (Internet, dicţionare electronice, baze de date etc.) / Les outils de recherche (Internet, dictionnaires électroniques, bases de données, etc.)

- *Conferinţă specială / Intervention spéciale*
  " Instrumentele de căutare - Resursele multilingve ale UE " / " Les outils de recherche Internet, dictionnaires électroniques, bases de données, etc. " [ro] [fr]
  Paulo Correia, DG - Département linguistique de langue portugaise

- *Un corpus paralel român-englez*
  Corina Forăscu, Facultatea de Informatică, Universitatea „Al.I.Cuza", Iaşi / Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română, Bucureşti

- *Resurse lexicale elaborate la Facultatea de Litere din Cluj-Napoca* [rezumat]
  Sanda Cherata, Universitatea Babeş-Bolyai

- *Baza de date terminologice a Centrului Naţional de Terminologie din Chişinău*
  Inga Druţă, director, Centrul Naţional de Terminologie

- *Baze de date lexicale: dicţionar morfologic, silabisitor, dicţionar de valenţe verbale*
  Ana Maria Barbu, CSII, Institutul de Lingvistică „Iorgu Iordan — Al. Rosetti"

Informatizarea limbii române (resurse informatice disponibile sau în curs de realizare)/ L'informatisation de la langue roumaine (ressources informatisées disponibles et en projet)

- *Conferinţă specială / Intervention spéciale*
  Sisteme de întrebare-răspuns în limbaj natural pentru spaţii de căutare

deschise
Dan Tufiş, director RACAI

- *Inteligenţa artificială — o tehnologie fiabilă pentru o infrastructură lingvistică* [rezumat]
Felicia Albescu, prof. univ.dr., director Institutul Naţional de dezvoltare Economică — Academia de Studii Economice

- *Realizarea inferenţelor textuale pe limba română*
Adrian Iftene, Research Assistant Drd., Facultatea de Informatică, Universitatea „Al.I.Cuza", Iaşi

- *Requêtes linguistiques sur alignements multilingues*
Sébastien Paumier, Dumitriu Marina, Université Paris-Est, Institut Gaspard-Monge

- *Construirea unui sistem de întrebare răspuns pentru limba română*
Adrian Iftene, Research Assistant Drd., Facultatea de Informatică, Universitatea „Al.I.Cuza", Iaşi

- *Portalul ConsILR - mijloc de informare asupra activităţilor de informatizare a limbii române* [rezumat]
Corina Forăscu, Facultatea de Informatică, Universitatea „Al. I. Cuza", Iaşi, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română, Bucureşti / Dan Cristea, Facultatea de Informatică, Universitatea „Al. I. Cuza", Iaşi / Ionuţ Pistol, Institutul de Informatică Teoretică, Academia Română, filiala Iaşi

- *Extraction de collocations monolingues et bilingues : application à la traduction*
Dan Ştefănescu, Alexandru Ceauşu, Amalia Todiraşcu, Ulrich Heid, Christopher Gledhill, François Rousselot

Sponsor de aur / Partenaire Or

Sponsor de argint / Partenaire Argent

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / Partenaire Bronze

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latină/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



Terminologie et travail en réseaux de langues

DEPECKER Loïc, Professeur, directeur de recherches, Université de Paris Sorbonne, Président de la Société française de terminologie

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Je suis très honoré de l'invitation à venir vous retrouver et très heureux d'être parmi vous. Dans cette conférence, je ferai part essentiellement des expériences de travail terminologique en réseaux de langues auxquels j'ai participé depuis 20 ans. Et des enseignements que nous pouvons tirer de ce travail en commun entre francophones et latinophones. Il s'agit principalement du RINT (Réseau international de néologie et de terminologie francophone) ; et de REALITER (Réseau international d'aménagement des langues latines), qui existe grâce à l'Union latine et auquel participe la Roumanie. Je terminerai par la question des outils de la terminologie, auxquels cette matinée est consacrée.

Pour définir brièvement la terminologie : c'est la science qui étudie les vocabulaires techniques et scientifiques. Aujourd'hui, son champ s'est étendu, en raison notamment des outils d'analyse du discours : la terminologie va au-delà des vocabulaires spécialisés et a tendance à englober les discours scientifiques et techniques.

Parmi les outils dont nous disposons pour travailler, il y a avant tout le **premier outil** qu'est la méthodologie. Pourquoi la méthodologie ? Parce que nous ne pouvons gérer les langues, surtout dans leurs dimensions techniques et scientifiques, qu'en commun, à plusieurs équipes de spécialistes. Et aussi parce qu'on a chacun une vision différente des langues. Il faut, pour travailler en commun :

- pouvoir s'entendre sur une conception semblable de la langue,
- pouvoir s'entendre sur ce que sont un terme, un synonyme, une définition, une catégorie grammaticale. Cela parait évident. Mais ce n'est pas le cas. Nous ne travaillons pas tous dans le même arrière-plan culturel, sous une même théorie, dans une même direction.

Généralement, soit on s'en tient à une théorie ; soit on va prendre ce qui nous intéresse chez Ferdinand de Saussure, Chomsky, Benveniste, Rastier, etc. C'est ce que nous faisons constamment en essayant d'être au plus près des avancées et des découvertes en linguistique, en sémantique, etc. Car c'est aussi par les linguistes que la théorie de la terminologie évolue et par là l'efficacité des méthodes et des applications. Cette approche plus proprement linguistique de la terminologie est récente, car la méthodologie en terminologie mise au point à l'ISO résulte d'un travail de plus de 50 ans sur la terminologie. À l'ISO, la terminologie a été et reste largement une science d'ingénieurs. Il y a deux grandes normes à regarder sur ce sujet de la méthodologie en terminologie :

- ISO 1087 : *Travaux terminologiques : vocabulaire* (2001)
- ISO 704 : *Travaux terminologiques : principes et méthodes* (2001)

Ces deux normes sont complémentaires :

- ISO 704 : *Travaux terminologiques : principes et méthodes* (2001) énonce les éléments de théorie de la terminologie. Cela va de la description de l'unité terminologique aux relations entre les concepts.
- ISO 1087 : *Travaux terminologiques : vocabulaire* (2001) formule la terminologie de la terminologie, c'est-à-dire qu'elle donne les définitions de toutes les catégories que nous employons pour mettre en place nos instruments de description de la terminologie. Ainsi, faut-il dire *notion* ou *concept* ? De quoi est constitué un concept ? De *caractères*, de

caractéristiques (*caracteristics* en anglais), d'autres choses ? Et l'*objet*, comment le décrire ? À chaque fois, on se situe dans une dimension différente. Cela a donné lieu à de très vastes débats. On voit donc qu'est inséparable de la méthodologie la théorie de la terminologie. C'est le **deuxième outil** dont je veux parler ce matin : la théorie.

De fait, les principes méthodologiques reposent sur une théorie. Sinon on ne fait rien. Et une théorie sans applications est vaine. On doit au moins considérer les trois éléments sur lesquels travaille la terminologie : les unités linguistiques, les concepts et les objets. Nous travaillons donc en terminologie sur au moins trois dimensions :

- la dimension linguistique
- la dimension conceptuelle
- la dimension de l'objet (dimension objectale).

On ne peut pas mélanger les trois, sinon on ne sait plus ce qu'on fait. Alors, est-ce que cela marche ? À l'évidence oui, car c'est ce schéma qui est appliqué depuis des années dans les entreprises.

Je disais tout à l'heure que la terminologie a tendance à englober les discours scientifiques et techniques. On rejoint ici la question fondamentale de la linguistique, à savoir celle du découpage des unités. On trouve donc dans les textes et les corpus, au moins (dans le domaine du bâtiment) :

- des unités terminologiques, exemple : *monteur de vitres* ;
- des unités phraséologiques, exemple : *poser doucement sur le sol* ;
- des unités traductologiques, exemple : *calcul et présentation des données climatiques*.

Enfin le **troisième outil** dont nous disposons pour faire de la terminologie, ce sont les outils logiciels. Ils constituent une discipline nouvelle, qui est l'application de l'informatique à la terminologie, qu'on appelle en français, selon un terme institutionnalisé par l'ISO, la terminotique. Je me souviens, quand nous commencions à travailler en 1987 dans le RINT (Réseau international de néologie et de terminologie franophone), les logiciels dont nous disposions étaient très hétérogènes. Et même avant, lorsque je travaillais au tout début des années 1980 dans le Réseau franco-québécois de néologie scientifique et technique, nous disposions d'un simple machine à écrire. Je ne suis pas sûr que cela ait beaucoup changé, car il reste difficile d'échanger des données, sauf si l'on travaille sous des logiciels propriétaires. On verra ce que diront sur ce sujet les intervenants qui viendront après moi.

Nous avons fait plusieurs sondages au sein des ministères français sur l'utilisation de leurs outils pour gérer la traduction et la terminologie. Là encore, l'harmonisation des outils n'est pas faite : chacun utilise un peu ce qu'il veut comme logiciel. Internet offre aujourd'hui des possibilités d'outils coopératifs pour le travail en réseau, sous un même langage qui est XML. On pourrait penser à des wikis, sites collaboratifs où chacun apporte ses données. La question reste cependant de pouvoir structurer ces données. C'est pourquoi nous avons élaboré un prototype qui permet à tout utilisateur de gérer ses données en ligne sur internet et de les échanger facilement dans un groupe de travail. Cela permet un cadre informatique commun, un travail collaboratif sur internet et une exploitation des données, par exemple leur impression, avec une sortie de qualité imprimeur. Cela évite des coûts et des restructurations informatiques très coûteuses et souvent impossibles, car beaucoup de chercheurs dans le monde travaillent sans équipe d'informaticiens gratuite et à proximité.

J'en arrive maintenant à ma **conclusion**. Elle puise dans les colloques scientifiques que nous organisons à la Société française de terminologie, à savoir :

- *La terminologie discipline scientifique* (2003)
- *Les néologies contemporaines* (2004)
- *Terminologie et sciences de l'information* (2005)
- *Terminologie et ontologie : descriptions du réel* (2006)
- *Terminologie et médias : la diffusion des termes* (2007)
- *Enjeux de la terminologie automobile dans la mondialisation* (2008).

L'idée reçue est que l'anglais est la langue mondiale. C'est ce qu'on appelle le « globish », et que j'appelle personnellement, dans le cadre des outils internet, le

« googlish ». Je connais même une entreprise qui m'a avoué que la langue de communication de l'entreprise est le « bad english ».

En fait, quand on regarde comment les choses se passent sur le terrain, on s'aperçoit que certes, la direction, les cadres en général, tendent à échanger entre eux en anglais. Mais sitôt que l'on va sur le terrain, que l'on voit travailler les ouvriers, les techniciens, les ingénieurs, on s'aperçoit que c'est la langue locale qui l'emporte. Beaucoup de nos entreprises commencent à prendre la mesure de cela. Je vois PSA Peugeot Citroën ou Renault développer des stratégies très intelligentes pour s'adapter aux marchés, en essayant de respecter les langues et les cultures de ces marchés. Il y a d'autres signes dans ce sens. Ainsi, on nous a contacté récemment pour travailler à un grand dictionnaire technique français-allemand de 100 000 termes. C'est dire qu'il y a un marché pour un tel produit.

Pourquoi pas aussi un grand dictionnaire technique français-roumain ?

Je vous remercie de votre attention.

Références bibliographiques

Depecker (Loïc), ***Entre signe et concept : éléments de terminologie générale***, Presses de la Sorbonne nouvelle, Paris, 2002, 200 p.

Depecker (Loïc) (dir.), *La terminologie : nature et enjeux*, *Langages,* n° 157, Larousse, Paris, mars 2005, 128 p.

ISO, Norme internationale ISO 639-1, Partie 1 : code alpha 2, *Codes pour la présentation des noms de langue,* (français et anglais), Afnor, février 2003, 37 p.

ISO, Norme internationale ISO 704, *Travail terminologique-Principes et méthodes* (français et anglais), 2e édition, Afnor, 15 novembre 2000, 41 p.

ISO, Norme internationale ISO 860, *Travaux terminologiques-harmonisation des notions et des termes,* Afnor, novembre 1997, 7 p.

ISO, Norme internationale ISO 1087, *Travaux terminologiques Vocabulaire, Partie 1 : Théorie et application,* (français et anglais), Afnor, février 2001, 42 p.

ISO, Norme internationale ISO 12616, *Terminographie axée sur la traduction,* Afnor, février 2003, 28 p.

ISO, Norme internationale ISO 15188, *Lignes directrices pour la gestion de projets de normalisation terminologique*, Afnor, octobre 2001, 16 p.

Société française de terminologie, Depecker (L.), Dubois (V.), Gresser (J.-Y.) (dir.), *La terminologie discipline scientifique*, actes du colloque tenu à l'École normale supérieure de la rue d'Ulm le 17 octobre 2003 sous le patronage du professeur Pierre Gilles de Gennes, prix Nobel de physique, col. Le savoir des mots, Société française de terminologie, Paris, 95 p.

Société française de terminologie, Depecker (L.), Dubois (V.) (dir.), *Les néologies contemporaines* (dir.), actes du colloque tenu à l'École normale supérieure de la rue d'Ulm le 15 octobre 2004 sous le patronage du professeur Pierre Gilles de Gennes, prix Nobel de physique, col. Le savoir des mots, Société française de terminologie, Paris, 123 p.

Société française de terminologie, Depecker (L.), Dubois (V.) (dir.), *Terminologie et sciences de l'information,* actes du colloque tenu à l'École normale supérieure de la rue d'Ulm le 2 décembre 2005 sous le patronage du professeur Pierre Gilles de Gennes, prix Nobel de physique, col. Le savoir des mots, Société française de terminologie, Paris, 128 p.

Société française de terminologie, Depecker (L.), Dubois (V.), Roche (Christophe) (dir.), *Terminologie et ontologies : description du réel,* actes du colloque tenu à l'École normale supérieure de la rue d'Ulm le 1er décembre 2006 sous le patronage du professeur Pierre Gilles de Gennes, prix Nobel de physique, col. Le savoir des mots, Société française de terminologie, Paris, 128 p.

Société française de terminologie, Depecker (L.), Dubois (V.) (dir.), *Terminologie et médias : la diffusion des termes,* actes du colloque tenu à l'École normale supérieure de la rue d'Ulm le 7 décembre 2007 sous le patronage du professeur Yves Chauvin, prix Nobel de chimie, Société française de terminologie, col. Le savoir des mots, Paris, 128 p. (actes à paraître).

Société française de terminologie, Depecker (L.) Dubois (V.) (dir.), *Enjeux de la terminologie automobile dans la mondialisation,* actes du colloque tenu au Comité des constructeurs français d'automobiles le 24 janvier 2008 sous le patronage du professeur Yves Chauvin, prix Nobel de chimie, Société française de terminologie, col. Le savoir des mots, Paris, 98 p. (actes à paraître).

Sponsor de aur / **Partenaire Or**

Sponsor de argint / **Partenaire Argent**

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / **Partenaire Bronze**

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latina/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



PRINCIPII DE ELABORARE A UNEI BAZE DE DATE TERMINOLOGICE MULTILINGVE ÎN DOMENIUL ECONOMIC

REZUMAT

Maria-Antoaneta Lorentz

Principiul reutilizării

Termeni cheie: bază de date terminologice, instrument pentru traducere, sistem de conepte, echivalent, echipă interdisciplinară, fiscalitate

Proiectul echipei noastre de cercetare îşi propune să stabilească principiile de elaborare ale unei **baze de date terminologice multilingve pentru domeniul economic** pentru a sprijini activitățile de documentare, traducere de specialitate, cercetare fundamentală şi aplicată, învățământ din ASE ca şi pentru specialiştii din întreprinderi.

Diferitele etape ale cercetării noastre îşi propun : - să definească domeniile economice în care numeroase concepte specifice acoperă realități în schimbare, diferit de la o țară la alta, ceea ce implică dificultăți în găsirea echivalenților în diferitele limbi europene (româna, franceza, engleza, spaniola, italiana); - să identifice sursele de informare şi documentare; - să armonizeze **competențele complementare ale membrilor echipei de cercetare** (**terminologi, economişti, informaticieni**); - să identifice principiile de delimitare a diferitelor domenii alese şi tipurile de relații semantice dintre diferitele lor concepte printr-un demers ştiințific confrom **standardelor internaționale**; - să identifice conceptele şi să stabilească sistemele de concepte ca şi tipurile de relații pe care le întrețin între ele; - să stabilească principiile de selecție a termenilor; - să analizeze conținutul fişei terminologice în funcție de obiectivele noastre de cercetare adaptate publicului căreia îi este destinată această bază de date.

Noutatea şi originalitatea proiectului constă în: orientarea teoretică şi practică privitoare la terminologie în cadrul ştiintelor limbajului printr-un **demers interdisciplinar coerent**, analiza şi adaptarea rezultatelor cercetării la domeniul **fiscalității**, pentru această etapă a proiectului, folosirea modelelor lingvistice în cercetarea limbajului economic şi adaptarea metodelor de reperare şi prelucrare a datelor la studiul domeniului ales. Stabilind ca **limbă de plecare limba română**, dorim să contribuim la **creşterea vizibilității limbii noastre** printre limbile latine şi nu numai. Impactul acestui proiect va trebui evaluat conform **calității produsului terminologic** conceput, baza de date terminologice multilingvă pentru economie, atât de necesară diferiților utilizatori pentru traducerea documentelor profesionale în condițiile noilor realități socio-economice româneşti şi europene.

Conf.dr. Maria-Antoaneta Lorentz
ASE – Bucureşti
Tel. 0040 21 224 69 46
0040 745 04 01 65

Semi
interna
„Instru
pent
asist
tradu

Gestionarul de terminologie
- un instrument flexibil şi eficient

DORINA CHIŞ conf.univ.dr - Universitatea „Tibiscus" din Timişoara - dorina_chis@yahoo.com
VALERIU DORIN CHIŞ – student, Universitatea „Politehnica" Timişoara - vchis@arcolite.com

UNIÓ LL
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

ACADEMIA ROMÂNĂ

1. Terminologie şi actualitate

- Societatea contemporană este una a informaţiei şi a comunicării. În toate domeniile cunoaşterea umană progresează într-un ritm fără precedent, iar transmiterea cunoştinţelor acumulate necesită precizie, nuanţare, claritate şi eficienţă.
- Pentru ca o limbă să poată asigura comunicarea specializată, ea trebuie să dispună de un lexic ştiinţific şi tehnic bogat, caracterizat prin claritate conceptuală şi eficienţă formală.
- Nevoia de a dispune de o terminologie sigură, standardizată, adaptată nevoilor de comunicare, este resimţită în toate sectoarele de activitate, lipsa ei putând genera confuzie şi incapacitatea de a participa la comunicarea specializată pe plan internaţional.
- Intensificarea fără precedent a relaţiilor economice, politice, strategice, culturale etc. între state, mai ales în condiţiile realizării unor noi comunităţi continentale, a condus la o creştere spectaculoasă a volumului de traduceri.

În aceste condiţii, capătă o deosebită importanţă dezvoltarea şi actualizarea vocabularului limbii noastre, precum şi elaborarea de instrumente lexicografice şi terminografice care să asigure cunoaşterea şi utilizarea lui eficientă.

2. Consemnarea rezultatelor activității terminologice

Activitatea terminologică şi terminografică beneficiază în prezent de resursele moderne, deosebit de eficiente, oferite de informatică. Cu ajutorul acestora se realizează astăzi consemnarea, clasificarea gestionarea, prelucrarea materialului lingvistic.

- Chiar dacă produsele lexicografice şi terminografice tradiţionale, pe suport de hârtie, sunt utilizate încă în mod curent, multe din ele dispun – în noile ediţii – şi de variante prezentate pe suport electronic (CD) sau direct pe Internet.
- De o complexitate sporită, băncile de date terminologice s-au dezvoltat şi s-au perfecţionat în timp, fiind utilizate pe o scară din ce în ce mai largă. Avantajele lor sunt multiple:
  - **Capacitatea** a băncilor electronice de date este impresionantă.
  - **Actualizarea** materialului lingvistic se poate face în mod constant: pe măsură ce apar elemente noi ce trebuie consemnate, acestea se pot adăuga, în timp ce materialul perimat este înlăturat; prezentarea poate fi modificată/îmbunătăţită, conform standardelor actuale şi nevoilor momentului, fără a fi necesar să se aştepte o nouă ediţie.
  - **Consultarea** băncilor de date, unilingve sau multilingve, este mult mai uşoară decât cea a unui dicţionar pe hârtie, de mari dimensiuni, cu atât mai mult cu cât navigarea permite trecerea de la o fişă la alta – pe baza trimiterilor şi legăturilor inserate – printr-un simplu clic.
  - **Conlucrarea** între specialişti, între grupuri care lucrează în mod independent, este nu doar posibilă, ci şi mult mai uşor de realizat şi mai eficientă decât ar fi prin alte mijloace. Fiecare din membrii grupului de lucru poate aduce modificări şi îmbunătăţiri materialului prezentat, care sunt imediat accesibile tuturor.

3. Apelul Uniunii Latine privind finanțarea creării unui software de gestionare a terminologiei

Uniunea Latină a lansat, în 24 iunie 2006, un **apel**

- **Obiectul apelului** : crearea unui software „open source", gratuit, de gestionare a terminologiei care să ţină cont de standardele actuale în materie de terminologie, ergonomic şi eficient în manipularea informaţiei.
- **Destinatari**: lingvişti şi terminologi, specialişti de limbă română (urmând ca, pe termen lung, lucrarea să fie tradusă şi în alte limbi, scop în care, la elaborarea ei, era necesar să fie prevăzute toate funcţionalităţile multilingve).
- **Motivaţie** : Demersul Uniunii Latine a fost determinat de aprecierea că, pentru limba română, nu sunt disponibile la ora actuală suficiente instrumente terminografice fiabile, standardizate, care să permită gestionarea şi armonizarea datelor şi să fie accesibile tuturor celor implicaţi în munca terminologică. Pentru că nu există întotdeauna posibilităţi materiale de achiziţionare a unor programe eficiente, dar costisitoare, pe care piaţa le pune la dispoziţie, numeroase terminologii se realizează utilizând aplicaţii ad-hoc, de multe ori incompatibile cu cele folosite de alte grupuri de lucru.

4. Documentare şi stabilirea caracteristicilor de bază

- **bibliografia** accesibilă în biblioteci şi pe Internet
- **bazele terminologice** :
  - Banque de données terminologique - Service de la langue française, de la Communauté française de Belgique, http://www.cfwb.be/franca/bd/bd.htm,
  - Base de terminologie du CILF (Conseil international de la langue française) http://www.cilf.org/bt.fr.html,
  - CRITER (Corpus du Réseau Interministériel de Terminologie), La délégation générale à la langue française et aux langues de France http://www.culture.gouv.fr/culture/dglf/terminologie/base-donnees.html,
  - EURODICAUTOM – actualmente IATE (Inter Active Terminology for Europe) - Bază de date a Comisiei europene http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load,
  - ILOTERM – Bază de date a Organizaţiei internaţionale a muncii, http://www.ilo.org/iloterm/ ,
  - IMF Terminology, Bază de date terminologice a fondului monetar internaţional, http://www.imf.org/external/np/term/index.asp?index=fra&index_langid=2,
  - Le Grand dictionnaire terminologique de l'Office québécois de la langue française, http://www.granddictionnaire.com/btml/fra/r_motclef/index1024_1.asp,
  - TERMROM - Baza Centrului Naţional de Terminologie, http://www.cfwb.be/franca/bd/bd.htm,
  - UNESCOTERM, Bază de date terminologice a UNESCO, http://termweb.unesco.org/ UNTERM,
  - United Nations Multilingual Terminology Database, http://157.150.197.21/dgaacs/unterm.nsf...

- **Consultări:** Biroul din Bucureşti al Uniunii Latine a organizat mai multe întâlniri, mese rotunde, consacrate consemnării şi prelucrării materialului lingvistic, întocmirii fişelor terminologice. Am putut constata cu aceste prilejuri că, în funcţie de instrumentele de lucru utilizate, de metodele abordate în realizarea terminologiilor şi în funcţie de publicul destinatar, structurile propuse erau diferite, chiar dacă mare parte din elemente se regăseau în mai toate variantele.
- **Precizarea obiectivelor principale:** Această constatare ne-a determinat să pornim de la premisa că gestionarul nostru trebuie să fie un instrument caracterizat prin:
  - **flexibilitate**: să pună la dispoziţie utilizatorului o structură ce respectă standardele internaţionale, dar, în acelaşi timp, să permită fiecăruia crearea unei/unor structuri noi care să răspundă nevoilor proprii.
  - **eficienţa:** pentru asigurarea ei era necesar ca modificarea structurii existente, ca şi crearea uneia sau mai multor structuri noi, să nu implice ajutorul unui specialist în informatică, ci să se poată realiza de la nivelul interfeţei – utilizator, de către toţi cei interesaţi. În aceeaşi idee a creşterii eficienţei gestionarului, am decis să oferim interfeţe în toate limbile romanice utilizate pe site-ul Uniunii Latine şi să dăm astfel posibilitatea de a se căuta şi obţine informaţii despre un termen într-o anumită limbă, chiar dacă termenul aparţine altei limbi.
  - **gratuitate:** trebuiau alese instrumentele informatice de lucru care să asigure caracterul gratuit al gestionarului.

5. Caracteristicile de bază ale băncii de date terminologice *GesTE*
5.1. Elemente de programare

- Programul este **compatibil** cu sistemele de exploatare acceptate pe piaţa românească. Operaţiile pe partea de server sunt efectuate de un script PHP care interoghează o bază de date MySQL.
- Bazată pe un software ***open source***, aplicaţia poate fi utilizată ***on line***, în sistem **Intranet** sau pe **calculatorul propriu**.
- Atât datele referitoare la structura bazei terminologice (felul în care sunt organizate câmpurile şi grupurile, listele şi intrările lor) cât şi cele terminologice sunt reţinute într-o **bază de date MySQL**, care mai cuprinde şi informaţiile despre utilizatorii care au acces (restrâns sau total) la aplicaţie.
- Tabelul datelor terminologice **nu are o structură fixă**, fiind modificat de program în funcţie de structurile care sunt create pe parcurs. Totuşi, fiecare câmp al oricărei structuri trebuie asociat cu o coloană a acestui tabel.
- Interogările MySQL se fac pe partea serverului de către un **script PHP**. Clientului îi sunt apoi trimise informaţiile în **format XHTML**. Totuşi, codul XHTML a fost separat de codul PHP cu ajutorul unui engine pentru template-uri.
- **Formatarea** datelor se realizează cu ajutorul unor fişiere .tpl care pot fi editate pentru afişarea datelor în orice format fără a modifica algoritmii programului.
- **Navigarea** prin paginile programului se face prin parametrii GET trimişi fişierului index.php. Astfel, nu sunt dezvăluite căi către alte fişiere ale programului, totul realizându-se intern. De asemenea, parametrii folosiţi pentru comenzi nu pot dezvălui în vreun fel structura bazei de date şi nu permit atacuri de genul *MySQL Injection*.

- La crearea unui **utilizator nou**, datele acestuia sunt înregistrate într-un tabel împreună cu un identificator. Referirea ulterioară la acel utilizator se face prin intermediul **identificatorului** creat la înregistrare. **Parola** folosită la autentificare nu este salvată în baza de date (nu poate fi recuperată). În schimb, programul înregistrează hash-ul md5 al parolei şi îl compară cu hash-urile parolelor introduse la autentificările ulterioare. Astfel, un acces la baza de date MySQL care nu foloseşte interfaţa aplicaţiei nu va dezvălui datele de autentificare ale utilizatorilor.
- Aplicaţia permite **exportul** şi **importul** datelor terminologice. Exportul presupune generarea unui fişier în format propriu (.gte) care conţine rezultatul unei căutări. Pentru import e necesar fişierul .gte exportat anterior precum şi o structură compatibilă cu datele importate. Atât importul cât şi exportul permit selectarea câmpurilor intrărilor terminologice asupra cărora se va opera. Astfel, se poate exporta doar o listă de termeni, sau doar termenii si definiţiile lor etc.

5.2. Standarde utilizate

Standardele folosite în elaborarea programului bazei de date GesTE au fost :

- ISO 1087 – Concepte utilizate în terminologie
- ISO 12618 – Principii de creare a bazelor de date terminologice
- ISO 12620 – Categorii de date în terminologie
- ISO 12200 sau MARTIF – *Machine-Readable Terminology Interchange Format* – Principii de conversie şi programare XML a băncilor de date terminologice.

5.3. Limbi de lucru

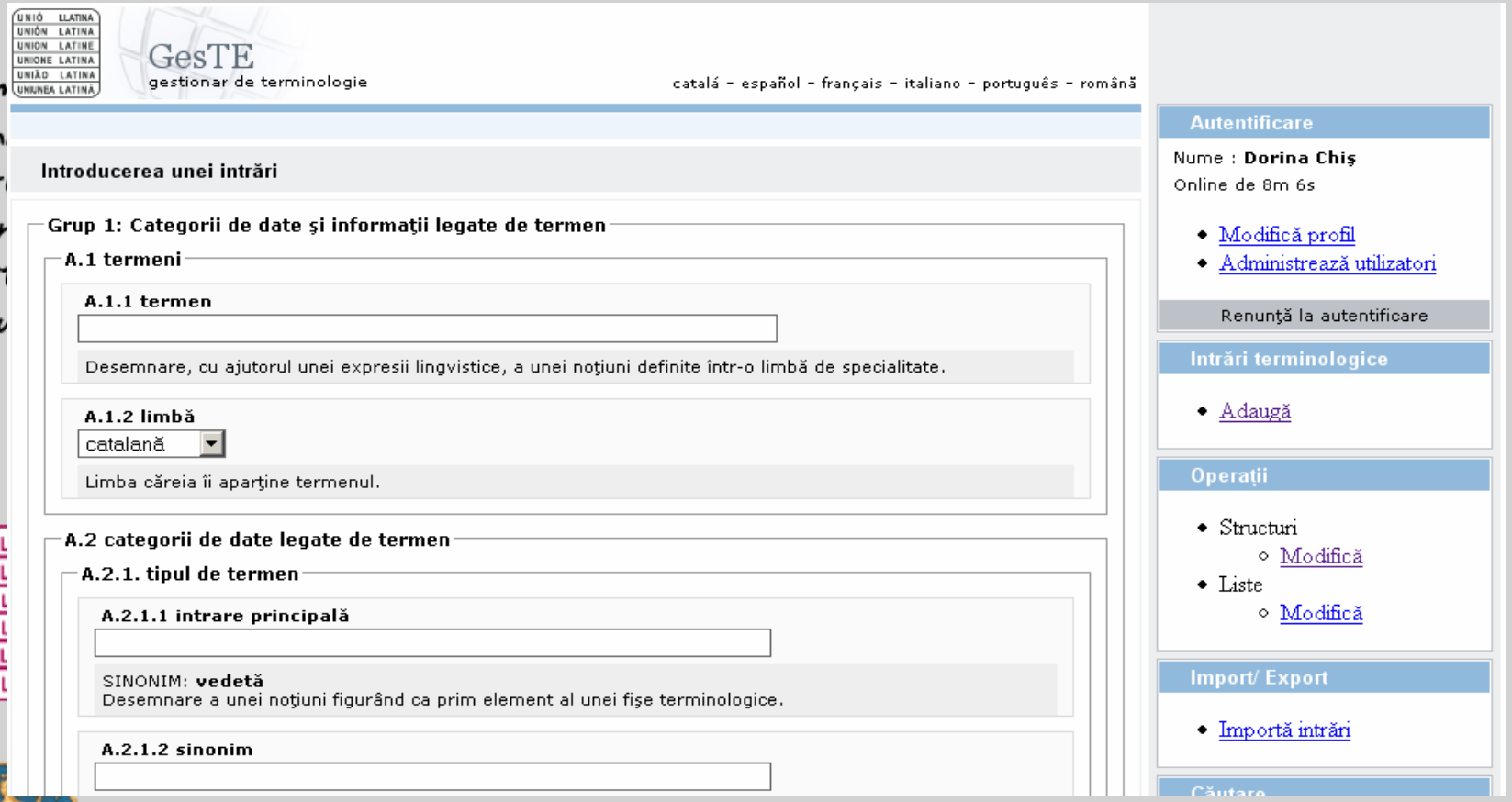
Programul bazei de date dispune de interfeţe în limbile:

- catalană,
- franceză,
- italiană,
- portugheză,
- română
- spaniolă

Biroul din Paris al Direcţiei de terminologie şi inginerie lingvistică (DTIL) al Uniunii Latine a asigurat redactarea textelor în alte limbi decât româna.

Utilizatorul poate alege limba de lucru pe care o stăpâneşte mai bine, şi va primi în această limbă toate informaţiile legate de un termen (definiţie, context, explicaţii...), chiar dacă termenul în cauză aparţine unei alte limbi.

5.4. Structura bazei de date



Structura implicită a bazei de date include categorii de informaţii cuprinse în standardul ISO 12620. A fost introdus, în plus, câmpul *limba căreia îi aparţine termenul,* tocmai pentru că există posibilitatea, evocată mai sus, de a alege pentru interfaţă o altă limbă decât cea căreia îi aparţine termenul în cauză.

Principalele categorii de date sunt grupate după cum urmează :

- **Grup 1 : Categorii de date despre termen şi informaţii legate de termen**
  A Termenul
  - Subgrup1 : Termeni
  - Subgrup 2 : Categorii de date legate de termen
  - Subgrup 3 : Echivalenţă
  - Subgrup 4 : Domeniu şi sub-domeniu
- **Grup 2 : Categorii de date legate de descrierea noţiunii**
  - Subgrup 5 : Descrierea noţiunii
  - Subgrup 6 : Relaţii ale noţiunii
  - Subgrup 7 : Poziţia noţiunii
  - Subgrup 8 : Nota
- **Grup 3 : Categorii de date administrative**
  - Subgrup 9 : Limbaj documentar
  - Subgrup 10 : Informaţii administrative

  B : Categorii de date bibliografice

5.5. Funcționalitate

Operațiile care se pot efectua cu ajutorul programului bazei de date GesTE sunt : căutare simplă/ avansată, adăugarea/ ştergerea sau modificarea unor câmpuri, crearea/ consultarea/ validarea/ suprimarea unor fişe, importarea şi exportarea datelor, administrarea bazei de date.

Căutarea simplă se poate realiza:

- folosind doar steluța * (se afişează toate intrările introduse) ;
- după un cuvânt, o sintagmă ;
- după primele litere ale unui cuvânt urmate de steluță * (se afişează toate cuvintele care conțin grupul respectiv de litere) ;
- după un grup de litere plasate între steluțe (omițând în felul acesta prima şi ultima parte a cuvântului) ;
- după autor (dacă îl cunoaştem) ;
- după echivalentul unui termen, de exemplu echivalentul lui *contrat* în germană **Vertrag** el trebuie pus însă între steluțe, având alte date înainte şi după el.

Căutarea avansată (folosind butonul *Opțiuni*), permite căutarea unui termen sau a unei expresii doar în unele câmpuri, pe care le selectează utilizatorul.

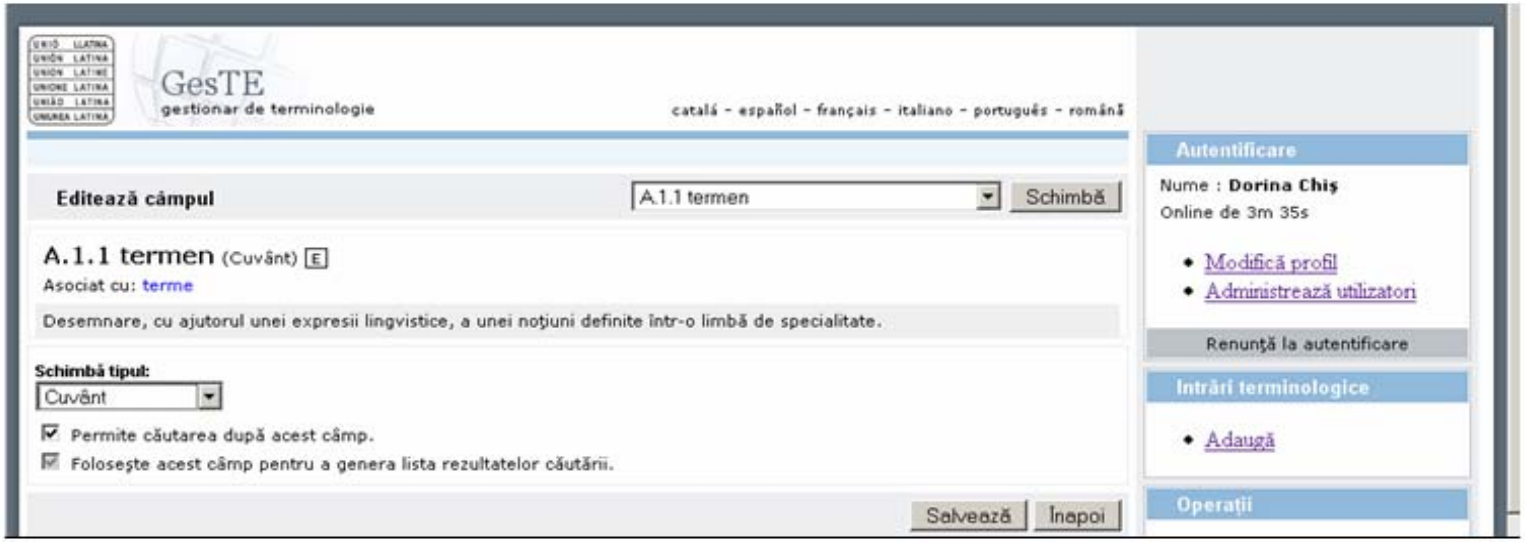
Pentru **căutarea cu filtre**, fiecare utilizator trebuie să-şi aleagă filtrele: din lista *Adaugă pentru câmpul*. Se alege apoi între butoanele **conține**/ **exclude** din lista care derulează şi se consemnează ce anume trebuie să conțină/excludă câmpul respectiv. Se adăugă apoi alt câmp, după acelaşi procedeu. Dacă se aleg, de exemplu, câmpurile *termen* şi *limba*, iar în acesta din urmă se înscrie *franceza*, la apăsarea butonului *Caută*, se vor afişa toți termenii în limba franceză existenți în baza de date.

Crearea grupurilor şi câmpurilor se realizează cu uşurinţă, folosind butoanele corespunzătoare şi urmând paşii indicaţi în ghidul de utilizare. Atât grupurile, cât şi câmpurile create, pot fi repoziţionate, modificate sau şterse, după ce au fost, în prealabil selectate, cum se poate vedea mai jos:



Sen
intern
„Instr
per
asist
tradu

Există şi posibilitatea de a edita fiecare câmp, pentru a efectua modificări în ceea ce priveşte descrierea lui, tipul ales iniţial, asocierea cu baza de date şi utilizarea lui în generarea fişei precum şi în căutare:

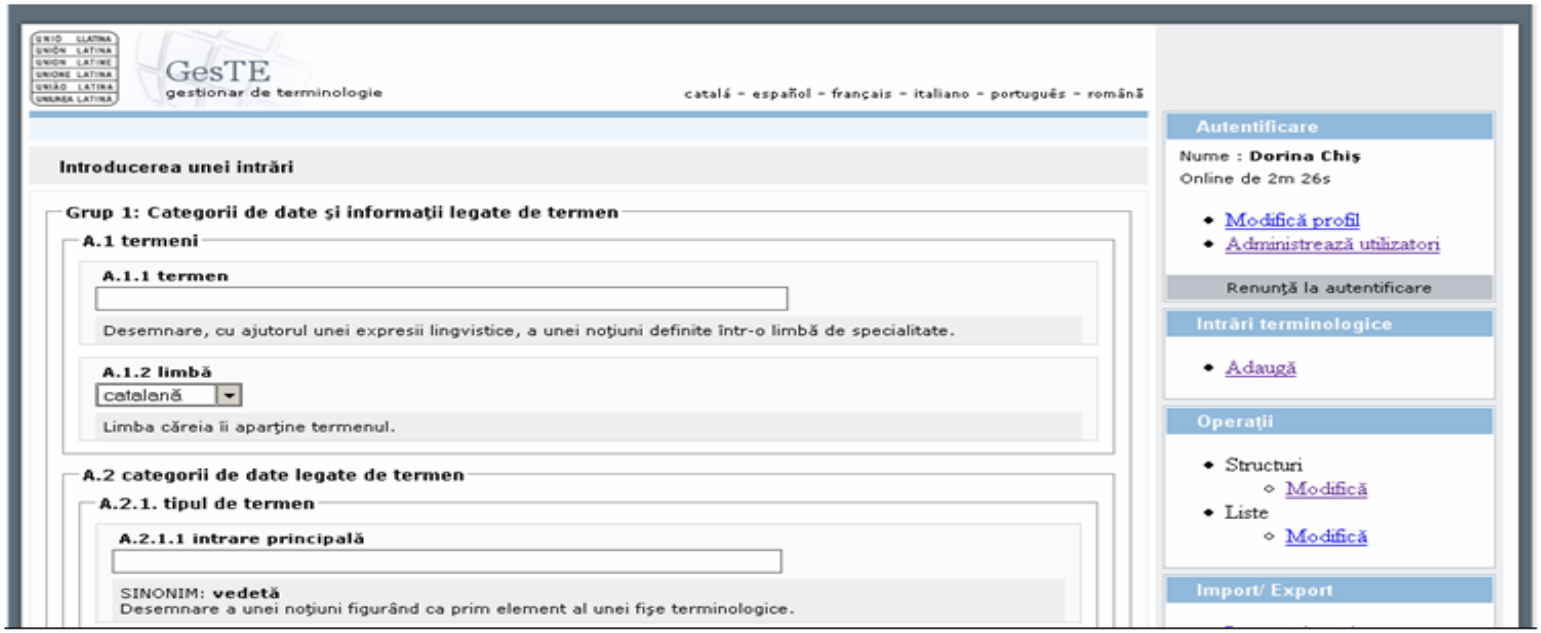


Sem
intern
„Instr
per
asist
tradu

Numele şi descrierile câmpurilor, ca, de altfel, fiecare cuvânt/fragment de text pot fi, de asemenea, editate şi apoi modificate prin adăugarea sau eliminarea unor elemente, prin reformulare. La fel şi traducerile pot fi editate şi apoi modificate după dorinţă:



Fiecare câmp dispune de o descriere clară, astfel încât confuziile să fie excluse şi completarea fişei terminologice să nu pună probleme. Aceasta este generată în conformitate cu structura activă (în cazul în care au fost create mai multe structuri, trebuie să fie selectată structura activă pentru sesiunea de lucru la momentul respectiv).



Sem
intern
„Instr
per
asist
tradu

Pentru a stabili lista **domeniilor** în care s-ar putea încadra termenii ce vor fi cuprinşi în bazele terminologice, am folosit un model de **ontologie** propus şi dezvoltat de Ingetraut Dahlberg, deja utilizat în Uniunea latină (la elaborarea băncii terminologice a Italiei, aşa cum este el prezentat în articolul Giliolei Negrini, *Quels systèmes conceptuels peut-on préconiser pour les banques de données terminologiques en ligne?*, publicat în revista Terminologies nouvelles, no.18, juin 1998, *Terminotique et documentation*, pp. 41-53).

GesTE este însoţit de un **Ghid**, tradus în toate limbile de lucru, care oferă utilizatorilor toate informaţiile necesare pentru o exploatare eficientă a acestui instrument de lucru.

UNIÓ L
UNIÓN L
UNION L
UNIONE L
UNIÃO L
UNIUNEA L

6. Concluzii

Gestionarul de terminologie *GesTE* pune la dispoziția celor interesați un instrument flexibil și eficient. El permite colectarea, prelucrarea, modificarea, stocarea, exportul și importul informațiilor. Fiecare utilizator are posibilitatea, de la nivelul interfeței de lucru, să adapteze acest instrument de lucru nevoilor proprii.

Realizarea lui a fost posibilă datorită inițiativei și suportului Uniunii Latine. El poate fi găsit pe pagina Direcției pentru Terminologie și Inginerie Lingvistică (DTIL) a Uniunii Latine (http://dtil.unilat.org/fr/index.htm).

Utilizarea acestui software poate contribui la armonizarea activității de gestionare a terminologiei și la facilitarea schimburile de informații între diversele grupuri de lucru.



CARTAGO : une terminologie large langues de l'enseignement électronique à distance, dans un contexte de co-élaboration multilingue de documents normatifs

HUDRISIER Henri, Maître de Conférences, Lab. Paragraphe, Université de Paris 8 -MSH Paris Nord, Liaison A de l'ISO/AUF
BEN HENDA Mokhtar, Maître de Conférences, Université Tunis et Bordeaux 3, Liaison A de l'ISO/AUF - Lab. Paragraphe, MSH Aquitaine

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Mots clefs *: normalisation, co-élaboration multilingue de documents, terminologie onomasiologique, modélisation terminotique, Génétrix, TICE (Technologie de l'Information et de la communication pour l'Enseignement), perte de domaines linguistiques, nomenclature, Agence Universitaire de la Francophonie, ISO*

Résumé :
On exposera les questions techniques (terminotiques), terminologiques, collégiales et institutionnelles ouvertes par la mise en place d'une terminologie onomasiologique, Cartago, indispensable dans un contexte de traduction très spécialisé ou plutôt de co-élaboration parallèle de documents normatifs (ou même quelquefois de la seule construction d'équivalent de vocabulaire ou de nomenclatures) : les technologies pour l'enseignement et la formation (ISO/IEC JTC1 SC36).
On voit (cf. note 1) que la co-élaboration d'un item de nomenclature, consiste donc non pas à traduire (ce qui présente un intérêt relatif en normalisation) mais à co-élaborer des collections de notions (concepts) dont on est par ailleurs capable, à travers une démarche onomasiologique structurée et hiérarchisée, de contrôler la (les) définition(s) consensuelle(s) et de donner en un autre niveau du modèle XML structuré, les termes le plus synonymes possible en n langues. On peut spécifier la différence locale par rapport à la (aux) définition(s) consensuelle(s) du concept dans un nombre ouvert de langues. Développer Cartago dans un grand nombre de langues (large langues) est éthiquement fondamental pour éviter chez ces communautés linguistiques des pertes de domaines linguistiques cruciaux pour les techniques de la transmission du savoir. .
Spécifier les conditions d'interopérabilité de plates-formes ou de ressources d'enseignement en ligne exige en effet de maîtriser (et non pas seulement de traduire) des équivalents très divers : en français seulement, classe de 6ème sera en effet synonyme de classe de 1ère selon que l'on est à Paris, à Genève ou à Tunis. C'est pourtant en fonction de l'équivalence correcte de ces notions qu'un système d'intelligence artificielle saura organiser et aiguiller vers les bons utilisateurs des bibliothèques de ressources pédagogiques.
Cartago est donc un outil déjà utilisé dans le contexte du SC36 qui organise en parallèle multilingue des terminologies normalisées élaborées dans le cadre du SC36. C'est aussi une Alliance d'experts qui a été initialisée pendant le SMSI (Sommet Mondial de la Société d'Information) à Tunis en 2005 par des experts de l'AUF (Agence Universitaire de la Francophonie), de l'Union Latine, de l'Université Ouverte de Corée et un expert australien.

« Une dizaine d'administrateurs de tout poil complétaient l'ensemble, qui s'intitulait National de l'Unification, ou, par abréviation, le CNU [1]. L'immeuble abritait encore quelques Inspecteurs généraux, anciens soudards retraités, qui passaient le plus clair de leur temps à ronfler aux réunions techniques, et ce qui leur en restait à parcourir la contrée sous le couvert de missions leur donnant un prétexte pour rançonner les adhérents dont les cotisations permettaient au CNU de subsister tant bien que mal. Afin

d'éviter les abus, le gouvernement, ne pouvant freiner d'un coup l'acharnement des Ingénieurs principaux Miqueut et Toucheboeuf à élaborer des Nothons, avait délégué, pour le représenter et superviser le CNU un brillant polytechnicien, le Délégué Central du Gouvernement Requin, dont la tâche consistait à retarder le plus possible l'aboutissement des Nothons...[...]...que penser d'un pays où l'on donne cent centilitres pour un litre et où un boulon garanti pour résister à quinze tonnes tient le coup sous une charge de 15000 kilos ? Heureusement, les professions intéressées, prenaient, avec l'appui du Gouvernement, une part prépondérante à la création des Nothons, et les établissaient de telle façon que des années devenaient nécessaires à leur déchiffrage : au de ce temps on en préparait une révision. »
Boris Vian [2]

Nous [3] *sommes très impressionnés d'être invités ici à Bucarest dans un colloque de traductologie. Qu'il soit clair dès le début que nous ne sommes ni traductologues, ni terminologues et encore moins traducteurs ou interprètes.*
Le petit savoir en terminotique [4] *que nous pourrons exposer ici, nous le devons à la fréquentation de terminologues dans le cadre du développement de Cartago et dans des instances de normalisation* [5].

Les instances de normalisation ont une importance grandissante dans la gouvernance technique du monde. Pour beaucoup de traducteurs professionnels présents ici la normalisation est une activité que vous avez du déjà croiser mainte fois car elle est souvent prescriptrice de traductions techniques... quoique les experts en normalisation préfèrent souvent « bricoler eux-mêmes » leur versions multilingues, ce que nous tenterons sinon de justifier, au moins d'expliquer.

Pour poser le problème nous partirons de deux concepts clefs, ceux de norme et standard et nous verrons (ce qui est source de nombreuses questions en matière de traduction et d'interprétations) que même les cultures linguistiques des langues européennes n'ont pas la même sémantique des termes ni le même découpage de conceptuel.

Nous mettrons ensuite l'accent sur la situation particulière que constitue l'élaboration multilingues de documents normatifs. Comme nous le verrons cette élaboration peut exister à 2 niveaux :

- Celui de la co-élaboration parallèle de documents normatifs en *n* langues par les experts eux-mêmes (le plus souvent, mais pas exclusivement dans les langues officielles de l'ISO : anglais, français et dans une moindre mesure en russe).

- Celui de l'adaptation/traduction dans toutes les langues possibles (en général ces traductions sont certifiées par une ou plusieurs instances nationales ayant cette langue comme langue nationale). Dans de très nombreux cas pour devenir une norme nationale ces traductions devront être adaptées au contexte institutionnel du pays considéré.

Ce n'est qu'après avoir longuement exposé ce statut particulier du multilinguisme en normalisation et après avoir aussi présenté le contexte de la normalisation et de ses institutions que nous en viendrons aux terminologies qui sont fondamentales pour maîtriser ces situations multilingues. Nous présenterons Cartago (une base terminologique multilingue *large langues*) qui s'insère avec succès à ISO/IEC JTC1 SC36 [7]. Nous proposons ce terme de *large langues* volontairement accrocheur (construit sur le même patron linguistique que *large bande* (en anglais « *broadband* »: largement ouvert sur un grand nombre de fréquences et permettant de hauts débits.) pour bien signifier qu'à l'ère de la mondialisation il n'est plus temps de continuer à croire que des terminologies bi, voire quadrilingue continuent d'être suffisantes pour répondre aux besoins de la communication mondiale. Les technologies (notamment pour les terminologies sont maintenant susceptibles de s'ouvrir à un nombre extensif de langues (plusieurs dizaines dans un nombre maximum d'écritures. Nous ne prétendons pas que ce genre de projet soit facile à développer. Nous savons même qu'ils nous posent déjà, et qu'ils nous poseront dans le futur un maximum de problèmes tant techniques que conceptuels.

Différence sémantique en français entre *standard* et *norme*.

En français *norme* et *standard* sont des mots qui restent longtemps des quasi-synonymes. Ils sont d'ailleurs souvent confondus par le grand public. Cependant une distinction s'établit entre standard et norme dès lors que notre monde technique a été contraint de systématiser une réglementation législative de la conformité des produits et services. Pour ce qui est des francophones c'est le terme d'origine latine *norme* [8] qui s'est ainsi spécialisé par opposition au *standard* [9] qui reste lui réservé à la

dénomination de types, de processus de fabrication ou de services rationnellement produits à l'identique, mais sans qu'il y ait un processus institutionnellement identifié d'enregistrement de ces standards. Les États ou leurs regroupements sous forme de traités ou ententes se sont ainsi organisés en déléguant à des institutions nationales ou internationales de normalisation le soin d'animer l'élaboration et de rassembler des corpus de règles fixant les conditions de réalisation d'une opération, de l'exécution d'un objet ou d'un produit dont on veut unifier ou sécuriser l'emploi ou assurer l'interchangeabilité et l'interopérabilité.

De ce fait, la norme tire sa force de son caractère juridique, ce qui ne signifie aucunement que toute norme s'impose par la force de la loi. Une norme peut-être très longtemps négligée au profit d'un standard mieux adapté à un usage.

Toujours en français, le terme standard s'applique pour qualifier des règles élaborées soit dans le cadre du savoir faire interne d'un industriel, soit dans celui d'un consortium. « Au sens industriel, le standard est une méthode qui préconise l'utilisation d'un type déterminé, donné comme modèle, comme étalon, pour être sûr que les méthodes de fabrication et la qualité ne dégénéreront pas. Un standard est valable jusqu'à son remplacement par un autre mieux adapté. Ainsi le standard ne peut être un frein aux progrès, mais il permet d'avoir la certitude de ne pas revenir en arrière [10] ». On peut noter que la mondialisation des réseaux a aussi favorisé (dès le 19ème siècle mais plus encore maintenant avec l'Internet) la création de « standards négociés collégialement » (et non à l'intérieur d'une seule entreprise) qui dans certains cas s'imposent comme des quasi-normes: les recommandations du W3C par exemple.

Il est clair cependant qu'en français nous disposons de 2 mots, correspondant à 2 concepts distincts, voire 3 si on considère un standard collégial comme conceptuellement distinct d'un standard industriel.

En anglais par contre nous ne disposons que d'un seul concept [11] en fait générique des 2 (ou 3) concepts correspondant du français [12]. Nous nous trouvons donc pour des traductions techniques « anglais **<- ->**français » dans des situations de dissymétrie de la traduction terme à terme. En traduction « français **->** anglais » les 2 termes norme et standard se traduiront indifféremment par standard. Par contre le traducteur de l'anglais vers le français sera en position d'indécidabilité hors du contexte.

Quelques remarques sur des langues qui déclinent deux termes venant du latin *norma* et de l'anglo-normand *estandard*

Une situation comparable au français existe en espagnol *estándar / norma,* en italien *standard / norma,* en portugais *padrão / norma*, en neerlandais : *standaard / norm*, en allemand *Standard / Norm* ou en roumain *standard / normă.* Nous ne pouvons pas assurer par contre que les concepts correspondants à ces couples de termes correspondent exactement à la situation française. Nous pouvons cependant constater que les noms des instituts nationaux de normalisation lusophones, francophones, roumanophone, hispanophones, néerlandophones et germanophones [13] traduisent tous [14] le concept de norme ou normalisation par des termes dérivés du latin *norma*. Ce serait à vous tous, réunis ici, que nous poserions la question corollaire : qu'en est il pour ces langues de l'emploi (ou du non emploi) d'un terme distinct qui correspondrait au concept français de standard c'est-à-dire : standard industriel ou collégial à l'exclusion des normes enregistrées et éditées par un organisme officiel de normalisation ?

Variations sémantiques entre norme, standard industriel et standard collégial.

Fonder le *distinguo* entre norme et standard sur la notion de légitimité institutionnelle des instances de normalisation peut paraître très artificiel pour les non spécialistes de la normalisation. Bien évidemment distinguer un standard industriel d'une norme est assez simple. Le standard est très souvent protégé au sein de l'entreprise comme un secret de fabrication. Il permet à toute une gamme de produit (par exemple l'ensemble bureautique Windows), d'être interopérable entre eux et avec un grand nombre d'autres produit avec lequel Microsoft à choisi de créer des passerelles d'interopérabilité. Mais à l'évidence ce standard crée une situation de monopole d'ailleurs très souvent dénoncée. C'est en fait le standard qui donne à l'entreprise une capacité de différentiation positive de son offre commerciale, donc un avantage concurrentiel. Corollairement le standard, du fait de sa position de monopole industriel conforte l'économie du développement des innovations. Pour l'utilisateur il est le surcoût à payer, comme on paie en pharmacie le développement des nouvelles molécules [15]. Mais, le standard industriel constitue aussi un frein à la volonté de l'usager (qui peut être aussi un autre industriel, une collectivité territoriale) de

pouvoir déployer ses applications quand elles sont prisonnières de « solutions logicielles propriétaires ». Cette capacité d'ouverture, d'interopérabilité, seule la norme peut la lui donner car elle est par construction publique, ouverte et interopérable. En ce qui concerne le standard collégial, la distinction est beaucoup moins nette. Par exemple, on entend certains informaticiens s'inquiéter de ce que les recommandations du W3C ne soient pas d'office enregistrées en tant que normes. La raison est pourtant simple : les recommandations du W3C sont l'objet d'une négociation collégiale certes, mais limitée au métier et à la recherche en informatique. Les usagers et nombres d'Etats en sont exclus ce qui entraîne par exemple une prépondérance dominatrice des USA. Le W3C cherche aussi à imposer sa culture d'innovation informatique (des environnements de travail en perpétuels bouleversement) peu supportable par l'usager qui peut vouloir préférer un usage simplifié et sûr plutôt que de se sentir exclus face à une surenchère d'innovations permanentes qui lui est imposées [16].

En tant que professionnel de l'information nous pouvons fonder notre distinction sur un autre critère : la norme est toujours et exclusivement un document enregistré, édité. Celui-ci est en principe public [17]. La norme n'est pas un produit ou un service industriel : c'est uniquement un document administratif. Il est fréquent que plusieurs produits industriels correspondent à une norme unique. Dans ce cas il est quasi-obligatoire qu'ils aient été fabriqués chacun selon un standard industriel propre à l'entreprise productrice. Ces standards distincts peuvent ainsi se distinguer par des ajouts supplémentaires, des solutions techniques différentiés mais tous ces produits correspondent à la même norme [18]. C'est en cela que la norme se distingue du standard industriel : suite de protocoles et procédés, écrits ou non, par définition gardé secret par l'industriel ; ce dernier s'en servant comme avantage concurrentiel pour construire des produits ; quoique la logique du logiciel libre et ouvert brouille maintenant cette distinction. Par contre, c'est par abus de langage que beaucoup de personnes confondent : (1) le comité [19] qui produit la norme et qui est un groupe d'experts (2) le document normatif (dont il importe de distinguer les niveaux de finitude [20] et de révision) (3) la technologie correspondant aux spécifications de cette norme (4) les produits industriels conformes et labellisés comme tels.

Une question complexe mais déjà connu : l'élaboration parallèle de documents normatifs en plusieurs langues

Nous abordons maintenant les problématiques de la traduction, ou plutôt de l'élaboration parallèle de documents normatifs en plusieurs langues

Dans tous les cas on perçoit bien que la traduction de documents normatifs, et *a fortiori* l'interprétation des rencontres de commission de normalisation n'est pas une question triviale. C'est une question sur laquelle nous nous attarderons maintenant pour établir un lien d'usage entre la terminologie dans le contexte normatif et les métiers de la traduction et de la traductique.

Les diplomates, les juristes et de plus en plus en plus le monde des affaires connaissent bien l'épineuse question qui consiste à élaborer en parallèle, en plusieurs langues, des traités, des dispositifs juridiques internationaux, des contrats ou tout autre document.

La normalisation internationale constitue en fait une annexe technique de la diplomatie c'est pourquoi l'ISO ainsi que les grandes organisations de normalisation (IEC, UIT) sont toutes et rattachées aux Nations Unies et situées à Genève, ville bien connue des traducteurs et interprètes.

En fait dans le monde de la normalisation, le multilinguisme est à prendre en compte à de nombreux niveaux que je vais rapidement lister :

L'appréhension de concepts définis en consensus dans une ou plusieurs langues:

> L'appréhension de concepts définis en consensus est déjà indispensable pour des experts travaillant sur une norme dans une seule langue. La démarche doit être encore plus rigoureuse pour des experts appartenant à des communautés linguistiques disparates et qui décident de travailler dans une langue de travail commune (c'est souvent l'anglais dans ce cas de figure mais il y a de nombreux contre exemple [21]). Dans le cas des normes internationales à vocation mondiale, l'appréhension de concepts définis en consensus a évidement intérêt à être faite non seulement de façon unilingue (ou même bi ou trilingue), mais sur un large nombre de langue. C'est uniquement en confrontant les diversités du découpage des concepts, induites par les disparités des cultures et de langue que l'on peut être assurés de pouvoir rédiger ensuite des normes internationales capables de répondre à la multitude des diversités des usages partout dans le monde ; mais surtout, car c'est très important, que l'on sera

capable de localiser (de traduire et d'adapter) ces normes pour que les institutions, les entreprises et les utilisateurs puissent ensuite se les approprier dans tous les pays du monde qui les adapteront (ou les subiront).

C'est la raison pour laquelle nous militons pour la mise en place de terminologies large langue, seules à même d'éviter à certaines communautés à faible niveau d'ingénierie linguistiques de risquer des « pertes de domaines linguistiques [22] », avec les conséquences de sous-développement (dont le plus visible est la fracture numérique [23]), que cela peut induire dans certains pays du Sud.

Pour systématiser cette appréhension de concepts définis en consensus, les institutions de normalisation conseillent instamment aux animateurs de chaque instance de normalisation [24] de créer un groupe de travail spécifique chargé d'élaborer une terminologie du domaine. Notons cependant que tous les experts en normalisation dans tous les métiers qui peuvent être concernés sont diversement sensibles à la nécessité de normaliser les terminologies de leur domaine de spécialité (malgré les normes de l'ISO TC37). De ce fait les terminologies produites par l'ISO sont méthodologiquement et techniquement très disparates. La coordination interdisciplinaire de terminologies spécialisées n'est pas une question simple, par contre la mise en place de terminologies rigoureusement normalisées (pour ce qui est de leur méthode et de leur interopérabilité terminotique est une voie d'avenir, conforme aux avancées vers un véritable web sémantique. C'est aussi la seule façon connue de mettre en place un environnement TIC permettant de répondre aux défis de la mondialisation.

Dans le cas particulier du SC36, il serait souhaitable que le vocabulaire qui a été mis en chantier [25] soit conforme aux recommandations normatives de l'ISO TC37 [26]. Ce n'est pas le cas et c'est pourquoi le SC36 et Cartago ont déjà amorcé un processus de collaboration : nous développons plus longuement ce point dans la deuxième partie de cette intervention.

L'intercompréhension mutuelle des experts impliqués dans une démarche de normalisation :

Une fois ce premier niveau de compréhension des concepts acquis pour les experts, il s'agit qu'ils puissent communiquer à l'écrit comme à l'oral malgré leurs niveaux inégaux de maîtrise de la (ou des) langue(s) de travail. Contrairement aux documents de la communication quotidienne, littéraire ou même diplomatique les documents normatifs (comme les documents juridiques) sont très difficiles à comprendre non pas seulement parce qu'ils sont écrits dans une langue qui peut nous être étrangère, mais parce qu'ils sont écrit dans un style faisant une très large place au références, aux sigles, aux acronymes divers. Comme le langage juridique (et de façon beaucoup plus systématique que les document scientifiques et techniques) le rédacteur de norme doit utiliser un vocabulaire codifié, rarement explicité même quand il fait référence à des commissions de normalisation autres que celles qui produisent le document. Nous l'expliquerons plus loin dans cet article mais il faut comprendre que les experts en normalisation se réunissent essentiellement à 2 niveaux : (1) le niveau de leurs institutions nationales ou ils élaborent soit des normes qui reste nationales (donc souvent unilingues sauf le cas des pays multilingues), (2) le niveau international qui implique que tous les experts délégués à un niveau international travaillent bien sûr au strict niveau de l'ISO (et instituts similaires) mais aussi à leur niveau national (l'AFNOR, l'EANOR…) où ils rendent compte, discutent, collaborent avec leur niveau national pour que l'ensemble des communautés nationales d'experts contribue à la construction des normes internationales. On comprend de ce fait que ce niveau de travail complique encore les taches de traductions, de commentaires d'aide à la compréhension de textes en langue étrangère pour ces communautés nationales d'experts.
Pour qui a l'expérience de ce type de commission, il faut avoir le courage d'admettre que des décisions sur des normes quelquefois cruciales pour nos sociétés sont souvent prises par des assemblées d'experts qui pour certains d'entre eux maîtrisent parfaitement la compréhension des textes, mais pour d'autres « ânonnent leur traduction » tout en faisant semblant de comprendre. Je le répète, si ils ânonnent ce n'est pas parce qu'il ne comprennent pas les mots et les phrases dans lesquels sont rédigés les documents (souvent de l'anglais), mais c'est parce qu'à ce niveau de traduction ou d'interprétation (déjà compliqué pour eux mais qu'ils pourraient surmonter) vient s'ajouter

une avalanche de références, de sigles, d'acronymes (dont le pire est quelquefois qu'ils soient traduit dans la propre langue de l'expert lecteur [27]). Là encore, on comprend que les outils d'aide à la traduction et les terminologies électroniques sont indispensables.

Soyons en effet très direct par rapport aux risques supposés de manque à gagner pour votre activité professionnelle (la traduction et l'interprétation). Aussi lourds soient les crédits que nous attribuions à des tâches de traduction et d'interprétation, ils seront toujours très largement insuffisants pour couvrir tous les besoins d'une activité comme la normalisation. De notre point de vue il est clair qu'il faut plaider à nombre de niveau de décision de nos institutions pour que soient fait d'énormes efforts financiers pour impliquer toujours plus de professionnels dans des tâches de traduction ou d'interprétation pour la normalisation, mais : (1) la localisation (car s'en est une) des documents normatifs, doit se réaliser, comme la localisation des logiciels, en faisant coopérer un traducteur professionnel, des outils d'aide à la traduction et un expert normalisateur impliqué dans le processus normatif concerné. (2) les réunions d'experts sont souvent complexes, diplomatiquement délicates et deviendraient impossible si elles devaient fonctionner dans un environnement de traduction simultanée. D'autre part le nombre de communautés linguistiques engagées dans ces réunions internationales est tel qu'il serait irréaliste de l'envisager dans un nombre extensif de langue (ou même seulement avec les 3 langues officielle de l'ISO). D'autre part ré-insistons sur le caractère très spécialisé de l'activité qui entraînerait inéluctablement de très nombreux contresens, pire sans doute que la frustration de la non compréhension immédiate compensées (si l'expert est sérieux), par une traduction ultérieure à tête reposée dans laquelle s'impose encore l'importance des aides électroniques.

Le process d'élaboration des normes et les conséquences pour leur appréhension multilingue

Une fois résolu ce premier niveau de discussion entre experts, on abouti à des consensus [28] débouchant sur la rédaction puis l'édition de normes. Le processus d'élaboration des normes est très strictement réglementé tant au niveau national q u'international. Quand les discussions des experts sont globalement terminées, commencent les phases de rédaction provisoire du *Committee Draft*, CD), puis celle du *Final Committee Draft*, FCD), puis enfin de la norme expérimentale (*Draft International Standard*, DIS), puis de projet final de Norme internationale (*Final Draft international Standard*, FDIS) qui précède la mise en place au bout de 3 ans d'expérimentation d'une Norme Internationale. Heureusement pour la clarté des discussions entre les experts on ne traduit pas (à ce nous ayons pu voir), les sigles CD, FCD, DIS & FDIS. Eventuellement on les développe en français en leur laissant leur sigle d'origine anglaise. Ce niveau d'élaboration des normes est on la vu très codifié. Dans les niveaux préalables au CD, les éditeurs sont encouragés à demander des corrections ou aditifs techniques et rédactionnels, a u-delà, sauf problèmes graves ce seront surtout des corrections dites éditoriales (syntaxe, orthographe, modifications factuelles et locales). Nous rentrons à ce niveau dans des phases où des traducteurs pourraient être très utiles (comme seraient aussi utiles des outils d'aide électroniques à la traduction). Dans tous les cas seuls les experts éditeurs ont la responsabilité de la maîtrise d'une rédaction de ces documents.

Leur traduction contrôlée par expert normalisateur du domaine peut-être aussi très stratégique. Il faut savoir que dans les 3 langues ISO une telle traduction ne pourra être distribuée (par exemple en France par l'AFNOR que si elle a été relue, révisée et approuvé par la commission de normalisation concernée.

L a « communication marketing » des instances normatives avec les différentes communautés linguistiques ou nationales d'usagers des normes:

Des normes peuvent être éditée et même adoptées, elles ne sont rien si les utilisateurs ne se les approprient pas. Bien sûr certaines normes (notamment celles qui ont trait à la sécurité) peuvent être imposées par la loi. Nombres de normes (notamment en matière de communication) sont là pour faciliter les échanges, assurer l'interopérabilité mais elles n'ont pas force de loi.

Cependant, un Etat peu imposer légalement certaines normes pour des raisons de sécurité, de santé publique ou même pour s'assurer de la

qualité d'un type de produit ou de service sur son territoire.

Par contre dans nombre d'autre cas, la dissémination de la norme ne dépendra que de la capacité du métier concerné à s'approprier ces normes et à les intégrer dans le développement de leurs nouveaux produits et services.

C'est la raison pour laquelle les instances de normalisation suscitent très souvent la création d'un groupe de travail spécialisé : un « *marketing group* ». Au SC36 c'est le RG1 (*Rapporter Marketing group*). Pratiquement tous les SCs du JTC1 disposent ainsi d'un « marketing group ». Le groupe marketing à plusieurs rôles : il peut être prospectif et enquêter, par exemple pour connaître les besoins des usagers de la famille de normes ; mais son rôle fondamental consiste à conduire une politique de communication de l'instance normative pour produire et faire produire des documents permettant de vulgariser la famille de normes que produit l'instance normative (le SC36 dans notre cas). Ce travail est assez simple dans les pays qui ont pour langue nationale la (ou les) langue(s) de travail du groupe normatif. Il suppose cependant dès ce niveau un important effort de localisation des concepts par rapport aux cultures, aux institutions et aux dispositifs législatifs ou réglementaires locaux. Pour les pays d'une autre langue que la langue de travail, la question se complique beaucoup.

Cette communication marketing des normes est elle aussi une activité très délimitée pour laquelle des traducteurs professionnels (et/ou des équipements d'aide à la traduction) seraient utile

L'environnement mondial ou national de la normalisation et son mode de fonctionnement

Nous attardons longuement sur la normalisation non seulement pour expliquer, mais aussi parce que l'organisation hiérarchique et fonctionnelle de ces différentes commissions, comités, sous-comités et autres sous niveaux de groupe de travail constituent un univers complexe qui pose de très nombreux problèmes de traduction et d'adaptation au réalité linguistiques, culturelles, économiques, organisationnelles de chaque nation. La question des sigle et acronymes qui servent à désigner ces instances normatives, ainsi que d'ailleurs que celles des nombreux processus d'élaboration des normes (décrit ci-dessus) est importante et nous avons déjà souligné plusieurs fois combien ces sigles compliquaient très souvent la communication multilingue. Doit-on traduire ces sigles ou acronyme, ou au contraire (c'est notre position) partager en commun entre un maximum de langues ces formes abrégée pour éviter la babélisation totale entre les experts.

Là encore, nous faisons l'hypothèse que des outils terminologiques sophistiqués sont seuls à même de faire face aux enjeux de la mondialisation et de localisation des normes, d'où l'obligation de disposer d'une alliance comme Cartago pour aider à la maîtrise de ces questions dans le cadre de la normalisation des TICE.

Au niveau international les instances principales de normalisation sont l'ISO [29], la CEI (en anglais IEC [30]), l'UIT (en anglais ITU [31]), et d'autres instances mais qui sont spécialisées hors du champ de la normalisation des TIC (Technologies de l'Information et de la Communication). A un niveau géographique plus régional nous trouvons le CEN.

Au niveau national, la normalisation s'élabore grâce à des instituts nationaux, des NBs (National Bodies) dans le jargon de l'ISO. Ces NBs constituent des « miroirs » qui servent à relayer, à co-élaborer (notamment voter), à adapter, à discuter les normes internationales. Ces instituts nationaux sont aussi l'instance qui dans un pays organisent des normes nationales (le plus souvent en synergie mondiale mais quelquefois à l'écart, voire de façon contradictoires avec les normes internationales).

Le CEN occupe une position intermédiaire : il n'est pas considéré comme une instance nationale (par exemple il n'a pas le droit de vote à l'ISO) mais suivant en cela les directives européennes, une norme européenne adoptée au CEN s'impose de facto dans tous les pays européens et vient remplacer les normes nationales correspondantes préexistantes.

Attardons nous encore sur l'importante question du découpage institutionnel de la normalisation. Il ne peut être compris qu'autant qu'on considère l'histoire de la normalisation et l'entrée en lice progressive de nouvelles technologies électriques ou de communications qui ne pouvaient se développer si tous les acteurs du domaine (fut-ils concurrents) (1) ne développaient pas des réseau communs (électricité, télégraphe, radio et TV) (2) ne s'accordaient pas sur la normalisation de produits semi-finis (composants électriques et électroniques) à l'instar de leurs collègues de la métallurgie.

La fin du 19ème siècle et la première moitié du 20ème siècle fut ainsi l'époque de

création des grandes instances internationales et nationales de normalisation. La CEI [32] a été fondée en 1906, l'UIT [33] sous sa première appellation d'Union internationale du télégraphe est crée en 1865. L'ISO, elle aussi situé à Genève est fondée en 1947 pour achever l'organisation normative de la quasi-totalité des produits et services dans le monde. C'est le plus grand producteur et éditeur mondial de Normes internationales>et c'est aussi un réseau d'instituts nationaux de normalisation de 158 pays [34], selon le principe d'un membre par pays, dont l'ISO assure le Secrétariat central et la coordination d'ensemble.

Parce que le nom "Organisation internationale de normalisation" et « International Organisation for Standardisation» aurait donné lieu à des abréviations différentes selon les langues (IOS en anglais et OIN en français), ses fondateurs ont opté pour un nom court, universel: "ISO". Ce nom est dérivé du grec *isos*, signifiant "égal". Quel que soit le pays, quelle que soit la langue, la forme abrégée du nom de l'organisation est par conséquent toujours ISO.

Au niveau international la structure hiérarchique est relativement claire. L'ISO et l'IEC sont subdivisés en TCs (en anglais ISO *Technical Committee* en français AFNOR, CCN, NBN & SEE : *Comité Technique : CT*), eux même subdivisés en SCs (*Sub- Committee* [35] en français AFNOR, CCN, NBN & SEE [36] : *Sous-Comité SC*), eux même subdivisés en WGs (*Working Group*, en français AFNOR : *Groupe d'Expert*, *GE* par contre pour CCN, NBN & SEE : *Groupe de Travail, GT*).

La traduction des noms et sigles des différentes instances, sous-instances et sous-sous-instances qui composent une instance nationale (voire la traduction en langue nationale des sous instances de l'ISO de l'IEC ou de l'UIT) peuvent apparaître comme successible d'améliorer la compréhension entre des experts (ou du public spécialisé) non obligatoirement anglophone. En réalité ces traductions (et pire encore le développement des sigles entraînent rapidement des confusions.

En effet la structure hiérarchique des instances et sous instances des NBs devient vite très complexe (notamment parce qu'elle est rarement exactement similaire à la structure hiérarchique des instances internationales.

Le niveau international

Dans les instances internationales (ISO, IEC et UIT [37]), la séparation en TCs [TGs], SCs [SGs] puis WGs est clairement la même pour tous les experts du monde entier. Elle est relativement uniforme et découpe les activités soumises à normalisation en autant de TCs, etc…

L'ISO et l'IEC sont ainsi divisé chacun en environ 200 comités techniques (TC), tous chargé de normaliser un domaine particulier (Voir en annexe 1 des extraits des listes de comités).

On peut souligner l'importance d'un TC très particulier, le JTC1 (*Joint Technical Committee*) commun à l'IEC et à l'ISO qui a été créé pour rassembler en autant de SCs des instances chargé de normaliser (à l'exception des TC37 et 46 de l'ISO la quasi-totalité des instances normalisant les TIC [38].

Insistons bien sûr sur la structure du SC 36 "Informatique pour l'éducation, la formation et l'apprentissage". Il se découpe au niveau international en 7 WGs, un SWG (Spécial Working Group) et un RGn°1 (Rapportor Group).

Focalisons nous sur le premier Working Group nous concerne directement ici puisqu'il définit le vocabulaire (et non la terminologie) du domaine: l'ISO/IEC JTC1-SC36 WG1 (en forme courte SC36WG1, voire en contexte, WG1).

Ne serait-ce qu'à ce niveau organisationnellement univoque nous laissons à penser la grande confusion qu'introduit la seule déclinaison de ces sigles en français (et/ou en russe, la 3ème langue de l'ISO avec l'anglais). En français nous pourrions donc écrire comme équivalant : ISO/CEI JTC1-SC36 GE1. Notons que nous n'avons jamais vu traduire JTC1, ni ISO (par définition), et nous avons la chance que *Sub-Commettee* se développe en français par Sous-Comité.

Mais nous ne sommes ici qu'au tout début des confusions possibles. Jusqu'à présent nous n'évoquons qu'une dimension de complexité : celle dans laquelle nous pouvons décrire en n langues (et notamment dans les 3 langues officielles de l'ISO) un seul et même univers conceptuellement cohérent, celui des institutions internationales de normalisation découpées en TCs, SCs et WGs plus quelques sous découpages atypiques (RGs, Comités bis, SWG, Special Working Group…).
Nous avons donc une syntaxe de type : *ISO TCxxSCyyWGzz* ou *ISO/IEC JTC1 SCxx WGyy*. Au plus nous risquons des confusions de contexte dues à l'usage des sigles courts. Ainsi par exemple « SC 02 [39] » est évidement plus ambiguë énoncé par son

seul sigle que « SC 36 » ; qui court très peu de risque d'être confondu avec la subdivision d'un autre TC de l'ISO ou de l'IEC [40]. D'évidence l'énoncé JTC1 SC 02 ne court aucun risque d'être ambiguë.

Evidement tous les rédacteurs d'un document sont libres et on ne peut empêcher un expert du SC36 de rédiger un document dans lequel en contexte il référera au WG4 ou au RG1. L'inconvénient, c'est que sorti de son contexte un tel énoncé devient parfaitement incompréhensible.

Deux catégories de délégation dans une instance internationale de normalisation : les NBs et les Liaisons

On peut distinguer deux catégories de délégation dans une instance internationale de normalisation : les Etats, représentés par leur institution nationale de normalisation désignées sous le terme générique de NBs et les Liaisons.

Le principe des liaisons, introduit une deuxième catégorie de délégations parallèle à celle des NBs. Dans le jargon de l'ISO on définit globalement l'ensemble des délégations sous l'acronyme global de NBLO (*National Bodies & Liaisons*, NBLOs au pluriel). Cela comprend donc : les NBs P members, (participant, c'est-à-dire ayant payé leur cotisation) qui disposent d'une voix par pays ; les NBs O members (pays observateurs), qui disposent d'une voix consultative et les Liaisons qui disposent elles aussi de voix consultatives.

Les liaisons sont de plusieurs ordres :

- Des liaisons techniques internes c'est-à-dire établissant un lien de coordination avec d'autres SCs à l'intérieur du JTC1 : par exemple le SC29, 32, 37, etc… auquel est lié le SC36
- Des liaisons techniques externes c'est-à-dire.établissant un lien de coordination avec d'autres instances de normalisation extérieure au JTC1 : par exemple l'ISO TC46 ou l'ISO TC37 auquel est lié le SC36
- Des liaisons A et C qui sont établies avec des institutions de dimension internationales qui collaborent étroitement et participent à l'avancement d'une famille de norme : c'est le cas par exemple pour le SC36 de l'IEEE, d'ADL, d'IMS, du DCMI, de l'AICC et de l'AUF (on a noté que les 2 auteurs composent cette liaison A de l'AUF auprès du SC36.)

Des instituts nationaux de normalisation et les remarques que l'on peut déduire de leur dénomination

Nous avons cherché à orienter tout cet exposé non seulement pour expliquer un domaine d'activité, mais aussi pour le regarder du point de vue multiculturel et multilinguistique. Nous avons aussi orienté notre exposé pour mettre en évidence les difficultés de traduction. De ce point de vue, le passage du niveau international complexe mais conceptuellement homogène, se complique considérablement quand on considère la disparité des découpages nationaux.

On aurait pu supposer que l'ISO, comme instance fédératrice de la normalisation, aurait du savoir imposer (par construction pourrait-on dire) une certaine uniformité et univocité terme à terme de ses procédures et de ses structures institutionnelles. Il n'en est rien à l'évidence et l'histoire des organisations internationales, comme celle des nations, intriquée encore dans celle des langues nous apprend beaucoup sur la diversité du monde et des cultures.

Plutôt de décrire ces institutions normatives comme un univers évidement complexe nous préférons continuer de les découvrir en soulignant leur diversité historique, organisationnelle et de plus linguistique.
Somme tout le traducteur est confronté à cette diversité dans un univers qu'il aurait pu supposer lisse et sans accidents. Il se doit d'admettre définitivement que les normes établis par les télécommunications ou l'électricité et l'électronique constituaient un tel patrimoine déjà organisé qu'il aurait été stupide d'uniformiser pour l'intégrer à la structure nouvelle et désormais fédératrice de l'ISO.

Les noms des instituts nationaux de normalisation correspondent à l'évidence à des institutions très disparates quand à leur mode de fonctionnement [41] et au champ total de leur objectifs. Il est aussi nécessaire de prendre en compte l'échelle de granularité institutionnelle, en général proportionnellement adaptée à la dimension des états ou aux ressources économiques ou humaines d'un pays : un pays comme le Luxembourg par exemple a créé l'Organisme Luxembourgeois de Normalisation au

sein du SEE (Service de l'Energie de l'Etat) alors que le Pérou intègre la normalisation dans le même institut que la protection de la propriété intellectuelle.

La question devient vite très complexe lorsqu'on change d'échelle de granularité ; que l'on ne veut pas seulement chercher à traduire le nom d'instituts nationaux de normalisation [42], mais que l'on cherche des équivalents pour des Comités techniques, des sous-comités des groupes de travail qui constituent autant de commissions d'experts susceptibles de se réunir et de coopérer au niveau national et international.

L'obligatoire diversité organisationnelle des instituts nationaux

Chaque institution nationale de normalisation doit gérer parallèlement 2 niveaux distincts d'organisation :

> (1) Celui de la transmission à l'identique des commissions et comités internationaux (ISO, IEC, JTC1, IUT) et,
>
> (2) celui de l'organisation locale « en miroir » (selon le jargon des professionnels) de commissions, comités, sous-comités et autre groupe de travail.

Ainsi à l'AFNOR le comité miroir du SC36 s'est appelé la CN36 (Commission Nationale 36) et il s'appelle actuellement le GC36 (Groupe de Coordination 36). A un autre niveau la CGTI (Commission générale des Technologies de l'Information de l'AFNOR) est l'instance miroir du JTC1.

Du point de vue du traducteur il faut bien comprendre que l'un (la transmission à l'identique des commissions et comités internationaux) n'est évidement pas la traduction de l'autre (les comités miroirs). Les comités miroirs sont nationaux, ils correspondent à des groupes d'experts nationaux, dans lesquels se recrute le sous ensemble qui sera délégué dans les instances internationales. Il est donc logique que, potentiellement, l'on soit dans ce cas face à 3 termes et les sigles ou acronymes correspondants : les 2 termes anglais et français désignant l'instance (le comité, sous-comité) ISO, IEC ou UIT ; le terme français désignant l'instance miroir.

On est d'ailleurs très souvent gênés comme expert international pour en parler avec nos collègues des autres NBs car on ne dispose pas de mots en anglais (et *a fortiori* dans d'autre langues) qui désignerait ces miroirs. On dit donc tout simplement la « (sous)commission miroir ». Mais, dans certain cas, ce peut être très vague car chaque NB, du fait de sa taille relative, de sa structure institutionnelle, de son mode de mise en œuvre de la normalisation devra créer des découpages des miroirs adapté à sa réalité des ressources humaines locales.

C'est ainsi qu'à l'AFNOR le CG36 (miroir du SC36) est découpé en GE (Groupes d'experts qui ne peuvent pas dans tous les cas recouvrir les WGs correspondants au SC36. Si par simplicité de dénomination, nous essayons de calquer absolument la sous hiérarchie des WGs pour les GE nous manquerions d'experts. Ainsi le CG36GE4 correspond à l'identique au SC36WG4. Par contre CG36GE3 est en fait en charge du SC36WG3, 5&7. Il existe aussi un CG36GE-CDM (Course Description Metadata), non numéroté comme on peut le constater. CG36GE-CDM est le miroir d'une problématique d'étude mise en place au niveau du CEN, mais non présente à ce jour au SC36.

Construire des terminologies normalisées : une nécessité aux fondements de la maîtrise multilingue de la mondialisation numérique

Par méthode la quasi-totalité des instances de normalisation et plus particulièrement celles qui normalisent les TIC normalisent une terminologie de leur domaine.

Cette terminologie a pour objectif de clarifier la communication entre les experts pour être capable de rédiger les normes du domaine de façon cohérente : vérifier que l'on parle bien des mêmes choses (des mêmes concepts, et non pas seulement des mêmes mots) : des mêmes produits, mêmes composants, mêmes logiciels, mêmes services, mêmes processus, mêmes ressources, mêmes institutions, mêmes acteurs, mêmes rôles, etc. Cela doit être fait au niveau international dans 1, 2 ou potentiellement autant de langues de travail que les experts peuvent en maîtriser ensemble et cela doit être aussi vérifié avec les réseaux d'experts auxquels chaque expert international est lié : notamment la maîtrise en « miroir » avec les experts référents dans leur NBs ou Liaisons [43].

Cette terminologie a pour objectif second de constituer un référentiel pour les futurs

utilisateurs de la norme : s'assurer que comme les experts, ils comprennent bien aussi les mêmes choses (les mêmes concepts). La difficulté du second objectif réside en ce que ces utilisateurs finaux (a) ne maîtrisent pas nécessairement de façon experte la totalité des concepts proposés par les experts internationaux (b) ne maîtrisent pas toujours les langues de travail des experts internationaux (c) cependant ces utilisateurs finaux sont confrontés à des particularismes culturels, linguistique, institutionnels (dans le cas du SC36 à des situations pédagogiques singulières) qu'il s'agit de préserver dans les situations locales réelles où ils ont besoin de mettre en œuvre les normes considérées.

Dès lors il devient indispensable de s'entendre sur des méthodes de construction d'une terminologie (un référentiel sémantique) devient dès lors indispensable pour permettre l'interconnexion et l'interopérabilité des matériels et des réseaux ainsi que celle des ressources notamment pour qu'elles puissent circuler et interagir malgré la diversité des langues.

Le SC36 s'est ainsi conformé aux règles de la majorité des instances normatives et il a crée un groupe de travail : le SC36-WG1 : *terminologie* qui se restreint assez vite pour devenir un groupe *vocabulaire*. Il s'avère que pour des raisons diverses la normalisation de l'e-learning a surtout été amorcée par les « pionniers du domaines » essentiellement des industriels de *l'e-training* aéronautique, militaire et de sécurité auxquels se sont joint des enseignants européens des sciences exactes et expérimentales. Ces pionniers s'accommodent incomparablement mieux que les nouveaux entrants actuels d'un unilinguisme pragmatique (c'est le cas quasi à 100% dans le secteur aéronautique), d'une relative univocité des concepts des matières enseignées et d'autre part ce sont des domaines pédagogiques dans lesquels les débats sur la diversité des cultures pédagogiques sont moins présents que dans les arts ou les lettres.

Cette première prépondérance des pionniers s'estompe aujourd'hui (même s'ils représentent toujours une part importante du marché global), les cultures linguistiques autres qu'anglophones et européennes (notamment Corée, Chine et Japon) affirment d'année en année un dynamisme et une croissance importante. Le Sud-est asiatique pourrait à lui seul multiplier le marché de façon considérable, et pourtant ce n'est que très récemment que cette partie du monde commence à affirmer son droit à la diversité linguistique et culturelle dans le processus normatif [44].

L'évolution historique du SC36 qu'il serait hors sujet de relater en détail fait que nous disposons maintenant d'un groupe de travail produisant un vocabulaire à ce jour bilingue (anglais/français), co-élaborée en consensus dans les deux langues. Ce vocabulaire n'a pas vocation à être très important [45]. Par décision de l'instance hiérarchiquement supérieure au SC36 (le JTC1) ce vocabulaire devrait en principe être mise en commun avec tous les autres vocabulaires ou terminologies du JTC1 dans une base terminologique commune gérée par Termium [46].

De nombreux problèmes restent ouverts :

- Les experts des TICE ne sont pas forcément sensibilisés aux problèmes méthodologiques de la terminologie et de la terminotique. Il a été long et difficile pour les quelques délégués du SC36 qui étaient sensibilisés aux méthodes terminologiques de faire comprendre au groupe de travail vocabulaire (SC36-WG1) que les termes anglais n'étaient pas équivalent à des concepts universaux

- Il a été aussi difficile (face au groupe des professionnels et animateurs de la normalisation [47]) de faire comprendre que contrairement à la plupart des autres normes qui portaient jusqu'à ces dernières années sur des produits ou des services très techniques et très matériels, le SC36 appartenait (et même était emblématique) des nouveaux champs de normalisation beaucoup plus culturels et sociétaux.

- D'évidence les normes des TICE achoppent sur des domaines très politiques, sur des options éthique, sur le champ du linguistique [48]… C'est pourquoi les experts en normalisation (notamment les professionnels des institutions de normalisation) ont du mal à accompagner (voire à comprendre) cette mutation évolutive de la normalisation des produits matériels vers la normalisation de services sociétaux [49].

- La production de ces normes de service sociétaux exige (non plus la mise en place de terminologies sous forme de nomenclatures techniques [50]) mais de véritable terminologies (voire d'ontologie) beaucoup plus vastes et sophistiquées, référent à la diversité des cultures, des langues, des nations sans pour autant perdre de vue que tout projet normatif mondial exige des référentiels sémantiques universels permettant précisément l'interopérabilité des réseaux ou des échanges

de ressources.

- On comprend bien en effet que la terminologie du SC36 appartient à l'évidence

    - pour une part à des nomenclature techniques relevant du premier cas (système de gestion d'apprentissage, environnement d'apprentissage, métadonnées pour des objets d'apprentissage, système de gestion de contenu, environnement virtuel d'apprentissage…) d'évidence ces notions ne posent pratiquement aucun problèmes quand à leur découpage dans la diversité des langues : les TIC se sont construites presque partout selon le même découpage conceptuel [51].

    - Par contre le problème se complique quand on s'attaque aux termes généraux, apprentissage, éducation, formation ou même des termes comme *e-learning*, *blended learning* qui posent des problèmes de diversité de traduction et surtout d'appréhension conceptuelle très délicat. L'interopérabilité entre terme devient inextricable quand on décrit un apprenant ou un enseignant et ses rôles, des institutions éducatives, des niveaux d'apprentissage, des modes de diplomation.

- D'évidence il ne peut être question d'envisager que l'ISO, l'IEC, le JTC1 se substituent à l'UNESCO ou aux grandes agences mondiales impliquées dans la culture ou l'éducation pour ce qui est de la réalisation de grandes banques terminologiques ou de vastes terminologies qui serait susceptible d'accompagner l'organisation normative (notamment la capacité d'interopérabilité donc de circulation équitable) des ressources et des acteurs des TICE. Ces enjeux sont stratégiques on le sait, mais ce n'est pas le rôle direct des instances de normalisation. Il ne peut être de la responsabilité directe du SC36-WG1 de réaliser des terminologies de plusieurs milliers de termes dans un nombre extensif de langues pour assurer l'interopérabilité normative de ce domaine technologique.

- Après que le problème a été longuement discuté un nombre significatif d'experts sont maintenant persuadés que seule une approche privilégiant les concepts définis en consensus vers lesquels « pointent » des termes organisés eux-mêmes chacun dans leur langue, peut clarifier la question. C'est la méthode onomasiologique, informatisable grâce aux savoir faire terminotique ; c'est tout cela qui a été normalisé à l'ISO TC37 grâce à l'effort de plusieurs décennies de travail par les normalisateurs terminologues et terminoticiens.

- Cependant le SC36-WG1 continue sa première démarche en faisant progresser une liste bilingue de vocabulaire. La première édition (en cours de normalisation définitive sous la référence ISO/IEC 2382-36) fait déjà place au chantier d'une seconde édition augmentée. Cette démarche est de notre point de vue parfaitement légitime à condition (ce qui est le cas) que le SC36 envisage de collaborer en synergie avec une instance qui prenne le problème selon des méthodologies plus adaptées à l'aspect sociétal et culturel que constitue les TICE. C'est ce qui est actuellement en cours de mise en place avec Cartago.

>>>>> Suite

---

[1] Un tel sigle nous semble très pertinent par rapport aux questions de la sémantique polysémique des sigles (dans une même communauté de langue, mais pas exclusivement) que nous abordons ci-dessous. En effet dans le contexte académique français CNU se décline comme Conseil National des Universités. C'est l'instance collégiale qui évalue tous les universitaires, ce qui a donc des conséquences importante sur leur carrière.



[2] Vian (Boris), Vercoquin et le plancton, Paris, édition Gallimard, 1947. Boris Vian était ingénieur de l'Ecole des Arts et Manufactures (École Centrale). En 1943, il devient ingénieur à l'AFNOR où il participe à l'élaboration de normes (nothons). Vercoquin et le plancton est un roman à clefs qui est passé dans la légende de l'honorable institution de normalisation.



[3] Commençons par définir ce « nous » qui s'adresse à vous. Au-delà des deux auteurs, ce « nous » recouvre en fait un « nous collectif, celui d'une instance de normalisation » et plus précisément la Délégation de l'AUF auprès de l'ISO.

[4] La terminotique s'intéresse aux outils informatiques pour la collecte, la production, le traitement, l'échange et la diffusion des données terminologiques. Ces données terminologiques seront présentées sous la forme d'une base de données, d'un corpus. Définition in <www.owil.org/lexique/t.htm>

[5] Je pense notamment à Daniel Prado à l'Union Latine, à Laurent Romary et l'équipe du Loria à Nancy, à Elisabeth Blanchon (Université de Paris Jussieu) et à André LeMeur (Université de Rennes) a qui nous devons tous les développement Génétrix qui constituent le cadre méthodologique et support technique actuel de Cartago.

[7] ISO/IEC JTS1 SC36 (en sigle court JTC1 SC36 ou même SC36) : « Normes des technologie de l'information et la communication pour l'enseignement et la formation ». Tentons ici un exemple que nous voudrions exemplaire de l'exercice auquel nous sommes constamment confrontés : celui du développement bilingue dans un contexte de normalisation du sigle ISO/IEC JTC1 SC36 (en sigle court SC36).
ISO (originellement en 1946 *International Organisation for Standardization*, en français Organisation Internationale de Normalisation.
IEC (*International Electrotechnical Commission)* fondée au début du 20ème siècle le sigle a donc son équivalent français CEI (Commission électrotechnique internationale), cependant l'agglomération de sigle ISO/IEC JTC1 SC36 est déjà très ésotérique et on envisage de moins en moins d'écrire : ISO/CEI JTC1 SC36.
JTC1 (*Joint Technical Committee 1*, Comité technique commun n° 1).
SC36 (Sous-comité n° 36) là aussi les équivalents francophones existent théoriquement mais on imagine mal de multiplier pour chaque langue des équivalents de sigle et d'énoncé développé pour signifier des notions partagées par une communauté relativement confidentielle d'experts.

[8] *Norma(ae)* l'équerre, mot que l'on trouve chez Vitruve ou au sens figuré, la règle et la loi acception que l'on trouve chez Cicéron.

[9] Standard vient du même mot en anglais, mais le mot est d'origine anglo-normande (*estandard* : étendard)

[10] Grand Larousse encyclopédique$$ref$$, entrée « standard ».

[11] Le terme *norm* existe en anglais mais n'est jamais utilisé dans le contexte des institutions de normalisation. De nombreux experts ISO (américains, canadiens, britaniques, irlandais, australiens ou néo-zélandais) que nous avons interrogé sur ce point confirment la totale inadéquation du terme *norm* dans le domaine de la normalisation. Selon l'Oxford Dictionary: norm • noun (1) (the norm) the usual or standard thing. (2) a required or acceptable standard. —ORIGIN Latin *norma* 'precept, rule, carpenter's square'.

[12] Il va de soit qu'en « français non spécialisé » les termes normes et standards continuent à être utilisés comme des quasi-synonymes.

[13] Voir plus loin en note une liste de ceux-ci.

[14] On peut souligner que l'Italie emploie le terme *unificazione* même si existe en italien le couple de terme : *standard / norma*.

[15] Sui équivalent à des standards quand elles sont des spécialités exclusives d'un laboratoire, et qui deviennent à terme des « espèces de normes » quand elles peuvent être déclinées sous formes de médicament génériques.

[16] La plupart des technologies émergeantes ont cette maladie infantile (l'automobile qu'on ne pouvait conduire à ses début si on ne savait pas ouvrir le capot, l'écoute radio jusqu'au milieu des années 20 qui exigeait de l'auditeur qu'il sache « bricoler la galène ». L'informatique en réseaux, parce qu'elle reste au mains d'experts informaticiens qui veulent décider entre chercheurs et industriels (mais aussi le *Departement of Defense USA*) ne parvient pas à se stabiliser comme un marché mondial sécurisé de l'échange d'information. Ils n'admettent pas les délais (très longs) de mise en place des normes mais fondamentalement la plupart des Etats (hormis les USA qui en bénéficient) admettent de moins en moins l'irruption de cet espace de non droit, la menace toujours possible que les USA coupent les circuits, le risque quotidien que constitue l'insécurité de cet espace numérique en évolution technologique permanente et imposée.

[17] Certaines normes stratégiques pour les Etats peuvent être réservées (non publiques). C'est le cas par exemple de certaines normes produites par le JTC1 SC37 « Outils de contrôle biométriques ».

[18] Ce discours peut paraître abstrait et pourtant cela nous parait tout naturel pour certain produit. On comprend tous qu'une douille électrique (par exemple à baïonnette) doive répondre à des normes (IEC) pour des raisons de sécurité et d'interopérabilité entre fabricants. Cela n'empêche pas, on le constate, une prolifération des formes, des matériaux (métal, porcelaine, matière plastique), des solutions de câblage tout en restant conforme à la même norme.

[19] Ainsi MPEG se décline *Moving Picture Experts Group* c'est-à-dire un *groupe* d'experts.

[20] Nous verrons plus loin ces étapes CD, FCD, DIS, FDIS, avant d'aboutir à une Norme Internationale (ou nationale ce sont des étapes similaires) éventuellement révisable.

[21] Ce peut être par exemple l'espagnol pour des normes élaborées dans le cadre global de l'Amérique latine.

[22] On parle de perte de domaine (ou perte de fonctionnalité) lorsque, dans un domaine donné, on perd la possibilité de s'exprimer, de travailler ou d'enseigner dans sa langue, avec les conséquences culturelles que l'on imagine. Initialement identifié par les universitaires scandinaves, ce phénomène préoccupe actuellement les instances de recherche tout autant que celles d'aménagement linguistique, tant au Québec qu'aux Pays-Bas ou dans les pays scandinaves. C'est aussi, au niveau européen, l'un des axes de travail de la toute nouvelle Fédération européenne des instituts linguistiques nationaux.

[23] Didier Oillo et Bonaventure Mvé-Ondo, (sous la dir. de), Fractures dans la société de connaissances, ouvrage collectif in Hermès n° 45, CNRS éditions, Paris 2006.

[24] Une instance de normalisation est animée par un président (chairman), matériellement assisté par un Secrétariat général (en général soutenu par un NB qui en prend la responsabilité, dans le cas du SC36, c'est le BSI, *British Standards Institute*). Toutes les sous-instances (les WGs ou RG éventuels), sont elles mêmes animées par des Présidents de WGs ; s'ajoutent à cela des éditeurs de normes, les Présidents des différents NBs (et des Liaisons). L'ensemble constitue une sorte d'état major hiérarchisé qui anime la production d'une norme. Remarquons toutefois que ces animateurs doivent avoir pour soucie constant la recherche du consensus maximum (ce qui les distingue d'un état major militaire ou industriel) et le verdict constant des votes des NBs. Ces animateurs doivent aussi veiller constamment à ne pas être en contradiction avec les instances ISO, IEC, etc… dont ils sont des sous-instances et qui ont sur eux un contrôle lui aussi soumis aux mêmes règles consensuelles et à des processus de vote des Etats membres.

[25] ISO/IEC 2382-36 Technologies de l'information - Vocabulaire. Le titre même de cette norme limite bien la limite de l'exercice : le SC36WG1 (groupe dédié en principe à la terminologie ne produit en fait qu'un vocabulaire.

[26] Comité technique de l'ISO sur la terminologie les autres ressources langagières

[27] L'expert habitué aux sigles cités dans les travaux internationaux (très souvent des sigles anglophones) peut être totalement désorienté quand on lui propose un sigle décliné dans sa propre langue. De façon générale il est très difficile d'appréhender des sigles traduits tant en lecture qu'en écoute de texte étranger, par exemple ITC= ITC ; SMSI=WSIS).

[28] Les normes se construisent en privilégiant au maximum le consensus. Bien sûr elle sont approuvées par des votes majoritaire des Etats participants mais avant de passer aux votes formels bureaucratiquement organisés entre tous les NBs participants les présidents de groupes de travail concernés, doivent explorer toutes les possibilités d'amandements, de modifications issues de commentaires formellement proposés par les NBs, d'ajouts, pouvant permettre d'atteindre le consensus ou la quasi-unanimité. C'est pour cela d'ailleurs que tous les votes négatifs (entre les phases d'avancement de la norme CD à FCD àDIS) doivent être obligatoirement accompagnés d'un commentaire écrit argumenté motivant le vote négatif.

[29] Nous verront dans quelques lignes comment l'ISO a pris une position exemplaire pour ne pas risquer la babélisation de son sigle.

[30] Commission Electrotechnique Internationale (CEI) ou *International Electrotechnical Commission (IEC)* en anglais.

[31] UIT (Union internationale des télécommunications), en anglais ITU, *International Telecommunication Union*.

[32] L'IEC (CEI en français) est un organisme de normalisation traitant des domaines de l'électricité et de l'électronique. La CEI a été créée en 1906 et compte actuellement 67 pays participants. Depuis 1948 elle est située à Genève comme l'ISO.

[33] C'est la plus ancienne organisation internationale technique de coordination. En 1932 elle devient L'UIT (en anglais ITU). Elle gère aussi l'attribution des bandes de fréquences radioélectriques pour la communication hertzienne (IFRB : *International Frequency Registration Board*). Elle est rattachée aux Nations Unies ; son siège est à Genève.

[34] Nous donnons ici un échantillon des noms, abréviations et acronymes d'un certain nombre d'instituts nationaux en insistant sur ceux des pays de langues de langues latines : **Algérie:** IANOR Institut algérien de Normalisation, **Argentine:** IRAM Instituto Argentino de Normalización y certificación, **Belgique:** NBN Bureau voor Normalisatie - Bureau de Normalisation, **Brésil:** ABNT Associaçào Brasileira de Normas Ténicas, **Canada:** SCC/CCN Standard Concil of Canada/Conseil Canadien des Normes, **Chili:** INN Instituto Nacional de Normalización, **Colombie:** ICONTEC Instituto Colombiano de Normas Ténicas y certificación, **Costa Rica:** INTECO Instituto de Normas Ténicas de Costa Rica, **Equateur:** INEN Instituto Ecuatoriano de Normalización, **France:** AFNOR Association Française de Normalisation, **Italie:** UNI Ente Nazionale Italiano di Unificazione, **Liban:** LIBNOR Lebanese Standards Institution, **Luxembourg:** SEE Organisme Luxembourgeois de Normalisation a été créé au sein du SEE (Service de l'Energie de l'Etat), **Maroc:** SNIMA Service de Normalisation Industrielle Marocaine, **Pérou:** INDECOPI Instituto Nacional de Defensa de la Competencia y de la Protección de la Propiedad Intelectual, **Portugal:** IPQ Instituto Português da Qualidade, **Roumanie:** ASRO Asociaţia de Standardizare din România, **Espagne:** AENOR Asociación española de Normalización y certificación, **Suisse:** SNV Schweizerische Normen-Vereinigung/Association Suisse de Normalisation, **Tunisie:** INNORPI Institut National de la Normalisation et de la Propriété Industrielle, **Uruguay:** UNIT Instituto Uruguyao de Normas Ténicas, **Venezuela:** FONDONORMA Fondo para la Normalización y Certificación de Calidad, **Allemagne:** Deutsches Institut für Normung, **Autriche:** Instituut Österreichisches Normungsinstitut, **Bulgarie:** BDS Bulgarian institute for standardization, **Grande Bretagne:** BSI British Standard Institute, **Hollande:** NEN Nederlands Normalisatie, **USA:** ANSI American National Standard Institute.

[35] A l'UIT on les appelle TGs et SG (Technical Group & Sub Group).

[36] Comme signalé en note ci-dessus, ce sont respectivement, les institutions de normalisation de la France, du Canada (versus francophone), de Belgique et du Luxembourg.

[37] Pour être cohérant nous devrions dire soit CEI et UIT, soit IEC et IUT mais CEI est très rarement traduit ce qui fait que le sigle francophone est très peu connu, alors que le sigle français UIT est couramment employé dans la presse généraliste.

[38] Le JTC1, créé en 1987 par convention entre l'ISO et l'IEC est l'organe de référence pour la normalisation des Technologies de l'Information au niveau mondial. C'est un comité technique commun à ses deux parents (l'ISO et l'IEC), il réunit les compétences de l'ancien TC97 de l'ISO relatif aux logiciels (langages de programmation, codage de l'information, ...) et les compétences de comités techniques de la CEI en matière de matériels : microprocesseurs, imprimantes, par exemple. Les quelques 2000 normes publiées par le JTC1, quoique appelées souvent normes ISO, sont reconnaissables par leur numéro de nomenclature qui commence par les deux sigles ISO/IEC). Le JTC1 est composé aujourd'hui de 17 SCs actifs (voir annexe 1).

[39] ISO/IEC JTC1 SC2 : Jeux codés de caractères.

[40] Mis à part le JTC1 il ne doit pas exister de TC comportant un nombre aussi important de SCs.

[41] On peut cependant souligner que ces institutions ne sont généralement pas directement intégrées dans la structure des gouvernements d'un Etat proprement dit.

[42] Ce qui est assez simple et se réduit à une nomenclature de 158 institutions et de leurs sigles associés. Notons cependant le cas particulier des Etats qui traitent de la normalisation non pas dans une institution autonome exclusive mais dans un sous service d'une institution aux attributions plus larges (nous le signalons plus haut dans cet article l'exemple du Luxembourg (SEE, Service de l'Energie de l'Etat), du Pérou (Compétence et propriété intellectuelle) ou de la Tunisie (normalisation et propriété industrielle).

[43] Beaucoup de NBs et certaines Liaisons réalisent des terminologies ou des vocabulaires : dans le cas du SC36 on doit citer notamment : la GB, le Canada, la Corée, le Japon, la France, l'Allemagne… Pour ce qui est des Liaisons on peut citer l'AICC, IMS, IEEE qui ont tous des terminologies considérables sur les TIC (et aussi sur les TICE).

[44] Voir Henri Hudrisier, Société de connaissance, le paradigme de l'appropriation, ou encore Mokhtar Ben Henda, Les contradictions d'une politique de diversité culturelle, in Hermès 45 « Fractures dans la société de connaissances », Sous la dir. de Didier Oillo et Bonaventure Mvé-Ondo. CNRS éditions, Paris 2006.

[45] Quelques centaines de termes au maximum.

[46] C'est la base de données terminologique Termium du Bureau fédéral de la traduction canadien. Celle-ci a offert ses services au JTC1 pour collecter et devenir le lieu de référence de la terminologie de tous les sous-comités qui en dépendent (y compris la terminologie e-learning du SC36). La base Termium, qui gère un fonds très important de ressources terminologiques est cependant organisée prioritairement pour répondre aux besoins de traduction d'un pays bilingue. De ce fait elle propose aujourd'hui des terminologies en français et anglais. A moyen terme elle est prévue pour être opérationnelle en 4, ou 5 langues au plus.

[47] L'ensemble des personnes qui constitue un groupe de normalisation (par exemple le SC36) se distribuent en différentes catégories. Des experts du domaines (non obligatoirement formés à la normalisation, par ex. au S36, des pédagogues, des industriels informatiques, des éditeurs d'e-enseignement), des experts liaison (un expert MPEG, un expert de l'e-learning aéronautique…), mais aussi des professionnels employés par les instances de normalisation (ISO, AFNOR, BSI…). Ce sont ces professionnels de la normalisation ainsi que les experts les plus dynamiques et les plus expérimentés (qui ont participés auparavant à d'autres domaines de normalisation) qui constituent bien sûr le noyau dirigeant permettant la production effective de la famille de normes que doit éditer l'instance (ici le SC36).

[48] On n'apprend pas selon la même logique une langue alphabétique ou une langue idéographique ; et on sait que l'enseignement de la langue et de l'écriture sont aux fondements même de la transmission de savoir.

[49] Evolution qui est au cœur des questions posées par la gouvernance mondiale.

[50] Les normes plus strictement techniques (métallurgie, composants électroniques, etc.) posent très peu de problèmes de découpage conceptuel dans la diversité des langues. Les terminologies de ces domaines techniques s'accommodent dès lors très bien de la réalisation de listes de vocabulaire en anglais qu'il suffit de traduire terme à terme dans d'autres langues.

[51] A l'exception notable de certaines langues (minoritaires ou dominées) qui n'ont pas eu l'opportunité socio-éducative d'accompagner la modernité et qui ont été victimes de perte de domaines linguistiques plus ou moins vastes. Nous y reviendrons mais un des objectifs éthiques de Cartago est de tenter de freiner, voire de réparer pour partie ces déficits, dans un champ sémantique aussi stratégique que la transmission de savoir.

Sponsor de aur / Partenaire Or

STAR

Sponsor de argint / Partenaire Argent

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / Partenaire Bronze

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

Seminar Internațional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latina/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



Metode de analiză lingvistică a textelor în limba română pentru extragerea terminologică. Instrumente şi resurse

PEEV Luciana , CS I, SC Software ITC, Cluj
ŞERBAN Felicia, CS I, Institutul de Lingvistică şi Istorie Literară „Sextil Puşcariu"

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanţi / Exposants

Rezumat
Lucrarea prezintă aspecte legate de instrumentele, resursele lingvistice şi metodologii de analiză existente în sistemul TEZAROM care pot fi folosite cu succes în aplicaţii lingvistice (dicţionar morfologic, analizor morfologic, analizor fonologic, extractor de posibili termeni).

- ***resurse lexicale: lexicon de cunoştinţe lingvistice (morfologice, fonologice şi sintagmatice) cu posibilitatea de a alege un standard internaţional şi alinierea la acesta cu ajutorul generatoarelor;***
- ***lexicon specializat cu atribute sintagmatice;***
- ***metode pentru extragerea posibililor termeni;***

Metodele care sunt prezentate unifică cercetările din domeniul limbajului natural cu cele din domeniul terminologiei şi terminografiei, oferind terminologilor unelte pentru a le uşura munca în domeniu.

Lucrarea de faţă îşi propune să prezinte aspecte tehnice legate de resursele lingvistice şi metodologia de analiză a textelor utilizată în realizarea unor aplicaţii cu specific lingvistic de către grupul RoLingva de la Software ITC Cluj. Orice aplicaţie care are ca scop realizarea unei analize asupra unui text scris în limba română se bazează pe resurse lingvistice şi metode. Realizarea resurselor lingvistice necesită o muncă laborioasă prin care se modelează şi se formalizează diferite capitole ale gramaticii limbii române, fiind astfel create sisteme informatice complexe care pun la dispoziţia dezvoltatorului de soft instrumente puternice ce conţin baze lingvistice cu atribute gramaticale, precum şi metode de analiză.

Sistemul informatic TEZAROM se compune dintr-o bază lingvistică de cunoştinţe complexă, metode de analiză morfologică, fonologică, sintagmatică şi generatoare care asigură maleabilitatea sistemului şi puterea de modelare a elementelor lingvistice în conformitate cu cerinţele aplicaţiilor proiectate sau cu standardele altor resurse lingvistice. Caracteristica de maleabilitate a sistemului TEZAROM ţine seama de apariţia pe plan internaţional a unor noi standarde în domeniul tehnologiilor limbajului uman, de necesitatea interschimbului de informaţii între diverse sisteme lingvistice şi de alinierea cu uşurinţă la noi proiecte fără a pierde informaţia existentă, în favoarea acestora.

Sistemul informatic TEZAROM poate fi reprezentat astfel:

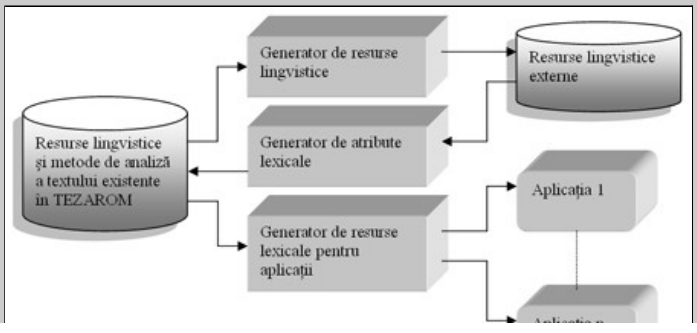




Figura 1

1. Resurse lingvistice şi metode de analiză

1.1. Trăsături ale tezaurului de cunoştinţe lingvistice

Studiile teoretice realizate de grupul RoLingva asupra morfologiei şi fonologiei au creat premisele elaborării tehnologiei lingvistice pentru limba română. Astfel, a fost realizat primul model formalizat al morfologiei şi fonologiei cu reguli şi metode care pot acoperi în totalitate sistemul flexionar al limbii române. Modelul, în afară de informaţiile referitoare la elementele morfologice, ortografice şi fonologice, păstrează informaţii incipiente despre atribute sintactice (de exemplu la verb informaţii despre tranzitivitate şi reflexivitate). Modelul morfologic respectă regulile gramaticale academice, ceea ce permite extinderea cu uşurinţă spre celelalte capitole ale gramaticii, reflectând faptul că diferitele capitole ale gramaticii se leagă între ele, existând o ierarhie a informaţiei. Modelul fonologic, de asemenea, respectă regulile larg acceptate şi se bazează pe un set de reguli şi algoritmi de determinare a acestora, iar în cazul analizei la nivel de text utilizează şi informaţiile morfologice caracteristice unui cuvânt, în vederea determinării corecte a părţii de vorbire (diferite părţi de vorbire, deşi sunt omonime, pot prezenta omografii). Desigur, determinarea corectă a părţii de vorbire se poate face cu condiţia ca analiza morfologică să fie urmată de o dezambiguizare a textului şi de o analiză, dacă nu sintactică, cel puţin sintagmatică. acest lucru poate sublinia necesitatea abordării analizei sintactice, respectiv sintagmatice.

Sistemul TEZAROM a fost conceput ca un sistem deschis pentru a putea acoperi cu uşurinţă şi alte domenii ale gramaticii limbii române. Acesta permite îmbogăţirea bazei lexicale cu atribute noi, care se pot lega cu uşurinţă de cele existente, având în vedere faptul că au fost respectate regulile în vigoare ale gramaticii şi permite, de asemenea, dezvoltarea metodelor cu reguli şi algoritmi noi, corespunzătoare noului capitol formalizat al gramaticii.

Modelele create au ţinut seama de faptul că limba română este o limbă care flexionează puternic şi permite realizarea de numeroase cuvinte noi prin sufixare şi prefixare, cuvinte, care la rândul lor, pot avea propria flexiune. Numărul formelor flexionate este mare şi practic, se poate pierde controlul asupra acestora, dacă se va face un simplu inventar a tuturor formelor însoţite de indicarea părţii de vorbire şi a categoriei gramaticale reprezentative pentru cuvânt. Din acest motiv, în tezaurul de cunoştinţe lingvistice s-a preferat obţinerea acestor forme, prin metode procedurale, pe bază de reguli, plecând de la o rădăcină condensată, liste de sufixe şi prefixe, şi o clasă flexionară părinte, care pot descrie în totalitate o intrare lexicală. Combinaţiile care se pot realiza astfel sunt foarte numeroase şi pot acoperi un lexic foarte bogat. Numărul de cuvinte care pot fi derivate din baza de cunoştinţe lingvistice este foarte mare, dacă descrierea la nivel de intrare lexicală este bine realizată. (prin derivare se obţin toate formele flexionare, prefixate sau sufixate). Capitolul dedicat formării cuvintelor din tezaurul de cunoştinţe lingvistice este la început, dar studii pentru realizarea unei tehnologii în acest sens au fost realizate şi în viitor se vor putea aplica. Efortul pentru îmbogăţirea inventarului lexical este mare în această manieră, dar mai redus decât cel al introducerii tuturor formelor flexionate, prefixate şi sufixate, cuvinte care de obicei nu se găsesc în dicţionare obişnuite. Pe baza cunoştinţelor lingvistice existente, pentru proiectul SIASTRO - *Sistem informatic pentru analiza sintagmatică a textelor în limba română. Fundamentare teoretică şi implementare* s-a generat un lexicon destinat îmbogăţirii tezaurului cu atribute sintactice necesare analizei sintagmatice la nivel de sintagmă.

1.2. Model morfologic

Modelul morfologic este aplicat pe o bază lexicală de cunoştinţe din sistemul TEZAROM şi acoperă în întregime morfologia limbii române. Fiecare intrare din baza lexicală corespunde unui cuvânt-titlu pentru care au fost introduse atributele morfologice care-l caracterizează. Pentru verificare regulilor care folosesc aceste atribute s-au proiectat metode care pot realiza analiza morfologică a unei forme flexionare sau sinteza pentru un cuvânt-titlu (paradigma cuvântului). Descrierea şi codificarea atributelor ţine seama de comportamentul rădăcinii şi a terminaţiilor în timpul flexiunii, având în vedere faptul că limba română este o limbă care flexionează puternic. Atributele utilizate de modelul morfologic sunt de două tipuri: statice şi dinamice. Atributele statice descriu o anumită caracteristică a cuvântului-titlu care nu se schimbă şi sunt valabile pentru toate părţile de vorbire. Din această categorie fac parte numele părţii de vorbire, genul la substantive, conjugarea la verbe etc. Atributele dinamice sunt specifice numai cuvintelor care flexionează şi sunt reprezentate prin clase flexionare.

Se numesc dinamice pentru ca la un moment dat pot primi o valoare ce corespunde pentru o anumită categorie gramaticală (mod, timp, persoană, număr - la verb - şi caz, număr, determinare - la nominale). Astfel, există clase care definesc variaţiile rădăcinii, poziţia accentului şi clase care definesc terminaţiile în timpul flexiunii.

Pentru a obţine la un moment dat o anumită formă flexionară, asupra claselor se aplică reguli care prin derivare determină structura unei anumite forme flexionare. Un cuvânt-titlu este reprezentat printr-o rădăcină şi o clasă flexionară. Rădăcina, la rândul ei, poate reprezenta variaţii, fie din cauza alternanţelor fonetice, fie din cauza temelor multiple.

O rădăcină care prezintă variaţii se numeşte rădăcină părinte, iar clasa corespunzătoare se numeşte clasa părinte şi se reprezintă astfel:

$Rad_{părinte} = r_1@a@vr_2 \rightarrow rad_1, rad_2 \ldots rad_n$

unde: @a = clasă accent
@v = clasă variaţie rădăcină
$Clasa_{părinte} = C_p \rightarrow clasa_1, clasa_2 \ldots clasa_n$

Derivarea rădăcinii $Rad_{părinte}$ şi clasei părinte $Clasa_{părinte}$ determină mulţimea de rădăcini reale împreună cu clasele lor flexionare.

Modelul morfologic definit are două metode care pot fi utilizate în diverse aplicaţii lingvistice, şi anume: 1) metodă de analiză morfologică şi 2) sinteză - generarea paradigmei.

Schematic, modelul morfologic se reprezintă astfel:

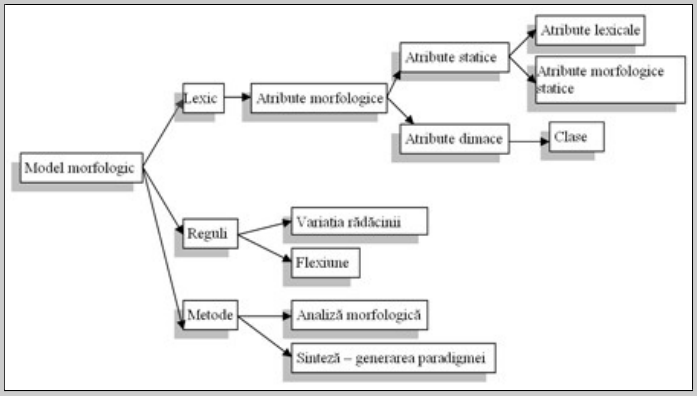


Figura 2

1.3. Model de silabisire

Formalizarea silabaţiei a sistematizat regulile conform normelor şi a ţinut seama de faptul că analiza morfologică utilizează dicţionare de fragmente de cuvinte (prefixe, rădăcini, sufixe, flective), iar între aceste fragmente pot apărea tăieturi silabice. Din acest motiv s-au definit o serie de reguli de silabisire la joncţiunea dintre fragmente. Regulile au fost definite astfel, încât şi excepţiile au putut fi modelate. Rezultatul este o silabisire 100% corectă. Desigur, se pot aplica şi algoritmi de silabisire, dar rezultatul lor nu este întotdeauna corect. Din acest motiv considerăm că modelul pentru silabisire pe care l-am creat este superior. Sistemul de silabisire creat în TEZAROM se aplică atât silabaţiei fonetice, cât şi celei morfologice. Modelul de silabaţie realizat în TEZAROM a arătat că la concatenarea a două fragmente există două puncte critice unde pot sau nu apărea tăieturi silabice. Aceste puncte sunt în interiorul fragmentului din stânga şi/sau la joncţiunea dintre fragmente.

$\ldots+@s_{i0}frag@s_{i1}ment_i@s_{i2}+@s_{j0}frag@s_{j1}mentj@s_{j2}+\ldots$

Interacţiunea codului din stânga fragmentului din dreapta sj0 determină apariţia tăieturii silabice în una sau în ambele poziţiile notate cu s_{i1} şi s_{i2}.

1.4. Model fonologic

La realizarea modelului „fonetic" s-a ţinut seama de fonologia limbii române standard şi de interacţiunea fonologiei cu morfologia şi silabaţia. Modelul fonologic realizat urmăreşte atingerea a două obiective:

- transcrierea fonetică a unui text scris în limba română în alfabetul internaţional IPA (*The International Phonetic Alphabet*);

- transformarea textului transcris fonetic în şiruri de sunete reprezentate prin foneme sau difoneme, preluate dintr-o bază vocalică, împreună cu parametri prozodici, cu scopul obţinerii formei sonore a textului din limba română.

Fonetica limbii române se bazează pe reguli foarte precise şi din acest motiv nu este necesară dezvoltarea unui dicţionar fonologic special, cum se întâmplă în limba engleză. Cuvintele care nu respectă regulile fonetice sunt în număr mic şi dintre acestea amintim: cuvintele care au în componenţă grafeme care prezintă o ambiguitate în recunoaşterea situaţiei fonetice, cuvintele străine intrate în limbă şi încă neacomodate, deosebiri de pronunţare între cuvintele din fondul vechi şi cuvintele neologice. Pentru acestea este necesară, înainte de transcrierea fonetică efectivă, o conversie *grafem - grafem,* prin care se realizează dezambiguizarea interpretării fonetice (de exemplu cuvinte care conţin litera *x*). Conversia *grafem-grafem* se poate face pe baza unei echivalenţe păstrate în baza lexicală de cunoştinţe.

Modelul fonologic a beneficiat de existenţa modelului morfologic, a informaţiilor de accentuare şi a modelului de silabaţie fonetică. Astfel, modelul morfologic poate transmite informaţii legate de categoria gramaticală şi poziţia accentului pentru un cuvânt analizat, iar modelul de silabaţie poate contribui la determinarea diftongilor, triftongilor şi a hiaturilor, procedeu necesar în determinarea vocalelor şi a semivocalelor. Cu aceste informaţii, modelul fonologic poate aplica regulile fonologice şi obţine transcrierea corectă pentru orice cuvânt.

Schematic, modelul fonologic se prezintă astfel:

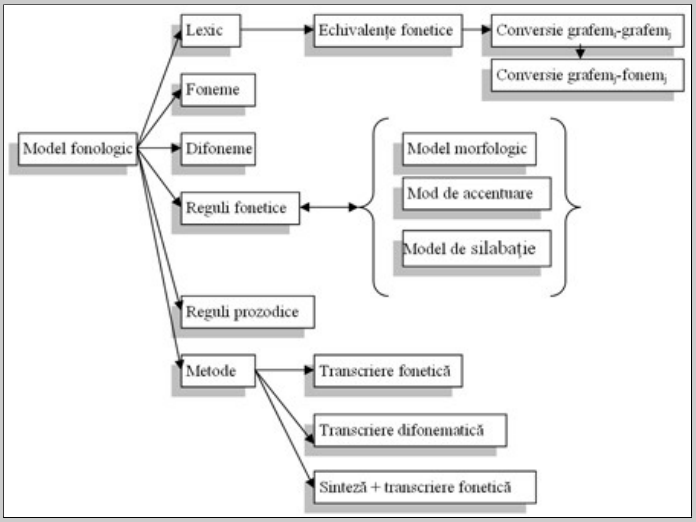


Figura 3

1.5. Model sintagmatic - resurse lingvistice: LEXICON

Resursele lingvistice cuprinse în LEXICON sunt preluate din TEZAROM şi urmează să fie îmbogăţite cu atribute sintactice. La preluarea bazei lingvistice şi proiectarea atributelor sintactice s-a ţinut seama de existenţa unor standarde pe plan internaţional şi de posibilitatea trecerii de la un standard la altul fără a fi nevoie de prelucrări majore ulterioare. La crearea Lexiconului din baza de cunoştinţe lingvistice existente în TEZAROM s-a ţinut seama, în principal, de standardul OLIF, care este un standard deschis conceput în anii '90 ca un mijloc de a asigura interschimbul de informaţii între aplicaţiile NLP (*Natural Language Proccesing*). Popularea bazei de date lexico-gramaticale s-a realizat prin intermediul unui *extractor de resurse lingvistice* care a preluat şi a prelucrat informaţia necesară din TEZAROM, convertind-o în atribute lexicale, morfologice şi fonologice, conform specificaţiilor de definiţie a noii baze de date. În paralel s-a realizat un produs informatic prin care utilizatorii pot să verifice atributele morfologice şi fonologice şi să completeze noile atribute sintagmatice. Este un produs intermediar care permite, pe măsura completării datelor, şi realizarea de note şi observaţii la nivel de cuvânt-titlu. Concepţia sistemului permite utilizarea lui individuală, paralelă de către mai mulţi utilizatori, în aşa fel, încât toate informaţiile cu care s-a îmbogăţit sistemul pe platformele individuale să fie reunificate în sistemul ţintă unic. Caracteristicile administrative individuale sunt şi ele păstrate, astfel în baza de date finală se va putea recunoaşte - pentru fiecare intrare în parte - autorul original al informaţiei de specialitate. Atributele introduse vor putea fi preluate în sistemul informatic TEZAROM prin intermediul unui *convertor de atribute*, astfel baza de cunoştinţe din sistem se va îmbogăţi cu noile informaţii. Pentru alinierea resurselor lingvistice la noi standarde,

sistemul va fi îmbogăţit cu generatoare de resurse lingvistice. De asemenea, pentru realizarea de noi aplicaţii lingvistice, în funcţie de complexitatea acestora, se vor crea noi generatoare de date.

Schema de dezvoltare a stemul informatic TEZAROM arată astfel:

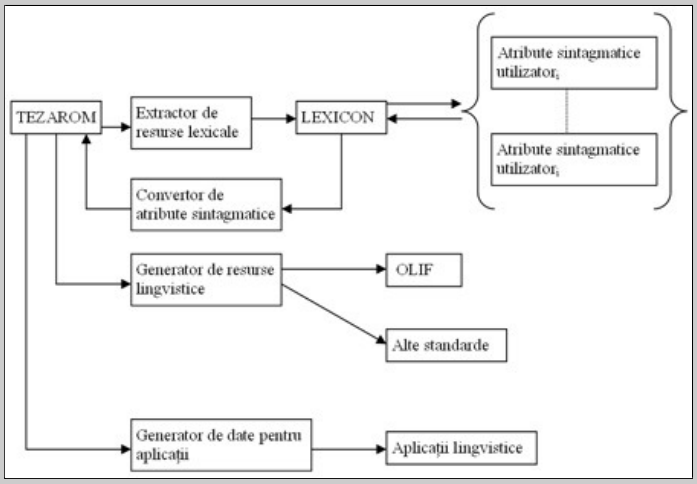


Figura 4

Avantajul sistemului constă în faptul că baza de cunoştinţe lingvistice este deschisă şi poate fi îmbogăţită cu uşurinţă oricând cu noi atribute definite clar, la acest nivel neexistând restricţii de memorie şi de timp, în schimb, aceasta nu va participa niciodată direct într-o aplicaţie lingvistică. Proiectarea aplicaţiilor lingvistice vor apela la generatoarele de date şi vor primi baze de date particularizate conform cu cerinţele lor. Nicio aplicaţie, indiferent cât de complexă este, nu are nevoie de toată informaţia înmagazinată în TEZAROM.

2. Aplicaţii realizate pe baza resurselor şi metodelor lexicale existente în TEZAROM

Modele gramaticale formalizate până în prezent, exceptând modelul sintagmatic care este în curs de realizare, au fost utilizate cu succes într-o serie de aplicaţii lingvistice cu caracter educaţional sau experimental. Mai jos prezentăm o parte din aplicaţiile realizate care există şi sunt funcţionale.

2.1. DMR – Dicţionar morfologic al limbii române

Dicţionarul morfologic al limbii române este prima aplicaţie bazată pe tehnologia realizată şi constituie prima formă informatizată a unui dicţionar fundamental al limbii române, oferind informaţii morfologice complete. Baza lexicală a dicţionarului conţine inventarul de cuvinte cuprins în *Dicţionarului explicativ al limbii române* (DEX), ediţia II-a, îmbogăţit cu cuvinte din *Dicţionarul de neologisme* şi DOOM, precum şi cuvinte uzuale de largă circulaţie în limba română literară actuală, dar absente din dicţionarele consultate.

Sistemul realizat poate răspunde corect la toate problemele legate de flexiunea cuvintelor (declinare, conjugare) accentuare, silabisire - atât morfologică cât şi fonetică - şi, de asemenea, poate da informaţii prin intermediul analizorului morfologic despre toate categoriile gramaticale ce corespund unei forme flexionate. În ceea ce priveşte flexiunea cuvintelor, s-a respectat norma pentru limba română stabilită de Academia Română prin lucrările de specialitate oferite de aceasta.

Produsul DMR este pe suport CD şi este destinat în primul rând celor care învaţă limba română, la diferite niveluri, dar şi celor care studiază aspecte morfologice ale stadiului ei actual. Este foarte uşor de utilizat, iar mediul de lucru, prin aspectul său, sugerează utilizatorului ce trebuie să facă pentru a obţine informaţia dorită.

Forma de prezentare a paradigmei este variată, în funcţie de interesul utilizatorului, şi este comandată prin intermediul unor butoane. Indiferent de afişare, se vizualizează întotdeauna accentul prin sublinierea vocalei tonice. În figura de mai jos este prezentat modul de vizualizare a unei declinări.



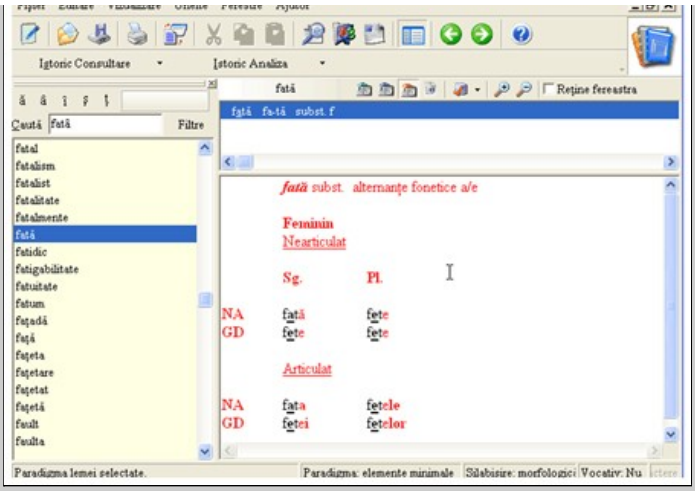


Figura 5

Sistemul este dotat şi cu un analizor morfologic la nivel de cuvânt. Această facilitate vine în întâmpinarea celor care învaţă limba română, şi nu numai, oferindu-le posibilitatea recunoaşterii formei de dicţionar a cuvântului. Figura 6 prezintă un exemplu.

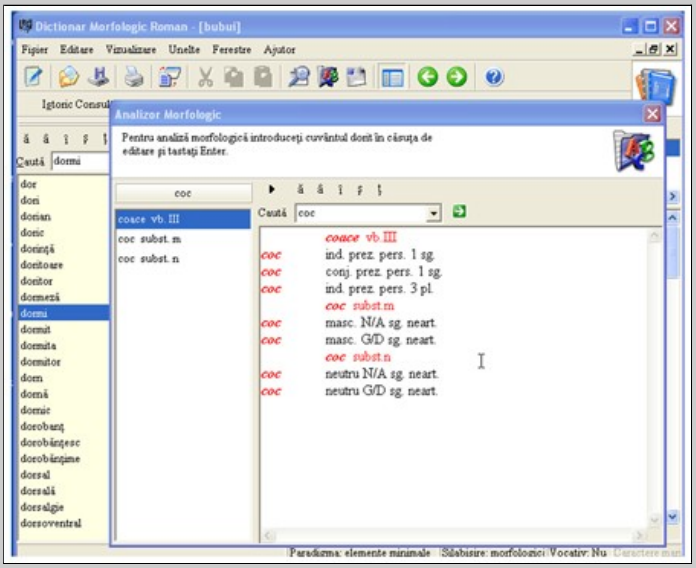


Figura 6

2.2. Analizor morfologic

Altă aplicaţie practică este cea a unui *analizor morfologic experimental* care poate analiza texte în format RTF şi pune în evidenţă paşii parcurşi, rezultatele fiind vizualizate într-un format hipertext. Realizarea analizei se face în doi paşi, şi anume: procesarea textului care transformă textul în elemente recunoscute de calculator şi analiza propriu-zisă. La procesare este posibil ca unele elemente necunoscute să dispară, iar altele să se transforme în şiruri de cuvinte recunoscute de calculator. Fluxul de date poate fi reprezentat astfel:

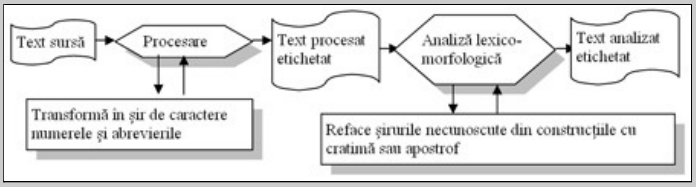


Figura 7

Procesarea textului identifică cuvintele din textul de intrare şi semnele de punctuaţie, transformând abrevierile, numerele şi semnele matematice în cuvinte pentru ca textul să poată fi analizat din punct de vedere lexico-morfologic. Fluxul de ieşire în urma procesării este un text etichetat. Analiza lexico-morfologică preia acest text şi-l prelucrează din punct de vedere morfologic, rezultând un nou text etichetat cu

informaţii morfologice. Avantajul utilizării rezultatelor intermediare în format hipertext este acela al vizualizării datelor intermediare şi depistarea analizelor eronate înainte de a se trece la proiectarea aplicaţilor comerciale. Prezentăm mai jos un exemplu de analiză morfologică aplicată unui text.

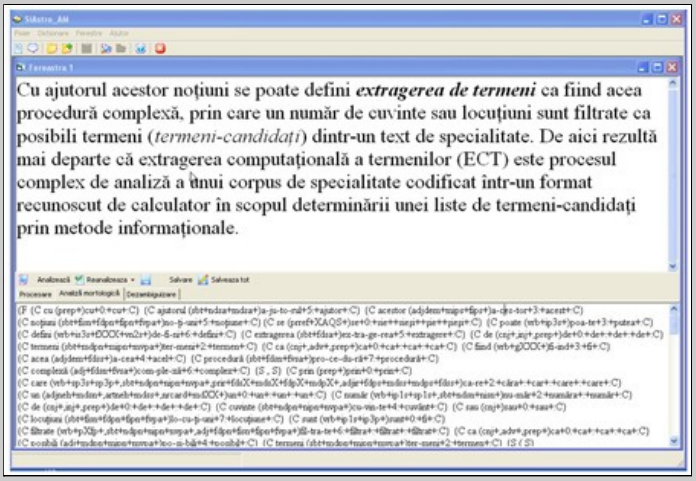


Figura 8

În fereastra de jos sunt date rezultatele analizei morfologice în format hipertext. Există şi un pas de dezambiguizare a textului care poate preceda paşii de analiză fonologică sau sintactică. Rezultatele dezambiguizării se pot vizualiza în panoul dedicat acestui proces.

2.3. Analizor fonologic

Altă aplicaţie practică este cea a unui *analizor fonologic experimental* care poate să analizeze texte în format RTF şi să pună în evidenţă paşii parcurşi, rezultatele fiind vizualizate într-un format hipertext. Realizarea analizei fonologice pleacă de la textul etichetat rezultat în urma analizei lexico-morfologice după dezambiguizare lui. Aplicaţia are rolul să pună în evidenţă superioritatea unui sistem bazat pe cunoştinţe lingvistice formalizate faţă de sistemele bazate exclusiv pe semnale fizice şi poate servi cu succes unor experienţe interesante în rafinarea metodelor şi algoritmilor propuşi în mod teoretic, dar încă insuficient probaţi în cazurile reale. Este un model de laborator deosebit de interesant, menit să susţină justeţea abordării pe baze lingvistice a întregului proces de sinteză a vorbirii din texte scrise, primul pas semnificativ în direcţia realizării unui sistem real. Sistemul permite experimente în vederea obţinerii sintezei dorite din text prin remodelarea datelor obţinute în urma analizei contextuale. Se poate interveni asupra grupării cuvintelor în procesul de rostire, a pauzelor, a valorii accentului la nivel de cuvânt sau de enunţ, simulând în felul acesta rezultatele prozodice. Se pot simula mai multe experimente în paralel şi păstra rezultatele pentru o comparare ulterioară. Fluxul de date poate fi reprezentat astfel:

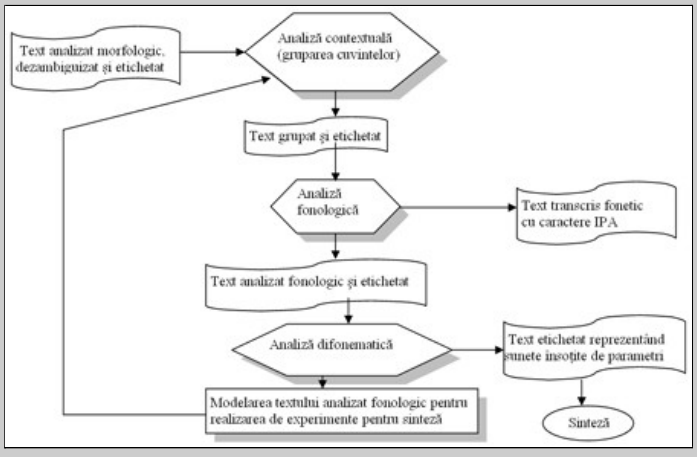
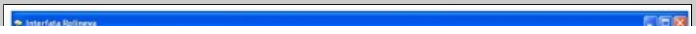


Figura 9

Un text analizat cu sistemul experimental şi transcris fonetic arată astfel:



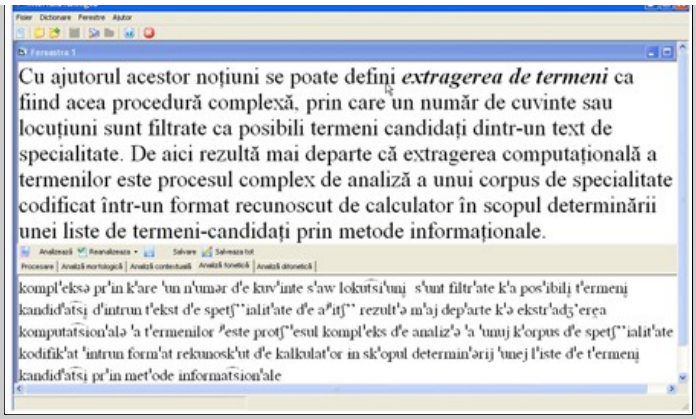


Figura 10

Modelul fonologic teoretic a fost aplicat, cu succes, asupra unei baze lexicale de aproximativ 75.000 de cuvinte-titlu care acoperă peste două milioane de cuvinte flexionate. Considerăm că verificarea aplicării modelului fonologic a fost făcută riguros şi rezultatele obţinute sunt edificatoare: *modelul fonologic răspunde întru totul transcrierii fonetice din limba română*.

2.4. Analizor sintagmatic

Analizorul sintagmatic este în curs de realizare. Rolul lui este de a studia algoritmii de recunoaştere a diferitelor sintagme şi de a semnala erorile înainte de a se trece la realizarea produsului propriu-zis de extragere a posibililor termeni. La realizarea lui se vor folosi modulele existente prezentate anterior: procesarea textului, analiza lexico-morfologică şi dezambiguizare, la care se va adăuga modulul de analiză sintagmatică.

În principiu un modul de analiză sintagmatică atribuie unui şir de intrare dat o descriere structurală relativă la o anumită gramatică (Hellwig (2002). Pentru aceasta sunt necesare:

- *un formalism gramatical*, care să permită descrierea gramaticii limbii tratate;
- *o gramatică* a limbii tratate descrisă în formalismul ales;
- *un algoritm* care să determine dacă un şir de intrare este în conformitate cu gramatica limbii şi, în caz afirmativ, - structura ataşată acelui şir.

Formal, analizorul sintagmatic poate fi o continuare a analizorului lexico-morfologic care preia textul de ieşire analizat şi etichetat, pentru a determina sintagmele pe baza gramaticii descrise. Rezultatul va fi un text analizat cu sintagme etichetate.

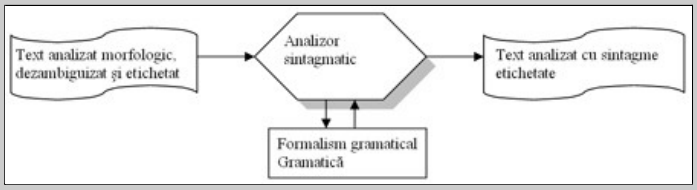


Figura 11

2.5. Extractor de termeni (EXTR)

În procesul de globalizare la care asistăm în prezent, transferurile internaţionale generează, pe lângă schimbul de produse, şi schimb de documente. Cantitatea de informaţii în format text a crescut vertiginos, iar traducerea documentaţiilor tehnice, legislative etc. dintr-o limbă în alta presupune nu numai cunoaşterea limbii, ci şi a domeniului, respectiv a termenilor de specialitate care îi aparţin. Aici intervine contribuţia terminologului care stabileşte termenii dintr-un anumit domeniu pe baza materialului studiat. Munca de specialitate pe care o desfăşoară în determinarea posibililor termeni este laborioasă şi presupune parcurgerea unor corpusuri mari în găsirea unor contexte semnificative pentru stabilirea unor definiţii relevante.

Extractorul de termeni vine în întâmpinarea terminologului, punându-i la dispoziţie instrumente care pot să-l ajute în găsirea termenilor, în stabilirea definiţilor şi a contextelor relevante pentru. Desigur, cunoaşterea termenilor este importantă şi în traduceri, dar stabilirea lor corectă poată fi făcută înaintea unui astfel de proces, în limba în care există documentaţia, în cazul de faţă, limba română.

Aplicaţia îi va permite utilizatorului să-şi creeze propriile proiecte pentru culegerea termenilor. În cadrul proiectului, utilizatorul va putea să definească structura termenilor, textele în care se va face căutarea, utilizarea sau nu a unor glose terminologice externe şi stabilirea unei liste de nontermeni. Executarea proiectului prin intermediul aplicaţiei va oferi ca rezultate: lista posibililor termeni, conform cu opţiunile stabilite de utilizator, şi contextele în care termenii apar în textele selectate în proiect. Pe baza rezultatelor obţinute, printr-un proces interactiv, terminologul va putea să selecteze din lista posibililor termeni pe cei corecţi, specifici domeniului studiat, precum şi contextele relevante care susţin definiţia stabilită. De asemenea, va putea să creeze glose terminologice, dicţionare terminologice, fişe terminologice în formate oferite de aplicaţie sau importate etc. Schematic, aplicaţia arată astfel:

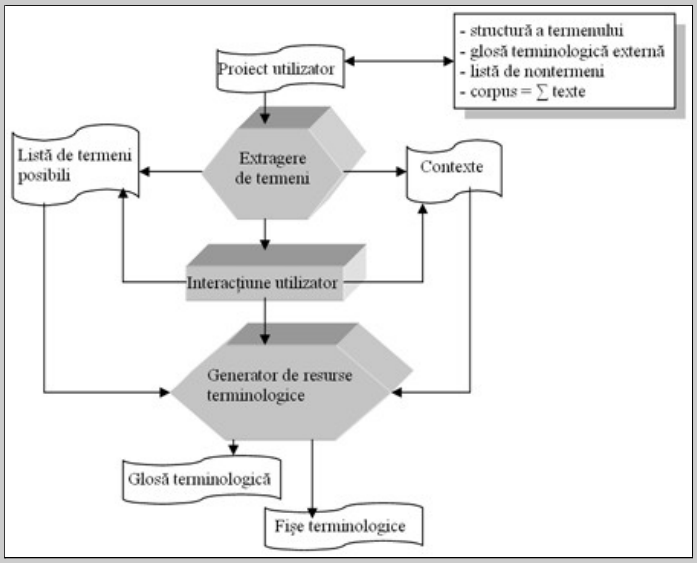


Figura 12

Aplicaţia pentru extragerea posibililor termeni urmează să fie proiectată şi va utiliza modelele şi conceptele existente în sistemul informatic TEZAROM. Acesta va pune la dispoziţia proiectanţilor aplicaţiei baze de date particularizate, reguli şi metode.

Concluzii

Sistemul informatic TEZAROM creat pentru analiza lingvistică a textelor scrise în limba română se bazează pe o bogată experienţă acumulată în domeniul cercetării limbii române prin mijloace electronice şi poate fi utilizat cu succes într-o mulţime de aplicaţii, ultima aplicaţie în curs de realizare fiind cea dedicată extragerii termenilor de specialitate din diverse corpusuri. Sistemul se fundamentează pe normele statuate ale limbii şi este pus de acord cu prevederile celor mai importante lucrări de gramatică şi fonetică-fonologie, fiind astfel conceput încât să ofere utilizatorului rezultatele obţinute până acum (precum flexiunea cuvintelor, silabaţia, pronunţare), dar să aibă şi posibilitatea de a se extinde asupra altor compartimente ale limbii, pe măsură ce lucrările progresează.

Instrumentele create până în prezent au abordat domeniul morfologiei (aplicaţii realizate - *Dicţionarul morfologic al limbii române* pe suport CD şi analizorul morfologic), precum şi cel al fonologiei (indicarea silabisirii atât fonetice, cât şi morfologice, şi a pronunţării corecte a cuvintelor prin transcrierea automată în IPA), fiind perfect funcţionale (aplicaţii realizate - analizor fonologic); următoarea deschidere s-a făcut spre identificarea sintagmelor prin modelarea elementelor de sintaxă (aplicaţie în curs de realizare - extragerea posibililor termeni de specialitate dintr-un text, printr-un sistem interactiv). Toate aceste instrumente, de maximă importanţă în acest moment când limba română a devenit limbă a Uniunii Europene. sunt destinate să faciliteze munca utilizatorului interesat de cercetarea sau de învăţarea limbii române.

CSI Luciana Peev

SC Software ITC Cluj SA din Cluj-Napoca
Tel.: 0264-587767
E-mail: luciana_peev@yahoo.com

dr. Felicia Şerban
CS I Institutul de Lingvistică
şi Istorie Literară „Sextil Puşcariu"
din Cluj-Napoca
Tel.: 0264-428 120
E-mail: feli_serban@yahoo.com

Referinţe bibliografice

Avram, A., 1958, «Semivocalele româneşti din punct de vedere fonologic», *Studii şi cercetări lingvistice*, p. 7-14.

Dascălu-Jinga, L., M. Teodorescu, A. Ulivi, 1988, *Antologia fonetică a limbii române,* Bucureşti, Institutul de Cercetări Etnologice şi Dialectologice.

Dicţionarul ortografic, ortoepic şi morfologic al limbii române, 1982/2005, ed. a 2-a, Bucureşti, Univers Enciclopedic.

Dicţionarul explicativ al limbii române,1975/1996, ed. a 2-a, Bucureşti, Univers Enciclopedic.

Draşoveanu, D. D., 1997, *Teze şi antiteze în sintaxa limbii române,* Cluj-Napoca, Clusium.

Gramatica limbii române, 1963, ediţia a II-a revăzută şi adăugită, vol. I-II, Bucureşti, Editura Academiei.

Gramatica limbii române, 2005, vol. I. *Cuvântul*, vol. II. *Enunţul*, Bucureşti, Editura Academiei Române.

Lombard, A., C. Gâdei, 1981, *Dictionnaire morphologique de la langue roumaine,* Bucureşti, Editura Academiei.

Marcu, F., 2004, *Marele dicţionar de neologisme*, Bucureşti, Editura Saeculum I.O.

Marcu, F., C. Maneca, 1978, *Dicţionar de neologisme*, ed. a 3-a, Bucureşti, Editura Academiei Române.

Peev, L., L. Bibolar, E. Jodal, 1996, «Un model de formalizare a morfologiei limbii române», Simpozionul internaţional „Limbaj şi tehnologie", Bucureşti.

Peev, L., L. Bibolar, E. Jodal, 1997, «A Formalization Model of the Romanian Morphology»,International Workshop „Speech and Computer" SPECOM'97, Cluj-Napoca.

Ribeiro, R., L. Oliveira, I. Trancoso, 2003, «Using Morphosyntactic Information TTS Systems: Comparing Strategies for European Portuguese Spoken Language Systems Lab» (electronic ed.).

Sfîrlea, L., 1970, *Pronunţia românească literară. Stilul scenic,* Bucureşti, Editura Academiei.

Stan, I. T., 1996, *Fonetica,* Cluj-Napoca, Editura Presa Universitară Clujeană.

Şerban, F., L. Peev, L. Bibolar, D. Bucerzan, 1996, «Banca de date a limbii române - fonetică şi fonologie», în D. Tufiş (ed.), *Limbaj şi tehnologie,* Bucureşti, Editura Academiei Române, pp. 157-160.

Şerban, F., D. Bucerzan, L. Peev, L. Bibolar, 1998, «Database of the Romanian Language», *Romània orientale,* 11, pp. 99-112.

Şerban, F., L. Peev, L. Bibolar, 2000, «La base de données de la Langue roumaine», *Terminometro: La terminologie en Roumanie et en République de Moldova,* hors-série, 4, Atlas-Clusium, pp. 40-42.

Şuteu, F., E. Şoşa, 1993, *Dicţionar ortografic al limbii române,* Bucureşti, Editura Vestala, Editura Atos.

Tătaru, A.,1999, *Dicţionar de pronunţare a limbii române,* ed. a 2-a, Cluj-Napoca, Clusium.

Vasiliu, E., 1965, *Fonologia limbii române,* Bucureşti, Editura ştiinţifică.

Hellwig, Peter, 2002 Natural Language Parsers - A "Course in Cooking"

http://www.cl.uni-heidelberg.de/kurs/ws06/ecl/etc/pars02.pdf

http://www.rolingva.ro

Sponsor de aur / **Partenaire Or**

Sponsor de argint / **Partenaire Argent**

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / **Partenaire Bronze**

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latina/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



eDTLR - Dicţionarul tezaur al limbii române în format electronic

RĂSCHIP Marius , Universitatea Al. I. Cuza Iaşi, Facultatea de Informatică
CRISTEA Dan, Universitatea Al. I. Cuza Iași, Facultatea de Informatică, Institutul de Informatică Teoretică, Academia Română, filiala Iaşi
FORĂSCU Corina , Universitatea Al. I. Cuza Iaşi, Facultatea de Informatică, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română, Bucureşti

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanţi / Exposants

1. Introducere

Iniţiat de Bogdan Petriceicu Haşdeu la sfarşitul secolului XIX şi continuat de Alexandru Philippide până în 1905, Dicţionarul Limbii Române a fost preluat ca proiect prioritar al Academiei Române sub colectivul condus de Sextil Puşcariu. Între anii 1913 şi 1949, sub titlul *Dicţionarul Academiei* (DA), se publică literele *A-C*, *D-De*, *F-K*, *L-Iojniţă*, iar după 1965, sub numele *Dicţionarul Limbii Române* (DLR), restul intrărilor (Haja et al., 2005). În cifre estimative, DA+DLR cuprinde, în 15.000 pagini, 175.000 de intrări şi 3.000.000 de exemple culese din 2.500 de referinţe bibliografice.

eDTLR este atât acronimul Dicţionarului tezaur al limbii române în formă electronică, cât şi numele proiectului naţional început în septembrie 2007 [1], cu o durată de 3 ani, al cărui scop e să aducă în formă de dicţionar electronic DA+DLR. Trei Institute de Lingvistică ale Academiei Române, din Bucureşti, Cluj şi Iaşi, alături de Institutul de Cercetări în Inteligenţă Artificială al Academiei Române, Bucureşti, Institutul de Informatică Teoretică al filialei Iaşi a Academiei Române şi Universitatea „Al. I. Cuza" Iaşi prin Facultatea de Litere, coordonate de Facultatea de Informatică a aceleiaşi Universităţi colaborează în cadrul acestui proiect la realizarea seriilor reunite ale dicţionarului şi crearea unui index în sursele bibliografice. Activităţile principale ale proiectului sunt:

- scanarea formei tipărite a DA+DLR şi recunoaşterea automată a caracterelor în fişierele scanate;
- corectarea *on-line* colaborativă a formei obţinute anterior;
- marcarea câmpurilor din intrările de dicţionar (sensuri şi definiţii, citate, explicaţii morfologice, etimologie etc.);
- scanarea surselor bibliografice, identificarea citatelor în surse şi crearea legăturilor dintre citate şi surse;
- crearea de instrumente software care să permită interogarea Dicţionarului în moduri variate.

Un tip special de acces va fi cel în sursele scanate: când un utilizator va face *click* pe un citat al unei intrări va primi un extras din sursa literară care înglobează citatul. Dimensiunea extrasului şi navigarea în surse va fi în conformitate cu legea drepturilor de autor şi a proprietăţii intelectuale.

În lucrarea de faţă sunt prezentate primele două activităţi din cele enumerate mai sus. O pagină scanată de dicţionar nu este altceva decât o fotografie digitală a paginii. Din aceasta se poate recunoaşte automat textul, dar în acest proces de recunoaştere optică a caracterelor (OCR - *Optical Character Recognition)* apar erori. Doar ultimele volume, tipărite după 1990, fiind „culese pe calculator" există sub formă de text electronic. Pentru a obţine o variantă apropiată de original într-un timp scurt şi pentru celelalte volume a fost lansat un *site* securizat pentru corectarea formei ce a rezultat din scanare şi recunoaşterea automată de caractere (Cristea et al., 2007). Site-ul funcţionează după principiul contribuţiilor benevole. Programul care gestionează baza

de date a site-ului distribuie spre corectare utilizatorilor segmente (10-12 rânduri) din cuprinsul dicţionarului într-o secvenţă aleatoare. După ce o secţiune e corectată de cel puţin doi utilizatori benevoli, aceasta urmează să fie verificată de un expert lexicograf. În continuare, pe forma corectată se pot extrage automat câmpurile intrărilor de dicţionar.

2. Scanarea şi recunoaşterea automată a textului

Volumele dicţionarelor au fost tipărite aproape de-a lungul unui secol şi există variaţii importante ale calităţii hârtiei şi tipăriturii. Dacă hârtia veche necesită atenţie şi grijă la manevrare, tipăritura recentă are mai multe defecte. Sunt evidente în unele cazuri deformări ale tamburului de imprimare, nepotriviri între cerneală şi tipul de hârtie, semi-transparenţe, impregnarea prea puternică etc. O mare parte din defectele de hârtie au fost eliminate prin folosirea unui scanner profesional.

Fotografiile digitale, în etapa de prelucrare a imaginii, sunt tăiate, îndreptate si curăţate de unele defecte cum ar fi marginile negre obţinute prin scanare. Timpul de procesare a fost redus prin automatizarea completă a acestor operaţii. Au fost folosite bibliotecile software şi programele: Leptonica [2], ImageMagick [3], unpaper [4].

Pentru recunoaşterea automată a caracterelor şi împarţirea în extrase de 10-12 rânduri au fost evaluate soluţii profesionale, comerciale sau gratuite şi open-source. Printre cele mai performante soluţii comerciale pentru recunoaşterea automată de caractere din documente scanate se numără ABBYY FineReader Engine SDK [5], Nuance Omnipage Capture SDK [6], IRIS iDRS [7]. Toate sunt biblioteci software, oferind şi funcţii de prelucrare a imaginilor. ABBYY FineReader Engine SDK se distinge prin opţiunea de antrenare pe noi seturi de caractere. Un alt motor comercial de OCR a fost Tesseract [8], dar a urmat alt traseu: a fost dezvoltat de Hewlett-Packard ca apoi să fie oferit gratuit şi open-source. Acum face parte şi este dezvoltat de o echipă de voluntari în cadrul proiectului OCRopus [9]. Dintre proiectele open-source s-a optat pentru Gamera [10], o bibliotecă software pentru recunoaştere de caractere, însoţită de programe şi interfeţe facile pentru antrenare.

Pentru antrenare (figura 1) au fost identificate tipuri de caractere şi alese pagini care să acopere cât mai bine mulţimea de defecte şi particularităţi întâlnite. Din aceste pagini au fost culese exemple pentru fiecare caracter în parte.

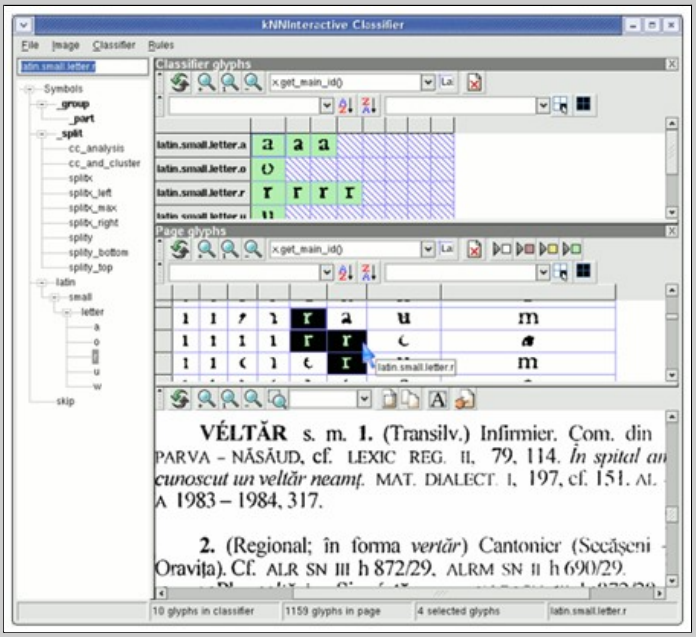


Figura 1. Antrenarea în Gamera

Pe setul de caractere latine, acurateţea cea mai bună se obţine cu IRIS iDRS, urmat îndeaproape de ABBYY FineReader Engine SDK. Gamera, deşi într-o primă antrenare nu a obţinut rezultate comparabile, poate fi integrat într-o arhitectură continuă de antrenare-corectură-validare, în care se pot reduce multe dintre greşelile frecvente.

3. Corectarea on-line colaborativă

Asupra textului recunoscut automat din paginile scanate ale dicţionarului se vor

efectua cel puţin două corecturi, în cascadă. Prima corectură este realizată de numeroşi voluntari care au cunoştinţe cel puţin de bază în utilizarea calculatorului. Un ghid de utilizare a interfeţei prezintă utilizatorilor principalele acţiuni şi operaţiile de corectare. În ultima corectură, experţi lexicografi preiau textul corectat de voluntari, cu mai puţine erori în medie decât varianta iniţială, textul recunoscut automat.

Pentru un acces facil, interfaţa de corectare e integrată într-un *site web* a cărui adresă este https://consilr.info.uaic.ro/edtlr. La primul acces, atât voluntarii cât şi experţii sunt invitaţi să se înregistreze. Accesele ulterioare se fac pe baza numelui de cont şi a parolei date la înregistrare. Conturile experţilor sunt validate de administratorul site-ului. După autentificare, voluntarii lucrează cu interfaţa de corectare naivă, iar lexicografii cu interfaţa expert.

3.1. Interfaţa de corectare naivă

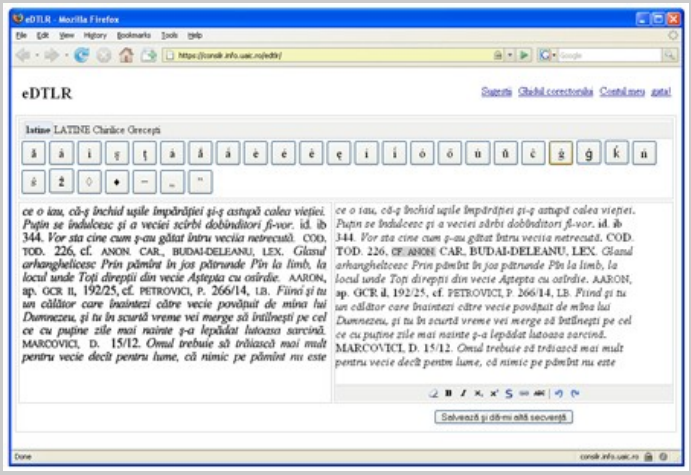


Figura 2. Interfaţa de corectare naivă

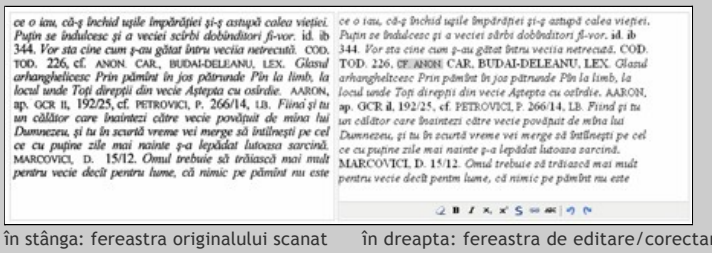
Interfaţa, ilustrată în figura 2, este constituită din două secţiuni orizontale, după cum urmează:

A. Secţiunea seturilor de caractere speciale include butoane pentru caractere **Latine, Chirilice, Greceşti** (implicit este afişat setul Latin)



Ultimele trei butoane din setul Latin sunt folosite pentru includerea unor caractere speciale.

B. Secţiunea editorului propriu-zis include două ferestre aşezate una lângă alta:


în stânga: fereastra originalului scanat în dreapta: fereastra de editare/corectare

Utilizatorul, fie benevol, fie expert, urmăreşte textul scanat, redat în fereastra stângă şi corectează, în fereastra din dreapta, greşelile care apar după operaţia de recunoaştere automată a caracterelor.

La subsolul ferestrei de corectare (figura 2) se află următoarele butoane:

- butonul de ştergere a formatării. Se utilizează pentru a lăsa porţiunea marcată de text fără nici o formatare;

- butonul de formatare cu caractere aldine (**bold**);

- butonul de formatare cu caractere cursive (*italic*);

- butonul de formatare cu caractere coborâte ($x_{subscript}$);

- butonul de formatare cu caractere urcate la umăr ($x^{superscript}$);

- butonul de formatare specifică a siglelor dicţionar. O siglă reprezintă o notaţie prescurtată a unei referinţe bibliografice.

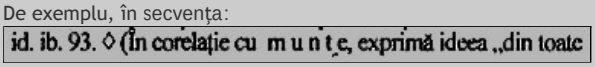
De exemplu, în secvenţa:

Vămşană bună. ALRM I/II h 248/61. *Văruşană dulce*. ib. h 248/80.

şirurile **ALRM I/II h 248/61** şi **ib. h 248/80** trebuie marcate ca sigle dicţionar (**ib.** reprezintă aici prescurtarea de la ibidem). Pentru a uşura prelucrările ulterioare, numerele trebuie încorporate siglelor. De asemenea, două sau mai multe sigle, separate ori nu prin semne de punctuaţie, trebuie să fie unite într-un singur marcaj de siglă. De exemplu, în secvenţa: **CATON, 81v/26, cf. POLIZU**, **CATON, 81v/26** şi **POLIZU** sunt sigle distincte, dar pentru a uşura lucrul, ele pot fi notate împreună. În momentul în care se aplică notaţia de siglă asupra unui şir, interfaţa realizează o transformare la caractere majuscule. În prelucrările ulterioare, siglele care sunt notate împreună, ca şi numerele de pagini din cadrul acestora, vor putea fi identificate şi separate automat.

- butonul de legare a caracterelor aparţinând unui singur cuvânt, care în original apar separate prin spaţii. Astfel de secvenţe pot apărea în redarea unor cuvinte titlu.

De exemplu, în secvenţa:



- şirul m u n te trebuie marcat drept cuvânt spaţiat.

- butonul de marcare a unor zone de text cu defecte în imaginea originală, care fac imposibilă corectarea sigură, sau a unor caractere care nu se regăsesc în seturile de caractere latine, chirilice ori greceşti. Marcaje de acest gen vor semnala lexicografilor că respectivele secvenţe vor necesita o atenţie deosebită, eventual prin confruntare cu mai multe surse tipărite, în vederea realizării unor corecţii în perfectă concordanţă cu originalul.

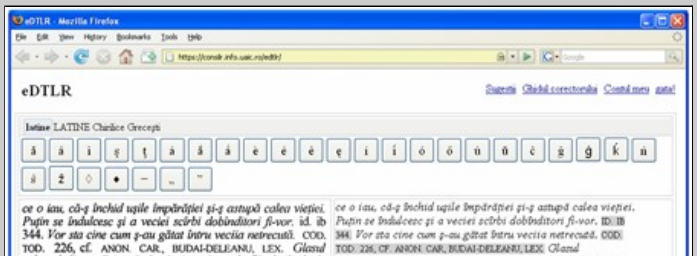
- butonul de revenire înapoi (undo).

- butonul de reluare a ultimei schimbări (redo).

3.2 Interfaţa de corectare expert

Rolul lexicografilor în procesul de corectare este dublu. În primul rând trebuie să valideze textul şi să rezolve ambiguităţile, atât în ceea ce priveşte ortografia textului cât şi formatarea acestuia. O formatare corectă a textului permite apoi codificarea adecvată a unor câmpuri speciale, precum cuvinte-titlu, sigle, etimologie etc. În al doilea rând experţilor li se cere să evalueze munca voluntarilor. Intenţia este de a promova voluntarii a căror contribuţie este de calitate.

Interfaţa expert, comparativ cu cea naivă, introduce în plus doar un singur grup de elemente care identifică autorul corecturii precedente şi afişează numărul de extrase corectate. Experţii lexicografi vor merge pe urma unui voluntar pentru a-l putea evalua. S-a optat pentru această metodă cu scopul de a stabiliza opinia expertului asupra voluntarului.



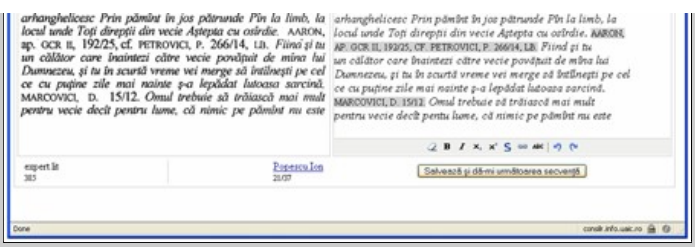
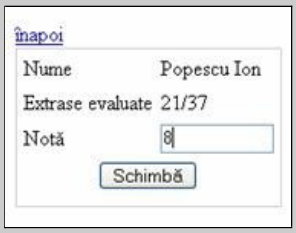


Figura 3. Interfața expert de corectare

Numele voluntarului face legătura cu o pagină de notare a paginii corectate de acesta:



După ce un expert corectează și notează contribuția voluntarului, se declanșează automat un proces de evaluare care adaugă o măsură obiectivă, dată de următoarele reguli:

- dacă la prima corectură un segment este salvat fără modificări, atunci există o mare probabilitate ca acest lucru să fie datorat grabei voluntarului și acesta este depunctat;
- dacă diferența numerelor de caractere dintre varianta salvată și cea primită pentru corectare este prea mare, atunci voluntarul poate fi suspectat de rea intenție și, prin urmare, este ușor depunctat;
- dacă diferența dintre varianta salvată de voluntar și cea salvată de expert este mică, atunci voluntarul este promovat.

Evaluarea finală a unui utilizator se obține combinând notele date de experții lexicografi cu evaluarea automată descrisă mai sus. Pentru a stimula activitatea de corectare benevolă, membrii proiectului eDTLR au inițiat campanii de promovare a acestuia. Se va încerca, totodată, și premierea corectorilor voluntari cu cea mai susținută și atentă activitate.

După corectare, secvențele de dicționar sunt integrate automat în forma integrală a dicționarului. Segmentarea și maniera de extragere aleatoare a secvențelor împiedică orice tentativă de recuperare a dicționarului în format electronic în această formă, încă nefinisată. Astfel, dacă cineva ar face 140.000 extrase, probabilitatea de a reconstitui o pagină formată din 12 segmente este de aproximativ 10 -55.

4. Concluzii

Realizarea eDTLR va permite în primul rând o viitoare unificare a formatelor celor două dicționare, DA și DLR. Printre altele, acest lucru va crea și posibilitatea corelării eDTLR cu alte resurse lingvistice, cum este, de exemplu, wordnetul românesc (Tufiș et al., 2008).

eDTLR va deschide noi perspective în dezvoltarea de aplicații de anvergură privind traducerea automată, dezambiguizarea semantică a cuvintelor, selecții de tipuri de intrări în vederea elaborării de noi dicționare specializate (tematice, etimologice, frazeologice, terminologice etc.). Mai mult, eDTLR creează premisele unei viitoare activități de actualizare a acestui adevărat dicționar tezaur al limbii române, în ritm cu însăși evoluția limbii.

Bibliografie

Dan Cristea, Marius Răschip, Corina Forăscu, Gabriela Haja, Cristina Florescu, Bogdan Aldea, Elena Dănilă (2007). *The Digital Form of the Thesaurus Dictionary of the Romanian Language*. In Proceedings of the 4th International IEEE Conference SpeD 2007, "Advances in Spoken Language Technology", C. Burileanu, H.N. Teodorescu (Eds.) Iasi, 10-12 Mai 2007, Romanian Academy Publishing House, ISBN 978-973-27-1516-1.

Gabriela Haja, Elena Dănilă, Corina Forăscu, Bogdan-Mihai Aldea (2005). *Dicționarul limbii române (DLR) în format electronic. Studii privind achiziționarea*, Editura Alfa, Iași.

Dan Tufiș, Radu Ion, Luigi Bozianu, Alexandru Ceauşu, Dan Ştefănescu (2008) *Romanian WordNet: Current State, New Applications and Prospects.* In Proceedings of 4th Global WordNet conference, GWC-2008, January 22-25 2008, Szeged, Hungary.

---

[1] https://consilr.info.uaic.ro/edtlr/wiki/

[2] http://www.leptonica.com/

[3] http://www.imagemagick.org/

[4] http://unpaper.berlios.de/

[5] http://www.abbyy.com/sdk/

[6] http://www.nuance.com/omnipage/capturesdk/

[7] http://www.irislink.com/c2-198/IDRS.aspx

[8] http://code.google.com/p/tesseract-ocr/

[9] http://code.google.com/p/ocropus/

[10] http://ldp.library.jhu.edu/projects/gamera/

Sponsor de aur / **Partenaire Or**

Sponsor de argint / **Partenaire Argent**

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / **Partenaire Bronze**

Comitetul Electrotehnic Român

Asociația Română de Terminologie

Seminar Internațional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latină/Union Latine
ISBN: 978-9-291220-37-3
2008



Terminologie et travail en réseaux de langues

MARIANI Joseph, LIMSI -CNRS & IMMI

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanţi / Exposants

Résumé

***Les technologies de la langue font l'objet d'un effort de recherche important depuis** de nombreuses années, à la hauteur de la difficulté des problèmes scientifiques à résoudre et des enjeux pour l'accès à **l'information et pour la communication.** Il est apparu nécessaire de **pouvoir disposer de ressources linguistiques et de** méthodes d'évaluation des performances pour pouvoir développer correctement des technologies qui soient utilisables dans les différents domaines **applicatifs** concernés. La France s'est dotée de 2002 à 2006 d'**un programme national, Techno-Langue, qui a eu pour objectif de produire** les données et outils nécessaires pour la langue française. **D'autres pays ont des programmes similaires pour leurs langues, sans qu'il y ait de** coordination pour partager les expériences acquises et les **ressources** produites. Le multilinguisme présente un enjeu majeur pour l'**Europe,** avec un double objectif : préserver les cultures des Etats-**Membres et faciliter la communication entre les citoyens de ces Etats. Dans une Europe qui** possède 23 langues officielles, sans compter les langues régionales, cela présente un coût très important. Les technologies de la **langue permettraient de** diminuer ce coût tout en offrant une plus grande étendue de services. L'investissement à y consacrer nécessiterait une **meilleure coordination des efforts de la** Commission Européenne et des **Etats-Membres, permettant la** réalisation de systèmes de recherche **d'informations inter-lingues, de** traduction assistée par ordinateur, voire **d'interprétation simultanée automatique intégrant plusieurs technologies** pour ces différentes langues.

1. Les technologies de la langue [Cole et al, 1998]

1.1. Domaines couverts

Les technologies de la langue concernent le traitement automatique de la langue écrite (analyse morpho-syntaxique et syntaxique, compréhension et génération de textes, résumé automatique, extraction de terminologie, recherche d'informations, système de réponses aux questions, traduction automatique ou assistée par ordinateur...) et de la langue parlée (reconnaissance et compréhension de la parole, transcription, synthèse vocale, dialogue oral, reconnaissance du locuteur, reconnaissance de la langue, traduction vocale [Mariani, 2002]...

1.2. Ressources linguistiques, évaluation et standards

Il apparaît à présent nécessaire, pour pouvoir développer correctement ces technologies, de disposer de ressources linguistiques, de méthodes d'évaluation et de standards.

Les ressources linguistiques rassemblent les corpus, bruts et annotés, les lexiques, dictionnaires, thésaurus... Elles sont nécessaires pour mener les recherches en linguistique, et elles sont essentielles pour effectuer l'apprentissage des systèmes de traitement automatique de la langue fonctionnant sur des méthodes statistiques.

L'évaluation des technologies se fait en particulier dans le cadre de campagnes d'évaluation. Elles consistent à comparer les performances de différents systèmes fondés sur diverses approches, sur des données communes, selon un protocole

commun. Ces évaluations comparatives apparaissent comme un indicateur de la qualité des recherches, à côté des classiques indices bibliométriques. Elles sont aussi un indicateur de l'avancée technologique. Elles permettent de s'affranchir des aprioris théoriques, le but étant d'obtenir les meilleurs résultats possibles, quelle que soit l'approche utilisée. Elles constituent donc un mélange de compétition et de coopération internationale, qu'on qualifie parfois du terme de « coopétition », et sont véritablement un nouveau mode d'accompagnement de la recherche technologique.

Enfin, il est nécessaire de pouvoir disposer de standards afin de pouvoir assurer la diffusion des ressources et des outils.

1.3. Besoin d'une infrastructure [Pierre, 2007]

Si la nécessité de ces ingrédients est à présent acquise, il apparaît aussi nécessaire de mettre en place une infrastructure pour pouvoir en disposer, ce qui implique un investissement important et une organisation. Concernant la production de ressources, il convient en effet de déterminer quelles sont les ressources nécessaires aux Recherches et Développements, d'assurer leur production et leur interopérabilité (et donc d'établir les standards), de valider leur qualité et de gérer leur diffusion (sous les aspects juridiques et commerciaux). Concernant l'évaluation des technologies, il convient, pour les chercheurs et les industriels, de choisir les tâches selon les besoins, de produire et de diffuser les données communes aux participants, de déterminer les métriques et les protocoles d'évaluation et d'organiser la présentation et la discussion des résultats.

1.4 L'exemple américain

L'exemple a été donné par la *Defense Advanced Research Agency* (DARPA) du Département de la Défense américain, qui a initialisé une telle infrastructure à partir de 1987, en s'appuyant sur le *National Institute of Standards and Technology* (NIST, alors encore *National Bureau of Standards* (NBS)) et en créant le Linguistic Data Consortium (LDC) en 1991 [Cieri et al., 2006], sur une décision du congrès américain. A d'abord été considérée la reconnaissance de la parole, puis la reconnaissance du locuteur et de la langue, et, dans le domaine de l'écrit, la compréhension de textes (*Message Understanding Conference* (MUC)), la recherche d'informations (TREC), avec des extensions aux réponses aux questions (Q&A) et aux documents vidéo (TREC Video), la traduction automatique et, récemment, la « distillation » d'informations, avec le programme GALE, qui inclut transcription vocale, traduction et extraction d'informations, avec l'anglais comme langue cible.

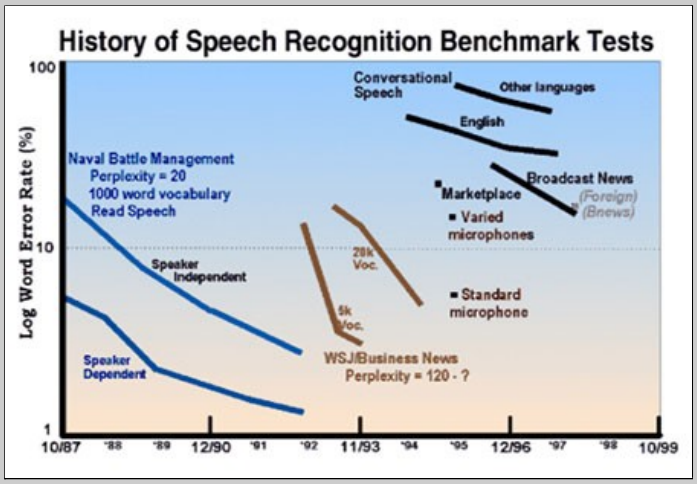


Figure 1 : Histoire de la reconnaissance vocale à travers les campagnes d'évaluation du NIST (document NIST).

La figure 1 donne une présentation de l'histoire récente de la reconnaissance vocale à travers les campagnes d'évaluation organisées par le NIST. Y figurent les résultats obtenus par le meilleur système à chaque campagne annuelle d'évaluation, sous forme de taux d'erreurs donnés en échelle logarithmique. Les tâches deviennent de plus en plus difficiles dans le temps, passant de la reconnaissance de textes lus à voix haute, pour un vocabulaire d'un millier de mots, pour un seul locuteur ou pour plusieurs, à la dictée vocale de textes de journaux, pour des vocabulaires de 5000 puis 20000 mots, à la transcription d'émissions radiodiffusées ou télévisées, pour des vocabulaires de taille indéfinie et à la reconnaissance de parole conversationnelle, sur le canal téléphonique, pour l'anglais ou pour d'autres langues. On voit que les résultats vont

s'améliorant sur chacune des tâches, ce qui est toujours rassurant pour les agences de financement qui voient ainsi que leur investissement a été utilisé à bon escient. Il est également intéressant de noter que lorsque ces campagnes ont été ouvertes à l'international en 1992, ce sont des laboratoires européens (Philips pour la reconnaissance monolocuteur et LIMSI-CNRS pour la reconnaissance multilocuteur) qui ont obtenu les meilleurs résultats, permettant ainsi de situer la recherche européenne du domaine sur l'échelle internationale. On peut aussi noter que lorsque les systèmes de dictée vocale ont atteint des performances acceptables au milieu des années 90, les produits utilisant ces technologies ont pu être commercialisées. A contrario, on note aussi que les performances sur la reconnaissance de parole conversationnelle sont restées à des taux d'erreurs élevés, justifiant le maintien d'un soutien public à la recherche malgré l'existence de solutions à des tâches plus simples.

Les activités américaines dans ce domaine ont porté initialement sur l'anglais américain, puis sur d'autres langues en fonction de leur intérêt géopolitique, mais la plupart des langues, dont le français, n'ont pas reçu un traitement comparable et il était impératif d'y pourvoir.

2. Le programme « Techno-Langue » (2002-2006)

2.1. Genèse [Mariani, 2008]

Ce programme fait suite à un rapport rédigé par un groupe de travail présidé par A. Danzin au sein du Conseil Supérieur de la Langue Française, dont le vice-président est le Délégué Général pour la Langue Française et les Langues de France, qui a été remis au Premier Ministre en novembre 2000.

Suite à ce rapport, une réunion interministérielle fut organisée en juin 2001, et trois actions furent décidées:

- Veille technologique et évaluation des outils de traitement de la langue française (pilotée par le Ministère en charge de la Recherche, en partenariat avec les Ministères en charge de l'Industrie et de la Culture)
- Développement des usages du traitement informatique de la langue française (pilotée par le Ministère de la Fonction Publique et de la Réforme de l'Etat, en partenariat avec le Ministère chargé de la Culture)
- Formation de professionnels en ingénierie documentaire (pilotée par le Ministère de l'Education Nationale)

La première de ces actions aboutit au programme Techno-Langue. Il fut demandé de l'articuler avec les instruments existants, et en particulier les Réseaux de Recherche et d'Innovation Technologiques (RRIT), permettant de financer des projets applicatifs menés en partenariat entre recherche et industrie, et le programme VSE du Ministère de la Recherche sur la veille stratégique. Techno-Langue a donc eu pour objectif d'assurer le soutien au développement de technologies de la langue génériques validées qui puissent alimenter des projets applicatifs qui en ont besoin, qui conservent pour leur part un financement global de l'ordre de 110 M€ annuels, venant des mêmes trois ministères partenaires.

2.2. Appel à projets

L'appel à projet fut organisé autour de 4 volets :

- Les ressources linguistiques : données écrites et orales (corpus, dictionnaires, bases terminologiques…) et outils de base de traitement de la langue (sous forme de Logiciels Libres)
- Les standards, afin de consacrer un effort pour améliorer la présence Française dans les organismes internationaux
- La veille technologique, en relation avec les actions alors en cours (dont le projet européen Euromap...)
- L'évaluation de technologies (campagnes d'évaluation, permettant la comparaison de systèmes), et d'applications (avec la production et la diffusion de kits d'évaluation)

L'organisation du programme comprenait un Comité de Pilotage, composé de chercheurs et d'industriels, et de représentants des RRIT (dans les domaines du logiciel, des télécommunications, de l'audiovisuel et du multimédia) et des

administrations (Ministères en charge de la Recherche, de l'Industrie, de la Culture et de la Défense), en recherchant un équilibre entre les chercheurs et les industriels, et entre les spécialistes du langage écrit et du langage parlé.

La durée maximale des projets était de 3 ans (2003 – 2005), les financements étant assuré pour chaque projet par l'un des trois ministères partenaires. La participation d'équipes étrangères dans les projets était fortement souhaitée, mais à la condition qu'ils apportent leurs propres financements.

2.3 Résultats de l'appel [Chaudiron, Mariani, 2006]

52 projets furent déposés, 28 retenus et 21 financés, regroupant 94 participants, dont 33 industriels, 40 laboratoires publics et 10 «autres participants» (Associations (AFCP, APIL…), Délégation Générale à l'Armement…). 11 laboratoires étrangers participèrent aux projets (Bell Labs (USA), NII (Japon), EPFL, LATL (Suisse), RALI (Canada)…).Le budget total du programme fut de 20 M€, dont 7,5 M€ d'aides publiques, sur 3 ans. Les projets retenus se placent dans les différents volets de l'appel.

2.3.1. Ressources Linguistiques (données)

- **Dictionnaires bilingues** (Français - Anglais, Allemand, Espagnol, Italien, Arabe),
- **Dictionnaires spécialisés** (aéronautique, sport, commerce, automobile…),
- **Dictionnaire de noms propres** (Français, Anglais, Espagnol, Néerlandais, Italien, Portugais, Serbe),
- **Corpus annotés** (français technique (corpus de normes AFNOR), journal Le Monde),
- **Corpus alignés** (36 ouvrages du 19ème siècle (Français, Anglais, Espagnol, Italien)),
- **Corpus de voix** d'enfants et de voix téléphonique de locuteurs de référence.

2.3.2. Ressources linguistiques (outils)

- **Boîte à outils de traitement de texte** : segmenteur, analyseur morphosyntaxique et syntaxique, résolution de co-références, extracteur de terminologie et d'entités nommées, résumé automatique…
- **Boîte à outils pour aider à la création de ressources sémantiques :** collecte et formatage de données, recherche d'informations sur le Web, éditeur d'ontologie terminologique, hiérarchisation de termes…
- **Boîte à outils de logiciels pour la reconnaissance du locuteur** (en licence LPGL)

2.3.3. Standards

- Pour l'écrit : Groupe miroir français du groupe ISO TC37-SC4 (ressources linguistiques)
- Pour l'oral : participation à divers groupes de travail (VoiceXML, CCXML (W3C), UNL, Aurora (ETSI), Forum SALT, IETF)

2.3.4. Veille technologique

- Mise en place d'un portail : http://www.technolangue.net

2.3.5. Campagnes d'évaluation [Chaudiron, 2007]

8 campagnes ont été organisée, 5 sur le traitement de l'écrit et 3 sur l'oral :

- EASY: Analyse syntaxique (Journaux, textes médicaux, littéraires, transcription de parole…) [Paroubek, 2006]

- ARCADE 2: Alignement de textes parallèles (phrases ou Entités Nommées (noms propres, noms de lieux…) , entre le Français et l'Anglais, l'Allemand, l'Italien et l'Espagnol, ou entre le Français et l'Arabe, le Mandarin, le Grec, le Japonais, le Persan et le Russe [Chiao et al., 2006]

- CESART: Extraction de terminologie, dans les domaines de la santé, de l'éducation, de la politique [Mustafa el Hadi et al., 2006]

- EQUER: Recherche d'informations, sous forme de Question/Réponse (corpus généraliste ou médical) [Ayache et al., 2006]

- CESTA: Traduction automatique (Anglais-Français et Arabe-Français), avec une étude spécifique des métriques d'évaluation (BLEU, NIST, mWER, X-score, WNM…) [Hamon et al, 2006]

- EVASY: Synthèse vocale à partir du texte (Conversion graphème-phonème (noms propres), Prosodie (intonation, rythme, intensité) et systèmes complets de synthèse à partir du texte) [Boula de Mareüil et al., 2006]

- MEDIA: Dialogue oral (avec production d'un corpus de dialogue « Magicien d'Oz » annoté (sur une tâche de réservation d'hôtels) et d'un dictionnaire sémantique, avec l'évaluation de la compréhension hors contexte et en contexte [Bonneau-Maynard et al., 2006]

- ESTER: Transcription de parole en vue de l'indexation automatique [Gravier et al. 2004]

2.3.6. La campagne ESTER

La campagne ESTER est exemplaire au regard de l'objectif recherché. La tâche traitait de la transcription sous forme écrite et de l'indexation automatique d'émissions de radio. Elle a permis de constituer un très large corpus pour le français (1600 h enregistrées, dont 100 h transcrites manuellement, représentant 1 Million de mots, pour 350 locuteurs), qui permet au français d'apparaître au deuxième rang mondial, derrière l'anglais américain, pour des données de ce type. Cela signifie également que de telles données n'existent pas pour la plupart des langues dans le monde, et donc qu'il n'existe pas de systèmes de transcription d'émission de radio fonctionnant correctement pour la plupart de ces langues. La campagne d'évaluation d'ESTER a rassemblé 13 participants, dont 3 industriels. Les tâches abordées concernaient la transcription (en temps réel ou différé), la segmentation (entre bruit ou musique et parole, et pour la séparation ou l'identification de locuteurs), et la reconnaissance d'Entités Nommées (directement sur la parole, ou sur le texte transcrit). Faute de temps, la tâche de détection et de suivi de thèmes pour l'indexation fut reportée.

Un atelier final fut organisé entre les participants en mars 2005. Un Kit d'évaluation fut produit et distribué, incluant les données de développement et de test, le logiciel de mesure des performances, ainsi que les résultats, afin que des équipes de recherche ou des industriels qui n'ont pas participé à la campagne d'évaluation puissent comparer leurs résultats a posteriori. Une partie des données d'ESTER fut utilisée dans la campagne d'évaluation EASY sur les analyseurs syntaxiques, afin de mesurer leurs performances sur de la parole transcrite. On pourrait ainsi mesurer l'influence d'un module d'analyse syntaxique sur la qualité globale d'un système de transcription et d'indexation. Vers l'amont, un atelier fut organisé par ailleurs pour les linguistes, afin de leur présenter les données et les outils disponibles, ainsi que les résultats obtenus, qui permettent d'identifier des verrous qui nécessiteraient des recherches fondamentales en linguistique. Vers l'aval, un moteur de recherche dans les émissions de radio fut développé dans le cadre du projet applicatif RNTL Audiosurf, regroupant les PME Sinequa et Vecsys, le LIMSI-CNRS et Radio France. Ce système peut être utilisé en libre service sur le portail Audiosurf (http://audiosurf.org/). Une suite de la campagne (ESTER 2) est en cours de lancement, avec un financement de la Délégation Générale à l'Armement (DGA). Les tâches abordées sont l'indexation automatique (détection et suivi de thèmes), non couverte dans ESTER, et la prise en compte des accents régionaux, des émissions de télévision et de la parole spontanée, telle qu'on la trouve dans les interviews, les débats, les plateaux télé…

3. Ouverture Européenne et internationale

Le besoin de technologies de la langue s'affirme de plus en plus pour permettre la communication, au niveau national ou international.

3.1. Les besoins sur la Toile mondiale

Les besoins liés au multilinguisme deviennent de plus en plus aigus sur la Toile mondiale, et on note avec intérêt la mise à disposition d'outils linguistiques par Google pour gérer la Toile mondiale, en traitant 11 langues, et 25 paires de langues. La traduction des notices techniques devient cruciale dans une économie qui s'est globalisée, et les erreurs d'interprétation ont des conséquences dramatiques qui ont récemment défrayé la chronique à plusieurs reprises.

3.2. Le programme indien TDIL

On note également avec intérêt le lancement d'un programme indien sur les technologies de la langue (TDIL (*Technology Development for Indian Languages*) : http://tdil.mit.gov.in/), qui constitue une des 10 priorités nationales pour l'informatisation de la société en Inde. 19 langues sont considérées dans ce programme : l'Anglais et 18 langues indiennes : Assamese, Bengali, Gujrati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Napali, Oriya, Punjabi, Sanskrit, Sindhi, Tamil, Telegu et Urdu. Les technologies traitées sont la traduction automatique, la synthèse et la reconnaissance vocales, les moteurs de recherche, la reconnaissance de caractères, les correcteurs orthographiques, les ressources linguistiques…

3.3. Les besoins liés à la construction européenne

Le multilinguisme apparaît comme un élément spécifique de la construction européenne, qui présente un enjeu économique, culturel et politique, avec une double nécessité :

- Celle de préserver les cultures des Etats Membres : On note par exemple une nette préférence pour accéder à des sites web dans la langue maternelle des utilisateurs (comme le montre une enquête pour 75% des allemands interrogés…). On peut souligner que 50% des citoyens Européens ne parlent qu'une langue, et que quand ils en parlent une deuxième, ce n'est pas toujours la même, et qu'à peine 3% des japonais parlent une langue étrangère.

- Celle de permettre la communication entre les citoyens des Etats-Membres : Aux 27 Etats-Membres de l'Union Européenne correspondent 23 langues « officielles », sans même parler des langues régionales, soit 506 paires de langues à traduire. La Commission Européenne employait en 2005 1650 traducteurs, qui ont traduit 1,4 Millions de pages cette année là. Le Parlement Européen consacre 30% de son budget annuel, soit 300 M€, aux questions liées au multilinguisme et emploie 500 traducteurs.

Le multilinguisme est une nécessité pour l'Europe, mais le coût qu'il entraîne est donc très important. Les technologies de la Langue permettraient de réduire ce coût, tout en apportant une meilleure réponse à la demande. Mais l'effort à leur consacrer est énorme, si l'on considère le nombre de technologies à développer pour le nombre de langues ou de paires de langues, et sans doute trop important pour la Commission Européenne seule. Il y aurait donc un grand intérêt à le partager avec les Etats-Membres, en parfait accord avec la notion de subsidiarité [Lazzari, 2006].

3.4. Les instruments de la coopération

Pour ce faire, il existe des instruments. Partant du constat qu'à peine 15% de la recherche européenne était financée par la Commission, le Commissaire Busquin proposa le concept d'Espace Européen de la Recherche (ERA), afin de mieux coordonner les efforts de recherche de l'Union Européenne. Cela conduisit au programme des ERA-Net dans le 6ème PCRD, ayant pour objectif de coordonner entre eux les programmes nationaux européens, et au programme ERA-Net+ dans le 7ème PCRD, qui permet également une coordination avec ceux de la Commission, et le financement de projets de recherche communs. Il existe également depuis longtemps la possibilité de soutenir des projets à travers l'Article 169 du Traité de l'Union Européenne, réunissant des efforts soutenus en commun de manière tripartite par la Commission Européenne, les Etats-Membres et les industriels. Cette procédure est cependant lourde, puisqu'elle demande un accord du Conseil Européen et du Parlement, et un seul programme de ce type a été financé jusqu'à présent sur les maladies infectieuse tropicales, pour un budget total annoncé de 600 M€. Les thèmes

évoqués pour le 7ème PCRD sont l'aide aux personnes dépendantes, le soutien aux PME, la Métrologie, la Recherche en mer Baltique… Il semble que seul le premier de ces thèmes ait été retenu jusqu'à présent.

3.5. Proposition de coordination des efforts européens

Les technologies de la Langue nous semblent bien adaptées au concept de l'ERA. La Commission Européenne pourrait avoir en priorité la responsabilité de veiller à assurer la coordination générale d'un tel programme (gestion, détermination des standards, évaluation des technologies, communication...) et le développement de Technologies de la Langue génériques, et chaque Etat-Membre aurait en priorité à assurer la couverture de sa ou de ses langue(s): production des ressources linguistiques qui sont essentielles pour le développement des technologies (corpus, lexiques, dictionnaires…) et développements ou adaptations des Technologies de la Langue tenant compte des spécificités de sa, ou de ses langue(s).

Le thème des technologies de la langue est actuellement dans le portefeuille des technologies de l'information et de la communication attribué à la Commissaire V. Reding. Depuis janvier 2007, un nouveau Commissaire a été nommé pour traiter spécifiquement du multilinguisme : Leonard Orban. Sous l'impulsion de son prédécesseur, J. Figel, qui avait également la charge de l'éducation, de la formation et de la culture, un rapport « Une nouvelle stratégie cadre pour le multilinguisme» a été rédigé en Novembre 2005. Suivant les recommandations de ce rapport, un Groupe de Haut Niveau sur le Multilinguisme a été créé en Novembre 2006. Une réunion ministérielle européenne sur le Multilinguisme est prévue en 2008, suivie de communications sur ce sujet devant le Parlement et le Conseil

L'intérêt pour ce thème est également exprimé par les Etats-Membres, comme on le trouve dans le« Mémorandum pour une Europe numérique » soumis par la France à la présidence finlandaise de l'Union Européenne en 2006, qui proposait comme un sujet de recherche spécifique prioritaire les «Technologies de la Langue pour une Europe Multilingue».

Ce thème est également présent dans de nombreuses initiatives européennes importantes. Dans celui de Bibliothèque Numérique Européenne, où la dimension multilingue est présente, mais qui aurait également besoin de passerelles interlingues, nécessitant de disposer de tels outils. L'Agence Européenne de Sécurité (ENISA) a proposé dans le domaine de l'Intelligence Economique la création d'une Plate-forme multilingue d'alerte et d'échange d'informations pour les Etats-Membres, qui aurait elle aussi besoin de technologies de la langue. L'Office de Brevets Européens a fait récemment l'objet de débat animé autour du nombre de langues dans lesquelles devaient être rédigés les brevets communautaires. Le coût prohibitif de la traduction a conduit à ne retenir que 3 langues, alors qu'un investissement dans le développement de technologies de la langue aurait pu permettre de considérer l'ensemble des langues, en réduisant les coûts et les délais…

3.6 Les technologies de la langue dans le 6ème PCRD

On note dans le 6ème PCRD un certain nombre de projets incluant une composante d'évaluation CLEF (Cross-Lingual Evaluation Forum), TC-Star, CHIL, AMI…), mais ces projets ne sont pas inscrits dans la durée, contrairement à la situation aux Etats-Unis [Mariani, 2002].

Le traitement du multilinguisme passe par un développement coordonné des technologies pour chacune des langues européennes. On en a l'illustration dans le Projet Intégré TC-Star [Lazzari, 2006] du 6ème PCRD, qui porte sur la traduction vocale, trois langues ayant été retenues (Anglais, Espagnol et Chinois). Un démonstrateur a été développé assurant la traduction vocale anglais-espagnol dans les deux directions, et portant sur les interventions au Parlement Européen. L'intérêt de cette application réside en la disponibilité de données de taille importante : enregistrements, transcriptions, traductions et interprétations par les interprètes et traducteurs du Parlement, permettant de construire les modèles statistiques correspondants. Il faut souligner qu'un tel démonstrateur a nécessité le développement d'un système de reconnaissance vocale et de synthèse de la parole pour chacune des deux langues, et de systèmes de traduction permettant le passage de l'une à l'autre. Il conviendrait à présent de généraliser ce démonstrateur à l'ensemble des langues officielles, et paires de langues, du Parlement.

3.7. Les technologies de la langue dans le 7ème PCRD

Dans le Programme "Cooperation" (ICT) du 7ème PCRD, les technologies de la langue figurent bien cachées dans le Pilier technologique «Simulation, visualisation, interaction, mixed realities», au sein du Challenge 2: Cognitive systems, Interaction, Robotics, qui ne comporte curieusement qu'un seul objectif, intitulé également "Cognitive systems, Interaction, Robotics ». Fort curieusement également, on trouve par contre ce thème en forte priorité dans le cadre de la partie « Coopération internationale », mais spécifiquement pour la langue arabe. Dans le programme portant sur les Infrastructures de recherche (ESFRI), le projet CLARIN a été mis en place qui porte sur la création d'une infrastructure pour les ressources et les outils linguistiques en appui aux Sciences Humaines et Sociales [Wittenburg et al., 2006]. Cela ne couvre donc pas le développement des technologies de la langue pour les applications qui en auraient besoin. Enfin dans le programme E-Content+, la mise en place d'un Réseau Thématique «Ressources linguistiques pour le traitement des langues» apparaît parmi les thèmes de l'appel 2007, qui concerne donc également les Sciences et Technologies de l'Information et de la Communication, mais les aspects liés au développement et à l'évaluation de technologies ne sont pas couverts, et le budget de 1 M€ annoncé ne permet pas d'envisager d'apporter la réponse technologique souhaitable à la hauteur des enjeux.

4. Conclusions

La France a pu mettre en place un programme national dans le domaine des Technologies de la Langue pour le Français, s'appuyant sur la production de ressources linguistiques, la détermination de standards, la veille technologique et l'évaluation. Sa pérennisation au-delà de ses trois ans d'existence et de financement demeure ouverte, et soulève la question de l'absence, en France ou en Europe, de l'équivalent d'une entité d'évaluation et de standardisation comparable au NIST.

Il apparaît une forte nécessité de disposer de Technologies de la Langue pour permettre le multilinguisme en Europe et dans le monde. C'est peut-être le thème le plus adapté entre tous à un effort partagé entre la Commission Européenne et les Etats-Membres, mais qui est malheureusement soutenu de manière insuffisante et non-coordonnée dans le 7ème PCRD, avec la crainte de devoir attendre la fin de celui-ci en 2013 pour le considérer à sa juste hauteur. La mise en place d'un groupe de haut niveau sur les technologies de la langue en soutien au multilinguisme pourrait aider à résoudre ce problème.

5. Références

Ayache C., Grau B., Vilnat A., 2006. EQueR: the French Evaluation campaign of Question-Answering Systems, LREC'06, Genoa, Italy, May 2006

Bonneau-Maynard H., Ayache C., Bechet F., Denis A., Kuhn A., Lefevre F., Mostefa D., Quignard M., Rosset S., Servan C., Villaneau J., 2006. Results of the French Evalda-Media evaluation campaign for literal understanding, LREC'06, Genoa, Italy, May 2006

Boula de Mareüil P., d'Alessandro C., Raake A., Bailly G., Garcia M., Morel M., 2006. A joint intelligibility evaluation of French text-to-speech synthesis systems: the EvaSy campaign, LREC'06, Genoa, Italy, May 2006

Chaudiron S., Mariani J., 2006. Techno-langue: The French National Initiative for Human Language Technologies (HLT), LREC'06, Genoa, Italy, May 2006

Chaudiron S. (éd.), 2007, « L'évaluation technologique dans le domaine du traitement automatique de la langue : les campagnes Technolangue », Hermès

Chiao Y., Kraif O., Laurent D., Nguyen T., Semmar N., Stuck F., Véronis J., Zaghouani W., 2006. Evaluation of multilingual text alignment systems: the ARCADE II project, LREC'06, Genoa, Italy, May 2006

Cieri C., Liberman M., 2006. More Data and Tools for More Languages and Research Areas: A Progress Report on LDC Activities, LREC'06, Genoa, Italy, May 2006

Cole R., Mariani J., Uszkoreit H., Varile N., Zaenen A., Zampolli A., Zue V., 1998. Survey of the State-of-the-Art in Human Language Technology, Cambridge University Press, http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html.

Gravier G., Bonastre J.F., Galliano S., Geoffrois E., Mc Tait K. and Choukri K., 2004. The ESTER evaluation campaign of Rich transcription of French Broadcast News.LREC'04, Lisbon, May 2004

Hamon O., Popescu-Belis A., Choukri K., Dabbadie M., Hartley A., Mustafa El Hadi W., Rajman M., Timimi I., 2006. CESTA: First Conclusions of the Technolangue MT Evaluation Campaign, LREC'06, Genoa, Italy, May 2006

Lazzari G. , 2006. Human Language Technologies for Europe. ITC IRST / TC-Star project report, http://www.tc-star.org, April 2006.

Mariani J., 2002. Are we loosing Ground to the US ? A contrastive analysis of US-EU research framework on HLT. http://www.hltcentral.org/page-975.0.shtml

Mariani J. (éd.), 2002, "Traitement Automatique du Langage Parlé: Reconnaissance", Hermès, juin 2002

Mariani J., 2008, "Research infrastructures for Human Language Technologies: A vision from France". Speech Communication journal, à paraître 2008

Mustafa el Hadi W., Timimi I., Dabbadie M., Choukri K., Hamon O., Chiao Y., 2006. Terminological Resources Acquisition Tools: Toward a User-oriented Evaluation Model, LREC'06, Genoa, Italy, May 2006

Paroubek P., Robba I., Vilnat A., Ayache C., 2006. Data, Annotations and Measures in the EASY Evaluation Campaign for Parsers of French, LREC'06, Genoa, Italy, May 2006

Pierre, J., 2007. La Langue au coeur du numérique, Délégation Générale à la Langue Française et aux Langues de France, http://www.dglf.culture.gouv.fr, Février 2007.

Wittenburg P., Broeder D., Klein W., Levinson S., Romary L., 2006. Foundations of Modern Language Resource Archives, LREC'06, Genoa, Italy, May 2006

Sponsor de aur / **Partenaire Or**

Sponsor de argint / **Partenaire Argent**

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / **Partenaire Bronze**

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latina/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



Systèmes de management de la terminologie et de la traduction assistée par ordinateur

STEFAN Veronica, Maitre de conférence,
Université Valahia du Targoviste
STEFAN Dumitru, Directeur general, ATS Targoviste

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Introduction

Bien que de long temps nous avons vu d'énormes progrès dans la manière dont nous communiquons, depuis le télégraphe Morse et jusqu'en utilisant les dernières mobile 3G avec la vidéo conférence, il y a une très importante barrière qui n'était pas encore à surmonter: la langue. Cette application permet de dissoudre cet obstacle en offrant un service automatisé bidirectionnel anglais-roumain. Son but est de traduire automatiquement non seulement les mots, mais l'ensemble des documents.

En combinant l'innovation technologique et le perfectionnisme linguistique, ATS a développé des outils d'aide à la rédaction et à la traduction qui sont des systèmes des outils de traduction anglais/ roumain indispensable pour tous les types d'entreprises, pour des usages professionnels sur Internet et Intranet. Nos principaux domaines d'intervention s'agir de traductions informatique, commerciale, juridique, financière, technique, industrie et médicale.

La solution sera disponible en tant que logiciel packagé et en tant que service en-ligne (sur abonnement). Offert en version Windows, le logiciel s'intègre harmonieusement aux logiciels de Microsoft dont Word et Excel, mais les dictionnaires et logiciels de langue sont aussi disponibles pour Palm, Pocket PC et Web.

La réalisation de ce système est d'une très nouvelle et stratégique importance dans la société digitale, en permettant l'accès vers e-government, e-business, e-work, e-learning, des entreprises et des communautés virtuelles.

1. Besoin du « traducteur » aujourd'hui. Traduction automatique

Aujourd'hui, la notion du «*traducteur*» ne défini pas seulement « *un traducteur experte* », mais aussi « *un utilisateur ordinaire* » qui doit accéder et employer les informations multilingue d'une société globale.

Indépendamment des méthodes de travail, les besoins des traducteurs en général sont identiques:

- terminologie pertinente (dictionnaires, glossaires, bases de données terminologiques);
- documents de référence (archives électroniques, textes alignés, etc.);
- recyclage de traductions antérieures (copier-coller à partir d'autres applications, archives électroniques, mémoires de traduction, etc.);
- assistance (prétraitement et finalisation des textes, outils de gestion de contenu).

La traduction automatique désigne au sens strict le fait de traduire entièrement un texte grâce à un ou plusieurs programmes informatiques, sans qu'un traducteur humain n'ait à intervenir. La traduction automatique est encore très imparfaite, et la génération de traductions d'une qualité comparable à celle de traducteurs humains relève encore de l'utopie.

On englobe parfois également sous ce terme la traduction assistée par ordinateur où la traduction peut se faire en partie manuellement ou de façon interactive avec la machine.

La traduction (entièrement) automatique, malgré ses faiblesses, peut rendre des services dans des domaines précis, comme par exemple celui de la veille technologique, où elle permet de prendre connaissance très rapidement et de façon superficielle de grandes quantités de textes. Depuis quelques années, elle connaît également un essor considérable sur le Web, puisque divers systèmes en ligne permettent de traduire automatiquement pages Web et textes brefs : à l'heure actuelle, cette technologie est appréciée du grand public car elle permet de déchiffrer de façon grossière le thème d'une page Web dans une langue totalement inconnue, et les principaux faits ou éléments d'information qu'elle contient. Pour simplifier la navigation, plusieurs moteurs de recherche comme Google, Altavista ou Yahoo! . La traduction automatique nécessite :

- des règles grammaticales, pour l'analyse du texte en langue source qui facilite la génération du texte en langue cible ;
- des règles stylistiques,
- des dictionnaires et des mémoires de traduction (des corpus alignés déjà traduits par des êtres humains).

Les systèmes de traduction automatique permettent de prendre connaissance très rapidement de grandes quantités de textes. La traduction automatique est encore très imparfaite, mais elle peut aider et faciliter la traduction manuelle de façon interactive avec la machine.

2. Le contenu du projet

Car les technologies de l'information jouent un rôle de plus en plus important dans le travail quotidien d'un traducteur, ATS a développé *un système intègre de management de la terminologie et de la traduction assistée par ordinateur* qui offre des solutions efficace a ces besoins. Par son projet, ATS fournit de *solutions de traduction par ordinateur:* systèmes et services de traduction automatique, dictionnaires et glossaires, mémoires de traduction, outils de forage de données, pour les traducteurs: le rédacteur professionnel, le scientifique, l'étudiant, l'internaute ou l'entrepreneur.

Le scope du projet est de faciliter l'accès aux ressources linguistiques, de réduire le travail (en essence, le travail répétitif) des traducteurs et d'établir la base de cohérence linguistique et méthodologique dans le domaine de la traduction.

Au moment, le projet offre des outils de traduction pour les langues roumain et anglais indispensables pour tous les types des traducteurs, pour des usages professionnels sur Internet et Intranet. Le system est modulaire et il peut-être développé pour toutes les langues, utilisant la technologie de mémoire de traduction et de la recherche terminologique. Nos principaux domaines d'intervention s'agir de traductions informatique, commerciale, juridique, financière, technique, industrie et médicale.

3. L'architecture de l'application

L'architecture du system est à 3 niveaux, respectivement une architecture client-serveur dans laquelle l'interface utilisateur, le traitement des données et le stockage sont considérés comme trois éléments distincts. L'intérêt de la chose est évidemment la modularité, ainsi que le support de plates-formes multiples. L'architecture est partagée entre :

1. Un client, c'est-à-dire l'ordinateur demandeur de ressources, équipée d'une interface utilisateur (qui peut-être un navigateur web) chargée de la présentation ;
2. Le serveur d'application, chargé de fournir la ressource mais faisant appel à un autre serveur
3. Le serveur de données, fournissant au serveur d'application les données dont il a besoin.



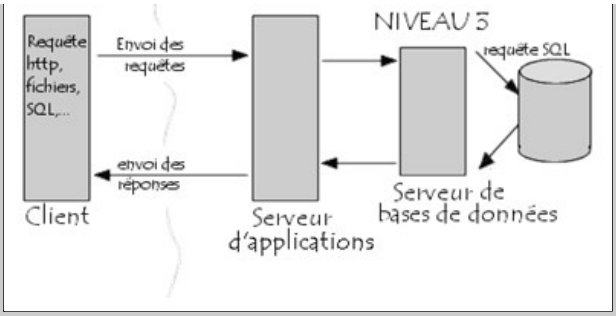


Figure 1
L'architecture du projet

L'architecture à trois niveaux permet :

- Une plus grande flexibilité/souplesse ;
- Une sécurité accrue car la sécurité peut être définie indépendamment pour chaque service, et à chaque niveau ;
- De meilleures performances, étant donné le partage des tâches entre les différents serveurs ;
- La solution est disponible en tant que logiciel packagé et en tant que service en-ligne (sur abonnement).
- Offert en version Windows, le logiciel s'intègre harmonieusement aux logiciels de Microsoft dont Word et Excel, mais les dictionnaires et logiciels de langue sont aussi disponibles pour Palm, Pocket PC et Web.

4. Les outils components du logiciel

En combinant l'innovation technologique et le perfectionnisme linguistique, le logiciel compris *des outils d'aide à la rédaction et à la traduction suivants:*

4.1. Dictionnaires et glossaires

Cet outil linguistique intègre sur la même page, dans le tableau de travaille, des dictionnaires et glossaires des langues roumaine et anglais, un dictionnaire des synonymes et un dictionnaire complexe *roumaine < > anglais*. Le dictionnaire bilingue est multidirectionnel incluant des fonctions grammaticales avancées telles que *la conjugaison* (dans tous les temps), *les accords en genre et en nombre*, etc.

Les dictionnaires sont ***généraux***, mais aussi ***thématiques***: alimentation, aéronautique, musique, construction, électronique, informatique, etc. incluant de termes anglais et roumain du vocabulaire industriel, scientifique et commercial, dans 50 domaines d'activité. Les dictionnaires et les glossaires sont en grande partie créés par des traducteurs qui travaillent dans un secteur spécialisé.

4.2. Mémoire traduction

Tout texte déjà traduit ne devrait plus faire l'objet d'une nouvelle traduction. La mémoire de traduction du logiciel est une base de données contenant des segments de texte en roumain et en anglais ainsi que l'équivalent de ces segments dans l'autre langue. Elle permet de pouvoir stocker des segments de phrase et de les réutiliser, formant en ce cas des outils de traduction assistée par ordinateur.

4.3. Systèmes et services de traduction automatique

Les systèmes de traduction automatique permettent de prendre connaissance très rapidement de grandes quantités de textes. La traduction automatique est encore très imparfaite, mais elle peut aider et faciliter la traduction manuelle de façon interactive avec la machine.

4.4. Correction

Pour une *correction linguistique*, nous contrôlons soigneusement votre texte pour ce qui est de l'orthographe, de la grammaire, de la ponctuation, de la logique, de l'interprétation, des omissions, des renvois et de la consistance de la terminologie.

Pour une *correction typographique*, le travail consiste à contrôler le formatage et la mise en page, la composition au carré ou en drapeau, la concordance des couleurs, la taille des caractères, la police, etc. La correction typographique est une condition requise pour le « bon à tirer ».

4.5. L'interface Web

Est utilisé par les clients pour la transmission électronique de leurs demandes de traduction.

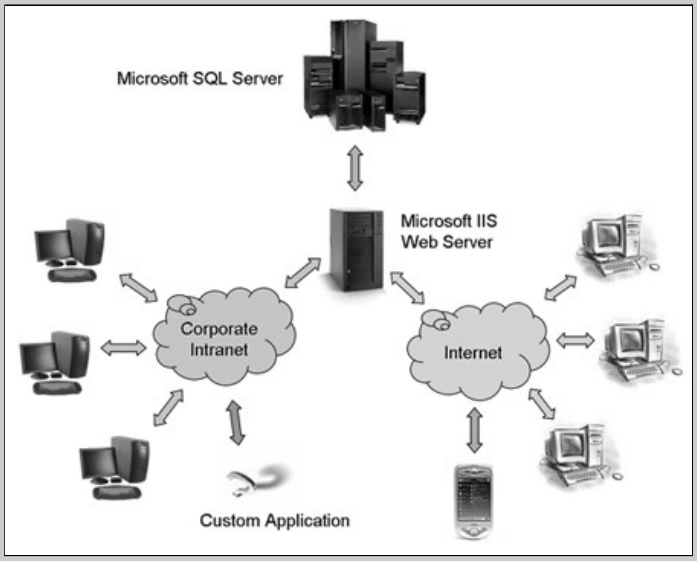


Figure 2
Les clients du projet

L'interface Web crée un dossier électronique contenant la demande, l'original à traduire et les documents de référence éventuels. Ce dossier est ensuite envoyé en bloc à l'ATS.

5. Les ressources linguistiques du projet

Les ressources linguistiques du projet contiennent :

- **terminologie** sous différentes formes (bibliothèques, accès généralisé à Internet etc.; les recherches terminologiques sont principalement effectuées par Internet;
- **des fragments de texte et des textes entiers** provenant des archives internes de la ATS ou de toute autre source gérer par Internet**;**
- **des mémoires de traduction** (des corpus alignés déjà traduits par des êtres humains) qui permettent un véritable partage des données;
- **la traduction automatique** utilisée comme outil d'aide, mais aussi à la traduction (des règles grammaticales pour l'analyse du texte en langue source qui facilite la génération du texte en langue cible et des règles stylistiques).

Les difficultés du logiciel sont diverses :

- *les ambiguïtés lexicales*, par exemple polysémie; La polysémie est la propriété qu'ont certains signes de la langue d'offrir plusieurs sens : « violon » désigne un instrument de musique, celui qui en joue, une prison. Aux unités polysémiques s'opposent les unités monosémiques (la plupart des termes scientifiques). Par exemple, beaucoup de termes scientifiques sont des mots de la langue courante empruntés pour un usage particulier.

- *les ambiguïtés grammaticales*: un même mot peut relever de deux catégories grammaticales différentes.

- *les ambiguïtés sémantiques*: selon le contexte, la traduction peut être différente. De même, les jeux de mots sont très difficiles à traduire (la poésie, ...).

- *les erreurs d'écriture* qui empêchent les logiciels de faire leur travail. Et il ne faut pas seulement penser aux fautes d'orthographe: les fautes de langue, même courantes, compliquent singulièrement la tâche.

- *sensibilité au contexte ;* les logiciels seraient capables de trouver des solutions différentes selon le contexte, soit qu'on le leur indique, soit qu'ils le trouvent automatiquement.

6. Le cycle de la traduction

Quand il commence à travailler, le traducteur lance un programme, qui segmente le texte source. Lorsque, dans la base de données, se trouve un segment source identique ou proche (le seuil de proximité étant déterminé par des options), le programme propose la traduction. Le traducteur a alors la possibilité d'accepter celle-ci, de l'adapter, ou encore de la refuser (et traduire lui-même). Le system compare le document qui doit été traduit avec les documents contenus dans les bases de données et détermine si le document ou des fragments du document on été traduit déjà.

Si le document n'a été pas déjà traduit, le system registre le document et généra ***un tableau de travaille*** qui contient :

- *un éditeur de textes alignés* avec le nouveau document et les documents de référence qui peut aider le traducteur ;

- *un outil automatise de recherche, de sélection et de remplacement de termes et de phrase ;l'accès a une bibliothèque électronique* qui comprise des dictionnaires, glossaires et bases de données terminologiques ;

- *une interface* qui permette de gérer les traductions.

- *un logiciel pour la transmission et la gestion électronique* des demandes de traduction.

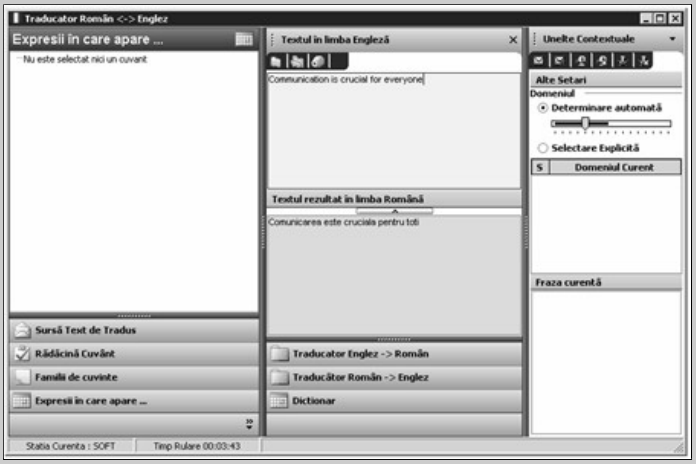


Figure 3
L'interface de l'application

Lorsque l'utilisateur introduit le numéro d'identification correspondant à la demande de traduction qu'il doit effectuer, cette interface lui donne accès à l'original ainsi qu'à tous les fichiers nécessaires pendant la traduction et, en particulier, aux documents de référence, documents prétraités, comparaisons de documents, traductions en cours, documents déjà archivés. Elle permet aussi de créer un «squelette de traduction» et inclut une fonction d'alerte et de note destinée à faciliter la communication entre les divers traducteurs travaillant sur un même projet.

TransArch est un système d'archivage électronique qui contient tous les documents entrés (essentiellement des originaux) et sortis (essentiellement des traductions) ;

L'interface propose divers critères de recherche (numéro du document, auteur, demandeur, titre ou contenu même du texte) et permet ainsi de retrouver n'importe quel document presque instantanément. Chaque document a sa propre notice, sorte de carte d'identité contenant toutes les informations pertinentes, dont les divers champs peuvent être affichés et servir de critères de recherche.

Les avantages de ces logiciels sont nombreux et évidents:

- transmission plus rapide ;
- intégration dans le système d'archivage ;
- disponibilité de l'original et des documents de référence sous forme électronique ;
- possibilité de traiter les documents avec les outils informatiques ;
- amélioration du cycle électronique de la traduction, etc.

Conclusions

Advanced Technology Systems fournit une gamme complète de solutions de traduction par ordinateur: systèmes et services de traduction automatique, dictionnaires et glossaires, mémoires de traduction et outils de forage de données pour le rédacteur professionnel, le scientifique, l'étudiant, l'internaute ou l'entrepreneur.

Cette application offre un service automatisé bidirectionnel anglais-roumain. Son but est de traduire automatiquement non seulement les mots, mais l'ensemble des documents.

Il fournira également un double sens anglais-roumain dictionnaire et thésaurus avec des définitions détaillées pour chaque langue en se développement sur une architecture ouverte, de sorte que les utilisateurs peuvent développer leurs propres applications qui s'interfacent avec le service. Le traducteur peut être utilisé pour traduire automatiquement les documents de n'importe quelle longueur.

Le texte peut être saisi manuellement ou importé des différents formats de fichier (texte, RTF, HTML ou Microsoft Word).

Le centre de la demande est le service web qui fait la traduction. La saisie de texte peut provenir de différents domaines, comme la documentation technique, des lettres, des news, etc. Le domaine (champ d'application) de ce document est déterminé automatiquement, mais peut aussi être spécifié manuellement à partir d'une liste de domaines prédéfinis.

Un dictionnaire peut facilement être utilisé pour rechercher des mots ou des expressions.

Les résultats de recherche sont filtrés automatiquement que l'utilisateur entre le mot ou l'expression désirée. Il contient des définitions détaillées sur les mots, y compris les multiples significations, des mots et des expressions contenant le mot.

L'utilisateur a également la possibilité d'ajouter un mot dans le dictionnaire ou de modifier celles déjà présentes dans la base de données. Le client Microsoft Windows peut fonctionner en arrière-plan de l'application et de traduire automatiquement le nouveau texte ou le texte sélectionné qui est dans le clipboard, en utilisant une combinaison clavier / souris. Ainsi, l'utilisateur a un plus rapide et transparent accès à l'information.

Bibliographie

Théorie de la traduction et théorie de la terminologie, in L'Histoire et les théories de la traduction - Les actes, ASTTI - ETI, Berne - Genève, 1997, Bruno DE BESSÉ.

Grammar sharing techniques for rule-based multilingual NLP systems, M. Santaholma, in: Proceedings of NODALIDA 2007, the 16th Nordic Conference of Computational Linguistics, May 25-26 2007, Tartu, Estonia.

A model for context-based evaluation of language processing systems and its application to machine translation evaluation, Popescu-Belis A., Estrella P., King M. & Underwood N., in LREC 2006 (Fourth International Conference on Language Resources and Evaluation), Genoa, Italy, p.691-696.

A flexible Speech to Speech Phrasebook Translator, M. Rayner and P. Bouillon, in: "Proceedings of ACL-02 Workshop on Speech-to-Speech Translation: Algorithms and Systems", Philadelphia, 2002, pp. 69-76.

Word sense disambiguation as a WordNets' validation method in Balkanet, Dan Tufis, Radu Ion, Nancy Ide:: *LREC-2004: Fourth International Conference on Language Resources and Evaluation*, Proceedings, Lisbon, Portugal, 26-28 May 2004; pp.1071-1074. [PDF, 422KB]

http://www.sdl.com/fr/globalization-knowledge-centre/whitepapers/Global_Communication_to_the_21st_Century.asp

http://www.eurolingua.de/cms/website.php?id=/fr/index/service-de-traduction/management-de-projet.htm

http://www.tc-star.org/
http://www.ats.com.ro

Sponsor de aur / Partenaire Or

Sponsor de argint / Partenaire Argent

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / Partenaire Bronze

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latină/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



Pretraduction d'un texte de breve

BERTACCINI Franco , SSLMIT - University of Bologna - Italie
LECCI Claudia , SSLMIT - University of Bologna - Italie

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Dans le cadre de ce projet, on a décidé d'effectuer la prétraduction d'un texte de brevet.

Description du texte de départ

Le texte qu'on choisi pour ce projet est un brevet Européen intitulé *CO2-terminated rubber*, numéro de publication EP 1373323, déposé par Firestone Polymers LLC le 2 janvier 2004. Ce brevet revendique la fabrication d'un polymère de lithium carboxyle pressurable en balles pour la production de gomme.

L'organisation du texte suit le modèle typique des brevets. Il s'ouvre avec les pages de présentation avec le titre du brevet dans les trois langues de l'EPO (anglais, allemand et français), les données concernant la classification et la registration du brevet et un résumé (*Abstract*) de 62 mots qui explique le contenu général du brevet. Successivement, il y a la description en détail du procédé, divisée en sections plus spécifiques, comme par exemple le cadre de l'invention (*Background of the invention*) qui décrit l'état de l'art, avec des références à d'autres études, et qui met en évidence ses limites ; le résumé de l'invention (*Summary of the invention*), qui présente le contenu du brevet et les buts principaux de l'invention ; la description des formes de réalisation préférées (*Detailed description of preferred embodiments*), qui décrit en détail toutes les phases du procédé et les caractéristiques de l'invention en utilisant aussi des graphiques et des tableaux. Enfin, une section du brevet est entièrement consacrée aux revendications, qui en ce cas sont 10, dans les 3 langues de l'EPO, qui justement revendiquent le droit d'exclusivité sur l'invention.

Du point de vue terminologique, il y a une prévalence de termes de spécialité relatifs aux éléments et aux composés chimiques (*butadiene*, *tetramethylethylene diamine*, *acetonitrile*) accompagnés de termes appartenant au secteur général de la chimie (*polymer*, *additive*, *molecular weight*).

Ressources utilisées pour la prétraduction

Pour la prétraduction du brevet on a utilisé les ressources décrites ci-dessous.

Description des ressources

Le procédé qu'on a suivi se compose de différentes phases, chacune caractérisée par l'utilisation d'un instrument différent:

- le repérage des brevets et des relatives traductions;
- l'alignement des textes avec WinAlign;
- la création de la mémoire de traduction (TM : Translation Memory);
- l'extraction automatique de la terminologie avec MultiTerm Extract;
- la création d'une base de données terminologique;
- la prétraduction d'un brevet.

Description des instruments utilisés

Les instruments utilisés dans le cas de ce projet font partie de l'application SDL TRADOS 2007.

Les applications TRADOS contiennent différents instruments, chacun ayant un rôle particulier. Pour développer ce projet nous avons utilisé les instruments suivants :

- WinAlign;
- Translator's Workbench;
- MultiTerm Extract;
- MultiTerm.

WinAlign

WinAlign est une application qui permet d'aligner un texte et sa traduction. Ce programme analyse les deux textes, le divise en segments et relie automatiquement les segments correspondants. Le traducteur peut établir le paramètre de segmentation en choisissant entre la segmentation au niveau de paragraphe, de phrase complète ou de phrase mitoyenne (Melby, 2000: 347, 348) la possibilité de reconnaissance automatique de(s) segments correspondants de la part du programme grandit proportionnellement au niveau de segmentation. L'alignement se base sur l'hypothèse intuitive, mais réaliste, que *des phrases longues donnent lieu à des traductions longues et des phrases courtes donnent lieu à des traductions courtes* (Véronis, 2000: 4). En utilisant des algorithmes combinés entre eux, l'instrument analyse les textes à la recherche de points d'ancrage, par exemple la première et la dernière phrase, et en utilisant les soi-disant *cognates*, c'est-à-dire les occurrences graphiquement identiques out très semblables (Véronis, 2000: 5) comme par exemple les dates, les symboles, les acronymes, etc. L'alignement peut évidemment trouver une série d'obstacles et produire des liaisons erronées, causées, par exemple, par des problèmes de segmentation, par l'absence d'uniformité dans le formatage des textes ou encore par l'absence de correspondance 1 :1 (eCoLoRe . eContent Localisation Resources for Translator Training). Pour cette raison, la phase d'alignement requiert en tout cas la supervision de l'usager qui doit confirmer ou corriger les hypothèses d'alignement fournies par l'instrument.

Translator's Workbench

Translator's Workbench est l'environnement de travail principal de TRADOS et permet d'utiliser et de gérer les ressources linguistiques crées par l'usager.

Il permet surtout de créer une mémoire de traduction de deux manières différentes: à travers la traduction interactive ou à travers l'alignement de textes qui ont été déjà traduits. Dans le premier cas, le traducteur travaille directement dans l'environnement de TRADOS et donc sur un texte segmenté en mémorisant automatiquement les segments et les relatives traductions qui vont former la mémoire de traduction (TM). Dans le deuxième cas, le traducteur utilise des textes précédemment traduits qui sont alignés et importés dans une mémoire de traduction vide ou déjà existante. Les deux manières présentent des avantages, mais aussi des désavantages: la traduction interactive permet de produire une TM de bonne qualité puisqu'elle est directement contrôlée par le traducteur qui l'a crée ; toutefois elle est caractérisée, au moins au début, par une quantité plus faible de matériel; l'importation des alignements, au contraire, produit de grandes quantités de matériel, mais elle ne permet pas un contrôle aussi direct sur le matériel importé, surtout si on importe des textes traduits par d'autres personnes (Bowker, 2002: 111).

Translator's Workbench permet, en outre, de traduire un texte dans l'environnement de Word, en utilisant et en mettant à jour automatiquement une TM existante, et d'utiliser en même temps de possibles bases de données terminologiques qui peuvent elles aussi être mises à jour automatiquement. En installant TRADOS, en effet, l'environnement de Microsoft Word est modifié automatiquement et présente les deux barres contenant les instruments principaux de TRADOS et de MultiTerm.

Translator's Workbench peut aussi être utilisé avant de commencer le travail de traduction pour effectuer une analyse du texte de départ en le confrontant avec une mémoire de traduction existante. A travers un instrument d'analyse statistique, il est en effet possible d'obtenir un *logfile* contenant les données relatives aux segments déjà présents dans la TM, organisés selon les différents degrés de *match*, et d'autres données relatives au nombre total de segments à traduire, au nombre de lignes, de

mots, etc.

Une autre fonction particulièrement intéressante du Translator's Workbench est la fonction de prétraduction (*Translate*) qui permet de prétraduire un texte en utilisant une mémoire de traduction et une base de données terminologique. De cette façon on obtient un texte hybride qui contient seulement des segments ou des termes traduit automatiquement par TRADOS. La prétraduction constituera la phase finale du procédé suivi dans ce projet et sera réalisée en utilisant des mémoires de traduction et une base de données terminologique.

MultiTerm Extract

MultiTerm Extract est une application qui permet d'extraire automatiquement la terminologie d'un corpus de textes.

Selon les textes qu'on a à disposition, il est possible de créer des bases de données monolingues ou bilingues. Dans le premier cas, si on possède des textes dans une seule langue, MultiTerm Extract permettra d'extraire automatiquement de ces textes une liste de termes candidats. Dans le deuxième cas, en ayant à disposition un corpus de textes alignés, il sera possible d'extraire une liste de termes techniques associés à la relative traduction. Dans les deux cas, il faut se rappeler que MultiTerm Extract est un instrument d'extraction semi-automatique. Pour cette raison l'intervention du traducteur humain est toujours nécessaire pour confirmer ou modifier les résultats obtenus par l'extraction automatique. L'usager a, en outre, la possibilité d'établir des paramètres d'extraction selon ses propres exigences. Par exemple, il peut établir le nombre maximal ou minimal des mots qui composent le terme, le nombre maximal de traductions pour le même terme, le seuil de fréquence, et enfin le nombre minimal d'occurrences que le terme doit avoir à l'intérieur du corpus pour être considéré un terme candidat.

MultiTerm

MultiTerm est l'application pour la gestion terminologique dans l'environnement TRADOS. Il permet de créer et de gérer des bases de données terminologiques qui peuvent être utilisées aussi en phase de traduction avec une éventuelle mémoire de traduction. Les bases de données terminologiques sont formées par des fiches terminologiques que l'usager peut personnaliser selon ses propres exigences. Ces fiches peuvent, en effet, contenir différents champs avec toutes les informations relatives au terme principal: à partir des champs basilaires, le terme, sa ou ses traductions dans le cas d'une base de données multilingue, à d'autres champs secondaires comme la définition, le contexte, les sources, les variantes, les notes de grammaire, etc.

La base de données terminologique peut elle-aussi être réalisée de deux manières: soit avec l'insertion séquentielle et manuelle des fiches terminologiques, ou de façon automatique avec MulitTerm Extract. Dans le premier cas, les fiches peuvent être créées en utilisant la barre des instruments de MultiTerm présente sur Word.

L'utilisation d'un système de gestion terminologique tel que MultiTerm offre de nombreux avantages surtout en ce qui concerne la gestion et la mise à jour de la base de données, la possibilité de faire des recherches simples et spécifiques, la possibilité d'utiliser la base de données en phase de traduction, et en assurant un niveau de cohérence terminologique plus élevé.

Les phases du procédé

Dans ce paragraphe on va expliquer toutes les phases du procédé qu'on a suivi pour créer la mémoire de traduction et la base de données terminologique et pour effectuer la prétraduction d'un brevet en utilisant ces ressources de traduction.

Le procédé s'est développé dans les phases suivantes:

- repérage des brevets et des relatives traductions;
- alignement des textes avec WinAlign;
- création de la mémoire de traduction (TM . Translation Memory);
- extraction automatique de la terminologie avec MultiTerm Extract;
- création d'une base de données terminologique;
- prétraduction d'un brevet.

Repérage du matériel bilingue

La première phase concerne la récolte des textes qui formeront le corpus bilingue contenant les brevets et les relatives traductions.

Le corpus final des textes est constitué par 60 brevets en langue anglaise et par les 60 traductions en italien. Pour repérer les textes originaux des brevets on a utilisé les sites web European Publication Server (http://www.epo.org/patents/patent-information/european-patentdocuments/publication-server.html) et esp@cenet (http://ep.espacenet.com/) qui permettent d'effectuer des recherches en utilisant différents paramètres comme le *Publication number* (le numéro de publication du brevet), l'*Application number* (le numéro de la demande du brevet), la date de publication, le titre du brevet, le nom de l'inventeur, etc.

Tous les brevets appartiennent au secteur chimique et sont caractérisés par des sous-secteurs plus spécifiques comme, par exemple, les différentes typologies de fibres, les composés et les dérivés chimiques.

Alignement des textes

Après la création du corpus contenant les textes en anglais e celui des textes en italien, on a procédé avec l'alignement des brevets avec WinAlign.

Le projet d'alignement a été subdivisé en six sous-projets qui contiennent chacun 10 alignements (le numéro maximale permis WinAlign pour chaque projet).

Pour créer un projet d'alignement il faut sélectionner le couple de langues, en ce cas anglais et italien, et insérer les 10 couples de textes à aligner.

Maintenant on peut procéder avec l'alignement de chaque couple de textes.

WinAlign produit en quelques secondes l'alignement automatique des deux textes et le traducteur se trouve devant un écran qui présente l'hypothèse d'alignement à accepter.

En général, le procédé d'alignement des brevets a été plutôt rapide parce que les hypothèses d'alignement de WinAlign ne nécessitaient pas de beaucoup de corrections. Les problèmes principaux de l'alignement sont survenus à cause de fautes de formatage (par exemple le manque d'un point ou d'un espace entre une phrase et l'autre faisaient en sorte que les deux phrases semblaient une seule), la présence de tableaux et de listes formatés de façon différente ou simplement de fautes de segmentation.

Pour construire la mémoire de traduction il est toujours conseillé de segmenter le plus possible le texte pour faciliter la recherche de correspondances au moment de la traduction.

S'il s'agit de textes techniques, on applique une segmentation au niveau de la phrase, en utilisant le point comme élément de séparation. Dans le domaine de ce projet, en alignant des textes de brevet, on a décidé en quelques cas de diviser ultérieurement des phrases particulièrement longues en utilisant, par exemple, le point virgule comme élément de séparation.

Pour séparer deux phrases réunies dans le même segment on utilise la commande S*plit Segment*, pour réunir deux segments dans un seul segment on choisit *Join Segment*.

S'il n'y a pas de problèmes de formatage, l'alignement automatique est plutôt correct même parce que, dans le domaine des brevets, la traduction respecte parfaitement l'organisation et la subdivision interne du texte original.

Quand tous les segments sont correctement alignés, on termine l'alignement des textes à travers la commande *Commit All Units* et puis on procède avec l'alignement des fichiers qui suivent.

Une fois terminés tous les alignements d'un projet, on valide tous les couples aligné à travers la commande *Mark File Pair(s) as Finished*. Le projet peut être exporté dans une mémoire de traduction.

Création de la mémoire de traduction

Maintenant, il est possible de créer une mémoire de traduction en important les alignements effectués.

La première phase consiste dans la création de la mémoire de traduction. Au moment de la création il est possible de spécifier des caractéristiques de la mémoire, par exemple les langues, les champs visualisés avec les segments (date de création, nom de l'usager, dernière utilisation, etc.), la possibilité d'associer plus d'une traduction au même segment, etc.

Ensuite on procède avec l'importation des alignements, précédemment exportés de WinAlign en format .txt. Pour faire cela on ouvre la mémoire de traduction qu'on a crée et à travers la commande *Import* on importe les projets d'alignement. Même dans ce cas il est possible d'établir les paramètres d'importation.

Avant tout, il faut spécifier l'entité de l'importation, et puis il faut en établir la modalité dans le cas où on importe un segment qui est déjà présent dans la mémoire. Selon les exigences, il est possible d'établir différents types d'information:

- *Leave unchanged*: permet d'ignorer le segment présent dans le fichier d'alignement et de maintenir intacte la mémoire de traduction ;
- *Keep most recent*: permet de considérer valide l'hypothèse de traduction plus récente;
- *Keep oldest*: permet de considérer valide l'hypothèse de traduction moins récente;
- *Merge*: permet d'insérer de nombreuses propositions de traduction pour le même segment;
- *Overwrite*: permet de substituer les vielles propositions de traduction avec celle contenue dans le fichier d'alignement.

Dans ce projet, le choix d'une de ces options d'importation n'est pas relevant, vu que les alignements sont importés dans une mémoire de traduction vide.

À la fin de chaque importation, dans la barre inférieure de Translator's Workbench on peut visualiser le numéro d'unités de traduction importées (TU: Translation Unit).

Dans notre projet on a importé 25410 unités de traduction.

Pour contrôler l'état de la mémoire et obtenir les données de l'importation il est suffisant d'utiliser la commande *Properties* qui permet de visualiser une fenêtre résumant toutes les caractéristiques de la TM.

Extraction de la terminologie

Une fois crée la mémoire de traduction, on extrait son contenu en format .tmx, un format d'échange obtenu à travers l'exportation du contenu de TM. Ensuite, on passe à la création de la base de données terminologique à travers un instrument d'extraction semi-automatique contenu dans SDL TRADOS 2007, c'est-à-dire MultiTerm Extract.

Au moment de la création du projet d'extraction il faut établir une série de paramètres. Avant tout, on doit spécifier le type de projet: monolingue, bilingue ou compilation d'un dictionnaire. Dans ce cas, on sélectionnera le projet bilingue.

Ensuite, on indique le nom du projet, la position où on veut le sauvegarder et on choisit les langues, dans notre cas anglais et italien.

Puis on sélectionne les projets d'alignement à importer en format .txt.

Une fois terminée l'importation des alignements on procède avec l'extraction et, même dans ce cas, il est possible de configurer les paramètres d'extraction.

Nous avons fixé à 1 mot la longueur minimale du terme pour pouvoir obtenir les termes relatifs aux éléments chimiques, et à 5 mots la longueur maximale du terme de façon à obtenir même des occurrences de termes techniques relatifs à des compositions chimiques composés par plusieurs mots. Cela comportera parfois la présence de séquences de mots inutiles du point de vue terminologique, mais permettra de ne pas exclure des séquences chimiques longues qui, au contraire, peuvent être très intéressantes du point de vue terminologique.

Une fois établis ces paramètres, on peut procéder avec l'extraction automatique de la terminologie.

Dans ce cas, les termes repérés sont 22.173. À la fin de cette opération il y a une page qui permet au traducteur d'examiner et de corriger la terminologie trouvée par

MultiTerm Extract.

L'écran de travail de MultiTerm Extract est organisé en trois fenêtres :la première est la liste des termes et des équivalents, à partir de laquelle on peut valider ou éliminer des termes, la deuxième présente les données relatives au terme en évidence et la troisième tous les contextes dans lesquels le terme se trouve.

Pour confirmer un terme et sa relative traduction il est suffisant d'insérer un tick dans le carré correspondant. En outre, il est possible d'insérer de nouveaux termes à travers la commande *Add as New Term* et *Add as Translation*.

Pendant la phase de validation de la terminologie il est possible de contrôler le développement du travail à travers un philtre de visualisation qui permet de visualiser seulement les termes validés, non validés ou la liste complète des termes.

Création de la base de données terminologique

Après avoir obtenu la liste des termes finals on a procédé avec l'importation de la terminologie dans la base de données terminologique.

Le procédé d'exportation des termes est très simple et permet d'importer la terminologie dans une base de données vide ou existante.

Le projet d'extraction est exporté sous un format reconnaissable par MultiTerm et peut être effectué aussi dans des moments différents.

On sélectionne les langues et on termine le procédé d'exportation.

Une fois terminé le procédé avec MultiTerm Extract, le projet peut être importé dans la base de données terminologique ; en utilisant la commande *Import* de MultiTerm on commence la configuration de l'importation.

Après avoir établis les paramètres d'importation, on choisit le fichier d'exportation du projet à importer dans la base de données. Enfin, on commande l'importation, sous forme de fiches terminologiques, du projet d'extraction automatique.

Prétraduction d'un brevet

Après avoir créé une mémoire de traduction et avoir intégré la base de données terminologique on a choisi un brevet, appartenant au secteur de la chimie, auquel on a effectué une opération de prétraduction avec le but d'analyser les résultats produits par les deux ressources linguistiques.

Avant de commencer avec la prétraduction, le texte a été formaté selon le modèle demandé par le client, de façon à effectuer la prétraduction directement sur le texte qui sera consigné au client.

Pour effectuer la prétraduction du texte il est nécessaire d'ouvrir en même temps la mémoire et la base de données. Puis, il faut établir le degré de correspondance (*Match Value*) qu'on veut utiliser pour la TM aussi bien que pour la base de données. Dans ce cas on a décidé d'effectuer deux tentatives, la première avec un *Match Value* de 100% et la deuxième avec un *Match Value* de 90%. Une fois établis ces paramètres, il est possible de procéder avec la prétraduction à travers la commande *Translate* qui ouvrira la fenêtre de configuration.

TRADOS produira un texte hybride contenant les segments et les termes traduits écrits avec une autre couleur par rapport au texte général.

> The process includes polymerizing at least one conjudated diene in the presence of an organolithium iniziatore to a substantial completion. Thereafter, the polymerization process is terminated by the addition of biossido di carbonio.

TRADOS produit, en outre un *logfile* contenant les données relatives aux segements et aux termes traduits automatiquement.

Mémoires de traduction

On a utilisé en même temps deux mémoires de traduction: une mémoire générale, et une mémoire spécialisée appartenant au domaine de la chimie. Le but de la mémoire générale serait celui d'obtenir des résultats dans la traduction de la phraséologie typique du langage des brevets, c'est-à-dire, par exemple: *the present invention relates to*, *according to the present invention*, *it is an object of the present invention*,

ecc.

La mémoire spécifique du secteur chimique peut fournir des résultats aussi bien en ce qui concerne la phraséologie qu'en ce qui concerne le langage de la chimie.

Bases de données terminologiques

En ce qui concerne la partie relative à la terminologie, on a utilisé une base de données terminologique contenant 2582 termes techniques du secteur de la chimie qui ont été le résultat de l'extraction effectuée avec MultiTerm Extract.

La base de données terminologique représente la ressource la plus importante dans le domaine de cette prétraduction parce que l'attention du projet est principalement concentrée sur l'aspect terminologique de la traduction des brevets et aux possible résultats qu'on peut obtenir en utilisant un procédé d'extraction terminologique automatique.

Prétraduction

La prétraduction du texte s'est développée en deux phases: la préparation du texte de départ et la véritable prétraduction.

Préparation du texte de départ

Avant de commencer la prétraduction, il est nécessaire de confronter attentivement les deux textes originaux. Dans notre cas, le texte d'arrivée est parfaitement identique au texte de départ et donc il peut être transféré dans un modèle de formatage prédéfini.

La phase de préparation du texte prévoit, en outre, une activité de pre-editing, c'est-à-dire un contrôle attentif du formatage, de la ponctuation, etc. Dans notre cas, il a été nécessaire de modifier des caractères spéciaux relatifs au langage de la chimie (*CO2*, *P-COO-Li+*, *a-methyl styrene*).

Preuves de prétraduction

Une fois préparé le texte de départ, on peut procéder avec la phase d'editing, c'est-à-dire avec la prétraduction. A travers Translator's Workbench et MultiTerm il est possible de lancer la prétraduction en établissant le degré de correspondance désiré, notamment 100% et 90% *match.*

Match value 100%

Si on établit un degré de correspondance de 100%, TRADOS traduit seulement les segments et les termes qui résultent identiques à ceux contenus dans les mémoires de traduction et dans la base de donnés terminologique. Les résultats peuvent être, donc, peu satisfaisants du point de vue de la quantité, mais de haute qualité. Dans ce cas, la prétraduction du brevet en utilisant un *match* de 100% a produit des résultats assez satisfaisants, même si en présentant quelques problèmes. Au moment de la prétraduction, TRADOS crée automatiquement un *logfile* contenant toutes les informations relatives au procédé de prétraduction du texte.

Dans le *logfile* il y a toutes les données relatives au texte en question, c'est-à-dire le numéro total de segments (146), le numéro total de mots (2537), le numéro de segments traduits automatiquement (9), le numéro des mots contenus dans les segments traduits (19) et le numéro totale de termes traduits en utilisant la base de données terminologique (267).

Dans ce cas, on peut noter que, avec le pourcentage maximal de correspondance, les résultats sont numériquement assez contenus. Si on calcule le pourcentage de texte traduit automatiquement, on observe que, dans ce cas, en traduisant 267 termes sur un total de 2537 mots contenus dans le texte, TRADOS a traduit 10,5% de la terminologie, et en traduisant 9 segments sur un total de 146, TRADOS a traduit 6,2% des segments du texte.

Match value 90%

La deuxième épreuve de prétraduction, réalisée en utilisant un *match* de 90%, a le but d'observer dans quelle manière change le résultat obtenu, aussi bien du point de vue quantitatif que qualitatif. Dans ce cas, en effet, vu que les termes set les segments ne doivent pas être identiques à ceux contenus dans les ressources utilisées, on s'attend que TRADOS propose une traduction même dans les cas où il n'y a pas une correspondance complète, comme par exemple dans le cas de termes au pluriel.

Avec une correspondance de 90%, le numéro de segments traduits n'a pas changé, tandis que le numéro des termes traduits a augmenté (301) par rapport à la preuve précédente. Le pourcentage des termes traduits devient donc de 11,8%, avec une augmentation de 12,7% par rapport au texte traduit avec un *match* de 100%.

L'usage combiné de systèmes TAO et systèmes TA

Le texte prétraduit représente maintenant un point de départ pour le traducteur qui peut valider ou corriger les parties de texte traduites automatiquement et traduire manuellement les autres parties. Il peut aussi insérer le texte prétraduit dans un Traducteur Automatique, notamment dans le traducteur automatique du moteur de recherche de Google.

Après avoir établis ses propres langues de travail et avoir commencé la traduction automatique du texte, on effectuera un contrôle des résultats obtenus.

Dans notre cas, le texte prétraduit contient non seulement la terminologie prétraduite par TRADOS, mais aussi une traduction complète du texte effectuée par le TA de Google. L'objectif qu'on se pose est celui d'arriver à augmenter et perfectionner le plus possible les ressources linguistiques, de façon à obtenir des résultats meilleurs du point de vue qualitatif et quantitatif en diminuant, donc, le travail du traducteur qui pourrait être substitué par une simple activité de révision d'un texte qui est déjà en langue d'arrivé.

Sponsor de aur / Partenaire Or

STAR

Sponsor de argint / Partenaire Argent

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / Partenaire Bronze

Comitetul Electrotehnic Român

TermRom Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latină/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org

altalingua®

TRADOS®
Language Technology For Your Business



SDL TRADOS - Soluția completă pentru traducătorii profesionişti

DTIL

Bucureşti, 28 februarie 2008

altalingua®

TRADOS®
Language Technology For Your Business

Prezentarea companiei SDL TRADOS (1/2)

- Pionier şi lider de piață în tehnologie lingvistică
- 50 filiale în toată lumea
- Peste 130.000 licențe vândute
- Soluția TRADOS:
  - suportă virtual toate limbile
    inclusiv limbi asiatice
  - suportă toate formatele uzuale de fişiere
    HTML, XML, SGML aplicații MS Office şi formate DTP
  - integrabil în alte sisteme
    DMS, CMS, ERP, CRM



altalingua®

TRADOS®
Language Technology For Your Business

Clienți



Microsoft
SAP
DELL
ORACLE
hp
MOTOROLA intelligence everywhere
PEOPLESOFT
Sun microsystems
NORTEL NETWORKS
CISCO SYSTEMS
SYMANTEC.
Lucent Technologies Bell Labs Innovations
NOKIA CONNECTING PEOPLE

CREDIT SUISSE GROUP
European Investment Bank
Bloomberg.com
Dresdner Bank

SEAT
BUGATTI
ŠKODA AUTO
DAIMLERCHRYSLER
Audi

WAL★MART
Baxter
OLYMPUS
Bayer

BOWNE
Lionbridge

JOHNSON CONTROLS
JOHN DEERE

amnesty international
NATO
The European Commission

Sabre
Retek
Amway

altalingua®

TRADOS®
Language Technology For Your Business

Ce este SDL TRADOS?

- Instrument de traducere asistată de calculator
- Foloseşte principiile de funcționare specifice programelor de memorie a traducerii
- Include un set complet de aplicații complexe necesare activității de traducere



altalingua®

TRADOS®
Language Technology For Your Business

Ce oferă SDL TRADOS?

- Creşterea productivităţii prin fluxuri de lucru mai eficiente
- **Calitate** pentru toate limbile traduse / standardizare la nivel frazeologic
- Control şi analiză
- Timp mai scurt pentru formatare
- Economii de **Costuri** şi **Timp**



Exemplu de traducere automată

Text original de pe site-ul Trados:

„Read the official press release"

Traducere automată:

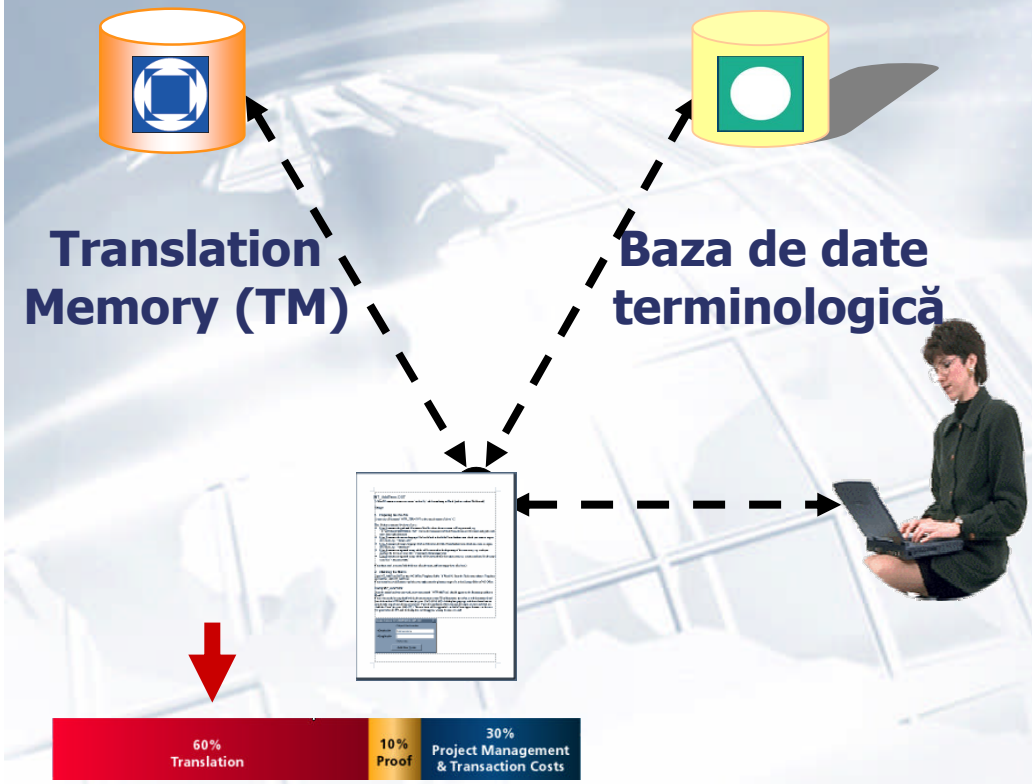
„Citeşte oficialul apasă demisie"

Această traducere nu a fost realizată cu TRADOS :-)

altalingua®

TRADOS®
Language Technology For Your Business

Interacțiunea a două baze de date



- Refolosirea propozițiilor deja traduse
- Afişarea ÎN TIMP REAL a termenilor
- Creşterea calității traducerilor

altalingua®

TRADOS®
Language Technology For Your Business

TRADOS, la nivel de propoziție



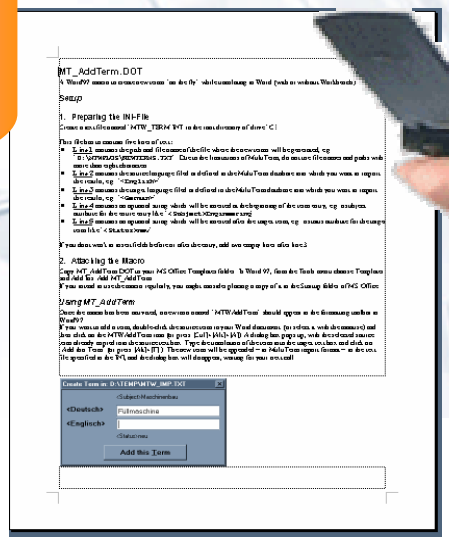
- Propozițiile sursă şi cele țintă sunt "legate" > unitate de traducere
- Construieşte o "Memorie" a Traducerilor: "Translation Memory (TM)"
- TM devine memoria limbajului unui produs sau a întregii companii

altalingua®

TRADOS®
Language Technology For Your Business

„Translation Memory" este o...







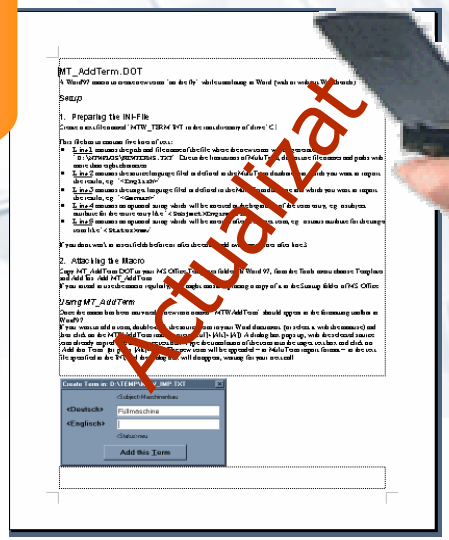
bază de date
ce creşte
şi
„învaţă"
permanent
de la
traducător

altalingua®

TRADOS®
Language Technology For Your Business

„Translation Memory" este o...





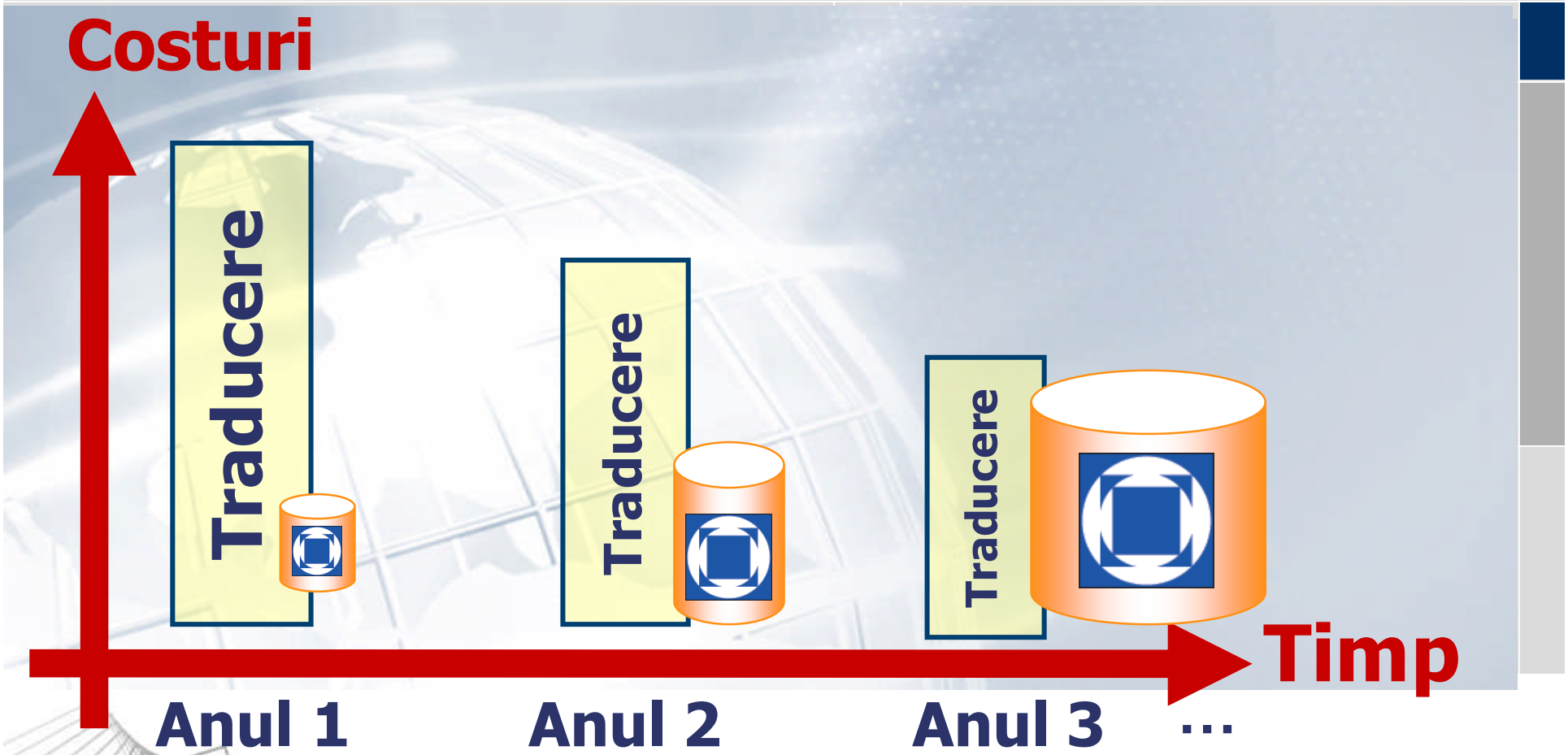


bază de date
ce creşte
şi
„aminteşte"
permanent
traducerile
existente

altalingua®

TRADOS®
Language Technology For Your Business

Efectele unui sistem TM



altalingua®

TRADOS®
Language Technology For Your Business

TRADOS, la nivel de cuvânt

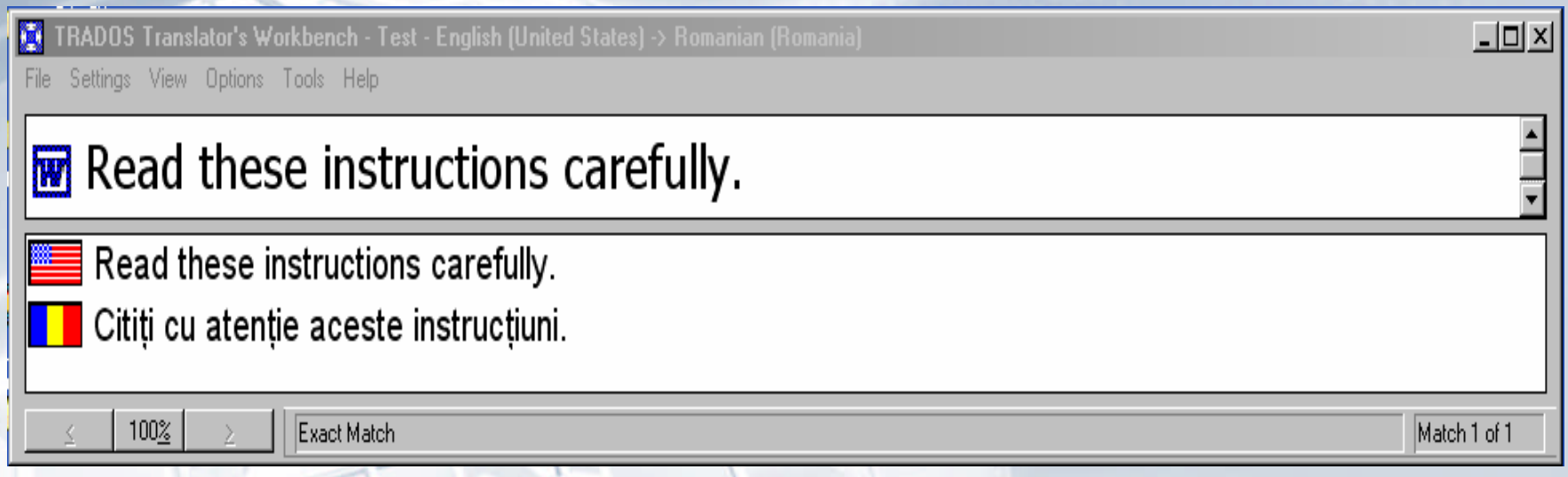
- Baze de date terminologice complet personalizate
- Interogarea bazei de date şi recunoaşterea termenilor din glosar se realizează în timp real
- Materiale accesibile întregii echipe prin Inter-/Intranet cu suport MultiTerm Online

TRADOS®
Language Technology For Your Business

Aplicații TRADOS: Translator's Workbench

- Aplicația principală a programului SDL TRADOS
- Gestionează memoria traducerii şi este conectată permanent la editorul de texte
- Traducătorul porneşte cu o bază de date goală, aceasta fiind populată cu traduceri în timp real, pe măsura realizării activității de traducere
- Odată ce utilizatorul întâlneşte fraze identice sau asemănătoare cu cele din segmentele memorate de Workbench, aplicația sugerează din memorie traducerea corespunzătoare

Exemplu: memoria unui text 100% identic

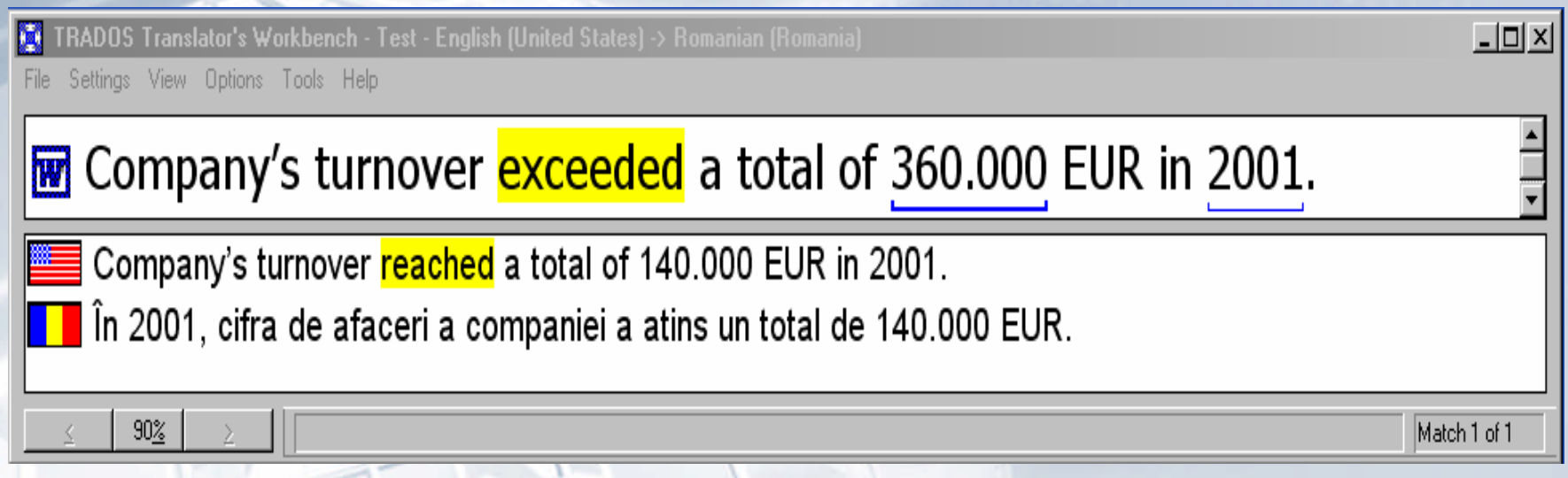


100% Match

altalingua®

TRADOS®
Language Technology For Your Business

Exemplu: memoria unui text "asemănător"



90% Match
"Fuzzy Match"

altalingua®

TRADOS®
Language Technology For Your Business

Funcţii suplimentare ale Translator's Workbench

- **Translate**
  funcţie folosită pentru traducerea automată a unor texte similare cu cele ce se regăsesc în memoria traducerii
- **Analyze**
  funcţie folosită la analiza/evaluarea cantitativă a noilor texte de tradus
- **Clean-up**
  funcţie folosită pentru etapa finală a traducerii, respectiv actualizarea memoriei de traduceri cu modificările finale şi generarea fişierului ţintă

altalingua®

TRADOS®
Language Technology For Your Business

Aplicații TRADOS: MultiTerm

- Aplicație ce gestionează bazele de date terminologice
- Are o interfață proprie
- Funcționează şi integrat, conectat la Translator's Workbench

TRADOS®
Language Technology For Your Business

Câmpurile glosarului MultiTerm

- Limba sursă şi cea țintă
- Referințe pentru traducerile respective
- Statusul validării termenilor
- Schițe şi grafice concludente

MultiTerm [Project Untitled.xdp]

Termbase Entry Search Tools Window Help

pa* | Deutsch | Englisch | Flags Layout

Demo (6)
- paletpac
- Palettenabnahmeplatz
- Palettenaufgabeplatz
- Paletteneinstellplatz
- Partnerunternehmen
- Pauschalwertberichtigung

UPU (14)
- Packwagen
- Packwagen mit Postabteil
- Paket des Land-/Seewegs
- Paket mit Wertangabe
- Paket mit zerbrechlichem Inhalt
- Paketmarke
- Palette
- Panafrikanischer Postverein

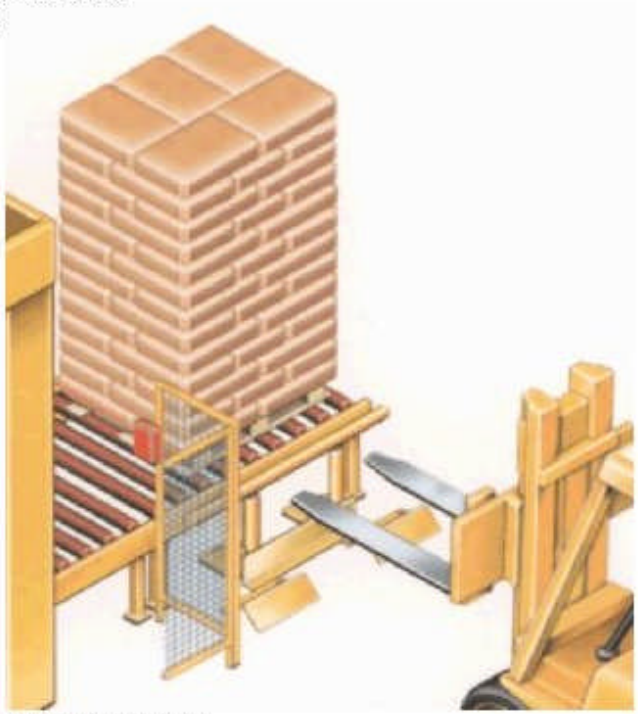
Hit List | Favourites | Project

- Palettenabnahmeplatz
- Palettenaufgabeplatz
- Paletteneinstellplatz
- Partnerunternehmen
- Pauschalwertberichtigung
- Periodenabgrenzung
- Periodenergebnis
- Periodenergebnis
- Pläne für Kapitalbeteiligungsleistungen
- Pläne für Leistungen nach Beendigung des Arbe
- Plankürzung
- Planvermögen
- Potentielle Stammaktie
- Präzision
- Preisrisiko
- Private Kunden

Demo

Entry number: 364

Graphic:



Fachgebiet: Maschinenbau
Erfassung: teilweise
Definition: Es handelt sich um den Platz, an dem Paletten oder Pakete mittels Gabelstapler von der Rollenbahn heruntergenommen werden.

Deutsch

Term: Palettenabnahmeplatz

Genus: m
Numerus: s & p
Wortart: Substantiv
Eintragsart: Wort
Endung: -(e)s / "e
Verwendung: Synonym
Erfasser: Landwehr

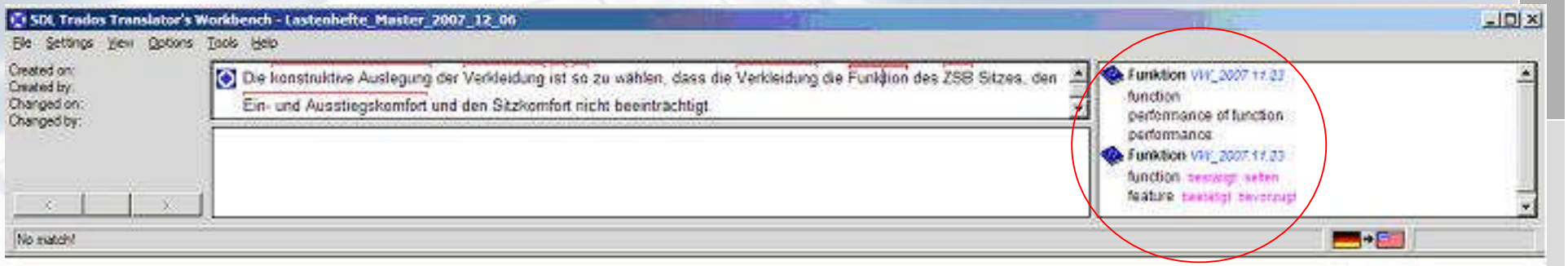
Palettenabnahmeplatz

Fertig

TRADOS®
Language Technology For Your Business

Active Term Recognition

- **Interogarea bazei de date terminologice în timp real, pe parcursul traducerii (conectat la Workbench)**



TRADOS®
Language Technology For Your Business

Aplicații TRADOS: TagEditor

- Aplicația are rol de editor de texte
- Cu TagEditor pot fi procesate textele ce provin din fişierele pachetului MS Office, programele specifice activității de DTP şi cele ce țin de HTML
- Funcționează integrat cu Translator's Workbench

TRADOS®
Language Technology For Your Business

Interfața TagEditor

- Texte editabile de către utilizator
- Texte needitabile, blocate sub forma ”tag-urilor” ce cuprind informațiile cu privire la tipul de fişier sursă şi formatarea textului



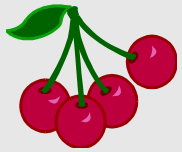
The **total eclipse** could not be seen well in Paris.



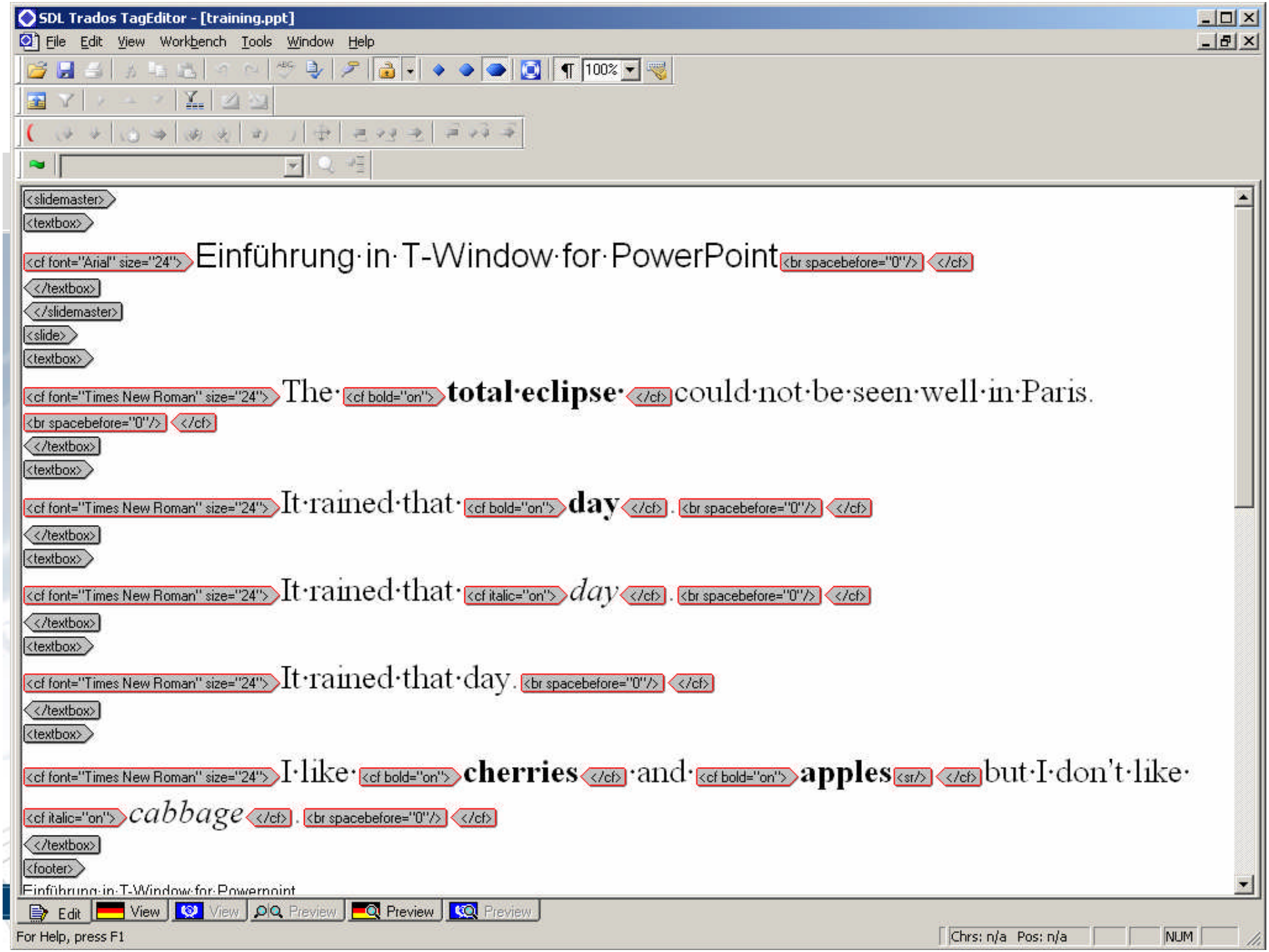
It rained that **day**.

It rained that *day*.

It rained that day.



I like **cherries** and **apples** but I don’t like *cabbage*.


SDL Trados TagEditor - [training.ppt]
File Edit View Workbench Tools Window Help
100%
<slidemaster>
<textbox>
<cf font="Arial" size="24">Einführung·in·T-Window·for·PowerPoint<br spacebefore="0"/></cf>
</textbox>
</slidemaster>
<slide>
<textbox>
<cf font="Times New Roman" size="24">The·<cf bold="on">total·eclipse·</cf>could·not·be·seen·well·in·Paris.
<br spacebefore="0"/></cf>
</textbox>
<textbox>
<cf font="Times New Roman" size="24">It·rained·that·<cf bold="on">day</cf>.<br spacebefore="0"/></cf>
</textbox>
<textbox>
<cf font="Times New Roman" size="24">It·rained·that·<cf italic="on">day</cf>.<br spacebefore="0"/></cf>
</textbox>
<textbox>
<cf font="Times New Roman" size="24">It·rained·that·day.<br spacebefore="0"/></cf>
</textbox>
<textbox>
<cf font="Times New Roman" size="24">I·like·<cf bold="on">cherries</cf>·and·<cf bold="on">apples<sr/></cf>but·I·don't·like·
<cf italic="on">cabbage</cf>.<br spacebefore="0"/></cf>
</textbox>
<footer>
Einführung·in·T-Window·for·Powerpoint
Edit View View Preview Preview Preview
For Help, press F1
Chrs: n/a Pos: n/a
NUM

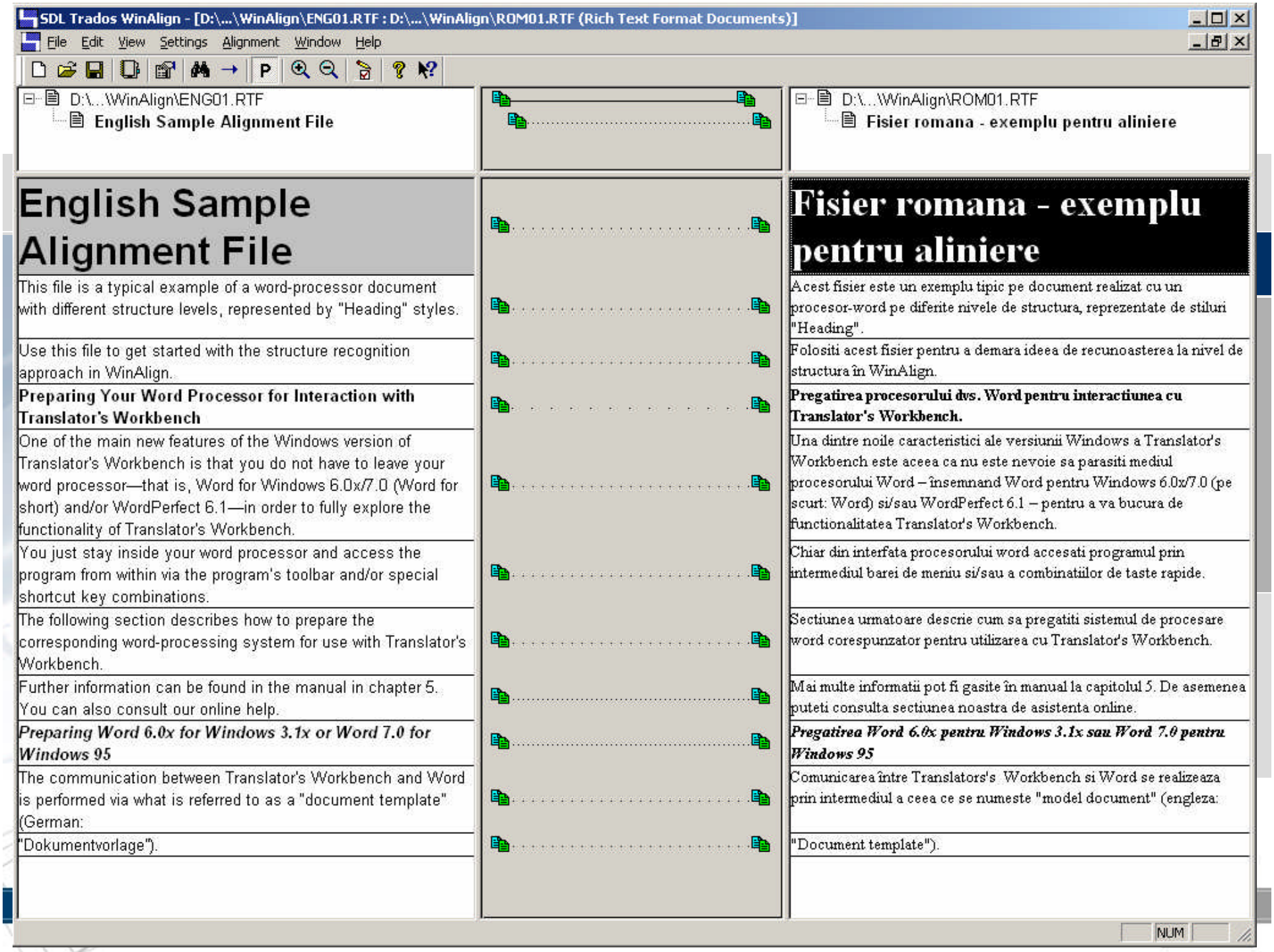
TagEditor QA plug-ins

- Spelling Checker, folosit pentru verificarea ortografiei
- Generic Tag Verifier, folosit pentru verificarea corectitudinii tag-urilor
- QA Checker, folosit pentru verificarea punctuației, integrității traducerii, consecvenței la nivel de propoziție, corectitudinii numerelor
- Terminology Verifier, folosit pentru verificarea concordanței terminologiei cu cea dintr-un glosar prestabilit

TRADOS®
Language Technology For Your Business

Aplicații TRADOS: WinAlign

- Aplicația WinAlign are rolul de a genera memorii de traducere din texte traduse fără TRADOS
- WinAlign ”aliniază” textele sursă şi țintă la nivel de propoziție, generând astfel unități de traducere ce sunt exportate la final ca memorie de traduceri Translator's Workbench
- Principiile alinierii textelor țin de formatarea similară a segmentelor din fişierele sursă şi țintă şi existența de numere şi date calendaristice identice în acestea


SDL Trados WinAlign - [D:\...\WinAlign\ENG01.RTF : D:\...\WinAlign\ROM01.RTF (Rich Text Format Documents)]
File Edit View Settings Alignment Window Help
D:\...\WinAlign\ENG01.RTF
English Sample Alignment File
D:\...\WinAlign\ROM01.RTF
Fisier romana - exemplu pentru aliniere
English Sample Alignment File
This file is a typical example of a word-processor document with different structure levels, represented by "Heading" styles.
Use this file to get started with the structure recognition approach in WinAlign.
Preparing Your Word Processor for Interaction with Translator's Workbench
One of the main new features of the Windows version of Translator's Workbench is that you do not have to leave your word processor—that is, Word for Windows 6.0x/7.0 (Word for short) and/or WordPerfect 6.1—in order to fully explore the functionality of Translator's Workbench.
You just stay inside your word processor and access the program from within via the program's toolbar and/or special shortcut key combinations.
The following section describes how to prepare the corresponding word-processing system for use with Translator's Workbench.
Further information can be found in the manual in chapter 5. You can also consult our online help.
Preparing Word 6.0x for Windows 3.1x or Word 7.0 for Windows 95
The communication between Translator's Workbench and Word is performed via what is referred to as a "document template" (German:
"Dokumentvorlage").
Fisier romana - exemplu pentru aliniere
Acest fisier este un exemplu tipic pe document realizat cu un procesor-word pe diferite nivele de structura, reprezentate de stiluri "Heading".
Folositi acest fisier pentru a demara ideea de recunoasterea la nivel de structura în WinAlign.
Pregatirea procesorului dvs. Word pentru interactiunea cu Translator's Workbench.
Una dintre noile caracteristici ale versiunii Windows a Translator's Workbench este aceea ca nu este nevoie sa parasiti mediul procesorului Word – însemnand Word pentru Windows 6.0x/7.0 (pe scurt: Word) si/sau WordPerfect 6.1 – pentru a va bucura de functionalitatea Translator's Workbench.
Chiar din interfata procesorului word accesati programul prin intermediul barei de meniu si/sau a combinatiilor de taste rapide.
Sectiunea urmatoare descrie cum sa pregatiti sistemul de procesare word corespunzator pentru utilizarea cu Translator's Workbench.
Mai multe informatii pot fi gasite în manual la capitolul 5. De asemenea puteti consulta sectiunea noastra de asistenta online.
Pregatirea Word 6.0x pentru Windows 3.1x sau Word 7.0 pentru Windows 95
Comunicarea între Translators's Workbench si Word se realizeaza prin intermediul a ceea ce se numeste "model document" (engleza:
"Document template").
NUM

altalingua®

TRADOS®
Language Technology For Your Business

Avantajele TRADOS pe scurt:

- refolosirea traducerilor existente
- gestionarea bazelor de date terminologice
- traducerea în echipă, lucrul în rețea
- traducerea fişierelor format DTP
- creşterea productivității la traduceri şi implicit scăderea costurilor
- controlul costurilor

Vă mulțumim pentru

interesul acordat **SDL TRADOS**!

Cristian Cerghedean

Coordonator Proiecte

alta**lingua** – reprezentant SDL TRADOS Desktop Technology

tel: +40 21 326.80.83

fax: +40 21 326.80.82

trados@alta-lingua.com

www.alta-lingua.com



Traducerea asistată de computer a corespondenței de afaceri engleze - române

MAXIM Victoria, Universitatea Tehnică a Moldovei

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Abstract:
articolul reprezintă o descriere a posibilităţilor de traducere semiautomată a termenilor economici din corespondenţa de afaceri engleză - română. A fost cules un corpus paralel de scrisori de afaceri din care au fost extraşi termenii şi stocaţi într-o bază de date terminologică. A fost elaborat un program care efectuează traducerea asistată de computer a sintagmelor nominale terminologice din scrisorile comerciale.

1. Introducere

Elaborarea şi dezvoltarea instrumentelor pentru procesarea automată a limbajului natural, vorbit într-o ţară anume sau în cadrul unui întreg grup de ţări, este extrem de importantă pentru această ţară în planul competitivităţii în domeniul ştiinţei şi al tehnologiei. Pentru a elabora asemenea aplicaţii, specialiştii din domeniul Inteligenţei Artificiale şi Computer Science ar trebui să posede instrumente şi mijloace adecvate pentru investigarea limbajului în vederea procesării lui automate. Unul din aceste instrumente este cunoaşterea profundă atât a lingvisticii computaţionale, cât şi a teoriei limbajului, în general.

Diverse comunităţi, inclusiv cele comerciale, ştiinţifice, educaţionale, îşi dezvoltă treptat propriile sisteme specializate de comunicare şi software. Pentru a fi operabile, majoritatea activităţilor, bazate pe limbajul natural, se referă la un set de tehnologii computaţionale.

Aceste tehnologii computaţionale includ:

- *Traducerea Automată*: traducerea unui LN în altul.
- *Regăsirea Informaţiei*: pentru localizarea textelor relevante undeva în spaţiul web, utilizând motoare de căutare gen Yahoo.
- *Recunoaşterea Vorbirii*: pentru convertirea cuvintelor vorbite în texte scrise.
- *Sumarizarea Automată*: pentru convertirea unui sau mai multor documente în mărimi manevrabile.
- *Înţelegerea LN*: limba este studiată, plecând de la cuvinte aparte (în formă grafică sau sonoră), şi ajungând la sens.
- *Generarea LN*: producerea răspunsurilor adecvate (corecte) în LN la intrări aleatorii (diferite).

Analiza limbajului comercial s-a transformat într-o filozofie privind organizaţiile comerciale şi sistemele vitale. O preocupare majoră este sensul informaţiei comerciale care le oferă acestor organizaţii un avantaj de supravieţuire, adaptare şi forţă strategică. Activitatea comercială din zilele noastre este inevitabil legată de sistemele tehnologice informaţionale. Prezenta lucrare se bazează pe premisa că este necesar un set de tehnologii sintactico-semantice pentru a înţelege şi a crea soluţii software, care să fie puse la dispoziţia întreprinderilor.

2. Analiza Limbajului Comercial sub aspect conceptual-ontologic.

Abordarea analizei limbajului comercial se efectuează cu ajutorul metodologiilor orientate la obiect şi construieşte un model de limbaj care specifică atât conţinutul, cât şi structura terminologiei utilizate în afaceri. Analiza limbajului comercial produce modele informaţionale, care sunt utilizate în comunicarea dintre întreprinderi sau organizaţii. Este o metodă de analiză a semanticii comerciale: sensul informaţiei din corespondenţa de afaceri.

Analiza limbajului comercial produce unele componente de modelare şi o documentaţie formală pentru a face accesibile aceste componente. Deoarece analiza limbajului comercial este, în esenţă, o analiză bottom-up a unui Corpus de terminologie comercială, procesul e orientat pe detalii. El începe de la un material lingvistic oferit sau bazat pe mediul de afaceri. Pentru a rezolva sau a reduce ambiguitatea limbajului specific domeniului, putem ataşa detalii la o structură de nivel înalt prin determinare de definiţii, aplicând modelele existente şi completând noi modele de abstractizare.

Pentru un model semantic complet al businessului este nevoie de două dimensiuni de bază: a). un lexicon de termeni de uz curent în domeniul de afaceri şi b). o ontologie de concepte, care ar ajuta la clasificarea sensurilor acestor termeni, descoperiţi în cadrul analizei limbajului comercial.

Următorul set de concepte rezultă din considerarea lucrurilor de care este legat orice tip de business:

- Oameni - inclusiv persoane şi organizaţii.
- Resurse - materiale, energie, abilităţi, bani, informaţii.
- Procese - evenimente, procese cap-coadă, funcţii, acţiuni discrete.
- Rezultate - produse şi servicii care au un motiv de a fi în business.
- Perioade de timp - conceptul standard de timp.

Acest set reprezintă doar una din numeroasele modalităţi de a diviza spaţiul conceptual la un nivel superior. Nu există o aşa-numită cale cea mai bună de a elabora un asemenea gen de clasificări. Iată un exemplu de relaţii tipice conceptuale: resursele sunt transformate de procesele generate de evenimente şi invocă funcţii, acţiuni discrete, decurgând din materiale şi informaţii. Oamenii şi organizaţiile joacă diverse roluri responsabile pentru diferite funcţii. Procesele creează rezultate, care la rândul lor pot fi resurse.

Conceptele oferite de o ontologie comercială generică alcătuiesc bazele acestei clasificări, însă ele vor trebui extinse prin concepte relevante şi eventual unice pentru un domeniu comercial particular. Un lingvist poate efectua clasificarea termenilor, însă un expert în afaceri trebuie să valideze această clasificare. Figura ce urmează este un exemplu al unui fragment dintr-o reţea semantică a conceptelor comerciale.

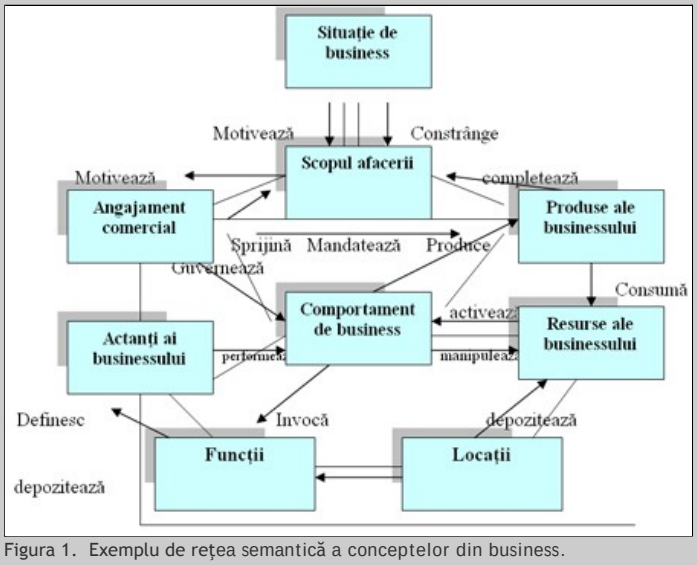


Figura 1. Exemplu de reţea semantică a conceptelor din business.

3. Analiza Limbajului Comercial sub aspect terminologic.

Elaborarea bazelor de date terminologice bilingve este facilitată în prezent prin

utilizarea programelor soft de extragere a terminologiei, cum sunt *XTS* [1] (Xerox), *Lexter* (Bourigault et al.,1999), *Nomino* (David et Plante ; Perron J., 1996), *TermPlus* (Ladouceur et al, 1996).

Dispunând de asemenea instrumente, mai multe echipe de cercetători s-au ocupat cu achiziţia terminologică mono- sau bilingvă din corpusul WEB pentru construirea BDT sau a lexicului specializat. Pe plan metodologic, majoritatea proiectelor au utilizat corpusuri paralele sau bilingve aliniate la nivel de fraze pentru a pune în corespundere unităţile terminologice extrase.

Pentru inventarierea automată a termenilor sunt utilizate modelele lingvistice şi modelele statistice. Cercetările întreprinse în ultimul deceniu au dat naştere unor metodologii care nu sunt nici pur lingvistice, nici pur statistice - ele sunt modele hibride.

Deoarece textele utilizate sunt specializate, lucrul asupra achiziţionării semiautomate a termenilor bilingvi a fost efectuat în următoarele etape:

- Adnotarea morfosintactică a textelor bilingve.
- Codificarea morfologică a substantivelor şi sintagmelor nominale susceptibile de a fi termeni.
- Extragerea acestor cuvinte şi expresii conform codificărilor.
- Validarea termenilor.
- Detectarea cuplurilor echivalente de traducere.
- Elaborarea regulilor de transformare (formalizarea unor categorii de reguli în scopul automatizării traducerii terminologiei din domeniul economic).

Nu există o corespondenţă perfectă între sintagma nominală şi unitatea terminologică. Însă multe sintagme nominale conţin termeni. Extragerea grupurilor nominale s-a efectuat în baza codificării morfologice. S-au determinat mai multe tipuri de sintagme: sintagme nominale bimembre, de ex.: *risc management, piaţa întreprinderilor, acord comercial;* sintagme nominale trimembre, de ex.: *producţie la comandă, legea creşterii costurilor, cambie fără acoperire;* sintagme nominale tetramembre, de ex.: *stoc de produse finite;* sintagme nominale polimembre, de ex.: *managementul afacerilor mici şi mijlocii,* care se pretează la următoare clasificare: sintagme din 5, 6, 7, 8 şi 9 componente. Sintagme din 5 componente constitutive, de ex.: *stoc în curs de transport*. Sintagme din 6 componente constitutive, ex.: *indicele preţurilor de vânzare a mărfurilor*. Sintagme din 7 componente constitutive, ex.: *gestiune cu cerere constantă la intervale egale*. Sintagme din 8 componente constitutive, ex: *criterii de definire a întreprinderilor mici şi mijlocii*. Sintagme din 9 componente constitutive, ex: *buget pentru activităţi de încasări şi plăţi în valută*.

4. Pregătirea corpusului şi decupajul termenilor.

Corpusul de scrisori de afaceri în limba engleză a fost parţial cules la computer, parţial scanat şi redactat. Apoi a fost tradus în limba română şi aliniat la nivel de paragrafe şi enunţuri. Adnotarea morfosintactică a textelor engleze a fost realizată cu ajutorul setului de etichete folosit de Centrul de Cercetări XEROX de la Grenoble pentru limba engleză. Pentru adnotarea morfosintactică a textelor româneşti în proiectul nostru a fost utilizat un set de 196 de taguri şi Dicţionarul Morfologic elaborat împreună cu colegul nostru, lector superior Artur Butucel, care conţine aproximativ 90 000 de forme ale cuvintelor. Adnotarea a fost efectuată cu ajutorul limbajului de programare PERL. Iată un exemplu de corpus adnotat morfologic:

```
<P>
O           vrea/Va--3s|un/Tifsr|=/Q|el/Pp3fsa
scrisoare   =/Ncfsrn
constă      consta/Vmip3s|consta/Vmis3s
dintr-      unknown
-o          un/Tifsr|=/Q|el/Pp3fsa
serie       =/Ncfsrn
de          =/S
elemente    element/Ncfp-n
,           comma
fiecare     =/Di3-sr|=/Pi3-sr
dintre      =/S
acestea     acesta/Dd3fpr|acesta/Pd3fpr
fiind       fi/Vmg
```

esenţială esenţial/Afsrn
sau =/C
contribuind contribui/Vmg
într- întru/Spsay
-o un/Tifsr|=/Q|el/Pp3fsa
măsură =/Ncfsrn
sau =/C
alta altul/Pi3fsr
la =/S
influenţarea unknown
destinatarului destinatar/Ncmsoy
. point

După cum se poate observa, unele cuvinte capătă mai multe taguri, fapt ce denotă ambiguitatea lor. Anume din cauza ambiguităţii morfosintactice a cuvintelor, adnotarea a fost corectată manual. Unele cuvinte nu au fost recunoscute (unknown), din mai multe motive: de exemplu, cuvintele scrise cu majuscule, abrevierile, cuvintele-lipsă din dicţionar, formulele, simbolurile sau, pur şi simplu, cuvintele ce au fost culese incorect. Alte două categorii formează cuvintele cu un singur cod, adică neambigue şi potenţiali termeni, şi cuvintele cu două sau mai multe coduri, adică ambigue.

Cuvintele ce aveau mai mult decât o etichetă au fost dezambiguizate manual. Apoi din Corpus au fost extrase sintagmele nominale susceptibile a fi termeni, au fost codificate şi alcătuite pentru ele reguli de transformare. În baza lor am alcătuit un glosar bilingv de termeni de afaceri.

5. Metode şi reguli.

A fost creată o listă de termeni, precum şi o analiză manuală a regulilor de transformare a termenilor englezi în cei româneşti. Plecând de la lista termenilor simpli întâlniţi, au fost stabilite 2 categorii:

1. Categoria de **isomorfisme** : adică de termeni morfologic identici în ambele limbi datirită utilizării lor internaţionale.

2. Categoria de **calchieri morfologice** cu transformări sufixale şi de accent : este vorba de termeni ce împărtăşesc radicalele greco-latine.

Iată nişte exemple din ambele categorii:

Termeni izomorfi.

Antidumping - politică comercială defensivă de contracarare a vânzării la preţuri mai scăzute.

Dumping - vânzarea în străinătate a unor preţuri mai scăzute decât cele practicate pe piaţa internă şi mondială. Modalitatea de concurenţă.

Broker - intermediar în acţiunile de bancă, mijloceşte încheierea contractelor fără a intra în relaţiile contractuale de durată, nici cu cumpărătorul, nici cu vânzătorul.

Clearing - operaţie de compensaţie globală a tuturor creanţelor şi angajamentelor dintr-o ţară faţă de alta.

Dispatching - sistem de repartizare a activităţilor de producţie.

Etc...

Calcuri lingvistice.

abrogare < lat. abrogare

acceptat < fr. acceptant

acciz < fr. accise pl. accize (taxe)

achita < fr. aquitter

achiziţie < fr. acquisition, It. acquisitio

acord < fr. accord, it. accordo

acreditiv < fr. accord, accreditif

Pentru termenii complecşi întâlniţi, regulile cele mai frecvente rezidă în transpoziţia sintactică, pentru care am obţinut 3 tipuri, şi anume: inversie sintactică izocategorială; inversie sintactică transcategorială şi inversia sintactică suprimând prepoziţia şi/sau determinantul. Putem vorbi despre inversie sintactică izocategorială atunci când categoria gramaticală a fiecărui membru al grupului nominal se păstrează pe parcursul transformării.

ADJ_N (în engleză) ® N_ADJ (în română), de exemplu,
current..account. **®** ***cont curent***

Se înţelege prin inversie sintactică transcategorială existenţa unei transformări a categoriei gramaticale a unui membru al grupului nominal al limbii-sursă în altă categorie gramaticală în limba-ţintă. De exemplu, numele englez se transformă în adjectivul românesc.

N_N(cap) ® N(cap)_ADJ, de exemplu,
acceptance market.. ® piaţă recepţionară.

Al treilea tip de inversie sintactică poate fi formulat în modul următor:

N2_N1 ® N1_prep_ N2, de exemplu,
accounting data ® documente de contabilitate

Am elaborat o serie de reguli de transformare din limba engleză în limba română. Iată câteva exemple:

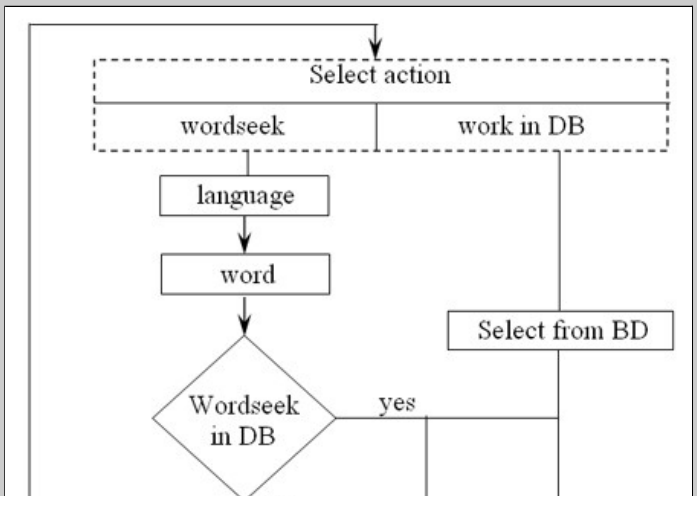
Adj + N ® N+Adj	
Current account	Cont curent
Absolute advantage	Avantaj absolut
Accessorial charges	Costuri accesoriale
N + N(cap) ® N(cap) + ADJ	
Acceptance market	Piaţă recepţionară
N2 + N1 ® N1 + prep + N2	
Accounting data	Documente de contabilitate.

6. Programul de Traducere Semiautomată a corespondenţei de afaceri engleze-române.

Programul nostru de traducere asistată de computer a fost elaborat în limbajul de programare Delphi-Borland 5 cu scopul de a crea un dicţionar bilingv al termenilor economici şi de a le oferi traducătorilor un mediu integrat cu diferite utilitare pentru a le simplifica lucrul.

Din cauza volumului mare de informaţie, programul se axează doar pe termenii ce conţin substantive, însă poate fi uşor adaptat şi pentru verb, de exemplu. Programul prezintă o secvenţă de evenimente: proceduri utilizate în cadrul programului: căutarea cuvântului în baza de date, accesul window la diferite nivele (căutare, editare, adăugare), editarea, editarea/adăugarea pentru mai multe părţi de vorbire.

Programul este elaborat în limbajul de programare Borland-Delphi 5, care oferă instrumente vizuale puternice de creare a aplicaţiilor, astfel interfaţa utilizator apare destul de rapid. De asemenea, datorită mai multor componente integrate, procesul de elaborare a aplicaţiilor este mult simplificat.



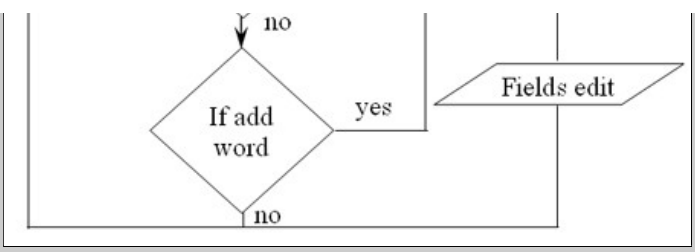


Figura 2. Algoritmul de editare a bazei de date.

Interfaţa programului. Pentru a facilita sarcina utilizatorului, el poate utiliza atât meniurile sarcinilor, cât şi butoanele de pa bara sarcinilor. Iată o interfaţă a dicţionarului care a fost creată cu scopul de a fi prietenoasă şi plauzibilă.

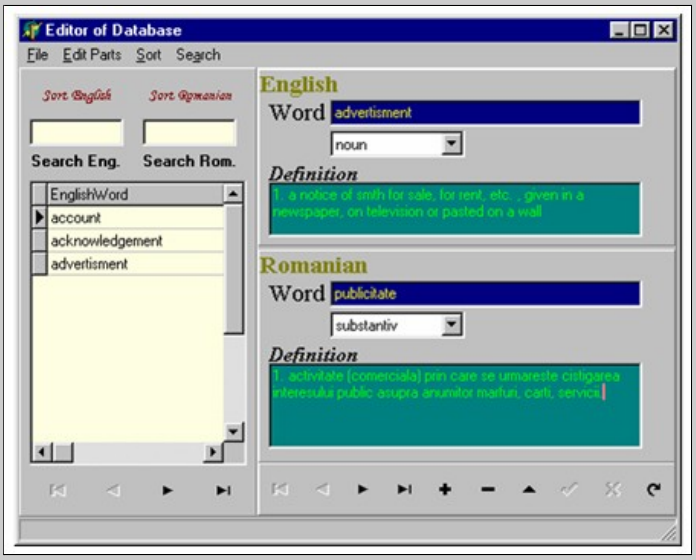


Figura 3. Interfaţa dicţionarului.

Pentru determinarea PDV incluse în BDT, utilizăm fereastra Editorul PDV, unde sunt incluse părţile de vorbire.



Figura 4. Editorul PDV.

Dacă un cuvânt nu a fost găsit în fereastra de dialog, apare următorul mesaj:



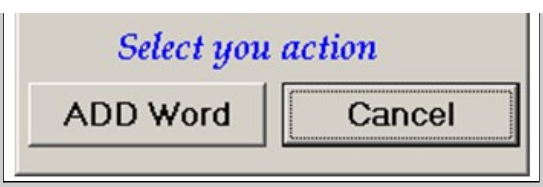


Figura 5. Fereastra rezultatelor căutării.

Aşadar avem opţiunea de a introduce un cuvânt nou sau să-l anulăm.

Şi totuşi, fereastra principală utilă pentru utilizatorul dicţionarului este ***Dictionary***, unde sunt date toate rezultatele traducerii şi explicaţiile de rigoare.

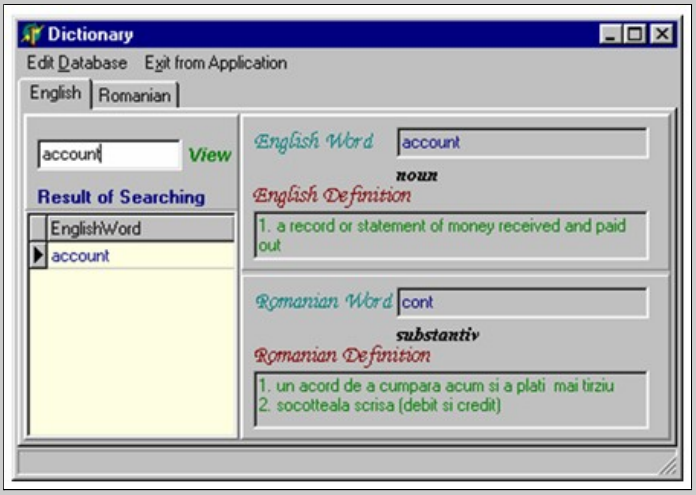


Figura 6. Dicţionarul.

Este important de menţionat faptul că limbajul Delphi este un instrument foarte util atât pentru începători, cât şi pentru programatorii experimentaţi. El combină formatul prietenos al Visual Basic cu viteza şi precizia lui C++, facilitând crearea rapidă a aplicaţiilor eficiente pentru Windows.

7. Concluzii.

Trebuie să recunoaştem, că programele cele mai evoluate existente pe piaţă sunt departe de a egala performanţa lingvistică a unui vorbitor nativ al unei limbi sau competenţa unui traducător profesionist. TA rămâne o tehnologie foarte complexă şi dificil de implementat, în măsura în care sensul unui cuvânt poate varia în funcţie de contextul utilizat (literar, ştiinţific, religios...), de nivelul limbajului practicat (specialist, începător), de numeroase referinţe socio-culturale. Din toate aceste motive, o traducere exactă sau cel puţin corectă şi adecvată necesită ca persoana sau programul să poată cuprinde (înţelege) contextul, structura şi regulile pentru cele două limbi implicate. Inginerii şi lingviştii se zbat asupra acestei probleme de mai multe zeci de ani, însă va fi nevoie de un efort enorm pentru ca să fie posibilă elaborarea/crearea unui sistem automat de traducere ale cărui rezultate să fie comparabile cu cele ale unui specialist uman din domeniu.

Referinţe bibliografice.

1. ALLEN, J., ***Natural language understanding.*** Benjamin / Cummings Publ., 2ns ed.,1995. Lucrarea poate fi consultată pe site-ul : http://www.uni-giessen.de/~g91062/Seminare/gk-cl/Allen95/al1995co.htm.

2. Frerot C, Rigou G., et Lacombre A. (2002), ***Approche phraséologique d'une extraction automatique de terminologie dans un corpus scientifique bilingue aligné***, // document disponibil pe pagina web http://www.inra.fr/Internet/Unites/UCD_Jouy/linguist/pub/approche.htm .

3. GELBUKH,A. ***Computational Linguistics. Models, Resources, Applications.*** Carte disponibilă la pagina web : http://cicling.org/gelbukh/clbook/Computational_Linguistics.htm.

4. GAUSSIER E., (2001), ***General considerations on bilingual terminology extraction,*** *Recent Advances in computational Terminology*, pp. 167-184, Amsterdam/Philadelphia: John Benjamins Publishing Company.

[1] http://www.rint.org/attrait

Sponsor de aur / **Partenaire Or**

Sponsor de argint / **Partenaire Argent**

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / **Partenaire Bronze**

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latina/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



TRANSITXV

REZUMAT

Principiul reutilizării

În traducere, orice document este o activitate cu totul nouă – dar oare chiar aşa este? Există porţiuni de text care sunt identice sau foarte asemănătoare cu ceea ce s-a mai scris şi publicat, adică au mai fost traduse cândva. Mai ales când este vorba despre versiuni noi ale produsului sau actualizări, documentaţia conţine informaţii noi ce trebuie traduse. Dar în acelaşi timp, există un grad mare de reutilizare a unor cuvinte şi propoziţii care se regăsesc în traducerile anterioare. Ideea de a reutiliza documentele şi traducerile existente (numite materiale de referinţă) formează baza unui sistem de memorie a traducerii (TMS). De peste 18 ani, STAR aplică acest principiu pentru a dezvolta şi perfecţiona TransitXV transformându-l într-un mediu de lucru profesional şi ergonomic, pentru proiectele de traducere în mai multe limbi.

Traducere fără frontiere

Deoarece suportă setul de caractere cu bit dublu Unicode, TransitXV se poate utiliza pentru editare în peste 180 de limbi diferite – lucru neegalat până acum de vreun alt competitor. Sunt suportate de TransitXV chiar şi limbi din Orientul Îndepărtat (chineza, japoneza) şi limbile cu scriere de la dreapta la stânga (araba, ebraica). Datorită interfeţelor moderne XML, transferului de date TMS şi a unui modul de import al bazei de date extrem de puternic, trecerea la TransitXV de la alte sisteme de memorie a traducerii n-a fost nicicând mai uşoară.

Standarde înalte de calitate pentru terminologia dvs.

Cu ajutorul unei interfeţe ce se poate configura, se poate importa cu uşurinţă terminologia existentă, stocată în liste şi baze de date, în sistemul complet integrat de gestionare a terminologiei din TransitXV, TermStar. Acest lucru înseamnă că funcţia de căutare în terminologia de specialitate dă rezultate chiar de la primul proiect de traducere.

Simplitate

Proiectele în mai multe limbi sunt pregătite într-o singură operaţiune – orice tip de document original, de exemplu FrameMaker, InDesign, AutoCAD, QuarkXPress, Word, SGML etc. se poateimporta în TransitXV. În acest proces, TransitXV separă automat de text informaţiile legate de structură. În timpul procesului de import al documentelor în TransitXV, acestea sunt pretraduse simultan folosind materialul de referinţă.

Comunicarea între echipele din locații separate şi îndepărtate nu mai reprezintă acum o problemă cu TransitXV – documentele ce necesită traducere şi dicționarele cu terminologie sunt comprimate de TransitXV şi trimise la traducătorii respectivi prin simpla apăsare a unei taste. Documentele traduse ale proiectului sunt expediate înapoi în acelaşi format comprimat. Acestea sunt apoi introduse automat în proiectul existent. Terminologia actualizată este introdusă în dicționarele respective.
Toate fişierele traduse sunt exportate printr-o singură comandă – textul tradus este reunit cu informațiile structurii. În final, TransitXV salvează automat documentele traduse pentru fiecare limbățintă într-o structură formată din dosare, identică cu cea a proiectului original.

Rapiditate

Fişierele de lucru şi cele pentru referință din TransitXV depăşesc rar dimensiunea de 10 KB. Prin urmare, memoria traducerii rămâne mică chiar şi după ani de traducere intensă. Avantajele constau în pretraducere mai rapidă, transfer imediat de fişiere către traducător, încărcare şi editare mai eficientă a fişierelor.

Calitate

Datorită structurii bazate pe fişiere, toate materialele de referință se pot gestiona cu ajutorul instrumentului File Manager. Acest lucru înseamnă că puteți selecta anumite fişiere din materialul de referință şi le puteți exclude cu uşurință dacă nu au calitatea dorită. Ca rezultat, memoria traducerii devine mai bună la fiecare utilizare. Şi nu trebuie să fiți un specialist în baze de date pentru a administra şi a întreține memoria traducerii.

Profitabilitate

În TransitXV puteți salva anumiți termeni şi prețuri pe care le-ați stabilit cu clienții, traducătorii şi cei care verifică traducerea şi le puteți utiliza pentru a efectua calcule la orice oră – în baza unor criterii definite individual şi în orice etapă a procesului. TransitXV poate genera rapoarte despre întregul proiect sau pentru subproiecte individuale în baza numărului de caractere, cuvinte, rânduri sau pagini sau despre prețuri. În plus, calculele dvs. se pot transfera automat către prețurile stabilite şi formularele de comandă. Se poate aplica şi un factor de extindere pentru limba țintă. În acest fel dvs. puteți vedea de la început şi la fiecare etapă intermediară a procesului traducerii care dintre proiecte sunt într-adevăr profitabile.

Cristina Zaharia
Director Executiv

Star Information Services & Tools
"Your Partner in Translations & DTP"

tel: +4 0729 883632
fax: +4 0232 212089
web: www.star-romania.com



LA TRADUCTION ASSISTÉE PAR ORDINATEUR EST-ELLE ANNONCIATRICE D'UNE MUTATION DU TRAVAIL DU TRADUCTEUR ?

RÉSUMÉ

Marie-Josée de Saint Robert

La traduction assistée fait l'objet de nombreuses controverses auprès des traducteurs et des gestionnaires de la traduction. Les mutations que laisse entrevoir le recours aux mémoires de traduction ou aux corpus de textes sont au centre de cette controverse. Les traducteurs à qui sont présentés des textes ou des parties de textes déjà traduits seraient tentés de les accepter sans possibilité de recherche ni d'amélioration tant serait grande la tentation des donneurs d'ouvrage de demander des délais de livraison de plus en plus courts et pour ce qui est des gestionnaires de la traduction de participer à une course de plus en plus effrénée à l'efficacité.

L'objet de cette communication est de montrer que les outils actuellement sur le marché de la traduction assistée par ordinateur ne sont pas identiques et que les résultats obtenus à ce jour ne permettent pas systématiquement de légitimer les craintes des traducteurs et des gestionnaires de la traduction. En effet, les outils d'alignement des phrases et de leur traduction hors contexte (les mémoires de traduction), d'une part, et les outils d'alignement de collections de textes intégraux en deux ou plusieurs langues (les corpus de textes), d'autre part, permettent de mesurer la part de créativité qui revient au traducteur au moyen de ces outils. Diverses opérations seront étudiées aux fins de comparaison de ces deux types d'outils : la validation des paires de langue ; la sélection de la traduction la plus adéquate en contexte ; le recours au « langage contrôlé » utilisé par les auteurs ; le contrôle de la qualité et de la cohérence des textes ; la transmission des pratiques de traduction.

Marie-Josée de Saint Robert
JIAMCATT
mjdesaintrobert@unog.ch
http://dcln042.un.org/QuickPlace/jiamcatt/Main.nsf/h_Toc/126415a5f0034b8485257274002d3ad1/?OpenDocument



Un Survol des Logiciels Trados, SDLX, Déjà Vu et WordFast de la perspective du traducteur freelan

POPA Adina, Traductrice Freelance

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Résumé

Le monde des outils d'aide à la traduction connaît une évolution extrêmement dynamique. Dans cette bataille commerciale, il est difficile pour le traducteur roumain freelance de se faire une idée des systèmes disponibles, de réussir les comparer avant de les acquérir.

La démarche du présent travail porte du générique vers le spécifique; commence par définir une mémoire de traduction, d'en expliquer le fonctionnement, les avantages, les inconvénients, les difficultés rencontrées dans la pratique. Vu le nombre considérable de logiciels TAO, notre recherche porte sur quatre logiciels de traduction assistée par ordinateur : Déjà Vu d'Atril Software, WordFast de Champollion WordFast, Trados et SDLX de SDL International.

Il y a une interdépendance entre les aptitudes à traduire et la maîtrise des outils techniques. Pour nous, les traducteurs, la partie technique peut représenter un piège dans l'activité de traduction.

Nous prendrons en compte tous les aspects : pragmatique (rapport prix / qualité, etc.), techniques (le type de logiciel, les formats de fichier traités, compatibilité avec d'autres logiciels, extensions des fichiers, le formatage des documents, etc.), professionnelles (les environnements de traduction proposés, la segmentation, la phase de prétraduction, les bases terminologiques, etc.) et de gestion (le décompte, etc.).

Nous avons choisi ce corpus, car nous utilisons les quatre logiciels dans notre travail journalier. Le corpus trilingue (français, anglais et roumain) extrait de notre activité de traducteur freelance représente une source d'exemples pour les observations et conclusions présentées.

Les aptitudes traductionnelles, donc les aptitudes à la recherche documentaire et terminologique, sont indissociables de la maîtrise des nouvelles technologies. Le traducteur roumain doit s'adapter et commencer d'utiliser la TAO, sans les ignorer ou les rejeter.

Nous pouvons parler d'une véritable mutation du métier de traducteur.

Introduction

Aujourd'hui, traduire, c'est non seulement trouver le mot juste, mais également utiliser le bon outil informatique. Il n'est pourtant guère éloigné de la réalité: tous les traducteurs sont susceptibles d'être concernés, tôt ou tard, par ces nouveaux développements. En ce XXI-ème siècle, la traduction est elle aussi devenue un art numérique. L'informatique a indubitablement beaucoup apporté aux traducteurs.

Dans cette bataille commerciale, il est difficile pour le traducteur freelance de se faire une idée des systèmes disponibles, de réussir les comparer avant d'acheter. Pourtant, un nome revient souvent sur les bouches des traducteurs, car il est utilisé par des institutions importantes [1] … nul n'ignore le produit Trados. Alors une première référence importante et nous sommes déjà influencés.

Un logiciel de traduction assistée par ordinateur (TAO) [2] est un corpus comportant des textes sources et leur traduction, segmentés pour que les segments, qui correspondent généralement au niveau de la phrase dans les textes sources, soient

mis en correspondance [3]. Il s'agit donc d'un corpus parallèle, créé à partir de traductions antérieures. De tels corpus permettent au traducteur de «recycler» la traduction [4]. L'intention n'est pas celle de remplacer le traducteur mais celle de l'aider. L'intérêt donc pour le traducteur, aussi bien que pour le donneur d'ouvrage, est de pouvoir réutiliser des traductions antérieures en gagnant du temps, de la cohérence et de la profitabilité (pour le donneur d'ouvrage).

Avec une mémoire de traduction remplie, la TAO permet donc un gain de temps, une comparaison avec les travaux précédents, le respect de la terminologie du client, l'accélération des recherches de références déjà traduites pour un même client, etc. Le traducteur est également certain de ne pas «oublier» de traduire un segment, car le logiciel ne passe au segment suivant qu'une fois que le premier a été validé.

Les logiciels TAO se révèlent très utiles pour la traduction de textes commerciaux, de documents légaux, scientifiques et surtout techniques. La mémoire de traduction n'est pas appropriée pour les travaux de traduction en littérature et autres textes créatifs.

La TAO demande d'abord un investissement en temps. Il faut apprendre les modes d'emploi du logiciel (qui peuvent avoir des centaines de pages) puis se familiariser avec son utilisation. L'inconvénient majeur au départ est donc qu'il faut forcément commencer pas un gros effort de remplissage du système avec des traductions déjà effectuées, qui serviront ensuite de références pour les traductions suivantes.

Il ne faut pas confondre logiciels de traduction automatique (TA) [5] et logiciels de traduction assistée par ordinateur (TAO). Cette ambiguïté est souvent exploitée par les sociétés qui commercialisent des logiciels de TA.

Le survol du traducteur roumain freelance

En consultant l'enquête réalisait par Imperial College London en 2006 [6] on voit que les logiciels utilisés par nous se situe sur les premiers quatre places (Figure 1 et 2).

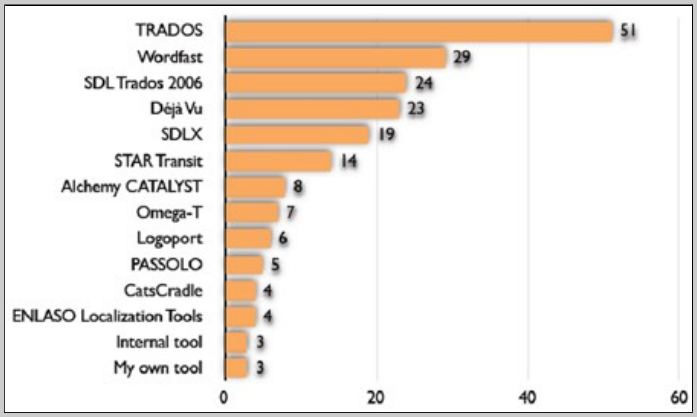


Figure 1. Les logiciels TAO les plus utilisés
Les chiffres sont des pourcentages

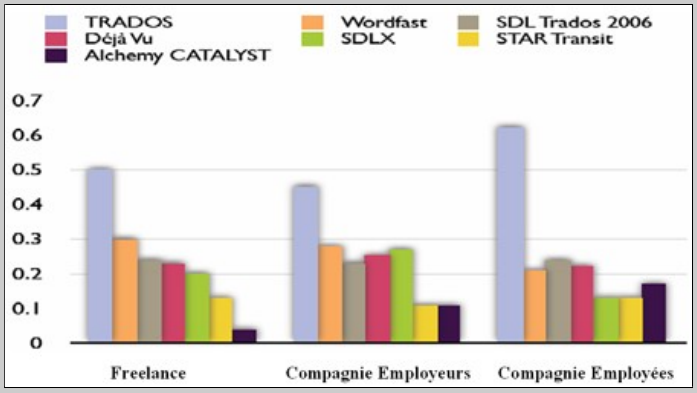


Figure 2. Les logiciels TAO et leur utilisateur

Trados, SDLX et Déjà Vu sont des logiciels informatiques puissants, robustes et technologiquement complexes conçus par des entreprises spécialisées. WordFast apparaît comme un logiciel artisanal, car il est conçu par un traducteur pour les traducteurs.

La première question qu'on pose est : Combien ça coûte ?

Les prix affichés par les trois compagnies (SDL International, Atril Software et Champollion WordFast) pour les quatre logiciels sont le reflet exact de cette différence de perspective. WordFast est un logiciel payant depuis 2002. Son prix est de € 250, mais pour des pays comme la Roumanie le prix descend à € 125. Une mise à jour du logiciel est gratuite pour une durée de 3 années. Le prix pour la version standard de Déjà Vu X est de € 490, la version professionnelle coûte € 990 et une licence workgroup pour le premier poste de traduction coûte € 2250. Une mise à jour coûte € 490. SDLX est vendu depuis cette année avec SDL Trados. La licence pour SDL Trados 2007 Freelance est de € 795, mais la plateforme Trados englobe aussi SDLPhraseFinder Freelancer Edition & MultiTerm 7 Extract Freelance ToolProof v2.02 (D) € 595 et des autres logiciels. Les mises à jour doivent être payées aussi (prix actuel pour décembre 2007). Pour mieux comprendre les différences de prix on peut consulter la Figure 3.

TRADOS 2007	WORDFAST	SDLX	DÉJÀ VU X
€ 755 licence pour SDL Trados 2007 Freelance La plateforme Trados (SDLPhraseFinder Freelancer Edition & MultiTerm 7 Extract Freelance ToolProof v2.02 (D) € 595 et autres.	- € 250, - € 125 pour la Roumanie	Maintenant se vend avec Trados 2007	€ 490 version standard € 990 version professionnelle € 2250 une licence workgroup pour le premier poste de traduction
Les mises à jour doivent être payées	- La mise à jour: gratuite pour 3 années	- Une mise à jour: €196 du SDLX 2005	Une mise à jour: € 490

Figure 3. Le prix des logiciels TAO

WordFast est un macro [7] qui s'installe dans MS Word. Il peut être installé aussi sur Linux et Mac, en temps que les autres logiciels non. Les trois autres logiciels sont grands et demandent un minimum de connaissances informatique pour les installer.

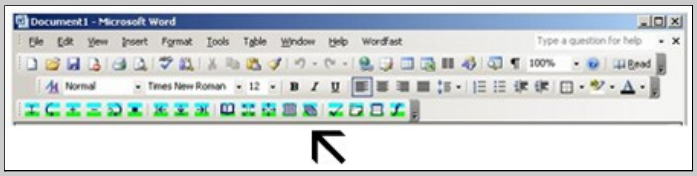
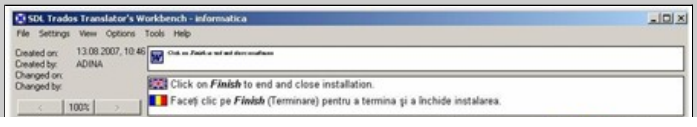


Figure 4. Une image avec les icônes de commande de WordFast.

Trados 6,5 intègre les fonctions essentielles réclamées par les utilisateurs pour accélérer la création de projet, la planification, la traduction et la relecture. Trados 6,5 [8] est une plate-forme constituée de six composants ou modules clés : Translator's Workbench, Trados MultiTerm, WinAling, TagEditor, T-Window for Clipboard et Filters. Maintenant SDL Trados 2007 englobe SDL Translator's Workbench, SDL MultiTerm, SDLX 2007, SDL PerfectMatch et SDL Trados Synergy 2007 Client. Donc, Trados est un outil d'aide à la traduction comprenant une mémoire de traduction, un outil d'alignement, un gestionnaire et un extracteur de terminologie, des outils de conversion de formats de fichiers et un système de gestion de projets plus particulièrement destiné à la localisation. WordFast, Déjà Vu et SDLX sont un seul outil (all-in-one) d'aide à la traduction.

Trados vous propose plusieurs environnements de traduction : Translator's Workbench (TWB), SDLX Edit et Trados Tag Editor. Déjà Vu X vous propose un seul environnement propre et WordFast travaille avec une seule fenêtre ouverte dans MS Word. L'un des désavantagés de Trados est le fait d'être contraint d'avoir deux applications ouvertes: dans la fenêtre de Word (voir Figure 5) ou du Trados Tag Editor (voir Figure 6) : il y a le texte source que l'on veut traduire et dans celle du Translator's Workbench il y a la mémoire de traduction (MT) créée auparavant où on cherche les phrases ou les segments que l'on veut traduire pour voir s'ils sont présents quelque part dans la mémoire. Il y a donc une interaction entre les deux fenêtres qui est impérative pour la correcte utilisation de la MT.



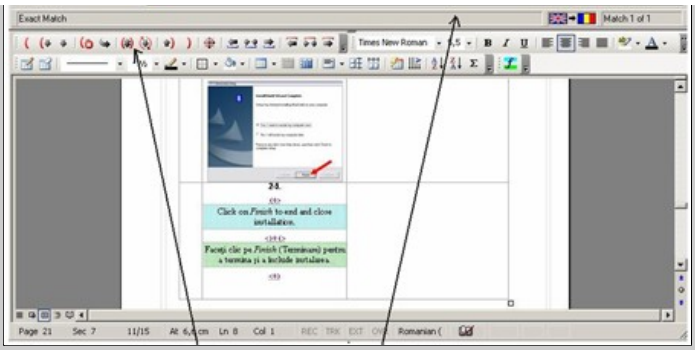


Figure 5. La fenêtre Word et la fenêtre Translator's Workbench ouvertes en même temps

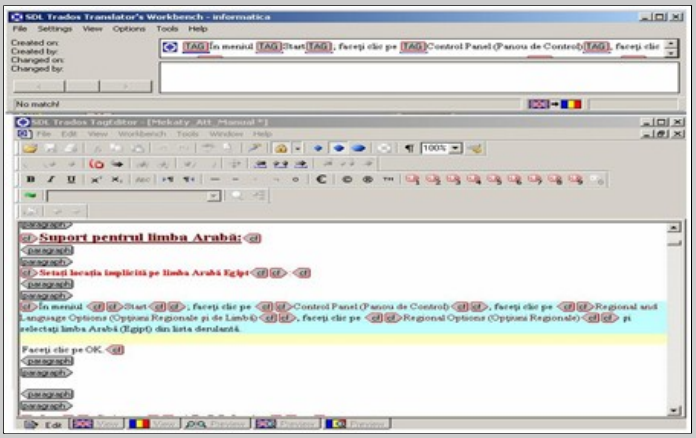


Figure 6. La fenêtre Trados Tag Editor et la fenêtre Translator's Workbench ouvertes en même temps

Les fabricants d'outils de mémoire de traduction utilisent le terme «segments» pour désigner les chaînes de texte stockées dans leur base de données. Au moment de la création de la base de données de MT initiale, tous les logiciels recourent aux mêmes algorithmes bien connus, utilisés aux fins de segmentation des documents de référence. Ces mêmes algorithmes servent également à la segmentation des nouveaux projets de traduction, aux fins de comparaison avec les phrases contenues dans la base de données de MT.

En utilisant WordFast ou Trados le formatage de documents (code de soulignement, caractère gras, etc.) est conservé dans le document traduit. Il ne traduit hors contexte comme dans le cas de Déjà Vu ou SDLX.

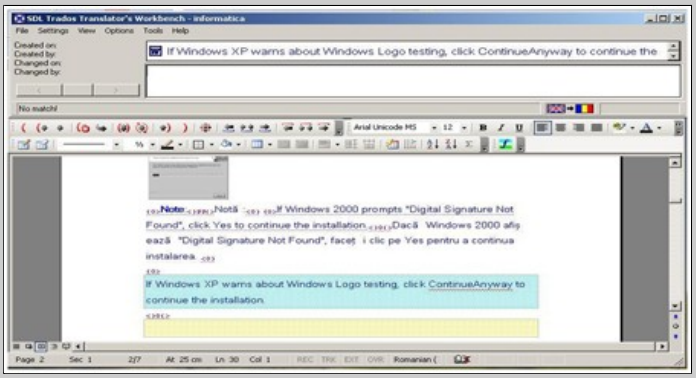


Figure 7. Traduction dans le contexte

Déjà Vu X et SDLX présentent la segmentation texte source – texte cible en miroir (Figure 8), en temps que les deux autres enchaînent les segments (à traduire et traduits) sur des fonds colorés. WordFast, aussi comme Trados, enverra les segments au logiciel de traduction au fur et à mesure des besoins. Déjà Vu et SDLX traitent en segment tous le documents avant de commencer la traduction. Trados ne supporte pas des segments vides tandis que WordFast supporte.



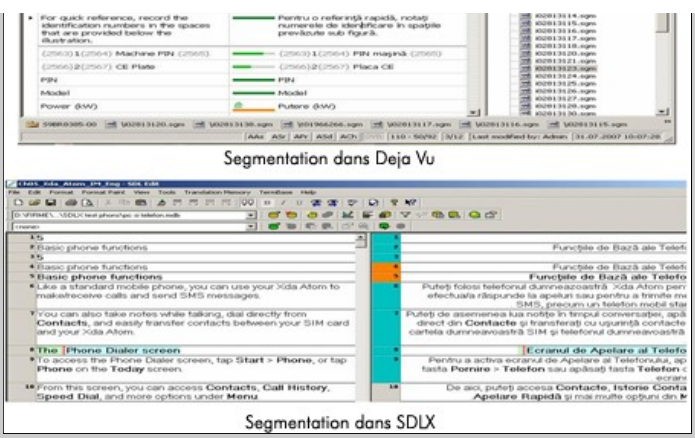


Figure 8. Segmentation en miroir

WordFast délimite les segments avec les mêmes délimiteurs de segments comme Trados. WordFast est compatible, au niveau du document traduit (segmenté, ou bilingue) avec Trados: un document traduit sous WordFast peut être nettoyé par Trados, et vice-versa.



Figure 8. Segmentation dans WortdFast et Trados

Les cellules correspondant aux différentes phrases peuvent être supprimées, fusionnées ou scindées. On part normalement du principe que la séquence des phrases source et cible est parallèle

La correspondance est marquée avec une couleur distincte. Dans le logiciel Trados Translator's Workbench (TWB) la correspondante est affichée et le traducteur peut la choisir ou non (voir Figure 5).

Pendant la phase de prétraduction le logiciel Déjà Vu X et SDLX proposent directement un résultat, la tâche réservée à l'utilisateur est la vérification d'une bonne traduction du point de vue contextuel. Dans le cas de SDLX une cellule apparaissant en gris indique un résultat final, une ligne colore entièrement en vert nous présent le résultat dans Déjà Vu X.

La similarité entre la phrase source et la phrase de la MT est exprimée en pourcentage. Le pourcentage d'analogie est le numéro qui apparaît entre le segment source et le segment cible. Trados Translator's Workbench fonctionne plus par analogie que par équivalence parfaite : il trouve dans la MT une ou plusieurs phrases similaires à la phrase source. Le même principe est valable pour WordFast. Dans les cas de SDLX et de Déjà Vu X le pourcentage est mentionné sur la barre qui sépare les deux champs (Figure 9).

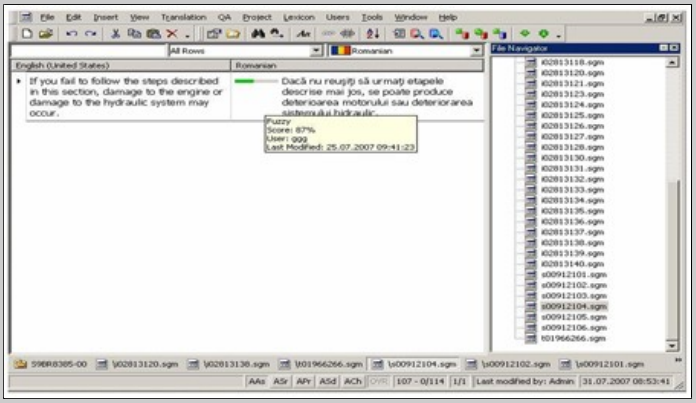


Figure 9. Affichage du pourcentage d'analogie dans Deja Vu

WordFast utilise tous les documents de format Excel, PowerPoint et Access, ainsi que les fichiers dits "taggés", c'est-à-dire préparés par RWS Rainbow ou Trados Stagger. WordFast peut être connecté à des logiciels de traduction automatique comme PowerTranslator ™, Systran™, Reverso™, etc.

Trados utilisez les formats de fichier les plus courants : traitez facilement tous les fichiers de projet, quels que soient les formats utilisés par vos clients : HTML, XML,

SGML, XLIFF, Interleaf/Quicksilver, Microsoft Word, Microsoft Excel, Microsoft PowerPoint, OpenOffice, StarOffice, Clipboard, Adobe PageMaker6.5, Adobe InDesign CS, Adobe FrameMaker 7, Adobe, ainsi que de nombreux nouveaux formats dont InDesign CS2, QuarkXPress 6.x et les fichiers texte délimités par une virgule ou une tabulation.

WordFast est compatible avec la quasi-totalité des environnements de traduction existants (Trados, IBM Translation Manager, SDLX, Déjà Vu, Transit), soit à travers la passerelle TMX, ou directement avec le format d'origine. On peut importer des fichiers Trados dans le Déjà Vu. Les fichiers WordFast (.txt) peuvent être importer en DVX en changent l'extension .TMX. WordFast lit directement TXT, XLS, MultiTerm, etc.

Pour reconnaître un document après son extension on doit savoir que l'extension des fichiers SDLX est « itd » et l'icône est un crayon. L'extension des fichiers Déjà Vu X est « sgm » et l'icône est un dossier sur lequel sont des drapeaux. L'extension du TWB est « rtf ».

Pour l'heure, seule la version anglaise de la documentation de Déjà Vu X existe, y compris dans l'aide en ligne. On a pu trouvé la documentation en français pour WordFast, Trados et pour SDLX.

WordFast ne compte pas les mots exactement comme le fait la fonction « Outils/Statistiques » de MS Word ou la fonction du Trados. En moyenne, le décompte de WordFast est de 5 à 15% supérieur à celui de MS Word, et de 5 % supérieur à celui de Trados selon la langue. Nous vous recommandons d'aborder le problème du décompte des mots avec votre client avant de commencer le travail ; n'oubliez pas non plus de préciser si le décompte sera fait sur le texte source ou cible.

Tous les quatre logiciels ont une base terminologique: Déjà Vu X a Lexicon, Trados a Multiterm, WordFast a +Tools et SDLX a TermBase.

Il y a une interdépendance entre les aptitudes à traduire et la maîtrise des outils techniques. Pour nous, les traducteurs, la partie technique peut représenter un piège dans l'activité de traduction. On trouve que Déjà Vu X demande un plus d'attention en ce qui concerne les codes et les codes incorporés. Il y a des règles qui doivent être respectées, comme : les codes ne peuvent pas être laissés à la fin des phrases, virgule ; l'icône avec le X rouge dans une ligne indique que pas tous les codes dans le segment source sont dans la section cible, ou vous avez trop de codes dans la fenêtre cible. Vous devez payer attention que les codes soient dans leur ordre et pour vérifier ça vous avez à votre disposition des combinaisons des boutons. Et à partir de ces règles on a des exceptions qui empêchent le traducteur de travailler normalement.

Les compagnies offrent la possibilité de suivre un cours, soit on-line, soit dans les centres spécialisés. Nous avons reçu, par exemple Déjà Vu ou SDLX, en même temps avec le projet à traduire. Alors on a eu d'apprendre utiliser le logiciel et de traduire en même temps.

La mémoire de traduction réduit le coût des traductions : le client paye à un tarif très préférentiel dû à la répétition des segments dans la traduction. Le logiciel de traduction peut automatiquement calculer le nombre de répétitions dans le document source ce qui permet une réduction précise du coût du projet.

Le problème du tarif pour les « 100% matches », les correspondances exactes, alimente le débat entre les traducteurs et les donneurs d'ouvrages. Les correspondances exactes ne sont en fait que de simples propositions faites par une machine aveugle. Les correspondances exactes (et même les répétitions) doivent par conséquent être soigneusement contrôlées – un facteur à ne pas sous-évaluer lors de la négociation du tarif.

La mémoire de traduction assure l'utilisation harmonieuse de la terminologie d'un texte ou d'un projet qui requiert des mises à jours régulières.

Des bases de données terminologiques, ainsi que des mémoires de traduction, peuvent être créées de façon spécifique pour vous, pour votre société tout en demeurant confidentielles.

Pour de projets plus amples en traduction, une équipe de traducteurs peut partager des bases de données qui sont mises à jour en temps réel. Cela permet des prestations plus rapides tout en respectant la cohérence et l'harmonie du texte.

Les clients peuvent fournir aux traducteurs leurs propres mémoires de traduction. Alternativement, ils peuvent fournir les traductions précédentes, avec leur texte source, qui peuvent être formatés en mémoire de traduction afin d'être utilisés dans des projets futurs. Si on parle des projets qui utilisent Déjà Vu ou SDLX et qui ont la prétraduction faite par le client, le dernier doit vous fournir la mémoire de traduction, le texte original et le document segmenté.

Nous, comme la grande majorité des utilisateurs, avons commencé à compléter la base au fur et a mesure du travail. Il y a d'autres traducteurs qui ont commencé par aligner de textes ou par importer de MT.

Un autre inconvénient, résulte du déchirement du texte en segments, qui mène parfois à un manque de cohérence, car la traduction s'effectue phrase par phrase.

Le défaut majeur du système est sans doute le fait qu'il ne reconnaisse ni les significations identiques de deux formes différentes ni les significations différentes de deux formes identiques. Le système ne fait qu'utiliser des tables de correspondance existantes, et rien d'autre. Il ne peut donc pas vérifier si le segment a le même sens dans un contexte différent, ni vérifier la cohérence du texte ou encore le style, la justesse de la traduction ou l'orthographe.

Conclusions

On peut conclure qu'une interdépendance s'instaure entre les aptitudes à traduire et la maîtrise des outils techniques. Les aptitudes à traduire, qui impliquent des aptitudes à la recherche documentaire et terminologique, sont en fait indissociables de la maîtrise des nouvelles technologies qui investissent l'activité. Disposer du matériel informatique adéquat est aujourd'hui un impératif pour tous les traducteurs (adresse e-mail, retranscription des textes effectués sur un fichier informatique, respect de la mise en page demandée, la terminologie et l'utilisation du logiciel donné ou demandé par le client, etc.). La maîtrise de l'outil informatique et des ressources qui lui sont liées devient une condition pour exercer le métier de traducteur, sans parler de la traduction des supports eux-mêmes. La maîtrise des outils techniques disponibles est donc une source d'amélioration, quantitative si ce n'est qualitative, des aptitudes à traduire. En fin de compte, le lien s'instaure à double sens car les aptitudes requises exigent peu à peu la maîtrise de nouveaux outils techniques et l'efficience qu'elle procure devient peu à peu la norme dans le métier. Le traducteur ne doit pas ignorer ces outils ni les rejeter, mais savoir s'y adapter lorsque cela est nécessaire.

Avec l'évolution constante et rapide des nouvelles technologies de l'information et de la communication, parions que notre étude ne constitue qu'un modeste rapport d'étape sur l'interaction des différents processus qui concourent à une véritable mutation du métier de traducteur.

Références

1. Arrouart, C, « Les mémoires de traduction et la formation universitaire : quelques pistes de réflexion », *Meta*, Volume 48, numéro 3, Septembre 2003;

2. ASTM International, *ASTM F2575-06 Standard Guide for Quality Assurance in Translation*, Book of Standards Volume: 15.11, http://www.astm.org/cgi-bin /SoftCart.exe /DATABASE.CART /REDLINE_PAGES/F2575.htm;

3. Bédard, C, « Translation memory seeks sentence translator… », Traduire, numéro 186, 2000;

4. Benis, M, « Review of Atril's Déjà Vu 2, The Happy Hoarder », *Translation Journal*, Volume 1, numéro 2, Janvier 1998;

5. Gerasimov, A, « An Effective and Inexpensive Translation Memory Tool », *Translation Journal*, Volume 5, numéro 3, Juliet 2001;

6. Gerasimov, A, « WordFast Windows version: Translation tool is user friendly », www.translation-services-usa.com;

7. Gow, Francie, *Metrics for Evaluating Translation Memory Software*. MA thesis. University of Ottawa, Canada, 2003;

8. Höge, Monika, *Towards a Framework for the Evaluation of Translators' Aids Systems*. PhD thesis. Helsinki University, Helsinki, 2002;

9. Falcone, S, « Translation Aid Software, Four Translation Memory Programs Reviewed », *Translation Journal*, Volume 1, numéro 2, Janvier 1998;

10. Gow, F, *Metrics for Evaluating Translation Memory Software*, Ottawa, ON Canada, 2003

11. Guadec, D, *Profession : traducteur - alias ingénieur en communication multilingue,* Maison du Dictionnaire, collection « Métiers des langues, langues des métiers », Paris, 2002 ;

12. Lagoudaki, Elina, *Translation Memory Systems: Enlightening users' perspective*, Imperial College Lonfon, 2006

13. Lebtahi, Y et Ibert, J, « Traducteurs dans la société de l'information : Évolutions et interdépendances », *Meta*, Volume 49, numéro 2, Juin 2004 ;

14. Outils d'aide à la traduction et cycle de travail, Direction Générale de la Traduction de la Commission Européenne Juillet 2005

15. Polguère, A, Bowker, « L. (2002): Computer-Aided Translation Technology: A Practical Introduction, Ottawa, University of Ottawa Press, Didactic of Translation Series, xx-185 p. », *Meta*, Volume 48, numéro 3, Septembre 2003

Site Web

12. www.atril.com

13. www.bridgeterm.com

14. www.fxm.ch

15. www.lisa.org

16. www.sdl.com

17. www.translationzone.com

18. www.wordfast.net

19. www.you-name-it.com/Englishpages/wordfast.htm

---

[1] La Direction Générale de Traduction de l'Union Européenne travail localement avec Translator's Workbench de TRADOS et/ou Word. Le Translator's Workbench (TWB) de TRADOS est le traducticiel intégré qui a été sélectionné à la suite d'un appel d'offres interinstitutionnel et adapté aux besoins spécifiques des institutions européennes. (consulter : *Outils d'aide à la traduction et cycle de travail*, Direction Générale de la Traduction de la Commission Européenne Juillet 2005, page 10).

[2] En anglais Computer - Aided Translation : CAT.

[3] Voir pour plus d'informations le site de l'agence FXM, www.fxm.ch.

[4] Consulter *Livre blanc de MultiCorpora*, Août, 2005

[5] La traduction automatique (TA) est traduire entièrement un texte grâce à un ou plusieurs programmes informatiques, dictionnaires, etc., sans qu'un traducteur humain n'ait à intervenir.

[6] Elina Lagoudaki, *Translation Memory Systems: Enlightening users' perspective*, Imperial College Lonfon, 2006,

[7] Un macro est un petit fichier rapide à télécharger.

[8] Nous parlerons de Trados 6,5 car on l'utilise dans l'activité quotidienne.

Sponsor de aur / **Partenaire Or**

Sponsor de argint / **Partenaire Argent**

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / **Partenaire Bronze**

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latină/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



Enjeux et perspectives de la traduction automatique dans le domaine juridique

DINCA Daniela, Université de Craiova

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Résumé

Le langage juridique est l'une** des langues de spécialité les plus complexes, singularisée par un ensemble de traits qui tiennent autant à **l'existence d'**un vocabulaire spécialisé qu'**aux** particularités de sa **structuration sur les plans stylistique, syntaxique et** sémantique. Sa **traduction d'**une langue à l'**autre est complexe elle aussi, car les** difficultés procèdent fondamentalement du caractère contraignant du **texte** juridique qui lui est attribué par la norme de droit. Traduire des **textes** juridiques signifie reconnaître les éléments juridiques et **linguistiques qui** ont façonné la norme de droit et les transposer dans **une autre langue et dans une autre culture.

*Dans le cadre esquissé, nous nous **proposons d'analyser les variantes de** traduction proposées par des logiciels de **traduction automatique en ligne (Reverso, Radins et Systran) pour des fragments du** Traité établissant une **Constitution pour l'Europe** **(anglais -** français) afin de les comparer avec **la variante officielle d'**un traducteur spécialisé.*

***Cette analyse nous permettra, d'une part, d'établir les différentes utilisations possibles d'un logiciel de traduction automatique dans le domaine juridique et, d'**autre part, de mettre en évidence les problèmes **de terminologie** et de syntaxe posés par ce type de traduction.*

1. Introduction

La traduction automatique désigne, au sens strict, le fait de traduire entièrement un texte grâce à un ou plusieurs programmes informatiques, sans l'intervention du traducteur humain. Avec un outil informatique performant, le lecteur dispose ainsi de la possibilité de transposer d'une langue à l'autre des fragments qui l'intéressent ou même des pages WEB.

L'objectif premier de la traduction automatique est donc de comprendre et/ou de se faire comprendre. En conséquence, un critère efficace est l'opinion du lecteur : « Le document traduit est-il exploitable pour le besoin que j'en ai et retranscrit-il le contenu du texte de départ ? »

Dans le cadre de la formation des traducteurs, l'enseignant est intéressé de voir dans quelle mesure les outils informatiques sont capables de remplacer le travail humain et surtout de comparer leur efficacité dans le domaine de la traduction spécialisée. Dans cette perspective, nous nous proposons d'analyser la pertinence des logiciels de traduction automatique dans le domaine de la traduction juridique.

Nous avons conçu notre travail en trois étapes :

1. la traduction automatique de la version anglaise d'un fragment du *Traité établissant une Constitution pour l'Europe* (Titre I) par trois logiciels de traduction distribués gratuitement sur Internet: Reverso (REV), Radins (RAD) et Systran (SYS) ;

2. la comparaison des trois versions de traduction automatique avec une version papier, officielle, réalisée par une équipe de traducteurs.

3. l'évaluation des fonctionnalités et des performances des trois systèmes de traduction automatique par rapport au traducteur humain.

Comme les paramètres pris en compte dans l'évaluation d'un système de traduction automatique sont la qualité, le temps et le coût, nous limitons notre analyse à la qualité de la traduction, qui est d'ailleurs le facteur décisif dans l'acquisition des logiciels de traduction.

Quand on pense à la qualité d'une traduction, le premier aspect qui vient à l'esprit est sa fidélité au texte de départ. Suivent après sa clarté ou son intelligibilité et, finalement, son style. Bien que les jugements portant sur ces aspects comportent inévitablement un côté subjectif, nous essayerons de formuler quelques hypothèses pertinentes concernant la qualité de la traduction automatique par rapport à la traduction humaine à deux niveaux : l'adéquation du vocabulaire et la couverture des structures syntaxiques.

2. Problèmes de terminologie

Le vocabulaire juridique comprend diverses catégories de termes, des termes spécialisés d'une technicité élevée mais aussi des termes qui proviennent de la langue courante, ayant un sens particulier dans un contexte juridique.

Dans la traduction des termes juridiques, le traducteur doit assurer leur passage d'une langue à l'autre, en tenant compte du système juridique en place parce que les termes du droit sont le plus souvent culturellement marqués, ce qui signifie que, derrière chaque terme, il y a toute une culture juridique en place. Comme source de documentation, il se sert des outils de plus en plus performants : lexiques de terminologie comparée, dictionnaires bilingues spécialisés, ouvrages généraux en droit, etc.

Si l'on compare les deux versions françaises (automatique et officielle), on peut se rendre compte que la machine préfère les structures parallèles, sans tenir compte des traits contextuels des mots, des restrictions sélectives concernant le déterminant sélecté ou même de la spécificité des termes dans la langue d'arrivée.

Au niveau du groupe verbal, dans le tableau ci-dessous, pour les verbes de la langue courante, les versions informatisées (1-3) proposent des solutions aléatoires, calquées sur le modèle anglais, par rapport à la version écrite, qui inclut des verbes plus appropriés au contexte où ils sont employés :

Version anglaise	Version française automatique	Version française officielle
(1) *build* a common future	REV : *construire* un avenir commun RAD : *construire* un avenir commun SYS : *établir* un futur commun	*bâtir* leur avenir commun
(2) the Member States *confer* competences	REV : les Etats membres *confèrent* des compétences RAD : les Etats membres *confèrent* des compétences SYS : les Etats membres *confèrent* des compétences	les États membres *attribuent* des compétences
(3) to *achieve* these objectives	REV : *réaliser* ces objectifs RAD : *accomplir* ces objectifs SYS : *atteindre* ces objectifs	*atteindre* ces objectifs
(4) tasks which *flow from* the Constitution	REV : tâches qui *coulent* de la Constitution. RA : tâches de réalisation qui *coulent* de la Constitution. SYS : tâches de mise en oeuvre qui *découlent* de la constitution.	missions *découlant* de la Constitution
(5) The Union *shall work for* the sustainable development of Europe (...)	REV : L'Union *travaillera* (marchera) pour le développement durable de l'Europe (...) RAD : L'Union *travaillera* pour le développement durable de l'Europe (...) SYS : L'union *fonctionnera* pour le développement durable de l'Europe (...)	L'Union *œuvre* pour le développement durable de l'Europe (...)

Parfois, la traduction directe peut conduire à la déformation du texte d'arrivée. Par exemple, dans (4), deux traducteurs automatiques proposent une version incompréhensible: « tâches qui *coulent* de la Constitution » (REV, RAD), SYS proposant pourtant une version rapprochée de la version humaine : « tâches de mise en oeuvre qui *découlent* de la constitution » (SYS) par rapport à « missions *découlant* de la Constitution ».

On peut également signaler le glissement de sens dans le couple *tâche* et *mission*, le premier terme signifiant « travail défini et limité, imposé par autrui ou par soi-même, à exécuter dans certaines conditions » par rapport au deuxième qui renvoie à « tâche confiée à une personne ou à un organisme ». Une fois de plus, le traducteur automatique ne peut pas sélecter le terme dans la langue d'arrivée sans se rapporter au texte de départ.

La créativité lexicale est exclue de la version automatique, qui en souffre beaucoup. Un tel exemple nous est fourni dans (5) par le verbe *work,* traduit par *travailler* dans les trois versions automatiques, mais qui a un correspondant différent dans la version écrite, où le verbe *œuvrer* ajoute comme sèmes : « agir au service d'une cause, travailler à une oeuvre considérée comme ayant de la dignité, de l'importance, de la noblesse ». (TLFi)

La même équivalence formelle entre les termes anglais et français se retrouve dans la classe du nom, aussi bien pour les noms de la langue commune que pour ceux de la langue spécialisée :

Version anglaise	Version française automatique	Version française officielle
(6) scientific and technological *advance*	REV : *l'avance* scientifique et technologique RAD : *l'avance* scientifique et technologique SYS : le *progrès* scientifique et technologique	le *progrès* scientifique et technique
(7) the *free movement* ofpersons, services, goods and capital	REV : le *mouvement libre* (gratuit) de personnes, des services, des marchandises et le capital (la capitale) RAD : Le *mouvement libre* de personnes, services, marchandises et capital. SYS : La *libre circulation* des personnes, des services, des marchandises et de capital	la *libre circulation* des personnes, des services, des marchandises et des capitaux
(8) the rule of *law*	REV : l'autorité de la *loi* RAD : la règle de *loi* SYSS : la règle du *droit*	l'État de *droit*
(9) International *law*	REV : *loi* internationale RA : *loi* internationale SYS : *loi* internationale	*droit* international
(10) Union *law*	REV : La *loi* d'union RAD : *Loi* d'union SYS : *Loi* des syndicats	Le *droit* de l'Union
(11) the *Law* of the Member States	REV : la *loi* des Etats membres. RAD : la *loi* des Etats membres SYS : la *loi* des Etats membres	le *droit* des États membres

Dans (6), pour « scientific and technological *advance* », deux versions automatiques emploient le terme : « *l'avance* scientifique et technologique » (REV, RAD) face à « le *progrès* scientifique et technologique » (SYS), version qui se superpose à celle officielle. Si l'on consulte le TLFi, pour le mot *avance*, on trouve la définition suivante : « action d'avancer ou son résultat; le fait d'être en avant de quelqu'un ou de quelque chose », par rapport à *progrès*, défini dans le même dictionnaire comme : « processus évolutif orienté vers un terme idéal. », ayant comme synonymes: « *amélioration, avancement, développement, perfectionnement* ». (TLFi)

Dans (7), le syntagme « free *movement* » est rendu par une forme inusitée dans « le *mouvement* libre (gratuit) » par rapport à la forme très usuelle « la libre *circulation* », compte tenu de ses traits contextuels qui incluent « le mouvement entre plusieurs

personnes ou entre plusieurs pays, ensemble des échanges, des transactions ». (TLFi)

Une autre preuve de non concordance lexicale nous est fournie dans les exemples de (8) à (11) par un nom très répandu dans le domaine juridique. Il s'agit du terme *law*, que le traducteur automatique transpose toujours par *loi*, sans tenir compte des contextes où ce mot apparaît et qui représentent des syntagmes figés : l'État de *droit, droit* international, le *droit* de l'Union, le *droit* des États membres.

Si l'on regarde la définition lexicographique de ces deux lexèmes, on a tout de suite l'explication : *law* enregistre, dans les dictionnaires anglais, comme première acception "the set of customs, statutes, or rules that govern a society » et ensuite « a rule, principle, or procedure, that demands formal or informal obedience » (OED), tandis que le français dispose de deux termes différents pour exprimer ces deux acceptions : *droit* défini comme « fondement des règles régissant les rapports des hommes en société, et impliquant une répartition équitable des biens, des prérogatives et des libertés » et *loi* qui renvoie à la « règle, prescription émanant de l'autorité souveraine dans une société donnée et entraînant pour tous les individus l'obligation de s'y soumettre sous peine de sanctions ». (TLFi)

Dans ce cas, on a de nouveau la confirmation que les mots polysémiques posent des problèmes au système automatique de traduction et que le contexte ne constitue pas une solution pour le choix du terme approprié. Le problème de la polysémie du langage juridique constitue d'ailleurs l'un des aspects les plus difficiles à surmonter, même pour les traducteurs professionnels.

Dans le domaine des déterminants du nom, plus précisément dans celui des adjectifs qualificatifs, les problèmes de transposition visent le choix du terme convenable et la place des constituants dans le groupe nominal :

Version anglaise	Version française automatique	Version française officielle
(12) *Internal* frontiers/market	REV : frontières *internes* RAD : *intérieur* les frontières SYS : *interne* frontières	frontières / marché *internes*
(13) *specific* provisions	REV : dispositions *spécifiques* RAD : provisions *spécifiques* SYS : dispositions *spécifiques*	dispositions *particulières*
(14) *legal* personality	REV : personnalité *légale* RAD : personnalité *juridique* SYS : personnalité *juridique*	personnalité *juridique*
(15) the *wider* world	REV : le monde *plus large* RAD : *le plus large* monde SYS : *le plus large* monde	le *reste* du monde
(16) Its *rich* cultural and linguistic diversity	REV : sa diversité culturelle et linguistique *riche* RAD : sa diversité culturelle et linguistique *riche* SYS : sa diversité culturelle et linguistique *riche*	la *richesse* de sa diversité culturelle et linguistique

Dans les exemples de (12) à (14), la traduction des adjectifs ne conduit pas forcément à la déformation du sens juridique. Entre *intérieur et interne, spécifique et particulier, légal et juridique*, il n'y a pas une trop grande différence.

Pour le spécialiste pourtant, *spécifique* est « qui s'attache à une espèce, qui la caractérise à l'exclusion de toute autre » (TLFi) par opposition à *générique,* tandis que *particulier* porte sur une disposition « qui ne concerne qu'une personne ou qu'une chose (ou un petit groupe de personnes ou de choses) » (TLFi) , c'est-à-dire qui s'applique aux membres de l'Union Européenne.

Dans (14), même si les deux adjectifs *légal* et *juridique* ont la même définition lexicographique, le syntagme figé *personnalité juridique* exclut celui de *personnalité légale*.

Un effet de moindre importance est obtenu par deux adjectifs de la langue courante dont la transposition pose des problèmes au traducteur automatique. Par exemple, dans (15), l'adjectif *wider* ne peut avoir comme correspondant que la version *plus large* qui, associé au nom centre *monde*, conduit à une ambiguïté qui ne peut pas trouver sa place dans le discours juridique : « le monde *plus large* » par rapport à « *le reste* du monde ».

Le deuxième adjectif est *rich* pour la traduction duquel le traducteur automatique

opère un changement de place, en le rejetant à la fin du syntagme, vu la présence de plusieurs déterminants pour le même nom. Le traducteur professionnel trouve une solution dont il se sert couramment en passant d'une langue à l'autre, la transposition, par laquelle l'adjectif devient le centre du syntagme nominal: « la *richesse* de sa diversité culturelle et linguistique ».

Pour conclure, on pourrait dire que, même si le traducteur automatique est programmé à traduire des textes dans des domaines spécifiques, ce qui veut dire que le vocabulaire approprié a été déjà inclu dans les dictionnaires du système, la traduction de la terminologie ne s'avère pas du tout si simple. Les problèmes relèvent des traits contextuels de chaque terme ou du contexte immédiat ou plus étendu où il figure et qui ne peuvent pas être programmés en totalité.

3. Aspects de la structuration syntaxique

Si le vocabulaire, selon les appréciations quasi unanimes des jurilinguistes, constitue le premier obstacle à la communication juridique, le second est *l'énonciation*, définie comme un « ensemble de marques formelles qui caractérisent linguistiquement l'émetteur et ses énoncés » (Sourioux&Lerat, 1975: 12).

En effet, les caractéristiques linguistiques des textes normatifs ne s'expliquent que par le fait que ceux-ci ont pour vocation de régler les conduites dans le corps social et que les conditions très particulières de leur énonciation entraînent un certain nombre de spécificités textuelles, syntaxiques, stylistiques, etc.

3.1. Le groupe nominal

Le groupe nominal présente dans le discours juridique des particularités spécifiques de structuration ou d'enchaînement des éléments qui le composent. Parmi les déterminants à plus haute fréquence, nous avons relevé des déterminants prépositionnels et des subordonnées relatives.

3.1.1. Le déterminant prépositionnel

Le choix de la préposition introduisant le déterminant prépositionnel est une provocation pour tout traducteur, car il ne dispose pas de règles lui permettant de choisir correctement parmi les multiples possibilités qu'il a à sa disposition. Les prépositions doivent assurer la cohésion du groupe nominal et elles ne coïncident pas dans les deux langues qui disposent, chacune, de ses propres moyens d'enchaînement des éléments.

Dans le corpus analysé, nous avons identifié des situations où il incombe au traducteur de choisir une seule préposition dans la langue d'arrivée pour remplacer deux prépositions synonymiques ou quasi-synonymiques de la langue de départ ou même de traduire mot à mot une locution prépositionnelle, situation qui conduit à une version non intelligible pour le lecteur :

Version anglaise	Version française automatique	Version française officielle
(17) the will (…) *to* build a common future	REV : la volonté (…) *pour* construire un avenir commun RAD : la volonté (…) *pour* construire un avenir commun SYS : la volonté (…) *pour* établir un avenir commun	la volonté (…) *de* bâtir leur avenir commun
(18) respect *for* human dignity	REV : respect *pour* la dignité humaine RAD : respect *pour* la dignité humaine SYS : respect *pour* la dignité humaine	respect *de la* dignité humaine
(19) any discrimination *on grounds of* nationality	REV : n'importe quelle discrimination *sur les raisons (terres)* de nationalité RAD : n'importe quelle discrimination *sur les terres de* nationalité SYS : n'importe quelle discrimination *pour des raisons*	toute discrimination *exercée en raison de* la nationalité

	de nationalité	

Dans (17) et (18), les prépositions anglaises *to* et *for* sont traduites en français par *pour*, tandis que le traducteur humain choisit, pour chaque cas, une autre préposition : dans (17) *de* à valeur de caractérisation et, dans (18), une structure génitivale à plus forte valeur particularisante.

L'exemple (19) pose énormément de problèmes aux traducteurs automatiques de sorte que deux des solutions proposées par REV et RAD sont presque aberrantes face à la solution humaine, où la locution prépositionnelle est même précédée d'un adjectif explicatif, capable de bien préciser le contexte de l'application de la discrimination : « toute discrimination *exercée en raison de* la nationalité ».

Pour transposer les structures du génitif qui sont rendues dans les deux langues par des moyens différents, le traducteur automatique a tendance à procéder à un remplacement machinal de la préposition anglaise *of* par *de*, sans se soucier si cela garde ou non la même valeur possessive dans la langue d'arrivée:

Version anglaise	Version française automatique	Version française officielle
(20) the values *of* respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights	REV : les valeurs *de* respect pour la dignité humaine, *la* liberté, *la* démocratie, l'égalité, l'autorité de la loi et le respect pour des droits de l'homme RAD : les valeurs *de* respect pour la dignité humaine, *la* liberté, *la* démocratie, l'égalité, la règle de loi et *de* respect pour les droits de l'homme SYS : les valeurs *du* respect pour la dignité humaine, *la* liberté, *la* démocratie, l'égalité, la règle du droit et le respect des Droits du Homme	les valeurs *de* respect de la dignité humaine, *de* liberté, *de* démocratie, *d'*égalité, de l'État de droit, ainsi que *de* respect des droits de l'homme
(21) an area *of* freedom, security and justice	REV : un secteur *de* liberté, la sécurité (le titre) et la justice RAD : une région *de* liberté, sécurité et justice SYS : secteur *de la* liberté, *de la* sécurité et *de la* justice	un espace *de* liberté, *de* sécurité et *de* justice
(22) eradication *of* poverty and the protection *of* human rights	REV : l'extermination *de* pauvreté et la protection *de* droits de l'homme RAD : l'éradication *de* pauvreté et *de la* protection de droits de l'homme SYS : l'éradication *de* pauvreté et *de la* protection de droits de l'homme	l'élimination *de la* pauvreté et la protection *des* droits de l'homme
(23) States *of* Europe	REV : Etats *de* l'Europe RAD : Etats *de* l'Europe SYS : Etats *de* l'Europe	États *d*'Europe
(24) Europe's cultural heritage	REV : l'héritage culturel *de* l'Europe RAD : l'héritage culturel *de* l'Europe SYS : l'acquis culturel *de* l'Europe	patrimoine culturel *européen*

Dans (20) et (21), le génitif *of* est transposé par la préposition *de* à caractère plutôt caractérisant que déterminatifaussi bien par le traducteur automatique que par le traducteur humain, avec la remarque que, dans le cas de la transposition de plusieurs génitifs successifs, la version automatique ne respecte pas la cohésion du groupe nominal et introduit le deuxième déterminant par l'article défini. Si pour le traducteur automatique l'explication réside dans la difficulté de la contraction de l'article avec la préposition *de*, pour la version officielle, il s'agit d'une raison sémantique qui dicte le remplacement d'une structure de possession par une structure caractérisante à plus forte valeur particularisante.

Il est à remarquer pourtant que le traducteur SYS propose, dans ces deux cas, une

première occurrence du génitif français par la contraction de l'article défini avec la préposition *de*, forme qu'il n'emploie plus dès la deuxième occurrence du génitif dans (24), mais qu'il garde dans (25).

L'exemple (22) nous fournit un cas évident de mauvaise traduction automatique, vu que le génitif représente l'unique solution pour la structure traduite.

S'il s'agit de syntagmes figés à structure possessive, la traduction automatique est tout aussi viable que celle officielle, comme dans (23) et (24).

3.1.2.. La subordonnée relative

Un autre déterminant que nous avons relevé dans notre corpus est la subordonnée relative, structure qui alourdit les dispositions légales et qui sera évitée, autant que possible, par le traducteur humain :

Version anglaise	Version française automatique	Version française officielle
(25) European States *which* respect its values and are committed to promoting them together.	REV : les États européens *qui* respectent ses valeurs *et sont remis (commis) à la promotion d'eux ensemble.* RAD : les États européens *qui* respectent ses valeurs et se consacrent promotion d'eux ensemble. SYS : les États européens *qui* respectent ses valeurs et sont commis à les favoriser ensemble.	les États européens *qui* respectent ses valeurs et *qui* s'engagent à les promouvoir en commun.
(26) (...) a society *in which* pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail.	REV : (...) une société *dans laquelle* le pluralisme, la non-discrimination, la tolérance, la justice, la solidarité et l'égalité entre des femmes et des hommes prévalent. RAD : (...) une société *dans laquelle* le pluralisme, la non-discrimination, la tolérance, la justice, la solidarité et l'égalité entre les femmes et les hommes prédominent. SYS : (...) une société *dans laquelle* le pluralisme, la non-discrimination, la tolérance, la justice, la solidarité et l'égalité entre les femmes et leshommes règnent.	(...) une société *caractérisée* par le pluralisme, la non-discrimination, la tolérance, la justice, la solidarité et l'égalité entre les femmes et les hommes.
(27) The Union shall coordinate the policies *by which the Member States aim to achieve these objectives.*	REV : L'Union coordonnera la politique *par laquelle* les Etats membres aspirent à réaliser ces objectifs. RAD : L'Union coordonnera les politiques *par lesquelles* les Etats membres ont l'intention d'accomplir ces objectifs. SYS : L'union coordonnera les politiques *par lesquelles* les Etats membres visent à atteindre ces objectifs.	L'Union coordonne les politiques *des États membres visant à atteindre ces objectifs.*
(28) (...) this Constitution establishes the European Union, *on which* the Member States confer competences to attain objectives they have in common.	REV : (...) cette Constitution établit l'Union Européenne, *sur laquelle* les Etats membres confèrent des compétences d'atteindre des objectifs qu'ils ont dans commun. RAD : (...) cette Constitution établit l'Union Européenne, *sur laquelle* les Etats membres confèrent des compétences d'atteindre des objectifs qu'ils ont en commun. SYS : (...) cette constitution	(...) la présente Constitution établit l'Union européenne, *à laquelle* les États membres attribuent des compétences pour atteindre leurs objectifs communs.

	établit l'Union européenne, *sur laquelle* les Etats membres confèrent des compétences pour atteindre des objectifs qu'ils ont en commun.	
(29) the Member States confer competences to attain objectives *they have in common.*	REV : les Etats membres confèrent des compétences d'atteindre des objectifs *qu'ils* ont dans commun. RAD : les Etats membres confèrent des compétences d'atteindre des objectifs *qu'ils* ont en commun. SYS : les Etats membres confèrent des compétences pour atteindre des objectifs *qu'ils* ont en commun.	les États membres attribuent des compétences pour atteindre *leurs* objectifs communs.

En ce qui concerne le comportement du traducteur automatique, on voit bien que, dans (25), le pronom relatif simple ne pose pas de problèmes, sauf sa deuxième occurrence qui y est exclue. Quant au pronom relatif composé précédé d'une préposition, dans (26) et (27), le traducteur automatique a très bien choisi la préposition, sauf (28) où la polysémie de la préposition *on* bloque son choix en fonction du régime prépositionnel du verbe régi et conduit à une forme aberrante. L'exemple (29) témoigne des possibilités du traducteur automatique qui, même dans l'absence du relatif en anglais, trouve la bonne solution pour obtenir une version correcte de traduction.

Pour ce qui est du traducteur humain, on peut ressentir la liberté dont il jouit dans la transposition des subordonnées relatives. Pour éviter la «lourdeur » des dispositions légales, il exploite largement les possibilités offertes par d'autres déterminants du nom, tels que : les participes présent (27) et passé (26) ou l'adjectif qualificatif (29).

3. 2. Le groupe verbal

3.2.1. Les temps verbaux

Dans la transposition des temps verbaux, le traducteur automatique (30-33) est fidèle au texte de départ de sorte qu'une forme verbale finie (futur) ou non finie (participe présent, passé) est traduite par la même forme dans le texte d'arrivée :

Version anglaise	Version française automatique	Version française officielle
(30) *Reflecting* the will of the citizens and States of Europe (…)	REV : *En reflétant* la volonté des citoyens et des Etats de l'Europe RAD : *En reflétant* la volonté des citoyens et des États de l'Europe SYS : La volonté *se reflétant* des citoyens et des états de l'Europe	*Inspirée* par la volonté des citoyens et des États d'Europe
(31) It shall respect their essential State functions, including *ensuring* the territorial integrity of the State, *maintaining* law and order and *safeguarding* national security.	REV : Il respectera leurs fonctions essentielles d'état, y compris *l'assurance* de l'intégrité territoriale de l'État, le *maintien* de l'ordre public et la *protection* de la sécurité nationale. RAD : Il respectera leurs fonctions essentielles d'État, en incluant *le fait de garantir* l'intégrité territoriale de l'État, *le fait de maintenir* l'ordre public et *le fait de protéger* la sécurité nationale. SYS : Elle respectera leurs fonctions d'État essentielles, y compris *assurer* l'intégrité nationale de l'état, de la loi et de l'ordre de *maintien* et de	Elle respecte les fonctions essentielles de l'État, notamment *celles qui ont pour objet d'assurer* son intégrité territoriale, *de maintenir* l'ordre public et *de sauvegarder* la sécurité nationale.

	sauvegarder la sécurité nationale.	
(32) The Union *shall coordinate* the policies	REV : L'Union *coordonnera* la politique RAD : L'Union *coordonnera* les politiques SYS : L'union *coordonnera* les politiques	L'Union *coordonne* les politiques
(33) The Union *shall be open*	REV : L'Union *sera ouverte* RAD : L'Union *sera ouverte* SYS : L'Union *sera ouverte*	L'Union *est ouverte*

Si nous analysons la version écrite, les différences sont évidentes. La liberté du traducteur lui permet de remplacer, dans (30), le gérondif par un participe passé, un nom par un verbe à l'infinitif (31) ou même le futur par le présent (32-33). D'ailleurs, le présent est la seule forme verbale qui apparaît dans cette traduction. L'explication est claire : l'Union est une réalité qui a des tâches précises et qui agit au profit de ses membres.

3.2.2. Les déterminants du verbe

Au niveau des déterminants du verbe, on retrouve les mêmes erreurs de transposition des prépositions que celles signalées pour le groupe nominal. L'explication est, comme dans le cas précédent, la traduction centrée sur le mot et non pas sur le sens :

Version anglaise	Version française automatique	Version française officielle
(34) The union shall respect the equality of Member States *before* the constitution.	REV : L'union respectera l'égalité d'Etats membres *avant* la constitution RAD : L'union respectera l'égalité d'Etats membres *avant* la constitution SYS : L'union respectera l'égalité des Etats membres *avant* la constitution	L'Union respecte l'égalité des États membres *devant* la Constitution
(35) The Constitution and law adopted by the institutions of the Union in exercising competences conferred *on it*. (…)	REV : La loi d'union que la Constitution et la loi adoptée par les institutions de l'Union dans l'exercice de compétences *y ont conférée*. RAD : La Constitution et la loi adoptée par les institutions de l'Union dans les compétences faisant de l'exercice conféré *sur cela.* SYD : La constitution et la loi adoptées par les établissements de l'union en exerçant des compétences *ont conféré là-dessus.*	La Constitution et le droit adopté par les institutions de l'Union, dans l'exercice des compétences qui sont attribuées *à celle-ci.*
(36) Europe day shall be celebrated on 9 May *throughout the Union.*	REV : Le jour de l'Europe sera célébré le 9 mai *partout dans l'Union* (le Syndicat). RAD : Le jour d'Europe sera célébré le 9 mai *partout dans l'Union*. SYS : Le jour de l'Europe sera célébré le 9 mai *dans toute l'union*.	La journée de l'Europe est célébrée le 9 mai *dans toute l'Union.*
(37) shall exercise *on a Community basis*	REV : exerceront *sur une base Communautaire* RAD : exerceront *sur une base de Communauté* SYS : exercera *sur une base de la Communauté*	exerce *sur le mode communautaire*

Par exemple, la préposition *before* à valeur locative dans (34) acquiert une valeur temporelle dans les traductions automatiques : « L'union respectera l'égalité d'Etats membres *avant* la constitution » tandis que la version officielle donne une transposition correcte : « L'Union respecte l'égalité des États membres *devant* la Constitution ».

Dans le cas des prépositions polysémiques, telle que *on,* les traducteurs automatiques sont vraiment en difficulté de choix. Pour traduire (35), les solutions proposées par les trois moteurs de traduction automatiques sont aberrantes par rapport à la version écrite, une preuve de plus que le traducteur automatique est programmé pour traduire la structure de la langue de départ en fournissant des équivalents dans la langue d'arrivée. La traduction ne dépasse pas un simple transcodage effectué au niveau des hétéronymes.

Un autre exemple nous est fourni par (37) dont les solutions de traduction automatique sont fidèles à la langue de départ, ce qui conduit à une déformation du sens, tandis que la version officielle se sert d'une construction spécifique au discours juridique : « sur le mode communautaire ».

4. Conclusions

L'analyse des quatre versions met en évidence, une fois de plus, ce que Darbelnet (1979) soulignait quand il parlait du domaine juridique : les deux éléments du discours juridique d'importance égale pour la signification du message sont la terminologie et l'agencement des mots.

Il est rare qu'un système de traduction automatique traduise sans faute. La correction des erreurs peut impliquer des changements aux entrées lexicales existantes, quand il s'agit de problèmes de terminologie, ou la modification de la description linguistique qui traite des règles de grammaire quand il s'agit de structures syntaxiques. Les problèmes de traduction relèvent de l'importance accordée au mot et non pas au sens exprimé par ces mots. Le traducteur automatique ne peut pas capter le sens et le réexprimer selon la structure morphosyntaxique de la langue d'arrivée tandis que le traducteur humain peut exploiter les ressources de la langue d'arrivée au profit d'une traduction plus souple et plus appropriée.

Même si le lecteur peut comprendre le message, pour le spécialiste, qui interprète les mots et s'en sert à son tour tout en manipulant une terminologie spécialisée, les versions automatiques ne sont pas satisfaisantes. Elles ne lui offrent aucune garantie de la bonne transposition des termes et des constructions spécifiques, finissant par nuire même à la reproduction et à la transmission. Le vrai traducteur doit dépasser l'aspect lexical de la langue de départ pour s'accrocher à son aspect sémantique, car la compréhension du message est un élément indispensable de son travail.

En fin de compte, seule la personne qui puisse juger de l'efficacité du système de traduction automatique est celle qui connaît bien les vrais besoins du service de traduction, l'environnement du travail et les contraintes imposées par le contexte spécifique.

Bibliographie

CORNU, Gérard, 1990, *Linguistique juridique*, Montchrestien, Paris.

DARBELNET, Jean, 1979, ***Réflexions sur le discours juridique,*** Meta, vol. 24, no.1, p. 9-17.

KOUTSIVITIS G. VASSILIS, 1990, "La traduction juridique standardisation versus créativité", *Meta*, XXXV, nr. 1, p. 226-229.

SCURTU, Gabriela, 2007, *Linguistique juridique. Les structures langagières du discours normatif français,* Sitech, Craiova.

SOURIOUX, Jean-Louis & LERAT, Pierre, 1975, *Le langage du droit*, PUF, Paris.

Dictionnaires

*** *Oxford English Dictionary* (OED), http://www.oed.com/

*** *Le Trésor de la langue française informatisé* (TLFi), CNRS, http://zeus.inalf.fr/tlf.htm

Corpus

Treaty establishing a Constitution for Europe,
http://www.eu-oplysningen.dk/emner_en/fremtid/

Traité établissant une Constitution pour l'Europe. http://www.europnat.com/ce.pdf

Logiciels de traduction automatique

REVERSO (http://www.reverso.net/text_translation.asp?lang=FR&action=nev)

RADINS (http://www.radins.com/linker.php?id=5617)

SYSTRAN (http://www.SYSstran.fr/traduction)

Sponsor de aur / **Partenaire Or**

Sponsor de argint / **Partenaire Argent**

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / **Partenaire Bronze**

Comitetul Electrotehnic Român

TermRom
Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latină/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



Rezoluția anaforei pentru traducere

MORUZ Mihai Alex, Facultatea de Informatică, Universitatea „Al. I. Cuza", Iași - Institutul de Informatică Teoretică, Academia Română, Filiala Iași

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Abstract
Anafora este procesul de utilizare a unei entități lingvistice, cum ar fi un pronume, cu scopul de a referi o altă entitate. Rezoluția anaforei este procedeul prin care este determinată relația de coreferință dintre un referent și un antecedent în limbajul natural. Această lucrare discută utilitatea procesului de rezoluție a anaforei pentru sistemele de traducere automată și asistare a traducerii.

1. Introducere

Rezoluția anaforei este procesul prin care se determină dacă două expresii în limbaj natural se referă la aceeași entitate. Rezoluția anaforei este o problemă importantă in cadrul sistemelor de procesare a limbajului natural. Sistemele de Extragere a Informației (IE - Information Extraction), cum sunt cele construite în cadrul conferințelor DARPA (DARPA Message Understanding Conferences - MUC), au dovedit că rezoluția anaforei este o componentă atât de importantă a problemei de IE încât o sarcină separată pentru această problemă a fost definită și evaluată începând cu MUC-6 (MUC-6 1995).

Relațiile de coreferință, așa cum sunt descrise în MUC-6 (MUC-6 1995) și MUC-7 (MUC-7 1997) sunt definite astfel: o relație de coreferință denotă o identitate a referinței și este stabilită între două componente textuale numite elemente marcabile, care pot fi grupuri nominale articulate hotărât, grupuri nominale demonstrative, substantive proprii, apoziții, părți ale grupurilor nominale ce funcționează ca modificatori, pronume etc. Așadar, procedeul de rezoluție a anaforei descris în cadrul acestei lucrări este conceput pentru a trata grupuri nominale cât mai generale, și nu este restricționat la un anume tip de grupuri nominale, cum ar fi, de exemplu, pronumele. De asemenea, nu se fac nici un fel de restricții asupra posibilelor elemente marcabile candidat, ceea ce înseamnă că toate elementele marcabile, indiferent dacă sunt „organizație", „persoană" sau alte tipuri de entități, sunt luate în calcul. Capacitatea de a lega grupuri nominale care coreferă atât în interiorul frazelor cât și în afara lor este foarte importantă pentru înțelegerea și analiza discursului în general.

Această înțelegere a discursului poate fi folosită fie pentru îmbunătățirea sistemelor de traducere automată, prin creșterea preciziei de aliniere a lanțurilor coreferențiale. O altă utilizare ar fi îmbunătățirea unei aplicații de asistare a traducerii deja existente, cum ar fi Translator's Workbench. Translator's Workbench stochează informațiile lingvistice sub formă de "fișiere de memorare a traducerilor" (translation memory files), conținând perechi de segmente (enunțuri) bilingve - numite "unități de traducere" (translation units). În acest mod, unitățile de traducere sunt reținute în baza de date Workbench, putând fi reutilizate în diverse alte proiecte de traducere. Acest proces poate fi îmbunătățit prin crearea de translation units care conțin nu o secvență de text ci un lanț coreferențial determinat anterior. În momentul în care un element al lanțului coreferențial va fi tradus, toate celelalte elemente ale lanțului vor fi asociate cu traducerea sa.

Lucrarea va fi structurată după cum urmează: capitolul al doilea este o descriere a procesului de rezoluție a anaforei și a dificultăților atașate acestuia; capitolul al treilea este o descriere a potențialelor utilizări ale elementelor de rezoluție a anaforei pentru asistarea traducerii și pentru îmbunătățirea sistemelor de traducere automată; în ultimul capitol sunt descrise o serie de concluzii și de potențiale direcții ulterioare de integrare a procesului de rezoluție a anaforei în traducere.

2. Rezoluţia anaforei

În cele ce urmează voi descrie o serie de abordări pentru rezolvarea problemei rezoluţiei anaforei. Acestea pot fi clasificate în două categorii: abordări lingvistice (metode care se bazează cu precădere pe cunoştinţe lingvistice şi cunoştinţe specifice domeniului pentru care se doreşte rezolvarea anaforei) şi abordări bazate pe învăţarea automată (metode care pornesc de la analiza datelor şi nu de la cunoştinţe teoretice).

2.1. Abordări lingvistice

Abordările lingvistice sunt bazate pe cunoaşterea mecanismului de funcţionare a fiecărui limbaj şi au fost primele metode de recunoaştere a relaţiei de coreferinţă; unul dintre primii algoritmi este acela al lui Hobbs, publicat în 1986.

Algoritmul Hobbs

Algoritmul Hobbs (Hobbs, 1986) este una dintre cele mai vechi abordări pentru problema de rezoluţie a pronumelor. Acest algoritm se bazează pe arborele sintactic al frazei şi foloseşte constrângeri sintactice pentru rezoluţia pronominală.

Pentru determinarea antecedentului unui pronume, iniţial sunt luaţi în calcul doar candidaţii din cadrul aceleiaşi fraze (arborele sintactic al frazei în care se află pronumele luat în calcul este parcurs pe lăţime - breadth first - de la stânga la dreapta). In cazul în care există noduri de nivel mai înalt în fraza curentă, atunci sunt propuşi posibilii antecedenţi rezultaţi dintr-o parcurgere pe lăţime, de la stânga la dreapta, a fiecărui subarbore. Apoi, arborii sintactici ai frazelor precedente, în ordine cronologică, sunt examinaţi conform aceluiaşi procedeu pentru a determina posibili antecedenţi. În esenţă, algoritmul Hobbs preferă entităţile aflate în aceeaşi frază cu anaforul şi apoi entităţile din frazele imediat anterioare frazei în care se află anaforul (Elango, 2005). Mai mult, în funcţie de poziţia pronumelui în cadrul frazei, diferite entităţi ale unei fraze pot deveni mai importante pentru procesul de rezoluţie.

Abordări baazate pe reguli

Conform (Cristea, Dima 2001) şi (Cristea et al., 2002), pentru definirea unei abordări bazate pe reguli pentru rezoluţia anaforei, noţiunea de model de rezoluţie a anaforei este fundamentală. Conform (Cristea, Dima 2001), un astfel de model este descris ca având patru componente:

- O **mulţime de atribute principale**, care completează nivelul de proiecţie şi sunt apoi raportate la nivelul semantic (conform (Cristea et al., 2002), expresiile coreferenţiale sunt reprezentate pe trei niveluri: nivelul text sau nivelul de suprafaţă; nivelul de proiecţie, unde apar toate entităţile care pot să se afle în relaţia de coreferinţă, codificate ca **entităţi referenţiale** sau **RE** şi nivelul semantic, unde fiecare concept spre care referă una sau mai multe entităţi este codificat în mod unic în **entităţi dereferenţiale** sau **DE**);

- O **mulţime de surse de cunoaştere**, sau procesoare virtuale, al căror rol este acela de a extrage valorile atributelor definite în timpul parcurgerii incrementale a textului;

- O **mulţime de euristici**, sau reguli, care cooperează pentru a hotărî dacă o entitate referenţială introduce o nouă entitate de discurs, iar dacă nu, către care entitate de discurs deja existentă referă;

- O mulţime de reguli care definesc **domeniul de accesibilitate referenţială**, şi astfel stabilesc ordinea în care entităţile dereferenţiale sunt verificate.

Definirea unui astfel de model permite determinarea oricărui tip de relaţie anaforică, inclusiv relaţia anaforică intre două substantive, ceea ce îl face mult mai flexibil decât modelul propus de Hobbs, care funcţionează doar pe o submulţime a relaţiilor anaforice.

Acest tip de abordare este folosit si de (Strube, 1998)pentru rezoluţia pronumelor; el menţine o listă de entităţi, iar entitatea cea mai potrivită extrasă din acea listă este considerată antecedentul pronumelui curent (această căutare este efectuată după ce o serie de constrângeri, cum ar fi acordul în număr şi gen, sunt satisfăcute). Acest procedeu permite rezolvarea incrementală a pronumelor, ceea ce seamănă cu modul în care mintea umană rezolvă coreferinţe. Această abordare favorizează ca antecedenţi entităţile din aceeaşi frază cu pronumele de rezolvat.

2.2. Abordări bazate pe învăţare automată

Se consideră ca o metodă de rezoluţie a anaforei se bazează pe învăţare automată dacă acumulează cunoştinţe folosind un algoritm de învăţare şi date de antrenare. Această abordare necesită un corpus de dimensiuni relativ mici de documente de antrenare, care au fost adnotate la lanţuri de coreferinţe. Unul dintre avantajele unei astfel de abordări este acela că nu necesită o cunoaştere aprofundată a mecanismelor limbajului pentru care se face antrenarea; de asemenea, o abordare bazată pe reguli este foarte costisitoare ca timp şi mult mai dificil de gestionat, întrucât numărul de reguli necesare este mare. Mai mult, o abordare bazată pe reguli poate funcţiona doar pentru limbajul pentru care a fost dezvoltat, pe când una bazată pe învăţare automată poate fi adaptată oricărei limbi, cu condiţia existenţei unui corpus de antrenare. În cazul în care limbajul luat în calcul nu are un corpus adnotat la coreferinţă sau nu are disponibile o parte din instrumentele necesare preprocesărilor, o abordare bazată pe reguli poate fi preferabilă.

Antrenarea sistemului de rezoluţie a anaforei se realizează astfel: toate elementele marcabile din cadrul documentului de antrenare sunt determinate printr-o serie de module de procesare a limbajului (adnotare la parte de vorbire, determinare de grupuri nominale, dezambiguare semantică etc), aplicate succesiv, pe baza unei filozofii pipeline. După această preprocesare, sunt generate exemplele de antrenare, sub forma unor vectori de trăsături generaţi din perechi de elemente marcabile. Aceste exemple de antrenament sunt apoi oferite algoritmului de învăţare, care construieste un clasificator pentru determinarea relaţie de coreferinţă. Pentru a determina lanţurile coreferenţiale din cadrul unui document, sunt determinate toate elementele marcabile iar perechile de elemente care sunt candidate pentru relaţia de coreferinţă sunt oferite ca date de intrare pentru clasificator, care decide dacă cele două elemente coreferă sau nu.

Naive Bayes

O abordare statistică pentru rezoluţia anaforei a fost introdusă de (Ge et al., 1998); modelul probabilistic descris conţine o serie de trăsături sintactice şi semantice semnificative pentru rezoluţia pronominală, iar variabila aleatoare este antecedentul candidat pentru un pronume dat. Trăsăturile folosite sunt următoarele:

- Distanţa dintre pronume şi antecedentul candidat (sunt favorizaţi candidaţii aflaţi mai aproape);
- Structura sintactică, ce ajută la impunerea unor constrângeri de corectitudine;
- Gen, număr şi trăsătura de persoană. Aceste constrângeri (numite şi constrângeri de acord), pot fi implementate pe baza cuvintelor care apar în text.
- Numărul de apariţii. Grupurile nominale care apar în mod repetat sunt preferate.

Numărul de apariţii poate fi motivat prin teoria centrelor (Grosz et al. 1995), conform căreia un subiect continuat este cel mai probabil candidat pentru un pronume. O versiune modificată a algoritmului lui Hobbs este folosită pentru a calcula distanţa dintre pronume şi un antecedent propus, pentru a putea lua în calcul atât distanţa cât şi structura sintactică.

Arbori de decizie

Rezoluţia anaforei poate fi privită şi ca o problemă de clasificare de perechi. (Soon et al. 2001) adoptă o abordare bazată pe arbori de decizie. Problema de rezoluţie a anaforei este transformată într-o problemă de clasificare astfel: dată o pereche de elemente marcabile referă ele la acelaşi element? Un element marcabil poate fi un grup nominal sau un pronume, ceea ce generalizează problema de rezoluţie a anaforei dincolo de pronume. Toate elementele marcabile sunt determinate în pasul de preprocesare. Pentru învăţare este folosit un arbore de decizie, care învaţă o serie de reguli pe baza unei serii de trăsături calculate peste perechile de elemente marcabile. Trăsăturile folosite pentru antrenarea arborelui de decizie sunt descrise mai jos; este de remarcat faptul că aceste trăsături folosesc foarte puţină informaţie de natură sintactică:

- Distanţa: măsoară distanţa dintre două elemente marcabile în număr

de fraze;

- Acordul: o serie de trăsături care conţin informaţii cu privire la acordul în număr şi gen pentru o pereche de elemente marcabile;

- Tipul de element marcabil: o serie de trăsături care indică tipul elementului marcabil, mai precis grup nominal demonstrativ, grup nominal articulat cu articol hotărât, pronume, pronume reflexiv şi substantiv propriu;

- Acordul clasei semantice: această trăsătură verifică dacă clasele semantice a membrilor unei perechi de elemente marcabile se află în acord conform ierarhiei din WordNet;

- Trăsătura de alias: două elemente marcabile au trăsătura alias pozitivă dacă sunt formate din acelaşi nume, dacă unul dintre elemente are doar un nume de familie iar al doilea numele complet sau dacă un element este un acronim al celuialt;

- Trăsătura apoziţie: o pereche de elemente marcabile are trăsătura apoziţie pozitivă dacă elementul marcabil din pereche care apare al doilea în ordinea de suprafaţă a textului are funcţia sintactică de apoziţie faţă de primul element marcabil din pereche.

Exemplele pentru antrenarea arborelui de decizie sunt extrase automat din corpus: exemplele pozitive sunt formate din perechi adiacente de elemente marcabile aflate în relaţie de coreferinţă, iar exemplele negative sunt generate grupând perechi de elemente marcabile astfel: între două elemente marcabile despre care se ştie că fac parte din acelaşi lanţ coreferenţial şi care sunt adiacente, sunt luate ca exemple negative perechile formate din toate elementele marcabile aflate între cele două, împreună cu al doilea element din pereche.

(Ng, Cardie, 2001) propune o modificare la structura descrisă mai sus şi anume generarea de exemple de antrenare prin tratarea diferenţiată a grupurilor nominale şi a pronumelor. În plus propun extinderea mulţimii de trăsături prin adăugarea unui nou set de trăsături; mai precis, sistemul lor foloseşte trăsături care iau în calcul rolul gramatical al elementelor marcabile (subiect, complement direct etc) şi o serie de euristici. În ciuda adăugării acestor noi trăsături, precizia sistemului nu creşte semnificativ, iar aceasta se datorează fragmentării datelor.

(Ng, Cardie, 2002) o altă extensie a structurii descrise mai sus prin includerea unui nou clasificator care decide dacă un sintagmă nominală este anaforică sau nu. Pentru antrenarea acestui clasificator este folosit un model de tip *maximum entropy*, iar pentru determinarea coreferinţei este folosită aceeaşi structură a unui arbore de decizie.

(Zang et al., 2003) descrie o abordare bazată pe învăţarea prin competiţie; în loc să folosească perechi de tipul antecedent - anafor pentru a crea o mulţime de exemple de antrenare, autorii folosesc o pereche de posibili antecedenţi asociaţi unui anafor, astfel încât unul dintre antecedenţii candidat constituie un exemplu pozitiv, iar al doilea un exemplu negativ. Motivul este acela că modelele cu un singur candidat nu sunt suficiente pentru a învăţa noţiunea de coreferinţă; modelul bazat pe doi candidaţi, descris mai sus, este avantajat prin faptul că este antrenat să recunoască diferenţa dintre exemplele pozitive şi cele negative.

3. Rezoluţiei anaforei şi traducerea

Procesul de rezoluţie a anaforei permite păstrarea unei caracteristici fundamentale a discursului, mai exact coerenţa; în plus, pentru versiuni paralele ale unui text scris în limbi diferite, lanţurile coreferenţiale şi natura lor semantică nu se modifică semnificativ, proprietate utilă pentru traducere. Rezultatele procesului de rezoluţie a anaforei ar putea fi aplicat în două domenii ale traducerii asistate de calculator: traducerea automată şi memoriile de traducere.

Traducerea automată

În cazul traducerii automate statistice, una dintre marile dificultăţi este aceea de a alinia corect sintagmele corespunzătoare din limba sursă şi limba destinaţie. În mod obişnuit, această aliniere se face folosind corpusuri mari de texte paralele, din care se extrag parametrii statistici ai alinierii textului din limba sursă în limba destinaţie. Prin folosirea procesului de rezoluţie a anaforei, este posibilă reducerea acestui corpus şi scăderea costului computaţional asociat cu calculul parametrilor de natura

statistică.

Procedeul de folosire a relațiilor de coreferință pentru creșterea performanțelor de aliniere ale textului este următorul:

- Pentru fiecare dintre textele paralele determină lanţurile coreferenţiale şi semantica asociată fiecărui lanţ;

- Pentru fiecare lanţ coreferenţial al limbii sursă se încearcă asocierea unui element al acestuia cu un element al limbii destinaţie, pe baza interpretării semantice asociate. În momentul în care un element al unui lanţ din limba sursă este asociat cu un element al unui lanţ din limba destinaţie, atunci cele două lanţuri coreferenţiale sunt considerate versiuni paralele ale aceleiaşi entităţi semantice şi sunt aliniate.

Întrucât o mare parte a sintagmelor nominale ale text fac parte dintr-un lanţ coreferenţial, alinierea acestora poate reduce considerabil atât erorile de aliniere cât şi cantitatea de text aliniat necesară antrenării unui sistem de aliniere de text. Pentru această aplicaţie, un sistem de rezoluţie a anaforei bazat pe învăţare automată este mai eficient, întrucât este intractabil să se creeze sisteme bazate pe abordări lingvistice.

Memorii de traducere

O altă potenţială aplicaţie a rezoluţiei anaforei în domeniul traducerii ar fi îmbunătăţirea unei aplicaţii de asistare a traducerii deja existente, cum ar fi Translator's Workbench. Translator's Workbench stochează informaţiile lingvistice sub formă de "fişiere de memorare a traducerilor" (translation memory files), care conţin perechi de segmente (enunţuri) bilingve numite "unităţi de traducere" (translation units). În acest mod, unităţile de traducere sunt reţinute în baza de date Workbench, putând fi reutilizate în diverse alte proiecte de traducere. Acest proces poate fi imbunătăţit prin crearea de translation units care conţin nu o secvenţă de text ci un lanţ coreferenţial determinat anterior. În momentul în care un element al lanţului coreferenţial va fi tradus, toate celelalte elemente ale lanţului vor fi asociate cu traducerea sa.

De exemplu, pentru secvenţa de text de mai jos, determinarea principalului lanţ coreferenţial (Rupert Murdoch) şi completarea sa automată folosind traducerea conceptului ataşat lanţului coreferenţial poate ajuta traducătorul prin eliminarea unor activităţi repetitive sau prin dezambiguarea antecedenţilor unor pronume:

> Rupert Murdoch's News Corp. confirmed his interest in buying back the ailing New York Post. But analysts said that if he winds up bidding for the paper, he faces a no-win situation in the overcrowded New York newspaper market.

4. Concluzii

Această lucrare discută posibilitatea utilizării procesului de rezoluţie a anaforei pentru îmbunătăţirea sistemelor de traducere automată şi a memoriilor de traducere.

În ceea ce priveşte dezvoltările ulterioare, este necesară îmbunătăţirea rezultatelor obţinute de procesul de rezoluţie a anaforei. Sistemele de vârf pentru rezoluţia anaforei obţin rezultate de aproximativ 66%, ceea ce poate duce la erori de aliniere sau de traducere cauzate de erorile de preprocesare.

Bibliografie

Cristea, D., Dima, G.E., 2001. *An Integrating Framework for Anaphora Resolution.* Information Science and Technology, Romanian Academy Publishing House, Bucharest, vol. 4, no. 3.

Cristea, D. O. Postolache, G. Dima, C. Barbu, 2002, *AR-Engine - A Framework for Unrestricted Coreference Resolution*, Proceedings of LREC 2002, pp. 2000-2007. Las Palmas, Spain.

Elango, P. *Coreference Resolution: A Survey* Technical Report, University of Wisconsin Madison, 2005.

Ge, N., Hale, J., Charniak, E. *A statistical approach to anaphora resolution*. In

Proceedings of the Sixth Workshop on Very Large Corpora, 1998.

Grosz, B. J., Weinstein, S., Joshi, A. K. *Centering: a framework for modeling the local coherence of discourse*. Comput. Linguist., 21(2):203-225, 1995.

Hobbs, J. *Resolving pronoun references*. Readings in natural language processing, pages 339-352, 1986.

Ng, V. Cardie, C. *Improving machine learning approaches to coreference resolution.* Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02), 2002.

Ng, V. Cardie, C. *Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution.* Proceedings of the 19th International Conference on Computational Linguistics (COLING-02), 2002.

Soon, W. M., H. T. Ng, D. C. Y. Lim, 2001, A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics (Special Issue on Computational Anaphora Resolution), Vol 27, No 4, pp. 521-544.

Strube M. *Never look back: an alternative to centering.* In Proceedings of the 17th international conference on Computational linguistics, pages 1251-1257, 1998.

Yang, X., Zhou, G., Su, J., Tan, C. L. *Coreference resolution using competition learning approach*. In ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pages 176-183, 2003.

Sponsor de aur / **Partenaire Or**

Sponsor de argint / **Partenaire Argent**

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / **Partenaire Bronze**

Comitetul Electrotehnic Român

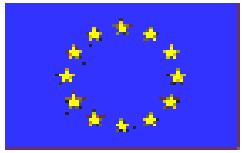
Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latina/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org

Direcţia Generală Traduceri





Comisia Europeană

Instrumentele de căutare

- Resursele multilingve ale UE -

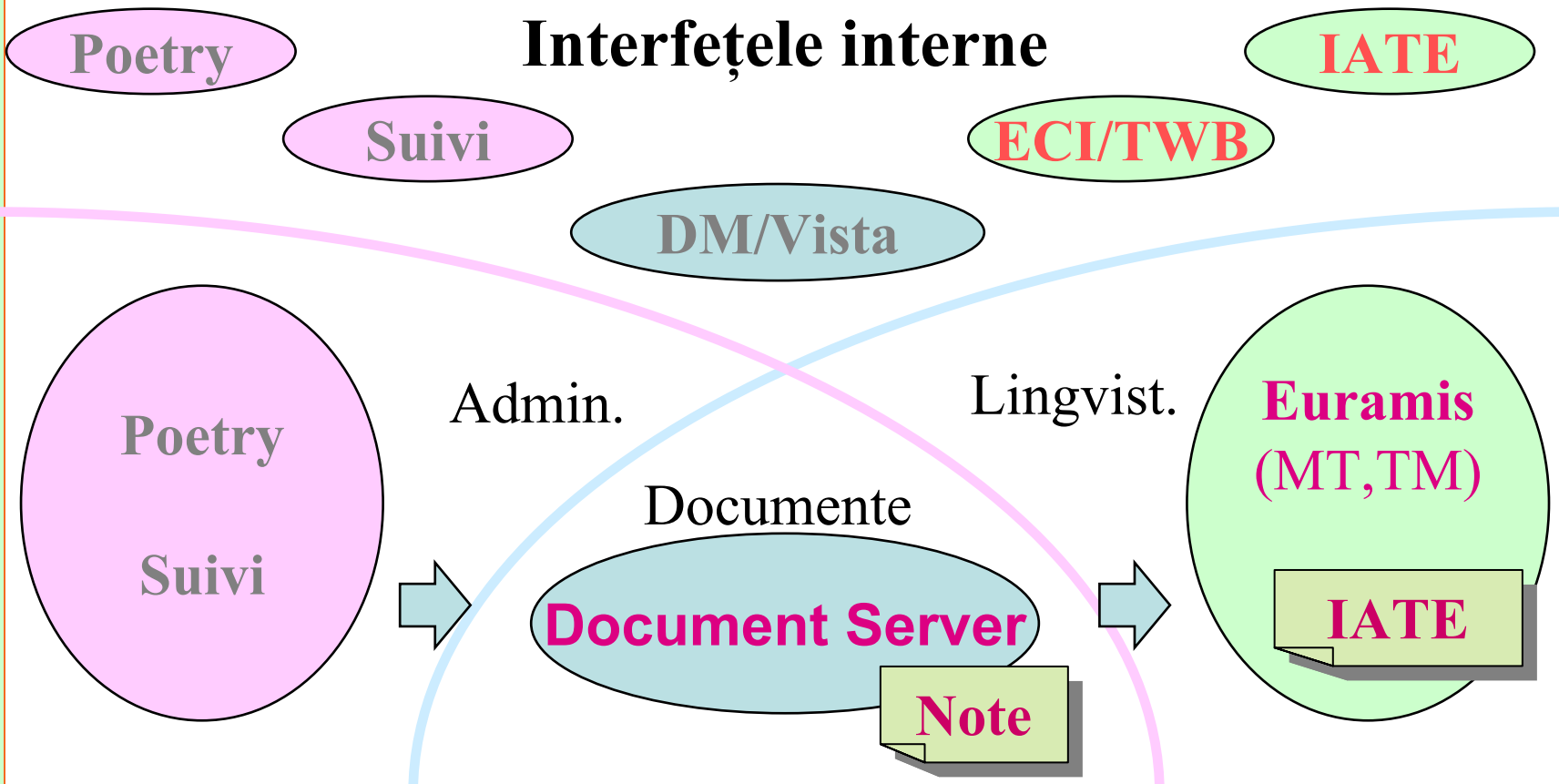
Paulo Correia
Departamentul de limba portugheză

Bucureşti, 29.02.2008
Seminar Internațional „Instrumente pentru asistarea traducerii”

Resurse

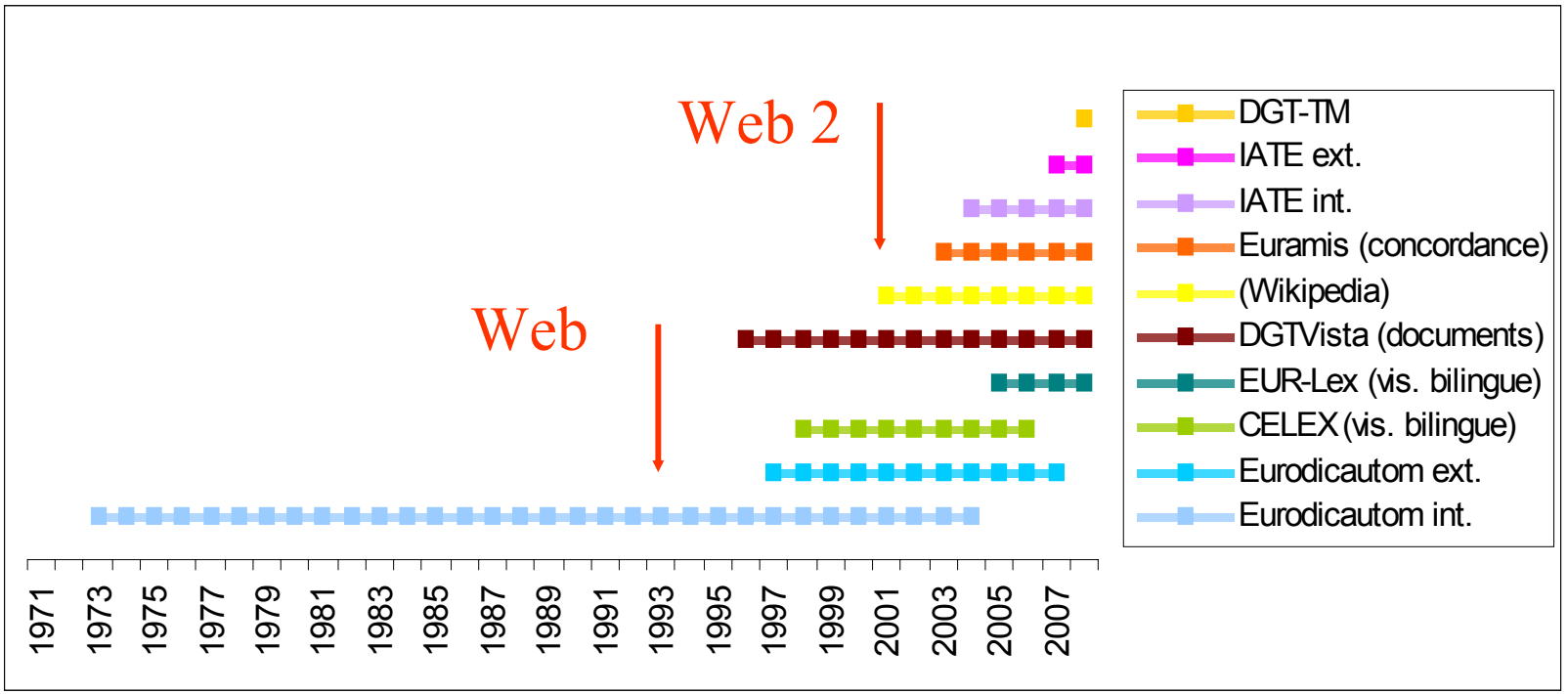
- Serviciile de traducere ale UE produc resurse multilingve (23 de limbi oficiale)
  - texte UE
  - memorii de traducere ale UE
  - baze terminologice ale UE

Informatizarea fluxului informațional



Direcţia Generală Traduceri

Interfețe – acces public



Accesul la resurse

- navigatoarele Web
- motoarele de căutare
  - …. ne observă !!!
  - indexarea celor mai profunde niveluri ale serverului **europa.eu**

1. Textele UE

- EUR-Lex
  - (CELEX - 1981-2004)
- Pre-Lex
- (Europa)
- ...

  - ocurențe în texte
  - conținuturi indexate de motoarele de căutare

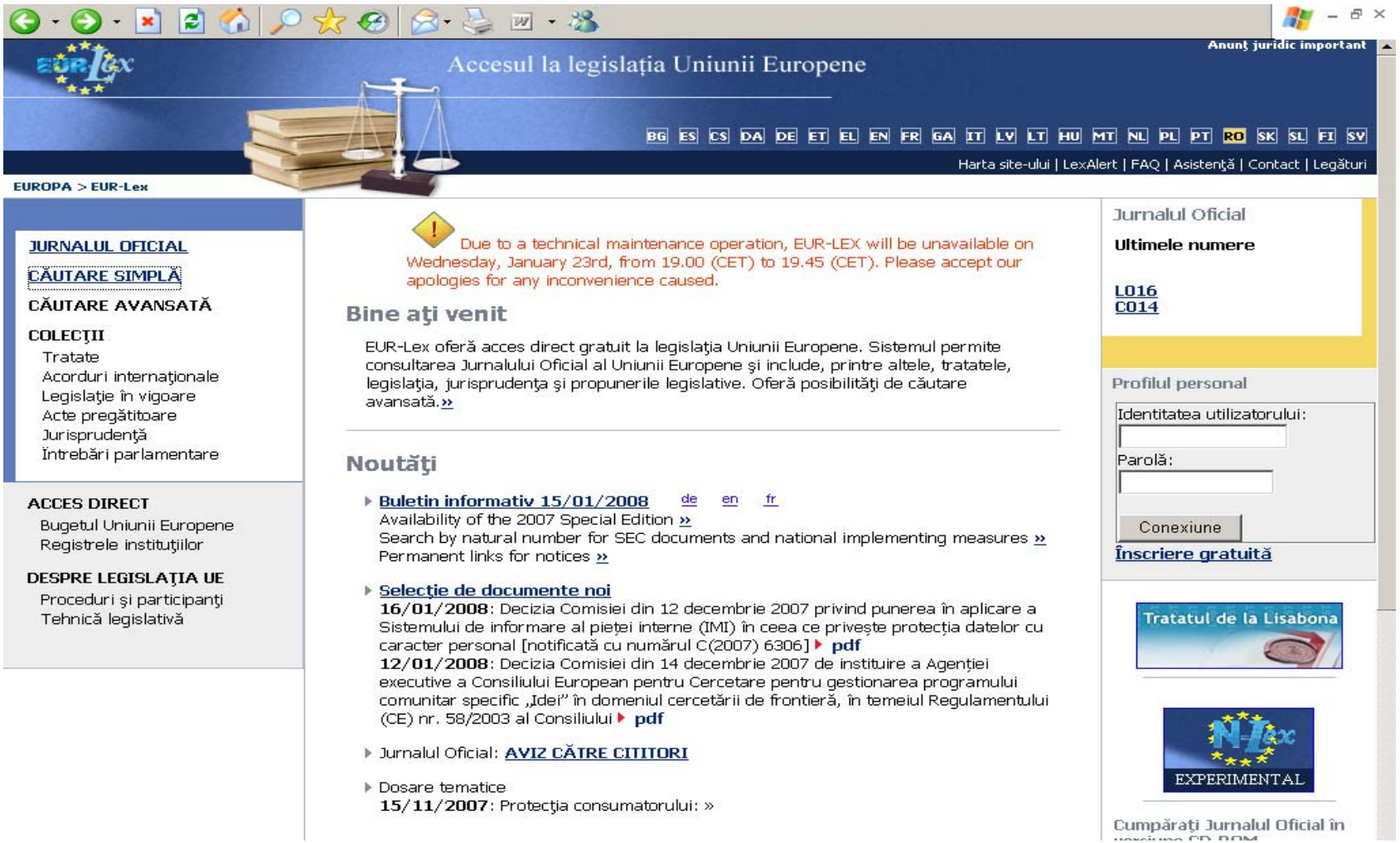
Direcţia Generală Traduceri

EUR-Lex

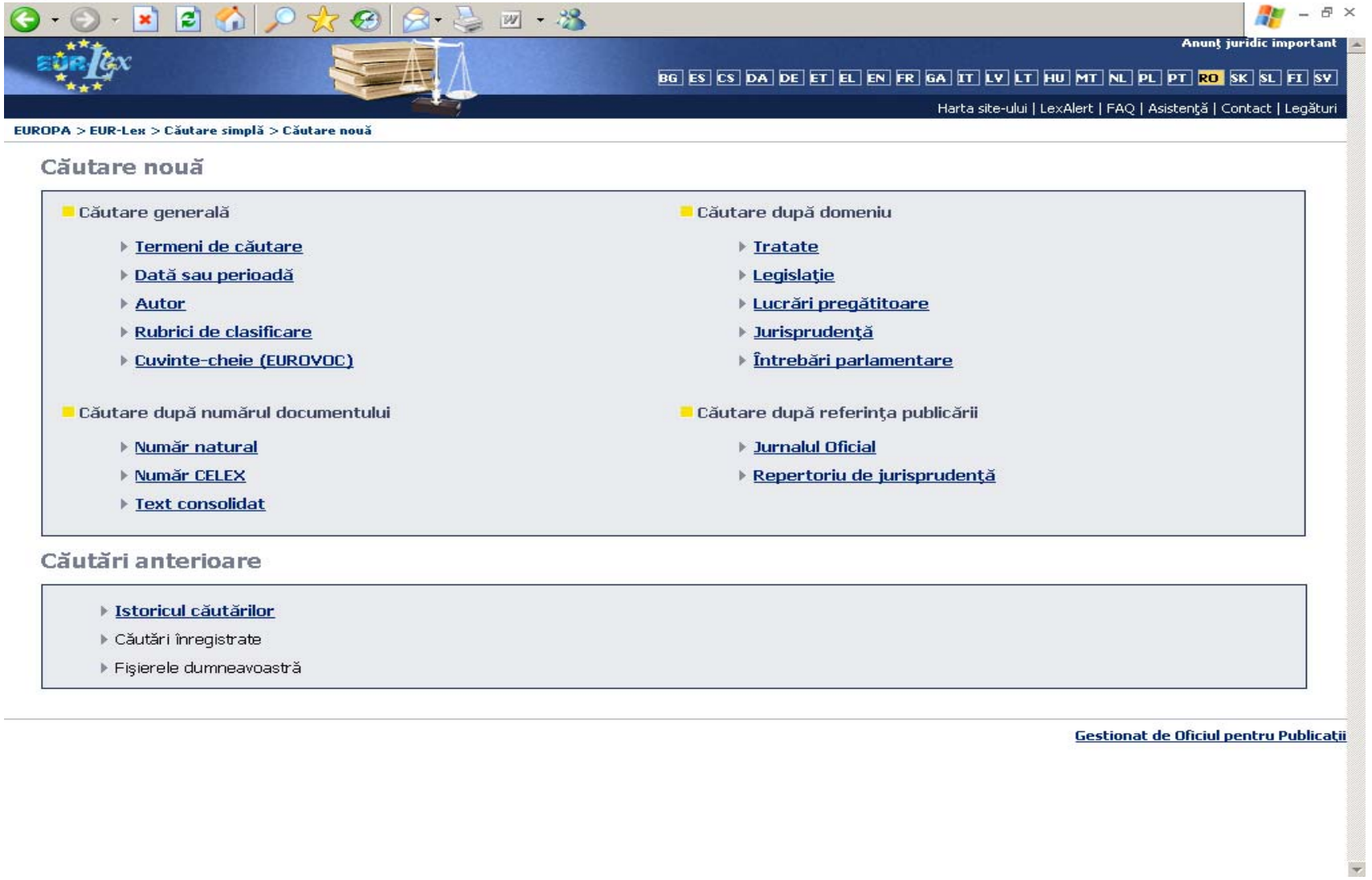
http://eur-lex.europa.eu/ro/index.htm

- Legislația Uniunii Europene
- Jurnalul oficial al Uniunii Europene
- … *inter alia*, tratate, legislație, jurisprudență şi propuneri legislative
  - aproape 2 milioane de documente
  - acces Internet gratuit
  - vizualizare bilingvă

Direcţia Generală Traduceri

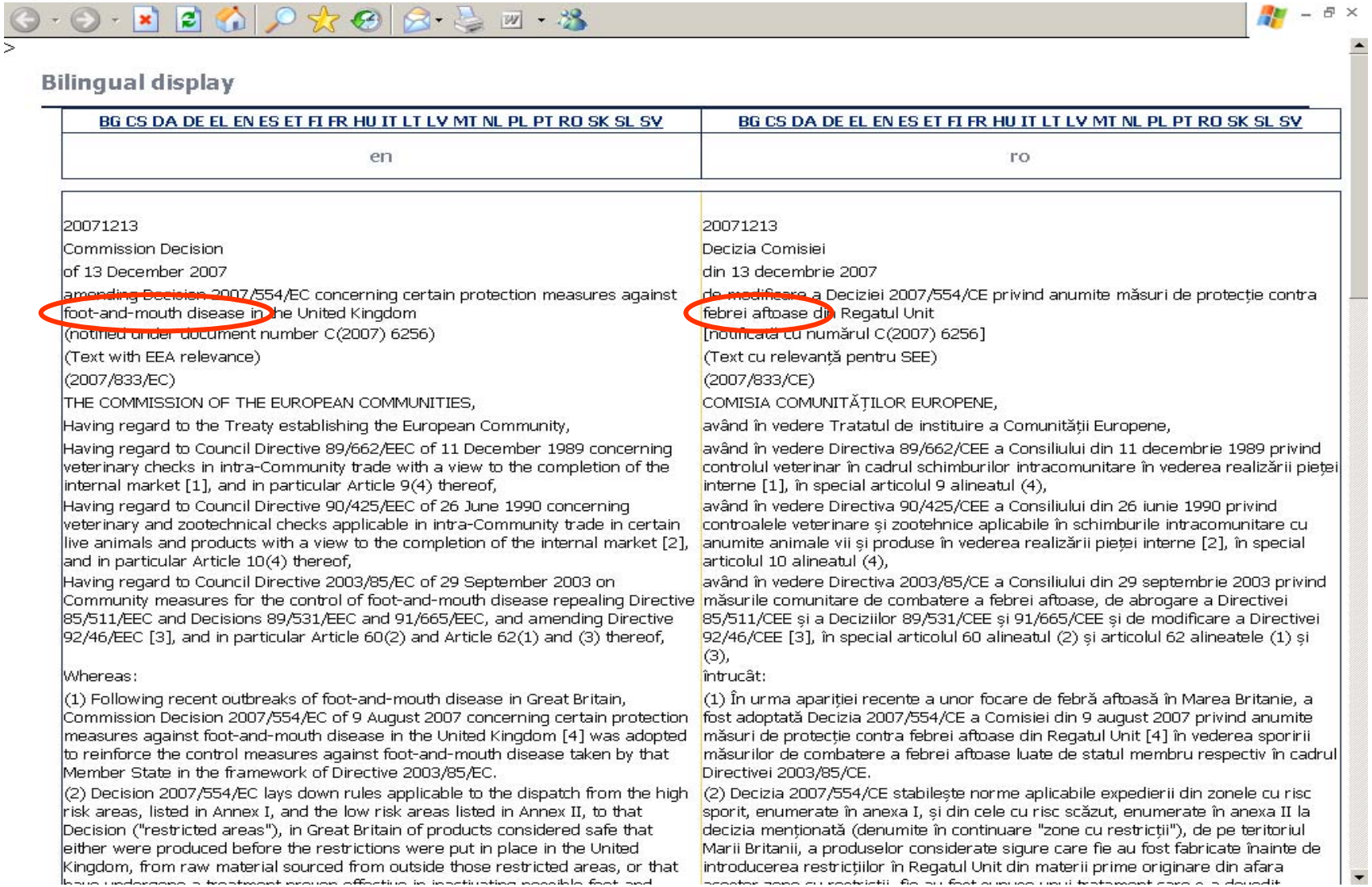


Direcţia Generală Traduceri



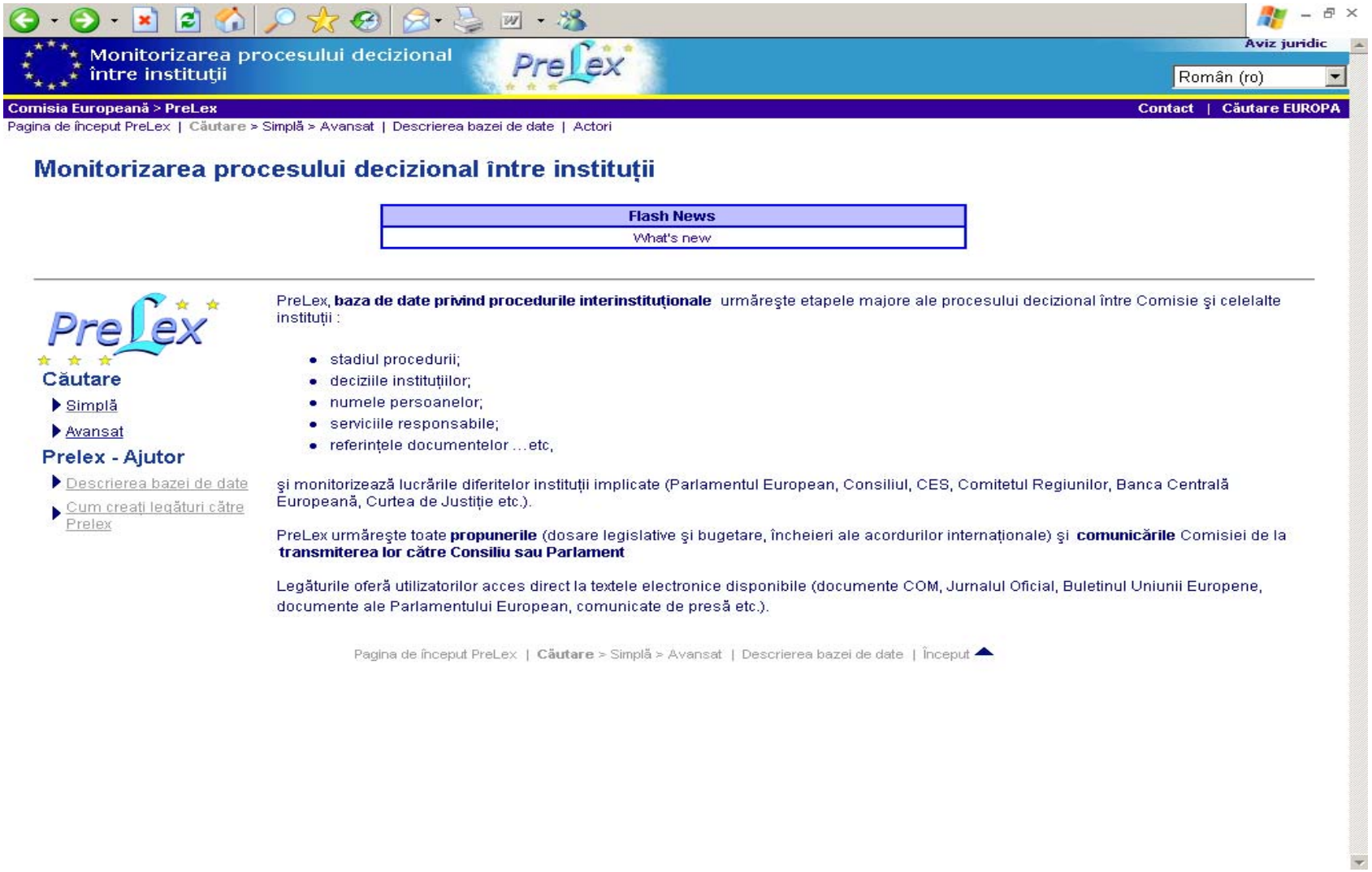
Direcţia Generală Traduceri



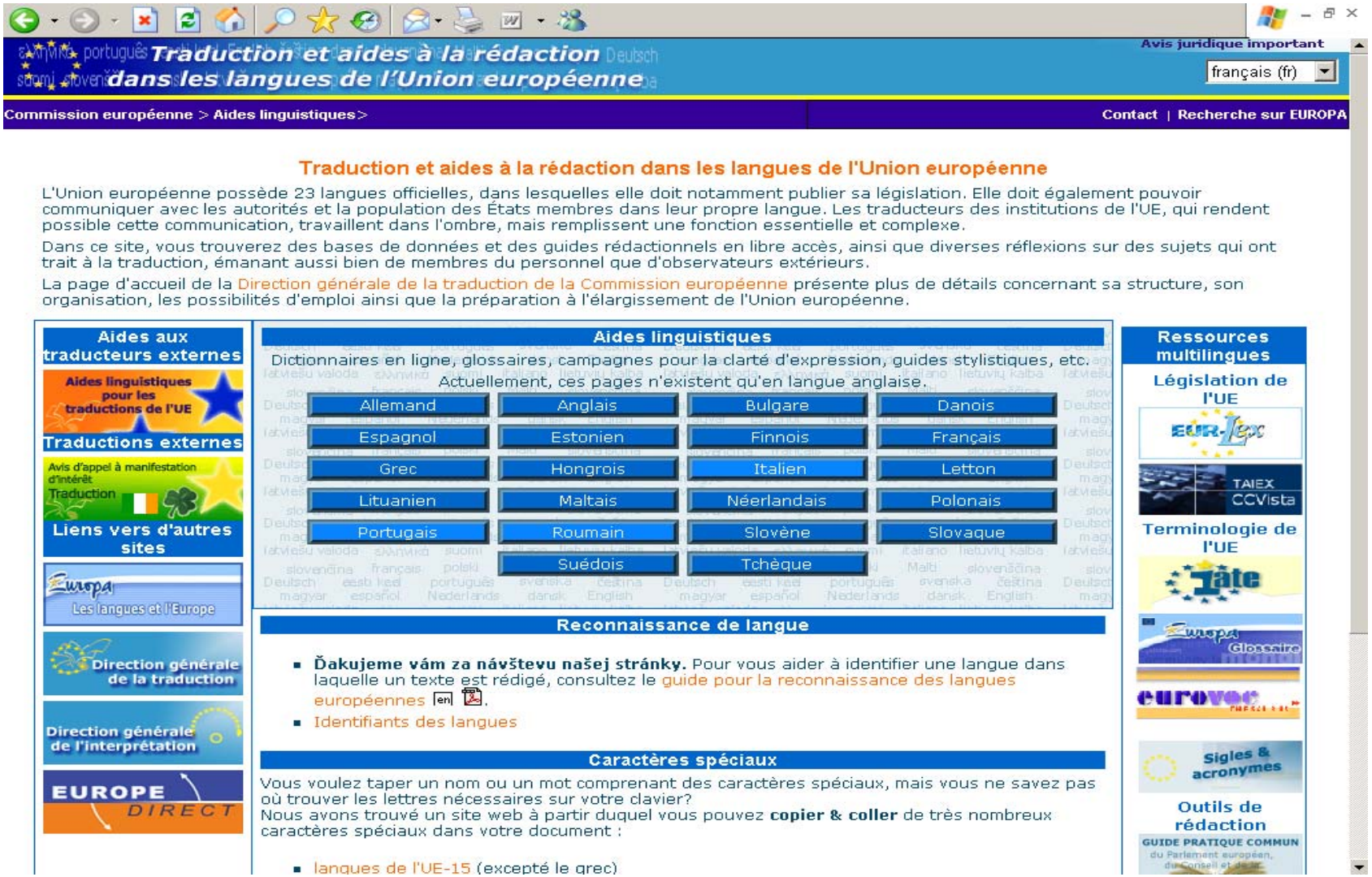


Bilingual display

BG CS DA DE EL EN ES ET FI FR HU IT LT LV MT NL PL PT RO SK SL SV	BG CS DA DE EL EN ES ET FI FR HU IT LT LV MT NL PL PT RO SK SL SV
en	ro
20071213 Commission Decision of 13 December 2007 amending Decision 2007/554/EC concerning certain protection measures against foot-and-mouth disease in the United Kingdom (notified under document number C(2007) 6256) (Text with EEA relevance) (2007/833/EC) THE COMMISSION OF THE EUROPEAN COMMUNITIES, Having regard to the Treaty establishing the European Community, Having regard to Council Directive 89/662/EEC of 11 December 1989 concerning veterinary checks in intra-Community trade with a view to the completion of the internal market [1], and in particular Article 9(4) thereof, Having regard to Council Directive 90/425/EEC of 26 June 1990 concerning veterinary and zootechnical checks applicable in intra-Community trade in certain live animals and products with a view to the completion of the internal market [2], and in particular Article 10(4) thereof, Having regard to Council Directive 2003/85/EC of 29 September 2003 on Community measures for the control of foot-and-mouth disease repealing Directive 85/511/EEC and Decisions 89/531/EEC and 91/665/EEC, and amending Directive 92/46/EEC [3], and in particular Article 60(2) and Article 62(1) and (3) thereof, Whereas: (1) Following recent outbreaks of foot-and-mouth disease in Great Britain, Commission Decision 2007/554/EC of 9 August 2007 concerning certain protection measures against foot-and-mouth disease in the United Kingdom [4] was adopted to reinforce the control measures against foot-and-mouth disease taken by that Member State in the framework of Directive 2003/85/EC. (2) Decision 2007/554/EC lays down rules applicable to the dispatch from the high risk areas, listed in Annex I, and the low risk areas listed in Annex II, to that Decision ("restricted areas"), in Great Britain of products considered safe that either were produced before the restrictions were put in place in the United Kingdom, from raw material sourced from outside those restricted areas, or that	20071213 Decizia Comisiei din 13 decembrie 2007 de modificare a Deciziei 2007/554/CE privind anumite măsuri de protecţie contra febrei aftoase din Regatul Unit [notificată cu numărul C(2007) 6256] (Text cu relevanţă pentru SEE) (2007/833/CE) COMISIA COMUNITĂŢILOR EUROPENE, având în vedere Tratatul de instituire a Comunităţii Europene, având în vedere Directiva 89/662/CEE a Consiliului din 11 decembrie 1989 privind controlul veterinar în cadrul schimburilor intracomunitare în vederea realizării pieţei interne [1], în special articolul 9 alineatul (4), având în vedere Directiva 90/425/CEE a Consiliului din 26 iunie 1990 privind controalele veterinare şi zootehnice aplicabile în schimburile intracomunitare cu anumite animale vii şi produse în vederea realizării pieţei interne [2], în special articolul 10 alineatul (4), având în vedere Directiva 2003/85/CE a Consiliului din 29 septembrie 2003 privind măsurile comunitare de combatere a febrei aftoase, de abrogare a Directivei 85/511/CEE şi a Deciziilor 89/531/CEE şi 91/665/CEE şi de modificare a Directivei 92/46/CEE [3], în special articolul 60 alineatul (2) şi articolul 62 alineatele (1) şi (3), întrucât: (1) În urma apariţiei recente a unor focare de febră aftoasă în Marea Britanie, a fost adoptată Decizia 2007/554/CE a Comisiei din 9 august 2007 privind anumite măsuri de protecţie contra febrei aftoase din Regatul Unit [4] în vederea sporirii măsurilor de combatere a febrei aftoase luate de statul membru respectiv în cadrul Directivei 2003/85/CE. (2) Decizia 2007/554/CE stabileşte norme aplicabile expedierii din zonele cu risc sporit, enumerate în anexa I, şi din cele cu risc scăzut, enumerate în anexa II la decizia menţionată (denumite în continuare "zone cu restricţii"), de pe teritoriul Marii Britanii, a produselor considerate sigure care fie au fost fabricate înainte de introducerea restricţiilor în Regatul Unit din materii prime originare din afara


Monitorizarea procesului decizional între instituţii
PreLex
Aviz juridic
Român (ro)
Comisia Europeană > PreLex
Contact | Căutare EUROPA
Pagina de început PreLex | Căutare > Simplă > Avansat | Descrierea bazei de date | Actori
Monitorizarea procesului decizional între instituţii
Flash News
What's new
PreLex
Căutare
Simplă
Avansat
Prelex - Ajutor
Descrierea bazei de date
Cum creaţi legături către Prelex
PreLex, baza de date privind procedurile interinstituţionale urmăreşte etapele majore ale procesului decizional între Comisie şi celelalte instituţii :
stadiul procedurii;
deciziile instituţiilor;
numele persoanelor;
serviciile responsabile;
referinţele documentelor ...etc,
şi monitorizează lucrările diferitelor instituţii implicate (Parlamentul European, Consiliul, CES, Comitetul Regiunilor, Banca Centrală Europeană, Curtea de Justiţie etc.).
PreLex urmăreşte toate propunerile (dosare legislative şi bugetare, încheieri ale acordurilor internaţionale) şi comunicările Comisiei de la transmiterea lor către Consiliu sau Parlament
Legăturile oferă utilizatorilor acces direct la textele electronice disponibile (documente COM, Jurnalul Oficial, Buletinul Uniunii Europene, documente ale Parlamentului European, comunicate de presă etc.).
Pagina de început PreLex | Căutare > Simplă > Avansat | Descrierea bazei de date | Început

Direcţia Generală Traduceri



2. Memoriile de traducere ale DGT

- DGT-TM
- …
  - ocurențe în contexte (unități de traducere – bi-texte)
  - conținuturi neindexate de motoarele de căutare

Direcţia Generală Traduceri

DGT-TM

http://langtech.jrc.it/DGT-TM.html

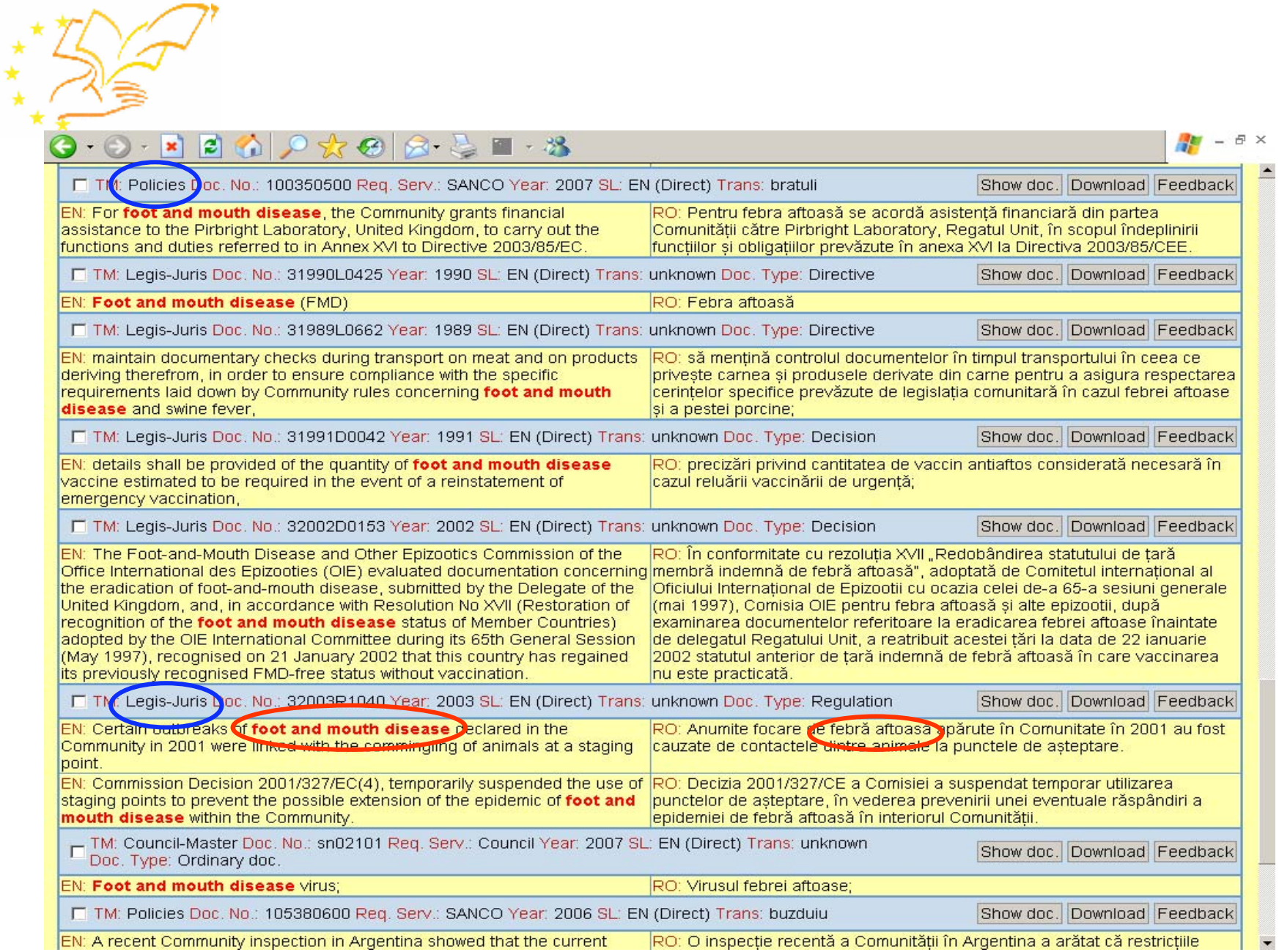
- din ianuarie 2008: memorii de traducere ale acquis-ului comunitar
  - baza Legis-Juris
  - aproape 1 GB

  - acces Internet gratuit
  - pentru descărcare
  - vizualizare bilingvă

DGT-TM

- **Memorii de traducere DGT ale achiului comunitar** (Legis-Juris)

limba	unități de traducere
engleză	2 187 504
franceză	1 106 442
bulgară	708 658
română	650 735



TM: Policies Doc. No.: 100350500 Req. Serv.: SANCO Year: 2007 SL: EN (Direct) Trans: bratuli — Show doc. | Download | Feedback

EN: For **foot and mouth disease**, the Community grants financial assistance to the Pirbright Laboratory, United Kingdom, to carry out the functions and duties referred to in Annex XVI to Directive 2003/85/EC.

RO: Pentru febra aftoasă se acordă asistență financiară din partea Comunității către Pirbright Laboratory, Regatul Unit, în scopul îndeplinirii funcțiilor și obligațiilor prevăzute în anexa XVI la Directiva 2003/85/CEE.

TM: Legis-Juris Doc. No.: 31990L0425 Year: 1990 SL: EN (Direct) Trans: unknown Doc. Type: Directive — Show doc. | Download | Feedback

EN: **Foot and mouth disease** (FMD)

RO: Febra aftoasă

TM: Legis-Juris Doc. No.: 31989L0662 Year: 1989 SL: EN (Direct) Trans: unknown Doc. Type: Directive — Show doc. | Download | Feedback

EN: maintain documentary checks during transport on meat and on products deriving therefrom, in order to ensure compliance with the specific requirements laid down by Community rules concerning **foot and mouth disease** and swine fever,

RO: să mențină controlul documentelor în timpul transportului în ceea ce privește carnea și produsele derivate din carne pentru a asigura respectarea cerințelor specifice prevăzute de legislația comunitară în cazul febrei aftoase și a pestei porcine;

TM: Legis-Juris Doc. No.: 31991D0042 Year: 1991 SL: EN (Direct) Trans: unknown Doc. Type: Decision — Show doc. | Download | Feedback

EN: details shall be provided of the quantity of **foot and mouth disease** vaccine estimated to be required in the event of a reinstatement of emergency vaccination,

RO: precizări privind cantitatea de vaccin antiaftos considerată necesară în cazul reluării vaccinării de urgență;

TM: Legis-Juris Doc. No.: 32002D0153 Year: 2002 SL: EN (Direct) Trans: unknown Doc. Type: Decision — Show doc. | Download | Feedback

EN: The Foot-and-Mouth Disease and Other Epizootics Commission of the Office International des Epizooties (OIE) evaluated documentation concerning the eradication of foot-and-mouth disease, submitted by the Delegate of the United Kingdom, and, in accordance with Resolution No XVII (Restoration of recognition of the **foot and mouth disease** status of Member Countries) adopted by the OIE International Committee during its 65th General Session (May 1997), recognised on 21 January 2002 that this country has regained its previously recognised FMD-free status without vaccination.

RO: În conformitate cu rezoluția XVII „Redobândirea statutului de țară membră indemnă de febră aftoasă", adoptată de Comitetul internațional al Oficiului Internațional de Epizootii cu ocazia celei de-a 65-a sesiuni generale (mai 1997), Comisia OIE pentru febra aftoasă și alte epizootii, după examinarea documentelor referitoare la eradicarea febrei aftoase înaintate de delegatul Regatului Unit, a reatribuit acestei țări la data de 22 ianuarie 2002 statutul anterior de țară indemnă de febră aftoasă în care vaccinarea nu este practicată.

TM: Legis-Juris Doc. No.: 32003R1040 Year: 2003 SL: EN (Direct) Trans: unknown Doc. Type: Regulation — Show doc. | Download | Feedback

EN: Certain outbreaks of **foot and mouth disease** declared in the Community in 2001 were linked with the commingling of animals at a staging point.

RO: Anumite focare de febră aftoasă apărute în Comunitate în 2001 au fost cauzate de contactele dintre animale la punctele de așteptare.

EN: Commission Decision 2001/327/EC(4), temporarily suspended the use of staging points to prevent the possible extension of the epidemic of **foot and mouth disease** within the Community.

RO: Decizia 2001/327/CE a Comisiei a suspendat temporar utilizarea punctelor de așteptare, în vederea prevenirii unei eventuale răspândiri a epidemiei de febră aftoasă în interiorul Comunității.

TM: Council-Master Doc. No.: sn02101 Req. Serv.: Council Year: 2007 SL: EN (Direct) Trans: unknown Doc. Type: Ordinary doc. — Show doc. | Download | Feedback

EN: **Foot and mouth disease** virus;

RO: Virusul febrei aftoase;

TM: Policies Doc. No.: 105380600 Req. Serv.: SANCO Year: 2006 SL: EN (Direct) Trans: buzduiu — Show doc. | Download | Feedback

EN: A recent Community inspection in Argentina showed that the current

RO: O inspecție recentă a Comunității în Argentina a arătat că restricțiile

Direcţia Generală Traduceri

3. Baza de date terminologică a UE

- IATE
  - *Inter-Active Terminology for Europe*
  - resurse terminologice (fişe terminologice)
  - conținuturi neindexate de motoarele de căutare

IATE

http://iate.europa.eu

- deschiderea oficială: 28 iunie 2007
  - din 2004: baza terminologică interinstituțională
  - din martie 2007: accesibilă publicului
  - aproape 1,5 milioane de fişe

  - acces Internet gratuit
  - vizualizare multilingvă
  - interactivitatea

Direcţia Generală Traduceri

IATE – definiție

- bază de date terminologice **interactivă** şi **multilingvă** pentru redactare, traducere şi interpretare în domeniile comunitare
- date cu **valoare adăugată** în raport cu alte resurse lingvistice

IATE – valoare adăugată

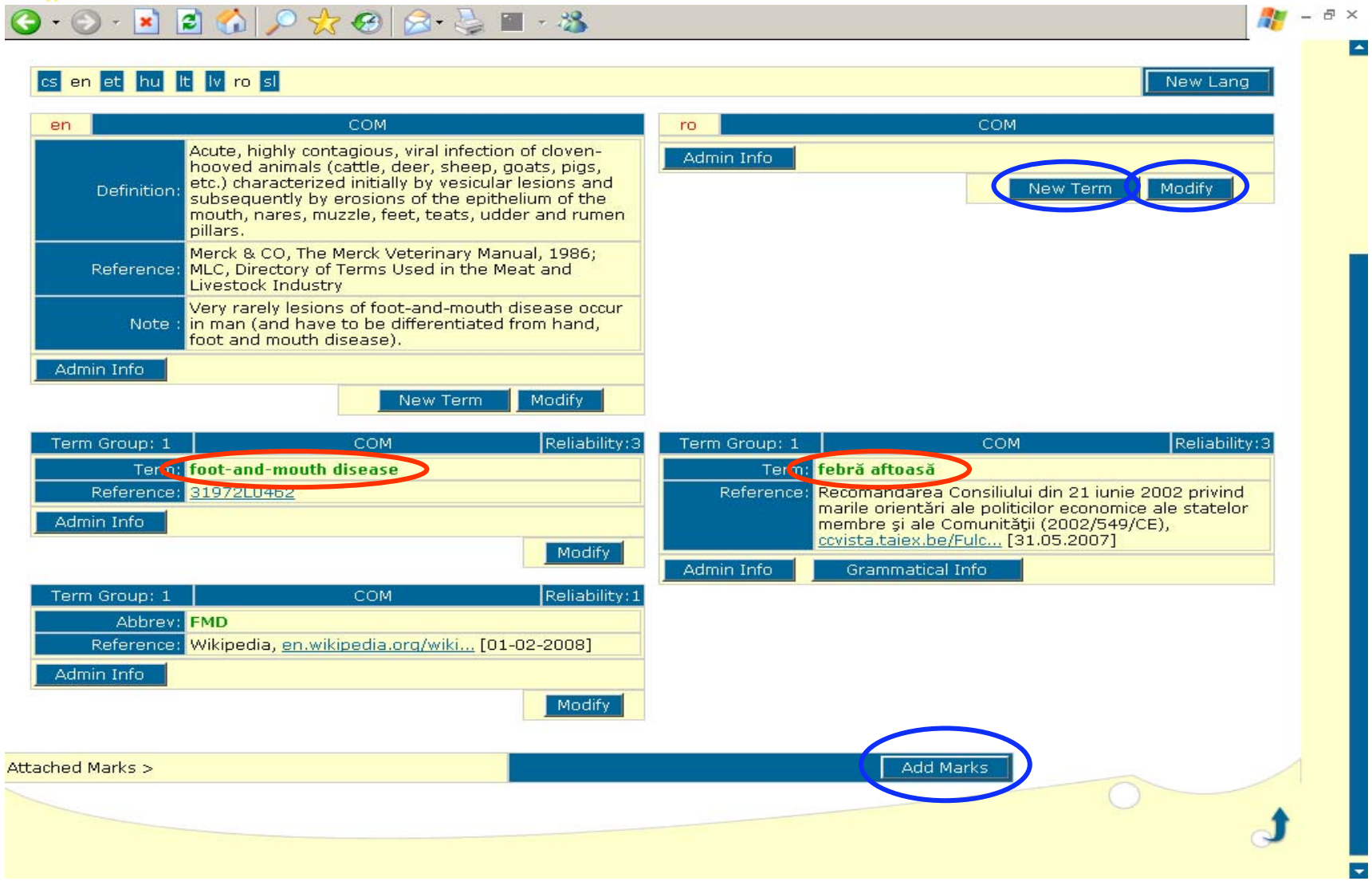
- multilingvism (în întregime)
- redondanță controlată
- concepte cu definiții
- referințe credibile
- …fiabilitate ridicată

- … **validare** de către terminologi
  - responsabili pentru noile conținuturi

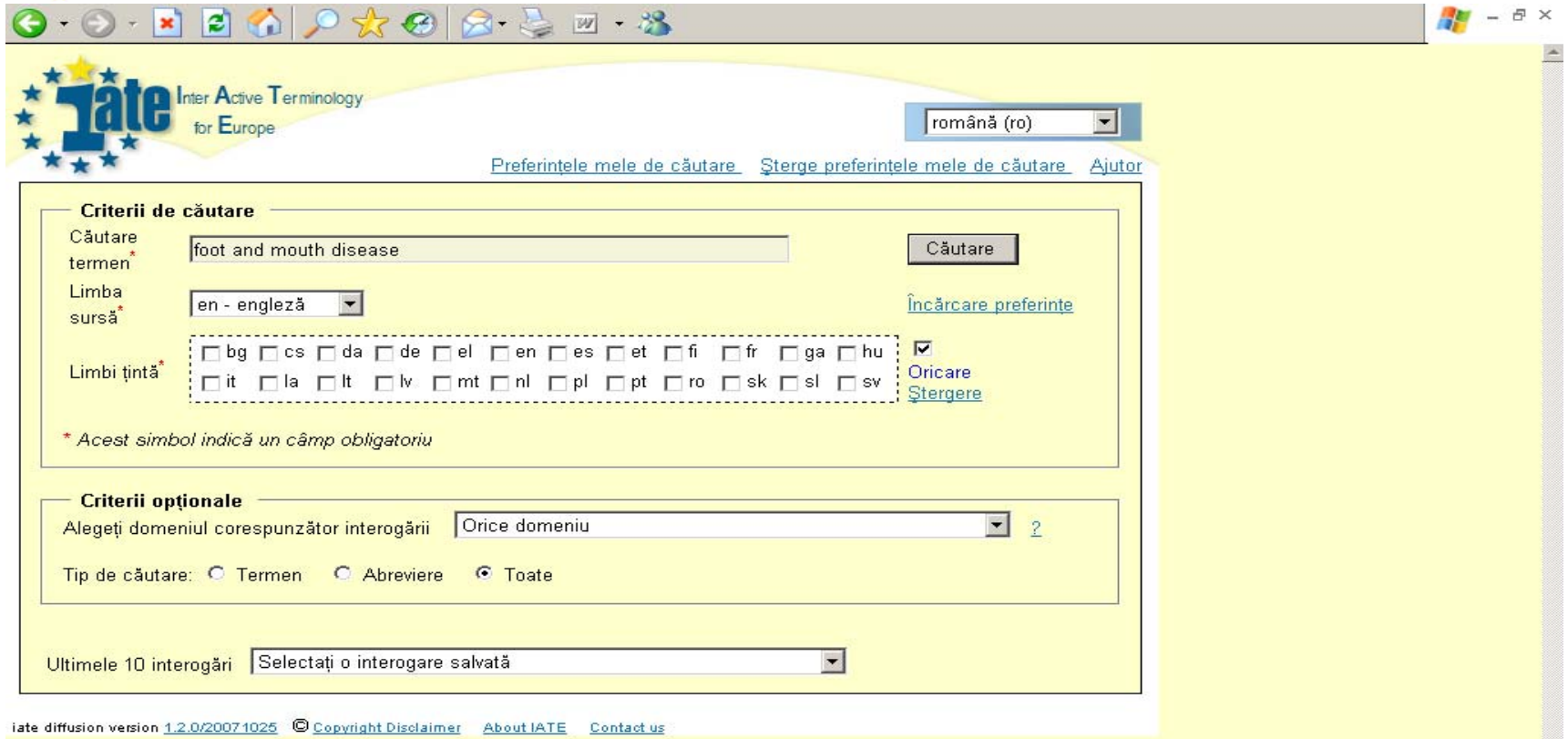
IATE - conținuturi

- soluții terminologice în limbile oficiale:
  - „jargon comunitar”;
  - limbaje de specialitate;
  - nomenclaturi;
  - ...
- multilingvism crescând

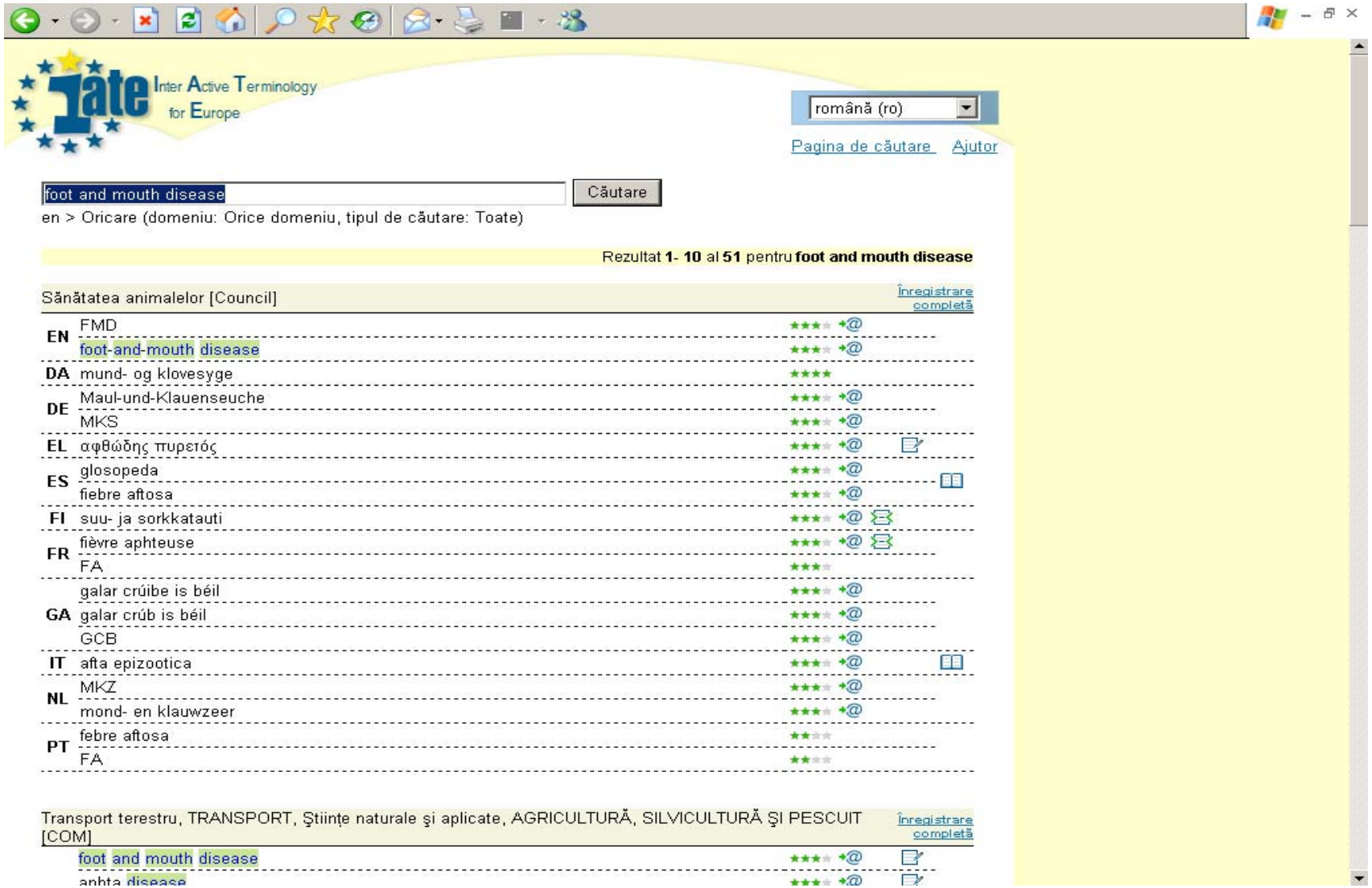
Direcţia Generală Traduceri



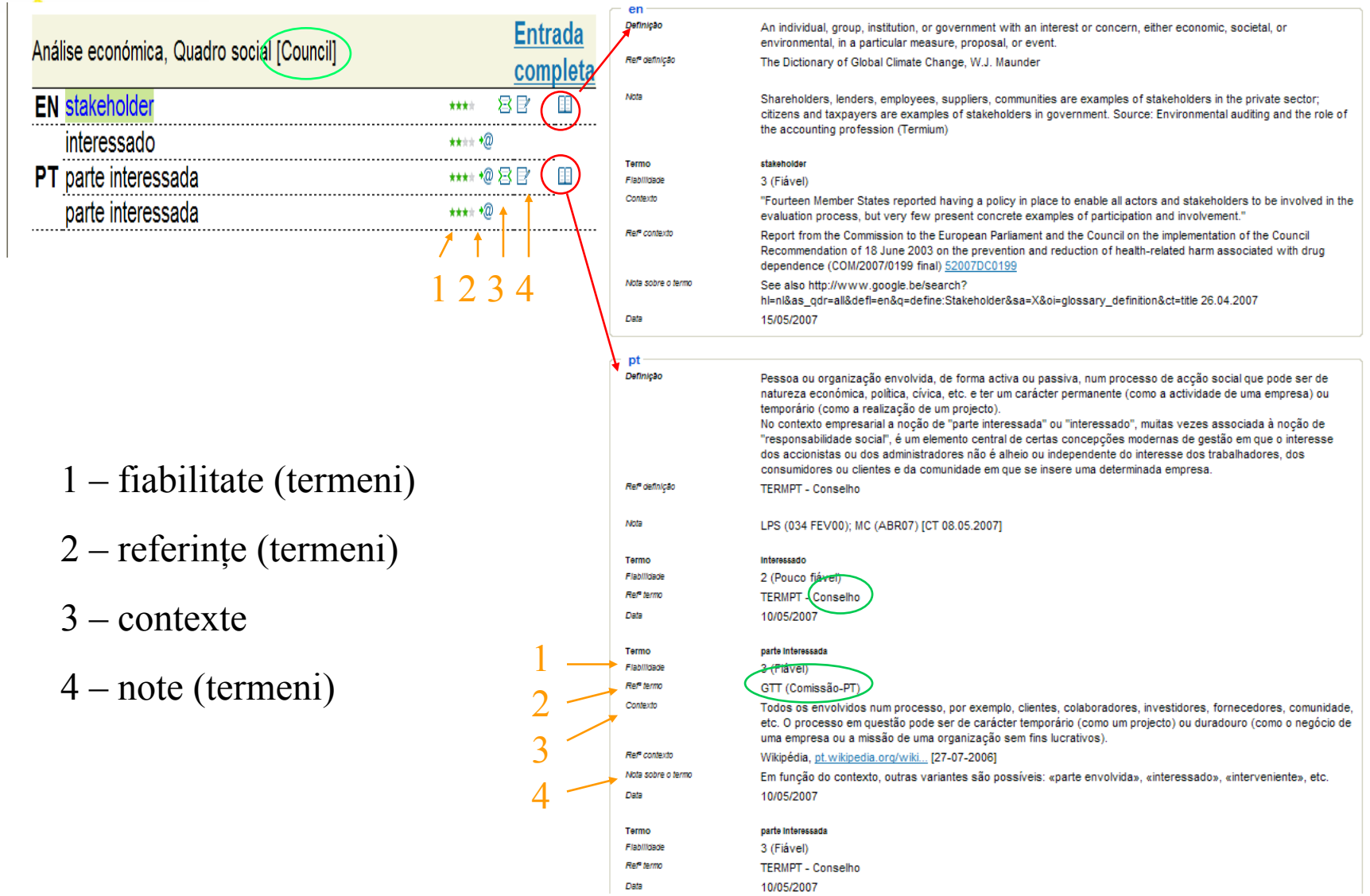
Direcţia Generală Traduceri



Direcţia Generală Traduceri



Direcţia Generală Traduceri



1 – fiabilitate (termeni)

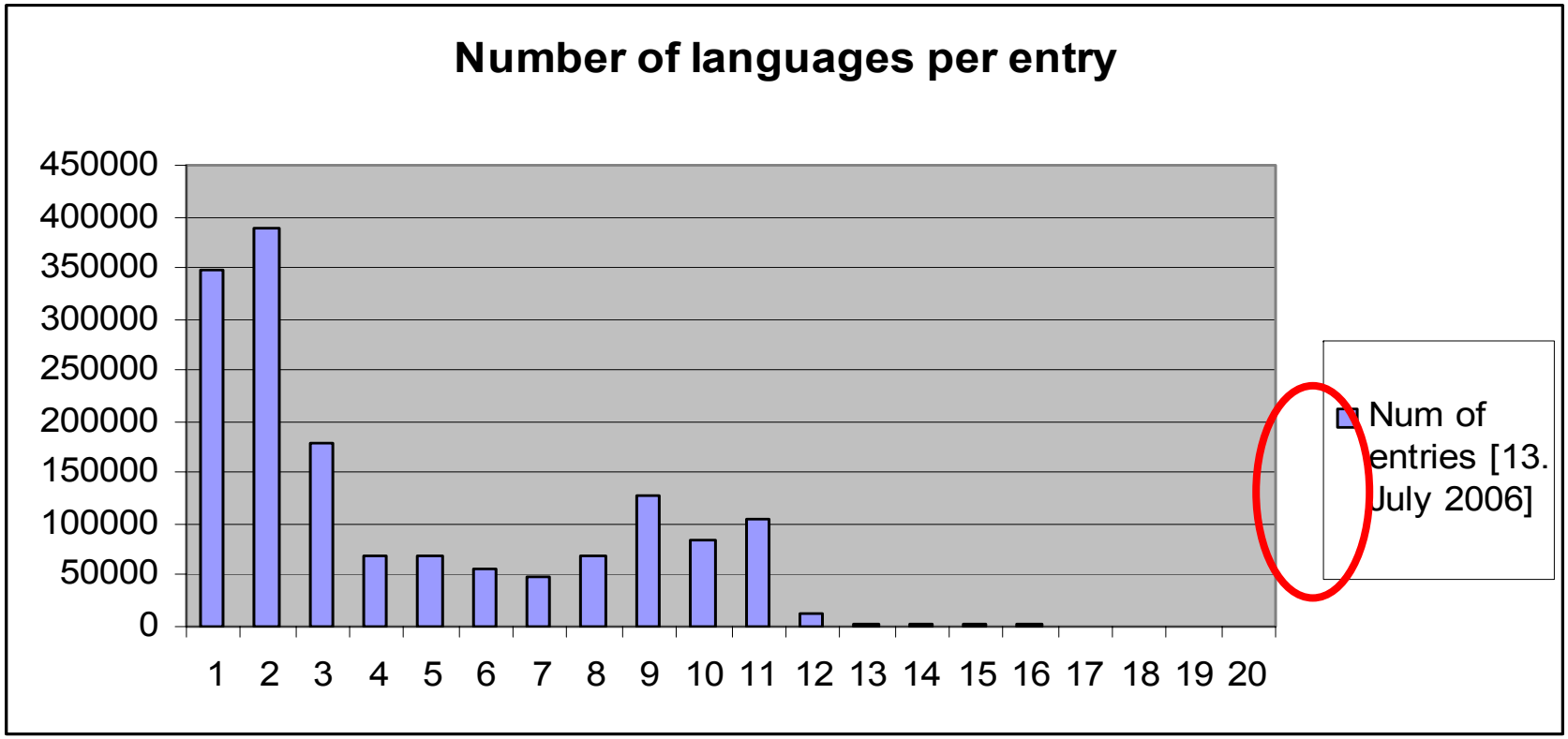
2 – referinţe (termeni)

3 – contexte

4 – note (termeni)

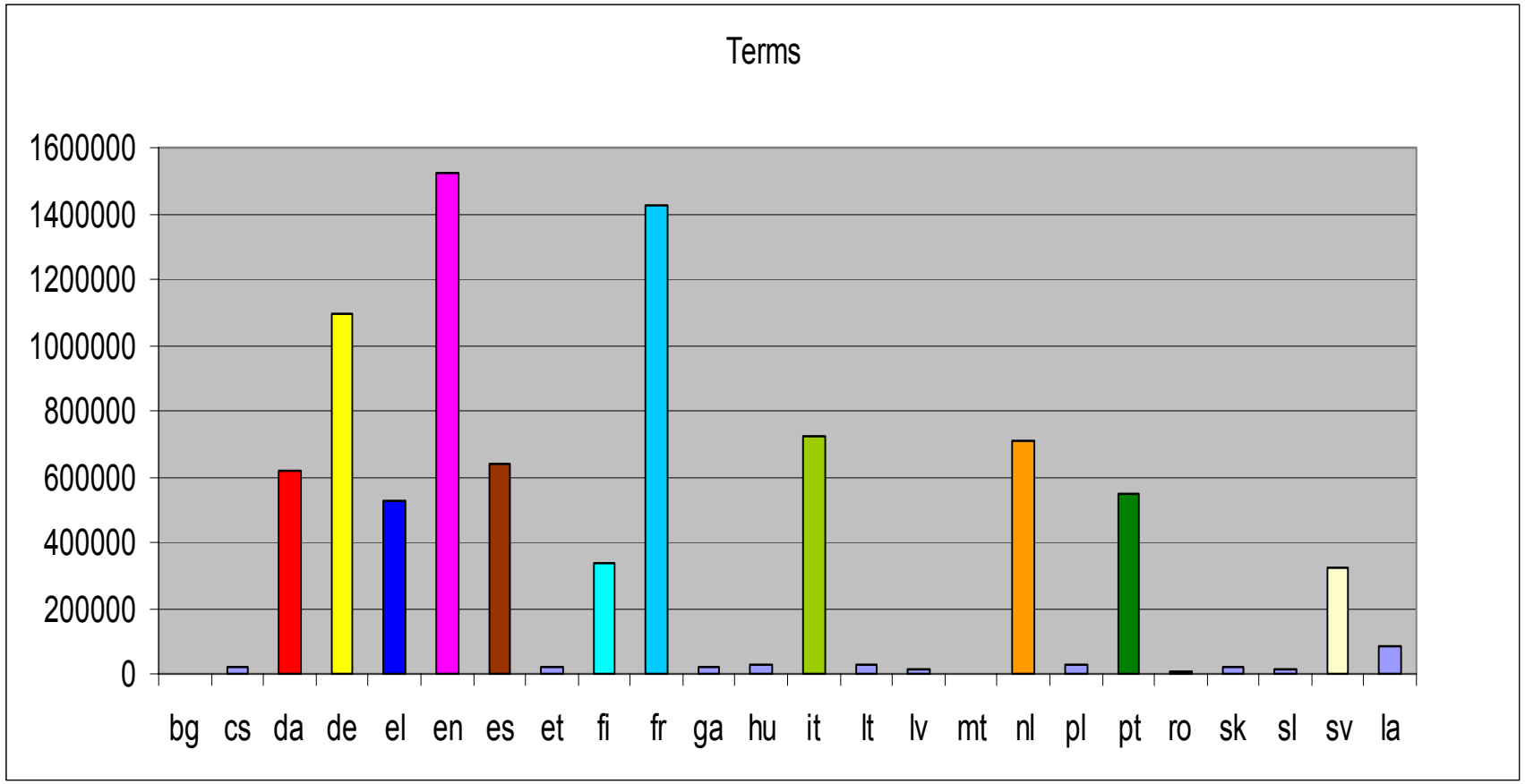
Direcţia Generală Traduceri

IATE – limbi după fişă



Direcţia Generală Traduceri

IATE- termeni după limbă

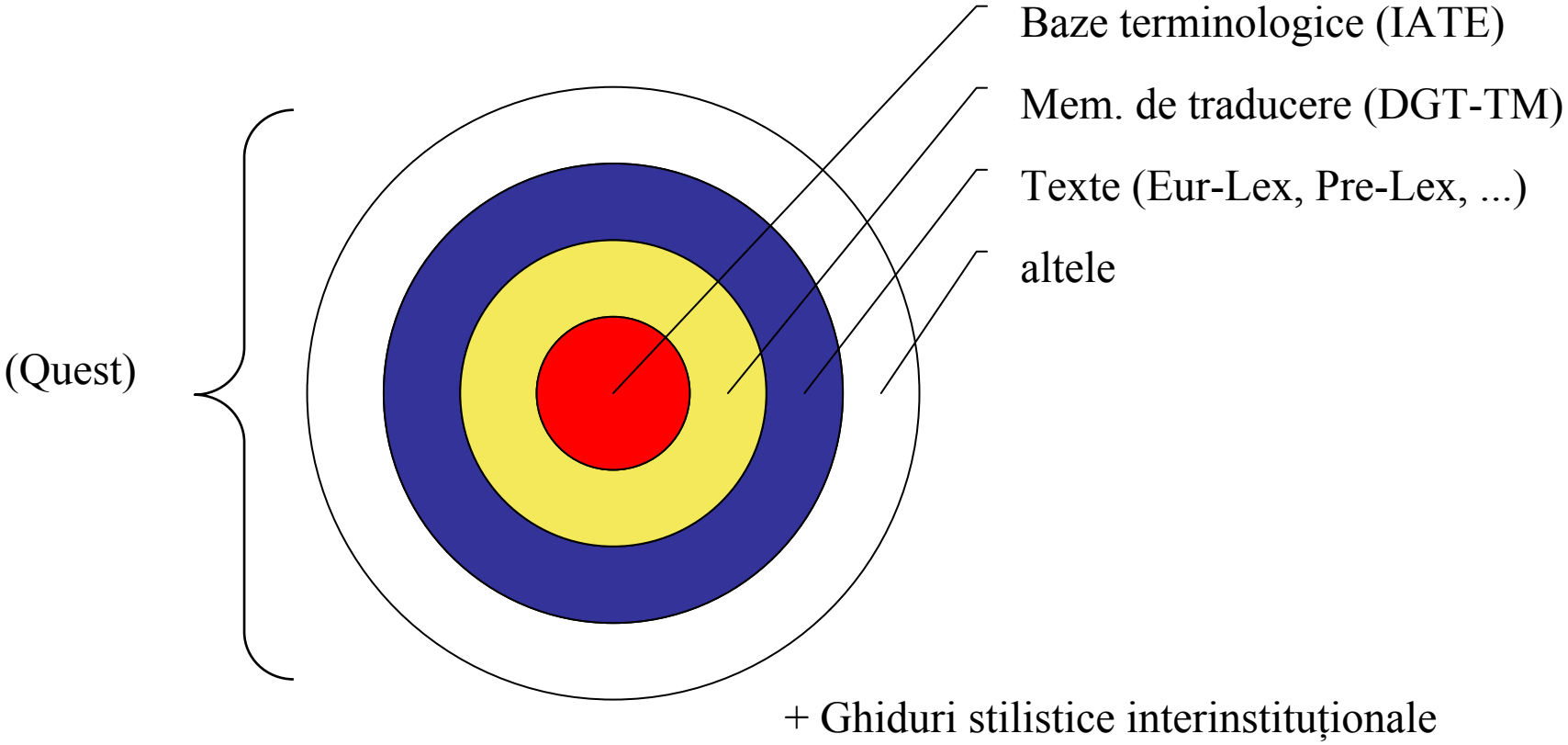


IATE - Consolidare

- **zgomot**
  - repetarea termenilor pentru anumite limbi (UE /15 țări)
- **liniște**
  - absența conținuturilor pentru mai multe limbi (UE /27 țări)

- … necesitatea consolidării conținuturilor
  - activitatea interinstituțională

Integrarea informației



Resursele multilingve ale UE

- orientare spre **serviciul public**
  - multilingvism garantat
  - calitate garantată
  - deschiderea instrumentelor spre un public larg

Direcţia Generală Traduceri

Paulo.Correia@ec.europa.eu

Direcția Generală Traduceri

Comisia Europeană

Les outils de recherche

Internet, dictionnaires électroniques, bases de données, etc.

Bucureşti, 29.02.2008
Séminaire international
«Les outils d'aide à la traduction»

Paulo Correia
Departamentul de limba portugheză

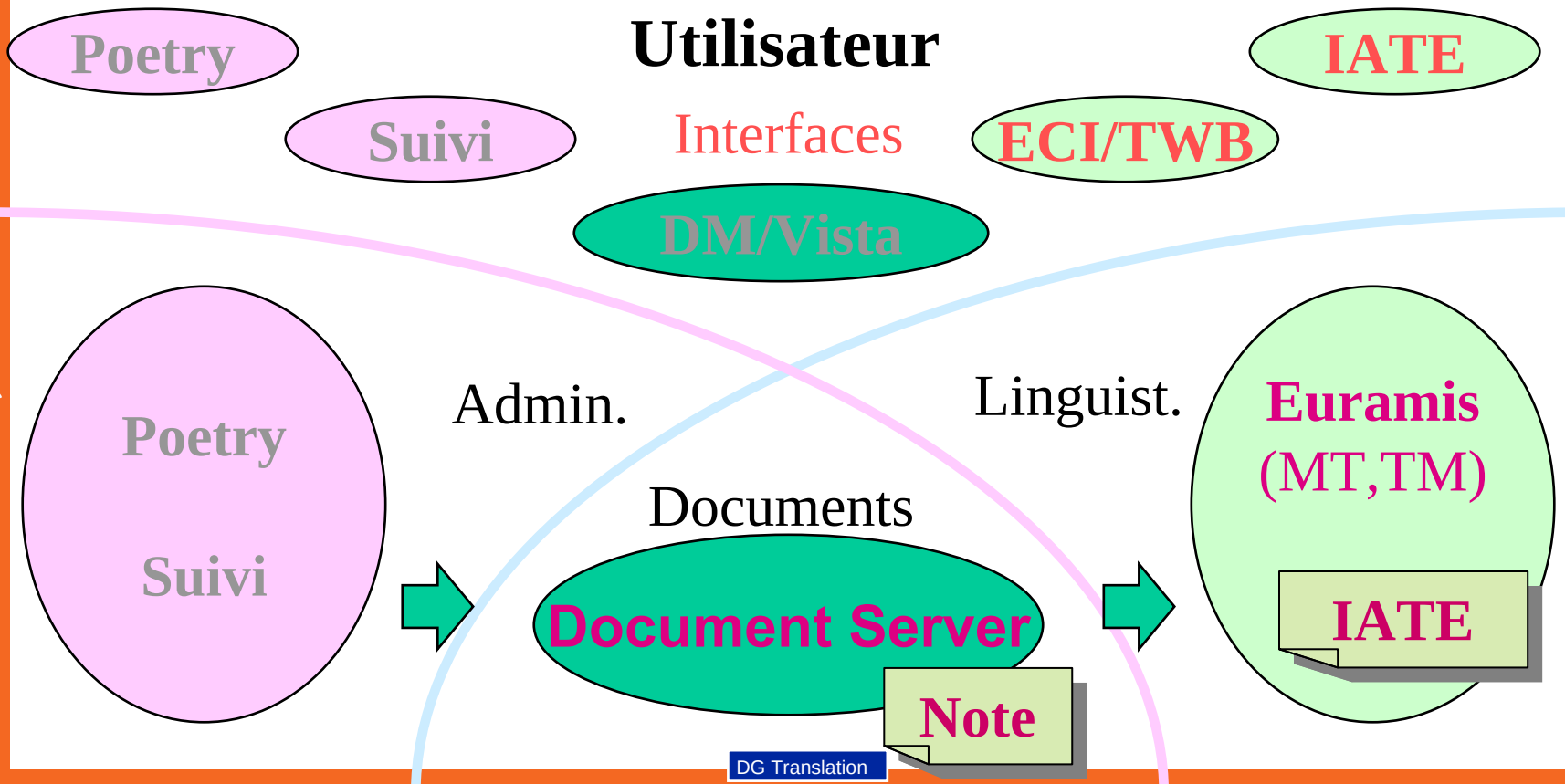
Les ressources

- Les services de traduction de l'UE créent des ressources multilingues
  - nos textes
  - nos mémoires de traduction
  - nos bases terminologiques

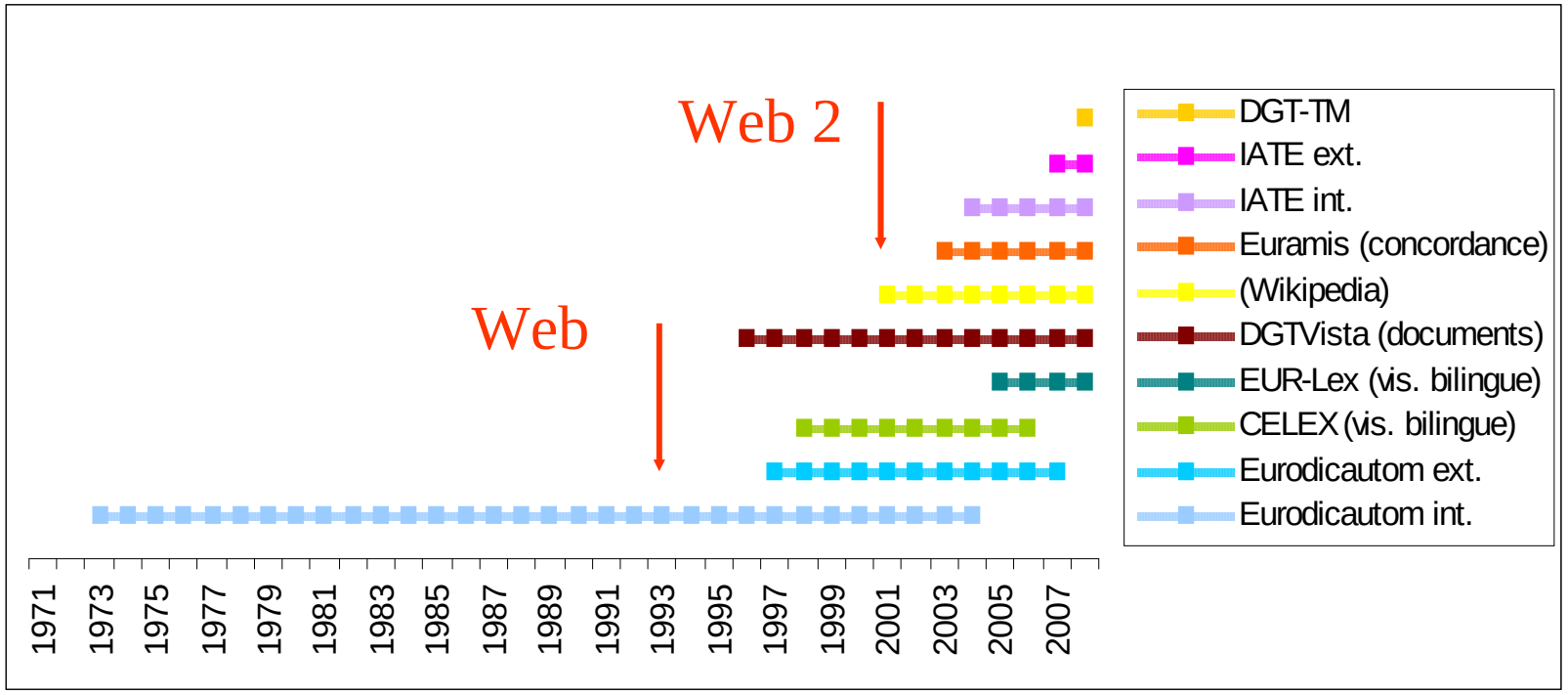
Les outils de recherche

- Informatisation du flux de documents dans les services de traduction de l'UE
  - interfaces internes
  - … interfaces externes

Flux d'information à la DGT



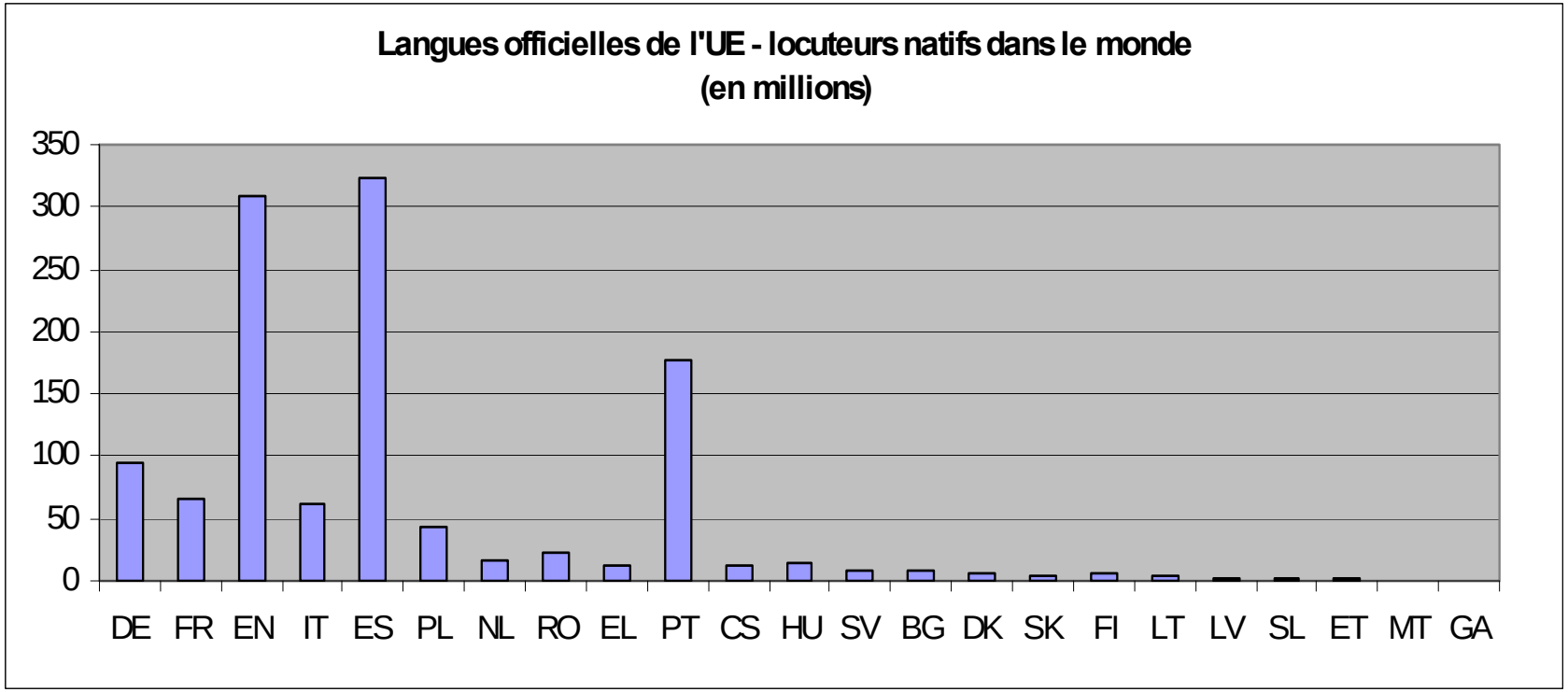
Les interfaces externes



L'accès aux ressources

- Les navigateurs Web
- … Google™ nous observe !!!
  - indexation des niveaux même les plus profonds des serveurs **europa.eu**

L'audience potentielle



1. Les textes de l'UE

- EUR-Lex
  - (CELEX - 1981-2004)
- Pre-Lex
- (Europa)
- ...
- occurrences dans les textes

EUR-Lex

http://eur-lex.europa.eu/ro/index.htm

- Législation de l'Union européenne
- Journal officiel de l'Union européenne
- … *inter alia*, traités, législation, jurisprudence et propositions législatives

  - accès Internet gratuit
  - contenus indexés par les moteurs de recherche
  - visualisation bilingue

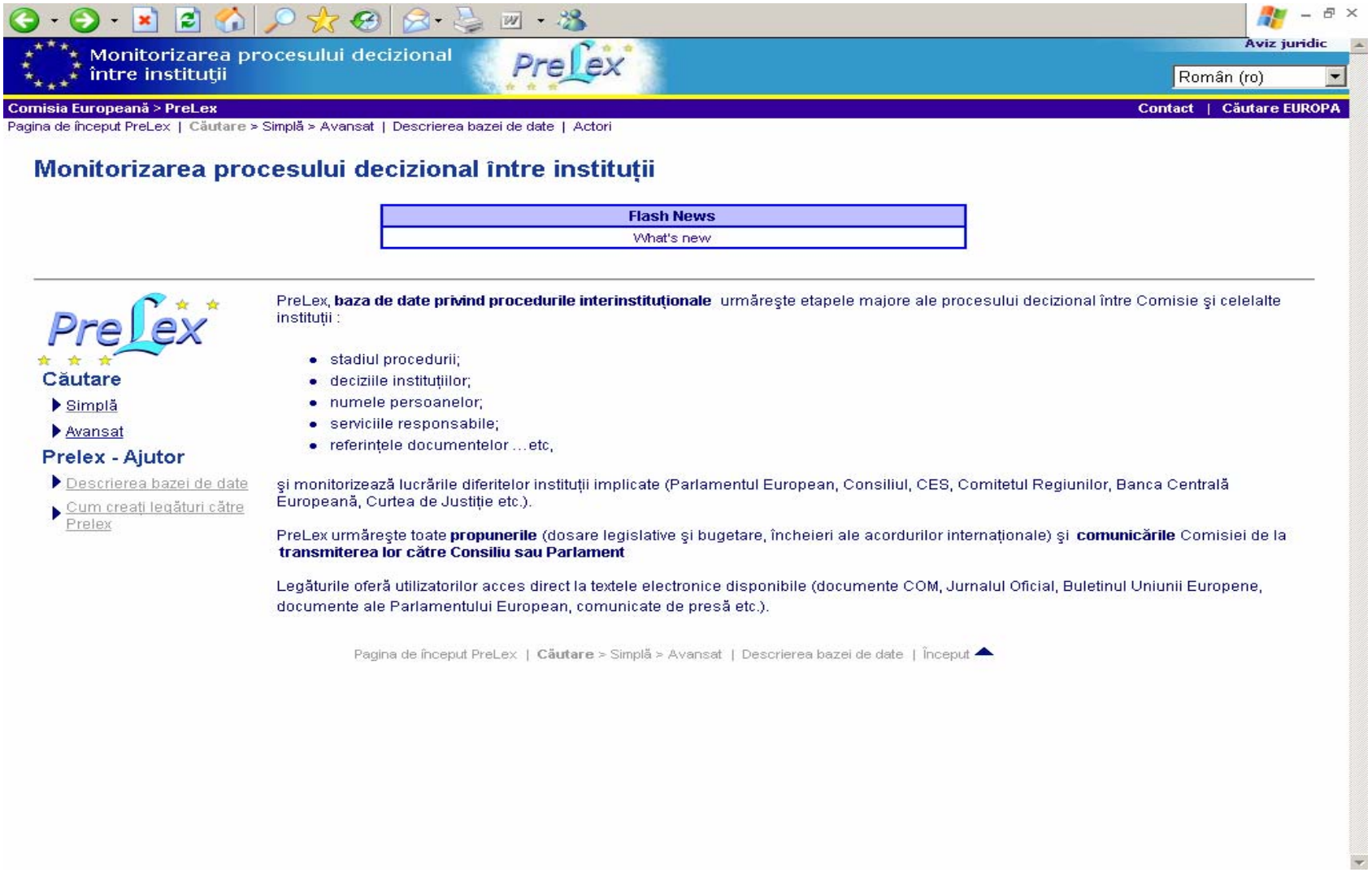

Accesul la legislația Uniunii Europene
Anunţ juridic important
BG ES CS DA DE ET EL EN FR GA IT LV LT HU MT NL PL PT RO SK SL FI SV
Harta site-ului | LexAlert | FAQ | Asistenţă | Contact | Legături
EUROPA > EUR-Lex
JURNALUL OFICIAL
CĂUTARE SIMPLĂ
CĂUTARE AVANSATĂ
COLECŢII
Tratate
Acorduri internaţionale
Legislaţie în vigoare
Acte pregătitoare
Jurisprudenţă
Întrebări parlamentare
ACCES DIRECT
Bugetul Uniunii Europene
Registrele instituţiilor
DESPRE LEGISLAŢIA UE
Proceduri şi participanţi
Tehnică legislativă
Due to a technical maintenance operation, EUR-LEX will be unavailable on Wednesday, January 23rd, from 19.00 (CET) to 19.45 (CET). Please accept our apologies for any inconvenience caused.
Bine aţi venit
EUR-Lex oferă acces direct gratuit la legislaţia Uniunii Europene. Sistemul permite consultarea Jurnalului Oficial al Uniunii Europene şi include, printre altele, tratatele, legislaţia, jurisprudenţa şi propunerile legislative. Oferă posibilităţi de căutare avansată. »
Noutăţi
Buletin informativ 15/01/2008 de en fr
Availability of the 2007 Special Edition »
Search by natural number for SEC documents and national implementing measures »
Permanent links for notices »
Selecţie de documente noi
16/01/2008: Decizia Comisiei din 12 decembrie 2007 privind punerea în aplicare a Sistemului de informare al pieţei interne (IMI) în ceea ce priveşte protecţia datelor cu caracter personal [notificată cu numărul C(2007) 6306] pdf
12/01/2008: Decizia Comisiei din 14 decembrie 2007 de instituire a Agenţiei executive a Consiliului European pentru Cercetare pentru gestionarea programului comunitar specific „Idei" în domeniul cercetării de frontieră, în temeiul Regulamentului (CE) nr. 58/2003 al Consiliului pdf
Jurnalul Oficial: AVIZ CĂTRE CITITORI
Dosare tematice
15/11/2007: Protecţia consumatorului: »
Jurnalul Oficial
Ultimele numere
L016
C014
Profilul personal
Identitatea utilizatorului:
Parolă:
Conexiune
Înscriere gratuită
Tratatul de la Lisabona
EXPERIMENTAL
Cumpăraţi Jurnalul Oficial în versiune CD-ROM
Start
8 Internet E...
10 Microsoft ...
2 Microsoft O...
Dicionário da Lí...
Microsoft Pow...
Microsoft Excel...
vademecum_PT
PT
19:04

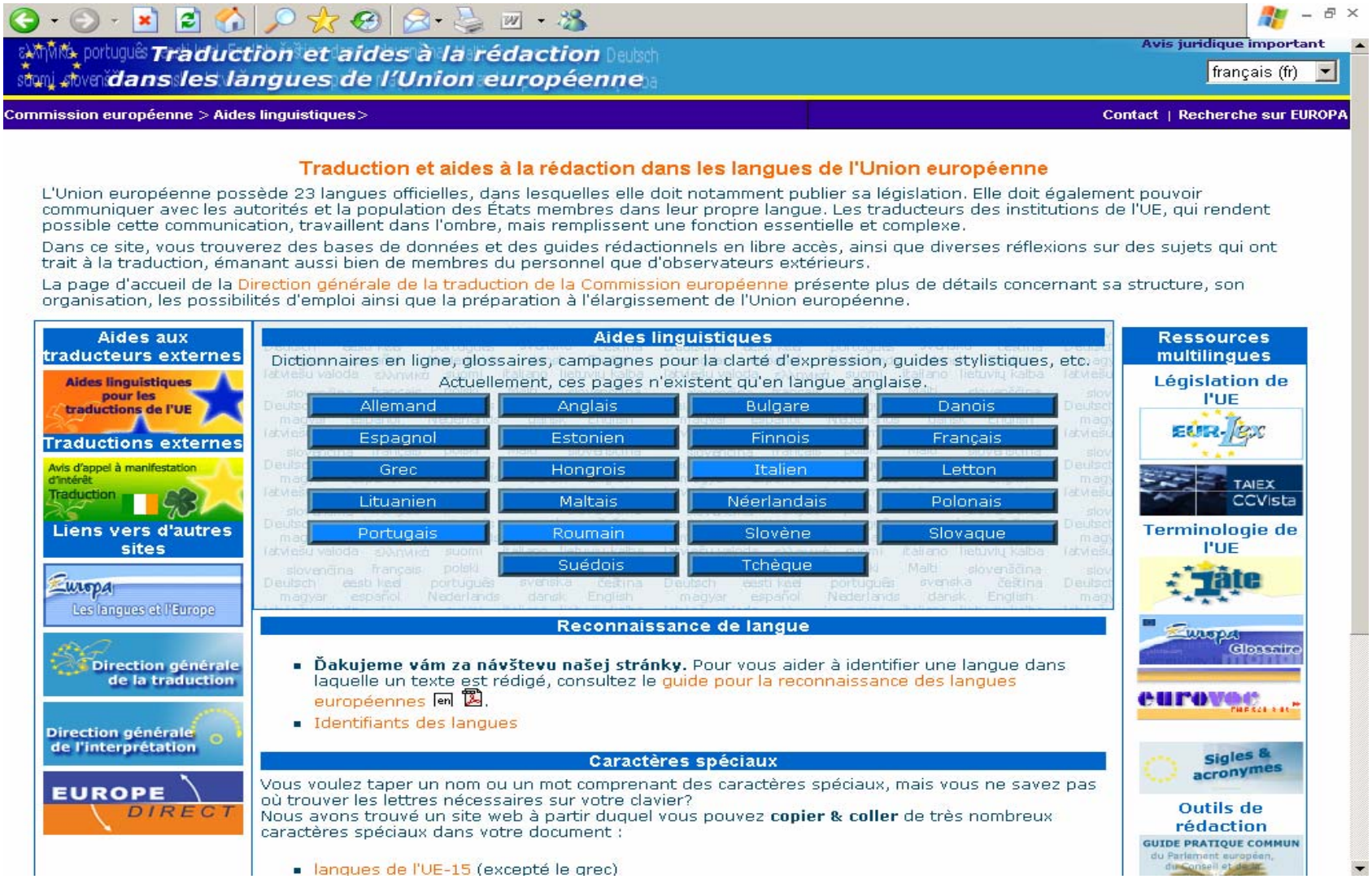

Anunţ juridic important
BG ES CS DA DE ET EL EN FR GA IT LV LT HU MT NL PL PT RO SK SL FI SV
Harta site-ului | LexAlert | FAQ | Asistenţă | Contact | Legături
EUROPA > EUR-Lex > Căutare simplă > Căutare nouă
Căutare nouă
Căutare generală
Termeni de căutare
Dată sau perioadă
Autor
Rubrici de clasificare
Cuvinte-cheie (EUROVOC)
Căutare după domeniu
Tratate
Legislaţie
Lucrări pregătitoare
Jurisprudenţă
Întrebări parlamentare
Căutare după numărul documentului
Număr natural
Număr CELEX
Text consolidat
Căutare după referinţa publicării
Jurnalul Oficial
Repertoriu de jurisprudenţă
Căutări anterioare
Istoricul căutărilor
Căutări înregistrate
Fişierele dumneavoastră
Gestionat de Oficiul pentru Publicaţii



Bilingual display

BG CS DA DE EL EN ES ET FI FR HU IT LT LV MT NL PL PT RO SK SL SV	BG CS DA DE EL EN ES ET FI FR HU IT LT LV MT NL PL PT RO SK SL SV
en	ro
20071213	20071213
Commission Decision	Decizia Comisiei
of 13 December 2007	din 13 decembrie 2007
amending Decision 2007/554/EC concerning certain protection measures against foot-and-mouth disease in the United Kingdom	de modificare a Deciziei 2007/554/CE privind anumite măsuri de protecție contra febrei aftoase din Regatul Unit
(notified under document number C(2007) 6256)	[notificată cu numărul C(2007) 6256]
(Text with EEA relevance)	(Text cu relevanță pentru SEE)
(2007/833/EC)	(2007/833/CE)
THE COMMISSION OF THE EUROPEAN COMMUNITIES,	COMISIA COMUNITĂȚILOR EUROPENE,
Having regard to the Treaty establishing the European Community,	având în vedere Tratatul de instituire a Comunității Europene,
Having regard to Council Directive 89/662/EEC of 11 December 1989 concerning veterinary checks in intra-Community trade with a view to the completion of the internal market [1], and in particular Article 9(4) thereof,	având în vedere Directiva 89/662/CEE a Consiliului din 11 decembrie 1989 privind controlul veterinar în cadrul schimburilor intracomunitare în vederea realizării pieței interne [1], în special articolul 9 alineatul (4),
Having regard to Council Directive 90/425/EEC of 26 June 1990 concerning veterinary and zootechnical checks applicable in intra-Community trade in certain live animals and products with a view to the completion of the internal market [2], and in particular Article 10(4) thereof,	având în vedere Directiva 90/425/CEE a Consiliului din 26 iunie 1990 privind controalele veterinare și zootehnice aplicabile în schimburile intracomunitare cu anumite animale vii și produse în vederea realizării pieței interne [2], în special articolul 10 alineatul (4),
Having regard to Council Directive 2003/85/EC of 29 September 2003 on Community measures for the control of foot-and-mouth disease repealing Directive 85/511/EEC and Decisions 89/531/EEC and 91/665/EEC, and amending Directive 92/46/EEC [3], and in particular Article 60(2) and Article 62(1) and (3) thereof,	având în vedere Directiva 2003/85/CE a Consiliului din 29 septembrie 2003 privind măsurile comunitare de combatere a febrei aftoase, de abrogare a Directivei 85/511/CEE și a Deciziilor 89/531/CEE și 91/665/CEE și de modificare a Directivei 92/46/CEE [3], în special articolul 60 alineatul (2) și articolul 62 alineatele (1) și (3),
Whereas:	întrucât:
(1) Following recent outbreaks of foot-and-mouth disease in Great Britain, Commission Decision 2007/554/EC of 9 August 2007 concerning certain protection measures against foot-and-mouth disease in the United Kingdom [4] was adopted to reinforce the control measures against foot-and-mouth disease taken by that Member State in the framework of Directive 2003/85/EC.	(1) În urma apariției recente a unor focare de febră aftoasă în Marea Britanie, a fost adoptată Decizia 2007/554/CE a Comisiei din 9 august 2007 privind anumite măsuri de protecție contra febrei aftoase din Regatul Unit [4] în vederea sporirii măsurilor de combatere a febrei aftoase luate de statul membru respectiv în cadrul Directivei 2003/85/CE.
(2) Decision 2007/554/EC lays down rules applicable to the dispatch from the high risk areas, listed in Annex I, and the low risk areas listed in Annex II, to that Decision ("restricted areas"), in Great Britain of products considered safe that either were produced before the restrictions were put in place in the United Kingdom, from raw material sourced from outside those restricted areas, or that	(2) Decizia 2007/554/CE stabilește norme aplicabile expedierii din zonele cu risc sporit, enumerate în anexa I, și din cele cu risc scăzut, enumerate în anexa II la decizia menționată (denumite în continuare "zone cu restricții"), de pe teritoriul Marii Britanii, a produselor considerate sigure care fie au fost fabricate înainte de introducerea restricțiilor în Regatul Unit din materii prime originare din afara


Monitorizarea procesului decizional între instituţii
PreLex
Aviz juridic
Român (ro)
Comisia Europeană > PreLex
Contact | Căutare EUROPA
Pagina de început PreLex | Căutare > Simplă > Avansat | Descrierea bazei de date | Actori
Monitorizarea procesului decizional între instituții
Flash News
What's new
PreLex
Căutare
Simplă
Avansat
Prelex - Ajutor
Descrierea bazei de date
Cum creați legături către Prelex
PreLex, baza de date privind procedurile interinstituționale urmăreşte etapele majore ale procesului decizional între Comisie şi celelalte instituții :
stadiul procedurii;
deciziile instituțiilor;
numele persoanelor;
serviciile responsabile;
referințele documentelor ...etc,
şi monitorizează lucrările diferitelor instituții implicate (Parlamentul European, Consiliul, CES, Comitetul Regiunilor, Banca Centrală Europeană, Curtea de Justiție etc.).
PreLex urmăreşte toate propunerile (dosare legislative şi bugetare, încheieri ale acordurilor internaționale) şi comunicările Comisiei de la transmiterea lor către Consiliu sau Parlament
Legăturile oferă utilizatorilor acces direct la textele electronice disponibile (documente COM, Jurnalul Oficial, Buletinul Uniunii Europene, documente ale Parlamentului European, comunicate de presă etc.).
Pagina de început PreLex | Căutare > Simplă > Avansat | Descrierea bazei de date | Început


Traduction et aides à la rédaction dans les langues de l'Union européenne
Avis juridique important
français (fr)
Commission européenne > Aides linguistiques>
Contact | Recherche sur EUROPA
Traduction et aides à la rédaction dans les langues de l'Union européenne
L'Union européenne possède 23 langues officielles, dans lesquelles elle doit notamment publier sa législation. Elle doit également pouvoir communiquer avec les autorités et la population des États membres dans leur propre langue. Les traducteurs des institutions de l'UE, qui rendent possible cette communication, travaillent dans l'ombre, mais remplissent une fonction essentielle et complexe.
Dans ce site, vous trouverez des bases de données et des guides rédactionnels en libre accès, ainsi que diverses réflexions sur des sujets qui ont trait à la traduction, émanant aussi bien de membres du personnel que d'observateurs extérieurs.
La page d'accueil de la Direction générale de la traduction de la Commission européenne présente plus de détails concernant sa structure, son organisation, les possibilités d'emploi ainsi que la préparation à l'élargissement de l'Union européenne.
Aides aux traducteurs externes
Aides linguistiques pour les traductions de l'UE
Traductions externes
Avis d'appel à manifestation d'intérêt Traduction
Liens vers d'autres sites
Europa Les langues et l'Europe
Direction générale de la traduction
Direction générale de l'interprétation
EUROPE DIRECT
Aides linguistiques
Dictionnaires en ligne, glossaires, campagnes pour la clarté d'expression, guides stylistiques, etc.
Actuellement, ces pages n'existent qu'en langue anglaise.
Allemand
Anglais
Bulgare
Danois
Espagnol
Estonien
Finnois
Français
Grec
Hongrois
Italien
Letton
Lituanien
Maltais
Néerlandais
Polonais
Portugais
Roumain
Slovène
Slovaque
Suédois
Tchèque
Reconnaissance de langue
Ďakujeme vám za návštevu našej stránky. Pour vous aider à identifier une langue dans laquelle un texte est rédigé, consultez le guide pour la reconnaissance des langues européennes
Identifiants des langues
Caractères spéciaux
Vous voulez taper un nom ou un mot comprenant des caractères spéciaux, mais vous ne savez pas où trouver les lettres nécessaires sur votre clavier?
Nous avons trouvé un site web à partir duquel vous pouvez copier & coller de très nombreux caractères spéciaux dans votre document :
langues de l'UE-15 (excepté le grec)
Ressources multilingues
Législation de l'UE
EUR-Lex
TAIEX
CCVista
Terminologie de l'UE
IATE
Europa Glossaire
euroVoc
Sigles & acronymes
Outils de rédaction
GUIDE PRATIQUE COMMUN

2. Les mémoires de traduction de la DGT

- DGT-TM
  - pour téléchargement
- ...

- occurrences dans des contextes (unités de traductions)

DGT-TM

http://langtech.jrc.it/DGT-TM.html

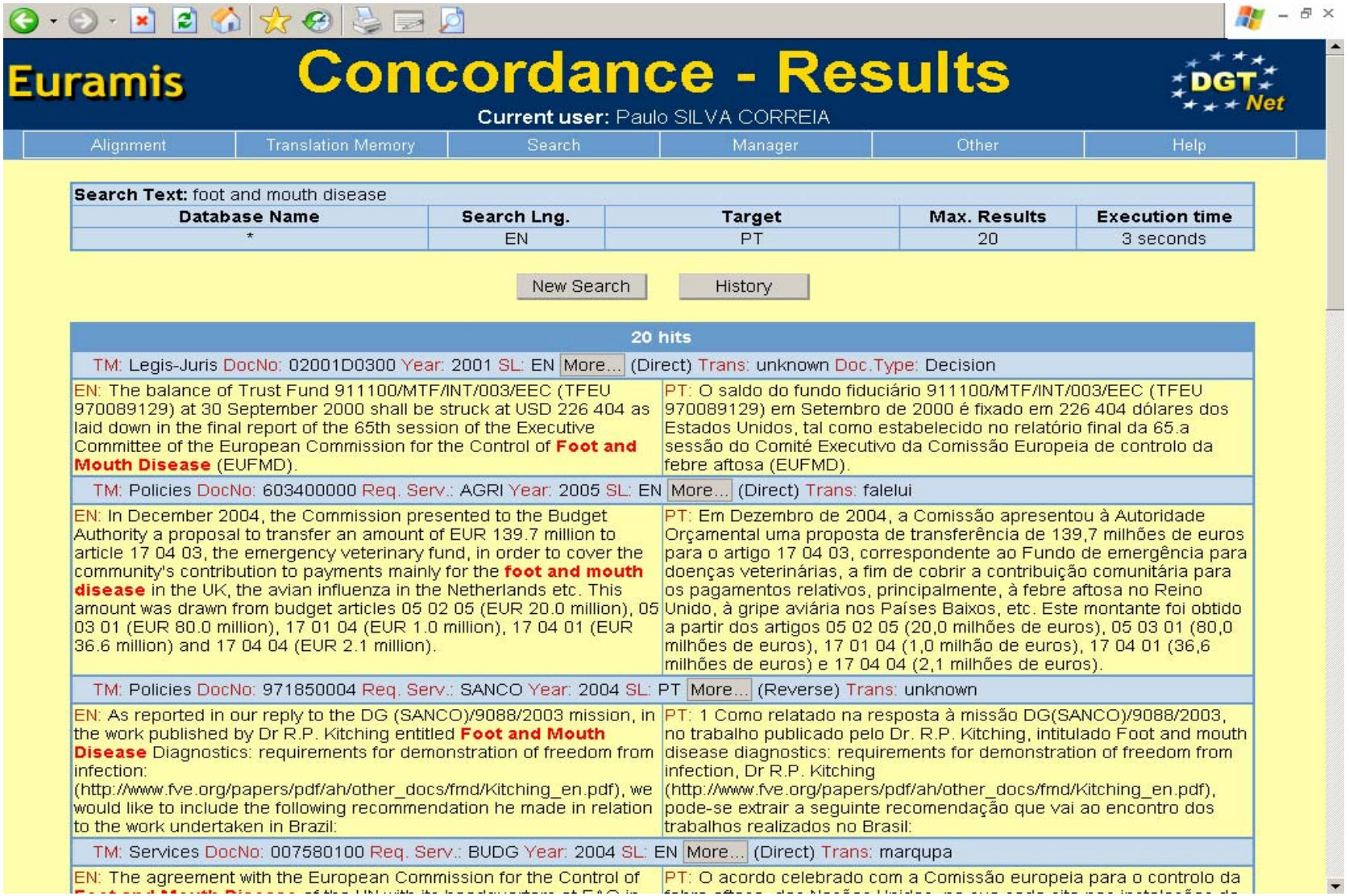
- dès janvier 2008: mémoires de traduction de l'aquis communautaire
  - accès Internet gratuit
  - contenus **non**-indexés par les moteurs de recherche
- occurrences en bi-textes

DGT-TM

- **Mémoires de traduction DGT de l'Acquis communautaire**

langues	unités de traduction
anglais	2 187 504
français	1 106 442
bulgare	708 658
roumain	650 735

- exemplo



Euramis **Concordance - Results**

DGT Net

Current user: Paulo SILVA CORREIA

Alignment | Translation Memory | Search | Manager | Other | Help

Search Text: foot and mouth disease

Database Name	Search Lng.	Target	Max. Results	Execution time
*	EN	PT	20	3 seconds

New Search | History

20 hits

TM: Legis-Juris DocNo: 02001D0300 Year: 2001 SL: EN More... (Direct) Trans: unknown Doc.Type: Decision

EN	PT
The balance of Trust Fund 911100/MTF/INT/003/EEC (TFEU 970089129) at 30 September 2000 shall be struck at USD 226 404 as laid down in the final report of the 65th session of the Executive Committee of the European Commission for the Control of **Foot and Mouth Disease** (EUFMD).	O saldo do fundo fiduciário 911100/MTF/INT/003/EEC (TFEU 970089129) em Setembro de 2000 é fixado em 226 404 dólares dos Estados Unidos, tal como estabelecido no relatório final da 65.a sessão do Comité Executivo da Comissão Europeia de controlo da febre aftosa (EUFMD).

TM: Policies DocNo: 603400000 Req. Serv.: AGRI Year: 2005 SL: EN More... (Direct) Trans: falelui

EN	PT
In December 2004, the Commission presented to the Budget Authority a proposal to transfer an amount of EUR 139.7 million to article 17 04 03, the emergency veterinary fund, in order to cover the community's contribution to payments mainly for the **foot and mouth disease** in the UK, the avian influenza in the Netherlands etc. This amount was drawn from budget articles 05 02 05 (EUR 20.0 million), 05 03 01 (EUR 80.0 million), 17 01 04 (EUR 1.0 million), 17 04 01 (EUR 36.6 million) and 17 04 04 (EUR 2.1 million).	Em Dezembro de 2004, a Comissão apresentou à Autoridade Orçamental uma proposta de transferência de 139,7 milhões de euros para o artigo 17 04 03, correspondente ao Fundo de emergência para doenças veterinárias, a fim de cobrir a contribuição comunitária para os pagamentos relativos, principalmente, à febre aftosa no Reino Unido, à gripe aviária nos Países Baixos, etc. Este montante foi obtido a partir dos artigos 05 02 05 (20,0 milhões de euros), 05 03 01 (80,0 milhões de euros), 17 01 04 (1,0 milhão de euros), 17 04 01 (36,6 milhões de euros) e 17 04 04 (2,1 milhões de euros).

TM: Policies DocNo: 971850004 Req. Serv.: SANCO Year: 2004 SL: PT More... (Reverse) Trans: unknown

EN	PT
As reported in our reply to the DG (SANCO)/9088/2003 mission, in the work published by Dr R.P. Kitching entitled **Foot and Mouth Disease** Diagnostics: requirements for demonstration of freedom from infection: (http://www.fve.org/papers/pdf/ah/other_docs/fmd/Kitching_en.pdf), we would like to include the following recommendation he made in relation to the work undertaken in Brazil:	1 Como relatado na resposta à missão DG(SANCO)/9088/2003, no trabalho publicado pelo Dr. R.P. Kitching, intitulado Foot and mouth disease diagnostics: requirements for demonstration of freedom from infection, Dr R.P. Kitching (http://www.fve.org/papers/pdf/ah/other_docs/fmd/Kitching_en.pdf), pode-se extrair a seguinte recomendação que vai ao encontro dos trabalhos realizados no Brasil:

TM: Services DocNo: 007580100 Req. Serv.: BUDG Year: 2004 SL: EN More... (Direct) Trans: marqupa

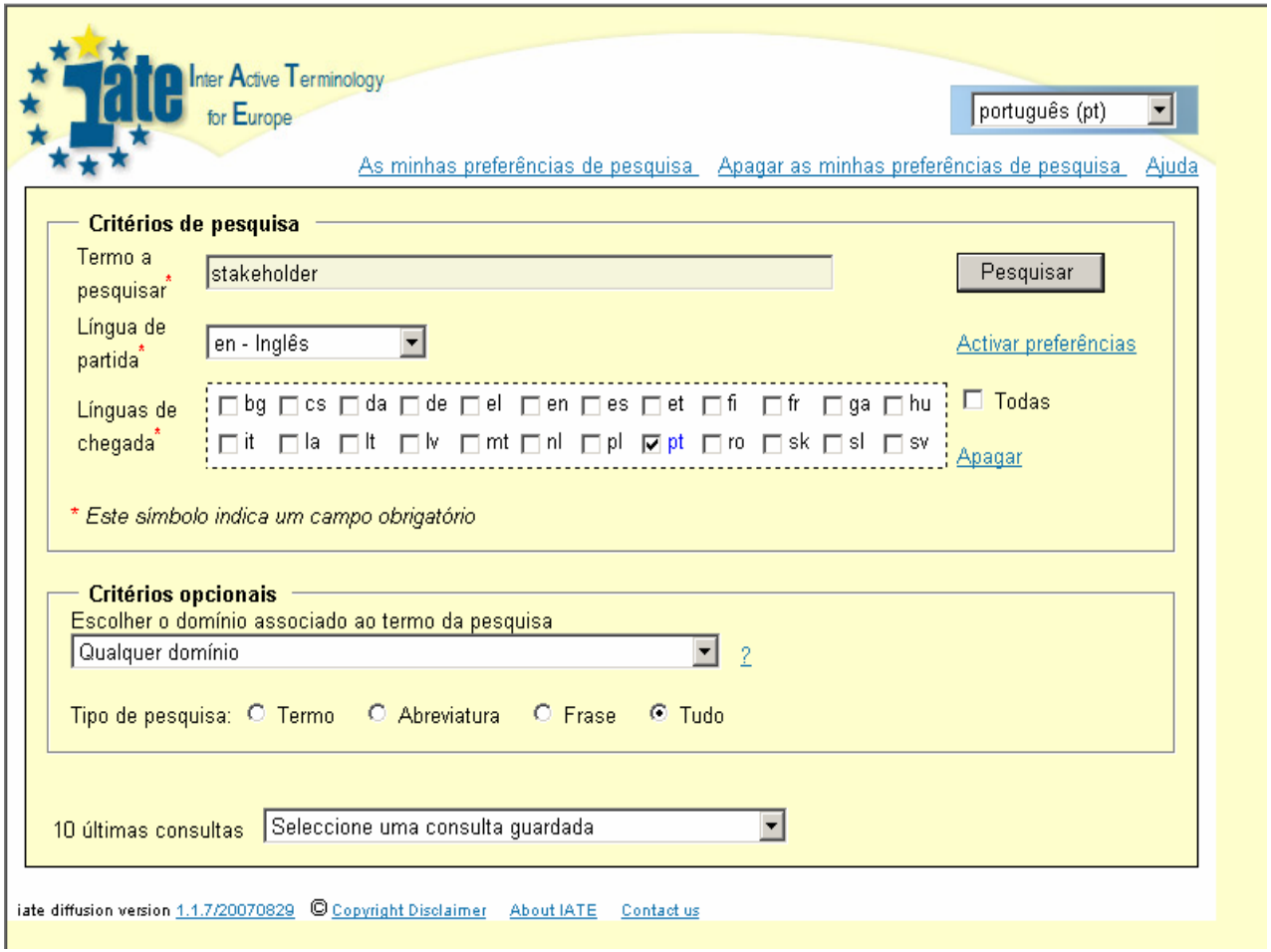
EN	PT
The agreement with the European Commission for the Control of	O acordo celebrado com a Comissão europeia para o controlo da

3. La base terminologique de l'UE

- IATE
  - accès Internet gratuit
  - contenus **non**-indexés par les moteurs de recherche

- fiches terminologiques

http://iate.europa.eu



IATE

Inter-Active Terminology for Europe

- dès 2004: base terminologique interinstitutionnelle
- dès mars 2007: ouverture au public (gratuit)

- ressource terminologique

IATE – définition

- base terminologique **interactive** et **multilingue** pour la rédaction, la traduction et l'interprétation dans les domaines communautaires
- données avec **valeur ajoutée** par rapport à d'autres ressources linguistiques

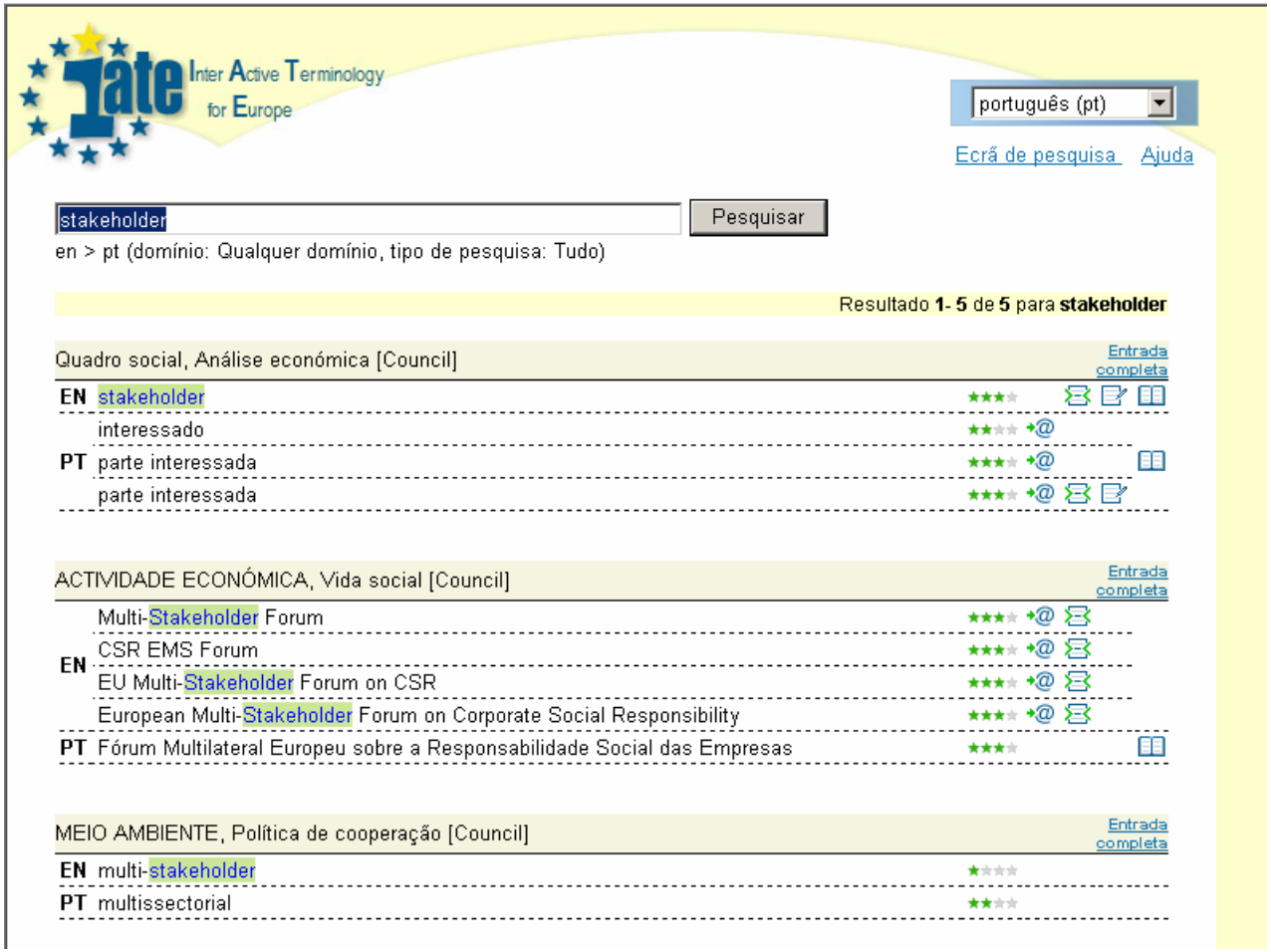
IATE – valeur ajoutée

- multilinguisme (plein)
- redondance controllée
- définitions des concepts
- références crédibles
- …fiabilité accrue

- … **validation** par des terminologues
  - responsables par les nouveaux contenus

IATE - contenus

- solutions terminologiques dans les langues officielles:
  - «jargon communautaire»;
  - langages de spécialité;
  - nomenclatures;
  - ...
- tendanciellement multilingue



Inter Active Terminology for Europe

português (pt)

Ecrã de pesquisa Ajuda

stakeholder Pesquisar

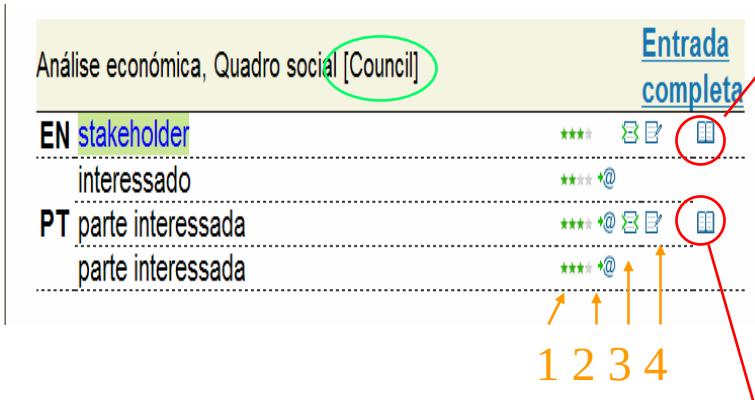
en > pt (domínio: Qualquer domínio, tipo de pesquisa: Tudo)

Resultado **1**- **5** de **5** para **stakeholder**

Quadro social, Análise económica [Council]		Entrada completa
EN	stakeholder	
	interessado	
PT	parte interessada	
	parte interessada	

ACTIVIDADE ECONÓMICA, Vida social [Council]		Entrada completa
	Multi-Stakeholder Forum	
EN	CSR EMS Forum	
	EU Multi-Stakeholder Forum on CSR	
	European Multi-Stakeholder Forum on Corporate Social Responsibility	
PT	Fórum Multilateral Europeu sobre a Responsabilidade Social das Empresas	

MEIO AMBIENTE, Política de cooperação [Council]		Entrada completa
EN	multi-stakeholder	
PT	multissectorial	



1 – fiabilité (terme)

2 – référence (terme)

3 – contexte

4 – note (terme)

en

Definição	An individual, group, institution, or government with an interest or concern, either economic, societal, or environmental, in a particular measure, proposal, or event.
Refª definição	The Dictionary of Global Climate Change, W.J. Maunder
Nota	Shareholders, lenders, employees, suppliers, communities are examples of stakeholders in the private sector; citizens and taxpayers are examples of stakeholders in government. Source: Environmental auditing and the role of the accounting profession (Termium)
Termo	stakeholder
Fiabilidade	3 (Fiável)
Contexto	"Fourteen Member States reported having a policy in place to enable all actors and stakeholders to be involved in the evaluation process, but very few present concrete examples of participation and involvement."
Refª contexto	Report from the Commission to the European Parliament and the Council on the implementation of the Council Recommendation of 18 June 2003 on the prevention and reduction of health-related harm associated with drug dependence (COM/2007/0199 final) 52007DC0199
Nota sobre o termo	See also http://www.google.be/search?hl=nl&as_qdr=all&defl=en&q=define:Stakeholder&sa=X&oi=glossary_definition&ct=title 26.04.2007
Data	15/05/2007

pt

Definição	Pessoa ou organização envolvida, de forma activa ou passiva, num processo de acção social que pode ser de natureza económica, política, cívica, etc. e ter um carácter permanente (como a actividade de uma empresa) ou temporário (como a realização de um projecto). No contexto empresarial a noção de "parte interessada" ou "interessado", muitas vezes associada à noção de "responsabilidade social", é um elemento central de certas concepções modernas de gestão em que o interesse dos accionistas ou dos administradores não é alheio ou independente do interesse dos trabalhadores, dos consumidores ou clientes e da comunidade em que se insere uma determinada empresa.
Refª definição	TERMPT - Conselho
Nota	LPS (034 FEV00); MC (ABR07) [CT 08.05.2007]
Termo	interessado
Fiabilidade	2 (Pouco fiável)
Refª termo	TERMPT - Conselho
Data	10/05/2007
Termo	parte interessada
Fiabilidade	3 (Fiável)
Refª termo	GTT (Comissão-PT)
Contexto	Todos os envolvidos num processo, por exemplo, clientes, colaboradores, investidores, fornecedores, comunidade, etc. O processo em questão pode ser de carácter temporário (como um projecto) ou duradouro (como o negócio de uma empresa ou a missão de uma organização sem fins lucrativos).
Refª contexto	Wikipédia, pt.wikipedia.org/wiki... [27-07-2006]
Nota sobre o termo	Em função do contexto, outras variantes são possíveis: «parte envolvida», «interessado», «interveniente», etc.
Data	10/05/2007
Termo	parte interessada
Fiabilidade	3 (Fiável)
Refª termo	TERMPT - Conselho
Data	10/05/2007

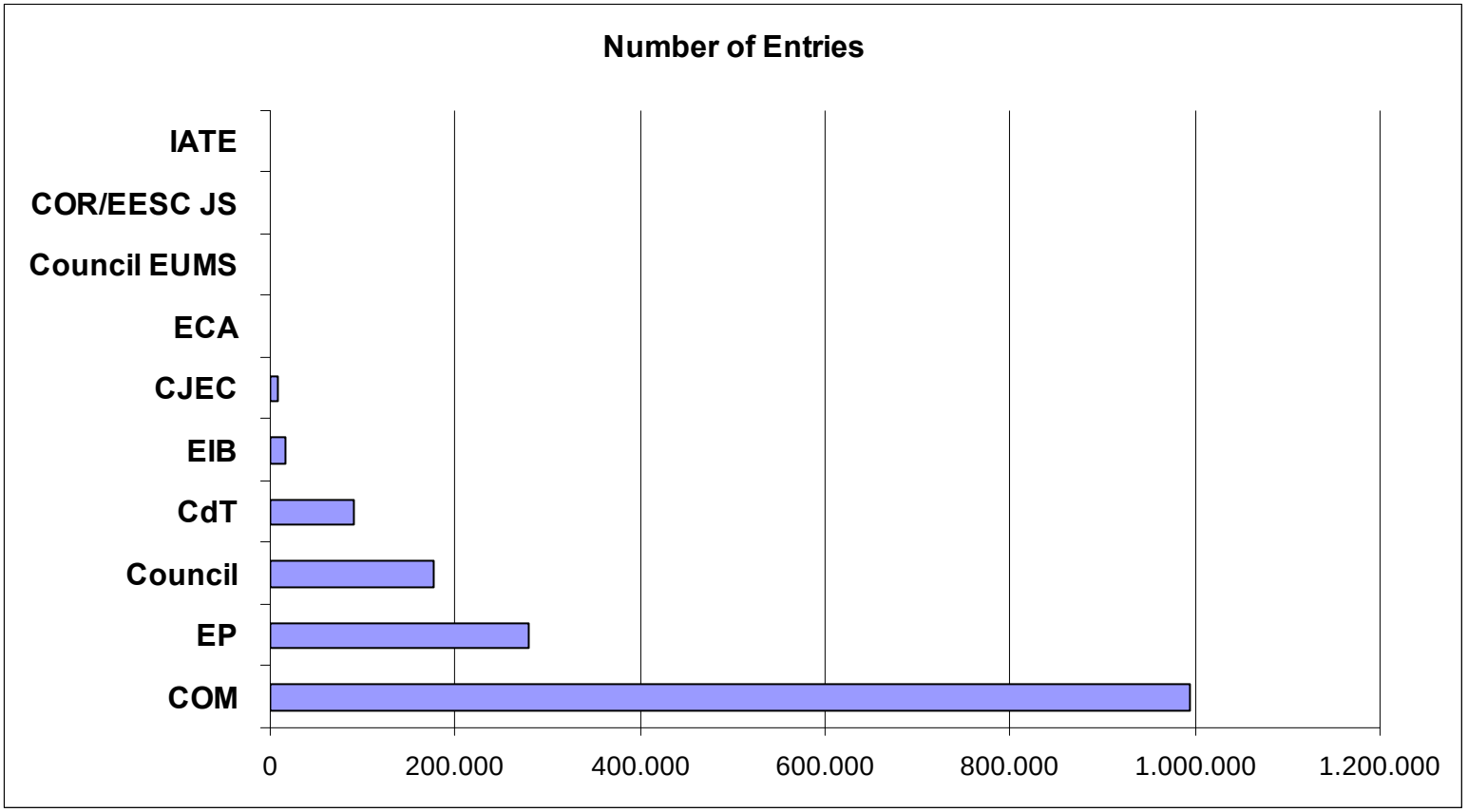
1 2 3 4

Termo	**interessado**
Fiabilidade	2 (Pouco fiável)
Refª termo	TERMPT - Conselho
Data	10/05/2007
Termo	**parte interessada**
Fiabilidade	3 (Fiável)
Refª termo	TERMPT - Conselho
Data	10/05/2007
Termo	**parte interessada**
Fiabilidade	3 (Fiável)
Refª termo	GTT (Comissão-PT)
Contexto	Todos os envolvidos num processo, por exemplo, clientes, colaboradores, investidores, fornecedores, comunidade, etc. O processo em questão pode ser de carácter temporário (como um projecto) ou duradouro (como o negócio de uma empresa ou a missão de uma organização sem fins lucrativos).
Refª contexto	Wikipédia, pt.wikipedia.org/wiki... [27-07-2006]
Nota sobre o termo	Em função do contexto, outras variantes são possíveis: «parte envolvida», «interessado», «interveniente», etc.
Data	10/05/2007

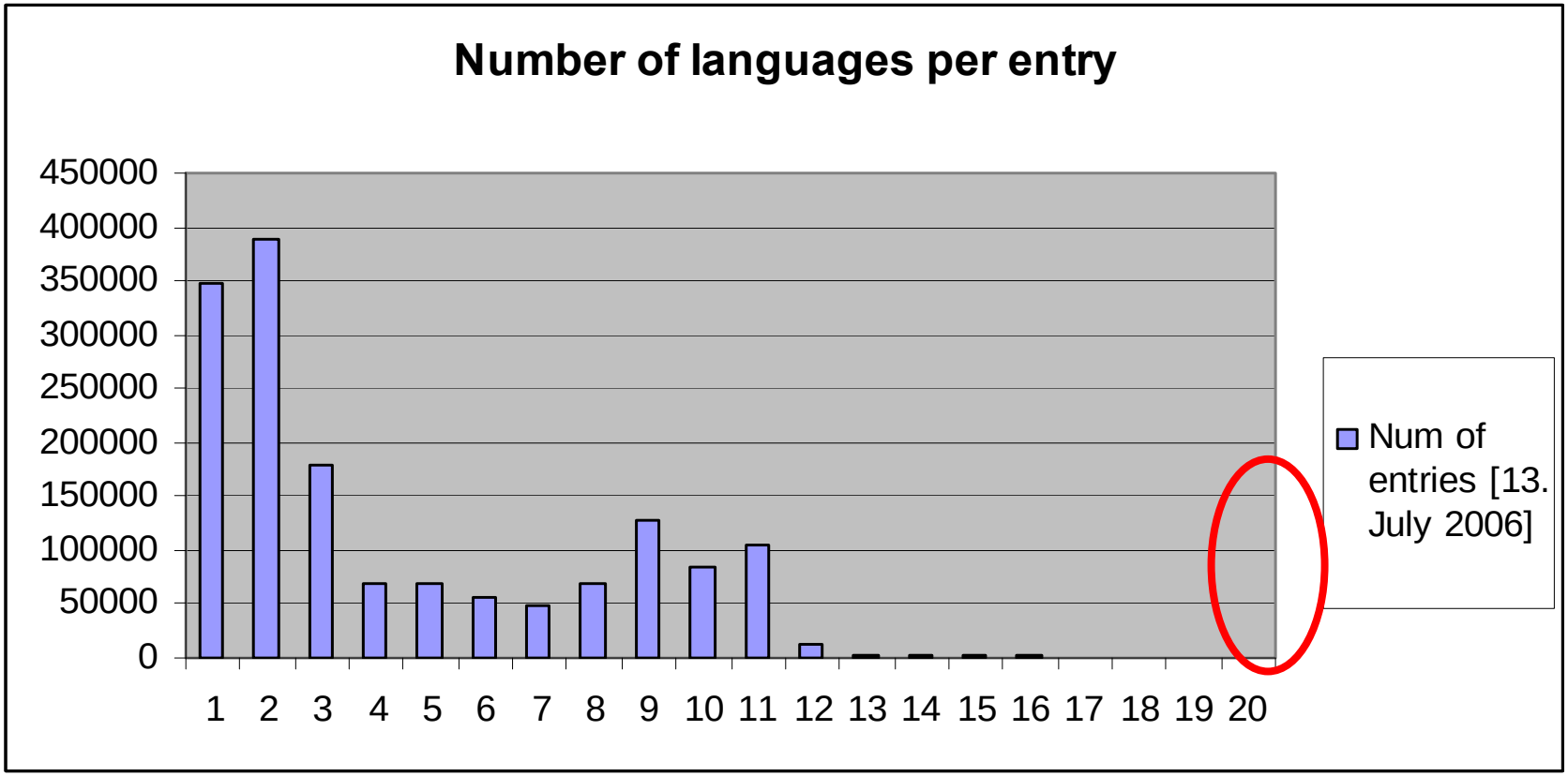
Fonte: Council IATE ID: 900023

Entrada seguinte

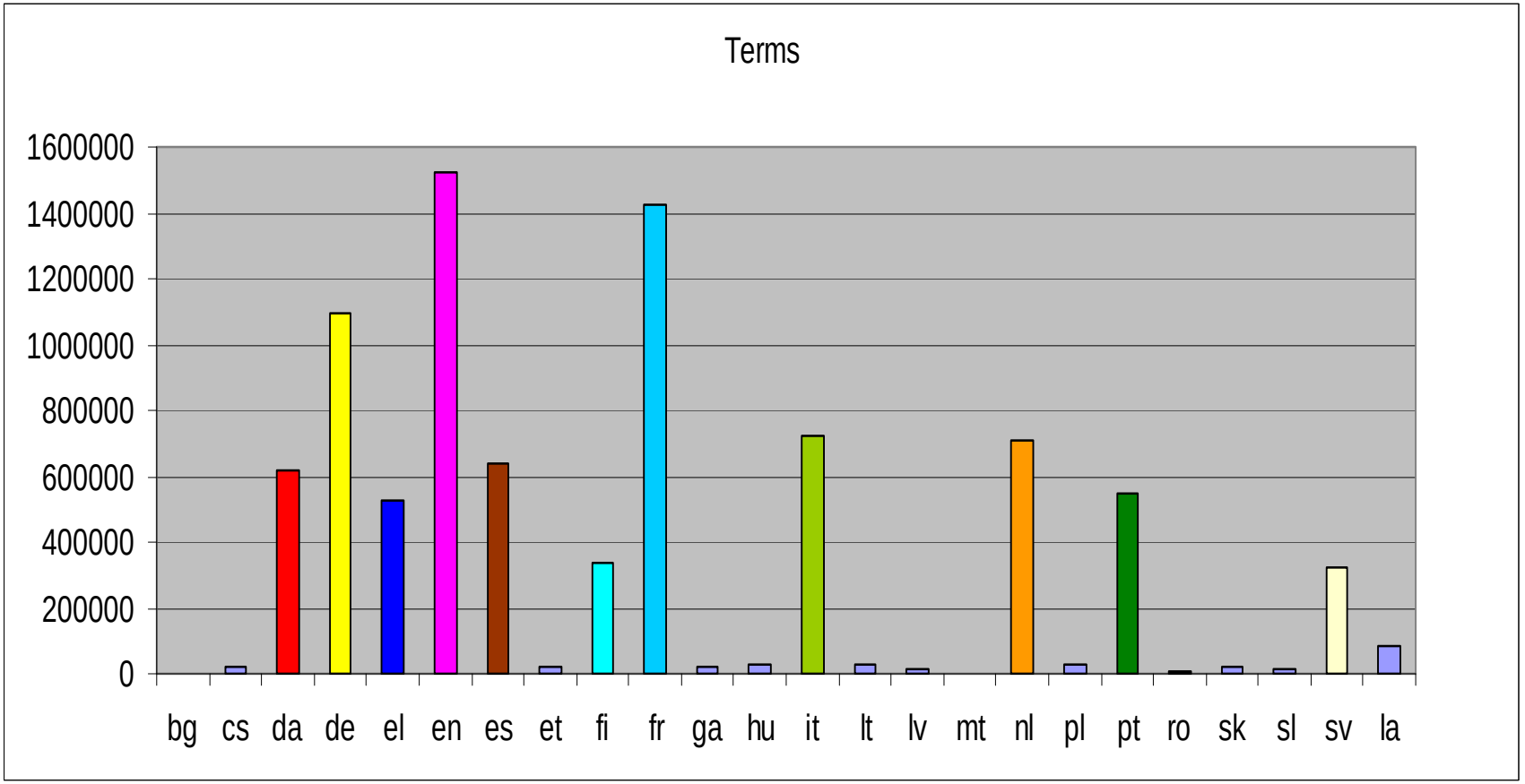
IATE – fiches par instituition



IATE – langues par fiche



IATE- termes par langues



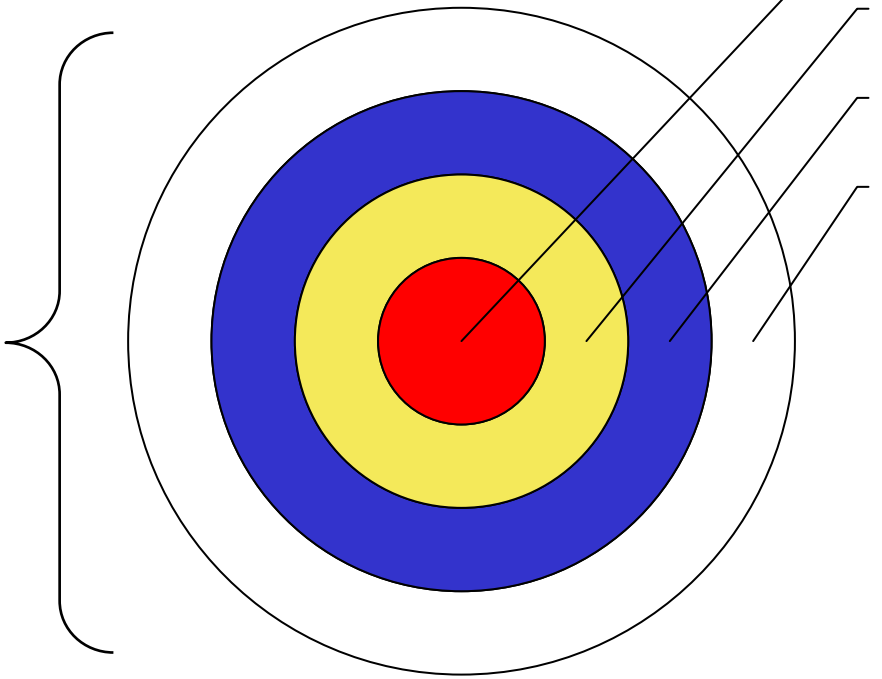
IATE - Consolidation

- bruit
  - répétition des termes pour certaines langues (UE à 15)
- silence
  - absence de contenus pour plusieures langues (UE à 27)
- … besoin de consolidation des contenus
  - activité interinstitutionnelle

IATE "Top 10" (2006)

- general approach
- progress report
- stakeholder
- governance
- accountability
- impact assessment
- benchmark
- policy debate
- financial framework
- law enforcement

La continuité de l'information



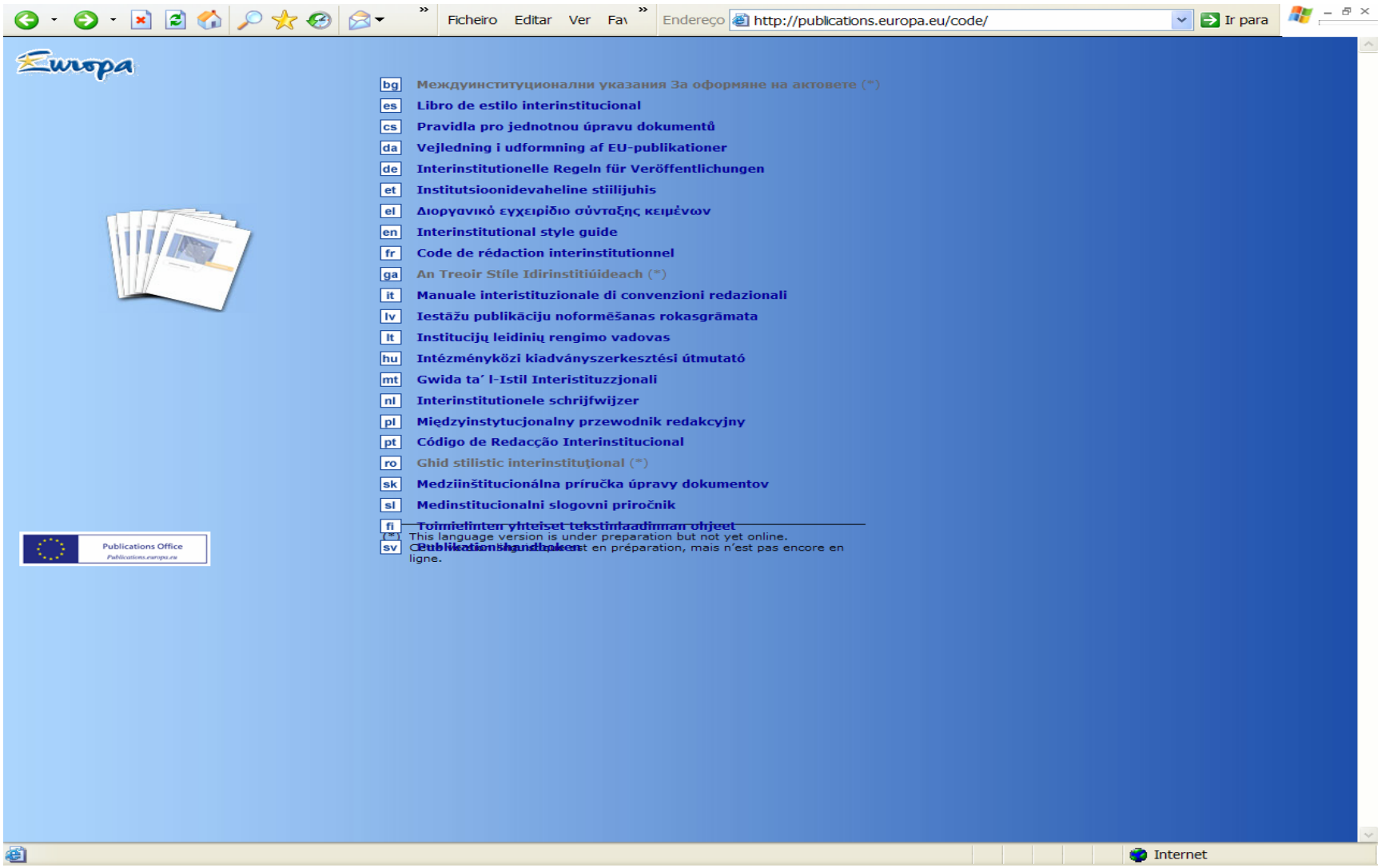
Terminologie (IATE)

Mém. de traduction (DGT-TM)

Textes (Eur-Lex, Pre-Lex, ...)

autres

+ Codes de rédaction interinstitutionnels


Ficheiro Editar Ver Fav
Endereço http://publications.europa.eu/code/
Ir para
Europa
bg Международни указания За оформяне на актовете (*)
es Libro de estilo interinstitucional
cs Pravidla pro jednotnou úpravu dokumentů
da Vejledning i udformning af EU-publikationer
de Interinstitutionelle Regeln für Veröffentlichungen
et Institutsioonidevaheline stiilijuhis
el Διοργανικό εγχειρίδιο σύνταξης κειμένων
en Interinstitutional style guide
fr Code de rédaction interinstitutionnel
ga An Treoir Stíle Idirinstitiúideach (*)
it Manuale interistituzionale di convenzioni redazionali
lv Iestāžu publikāciju noformēšanas rokasgrāmata
lt Institucijų leidinių rengimo vadovas
hu Intézményközi kiadványszerkesztési útmutató
mt Gwida ta' l-Istil Interistituzzjonali
nl Interinstitutionele schrijfwijzer
pl Międzyinstytucjonalny przewodnik redakcyjny
pt Código de Redacção Interinstitucional
ro Ghid stilistic interinstituțional (*)
sk Medziinštitucionálna príručka úpravy dokumentov
sl Medinstitucionalni slogovni priročnik
fi Toimielinten yhteiset tekstinlaadinnan ohjeet
sv Publikationshandboken
(*) This language version is under preparation but not yet online.
Cette publication est en préparation, mais n'est pas encore en ligne.
Publications Office
Internet

Paulo.Correia@ec.europa.eu



Un corpus paralel român-englez

FORĂSCU Corina, Facultatea de Informatică, Universitatea "Al.I.Cuza", Iași - Institutul de Cercetări pentru Inteligență Artificială, Academia Română, București

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

1. Introducere: Informație temporală în limbajul natural

Privind un corpus ca o colecție de documente selectate și ordonate conform unor criterii lingvistice stabilite, el permite punerea în evidență, informarea și fundamentarea unor teorii lingvistice specializate. Cum la ora actuală engleza este limba cu cea mai densă realizare de documente adnotate, ea este deseori utilizată ca sursă din care să se transfere adnotări specifice asupra altor limbi. Pentru limba română o serie de corpusuri paralele au fost deja create (Cristea, Forăscu, 2006), însă niciunul care să permită evidențierea informației temporale.

Teoria logicii temporale s-a dovedit extrem de relevantă încă din anii '70, mai ales în Inteligența Artificială. Posibilitatea de a identifica și analiza informația temporală este de mare importanță pentru multe dintre aplicațiile Prelucrării Limbajului Natural precum traducerea automată (prin transferarea și normalizarea referințelor temporale precum și mapările dintre timpurile verbelor), rezumarea multi-document, sistemele de întrebare răspuns, regăsirea și extragerea informației, etc.

Informația temporală este reprezentată în limbajul natural prin (Mani et al., 2005):

- expresii temporale exprimate prin grupuri nominale, prepoziționale sau adverbiale - ore (timp al zilei), date, durate: *acum trei ore, mai 1984, anii 90, 5 februarie 2007*, etc.; aceste expresii temporale referă timpul ca:
  - punct (moment): Am deschis ușa la *ora doisprezece*.,
  - interval: *Am fost plecat ieri.*
- expresii ce denotă evenimente exprimate în principal, pe lângă adjective, clauze predicative sau grupuri frazale prepoziționale, prin:
  - propoziții, mai exact prin centrul (eng. *head*) sintactic, anume verbul principal: *Ion a plecat la munte.*
  - grupuri nominale: *Greva va continua și în zilele următoare.*

Expresiile ce denotă timpul pot avea:

- referințe explicite (specificate), care referă la o intrare într-un sistem calendaristic / orar: *amiază, 11.10.2006 (midday, 11.10.2006)* ;
- referințe implicite (sub-specificate) - pot fi evaluate doar prin intermediul unui timp indexat: *anul viitor, săptămâna trecută, acum două ore*;
- referințe vagi (nespecificate, neancorate), care nu pot fi corelate cu un punct sau interval exact de timp: *după-amiază, în câteva săptămâni, acum câteva zile*.

Evenimentele exprimate prin verbe pot fi temporal ancorate:

- indirect, prin categoria morfologică a timpului și
- direct, prin modificatori adverbiali (adverbe de timp și frecvență, grupuri nominale și prepoziționale și clauze subordonate).

Pentru a codifica toate tipurile de expresii temporale, evenimente şi relaţii între acestea, a fost creat standardul TimeML (Pustejovsky et al., 2006), ale cărui fundamente s-au pus încă din 2002. Standardul reuneşte multe dintre eforturile anterioare de adnotare temporală, diferind de acestea prin separarea reprezentării evenimentelor şi a expresiilor temporale de legăturile de ancorare, ordonare sau dependenţă ce apar în texte.

Standardul TimeML defineşte 7 etichete:

- pentru evenimente şi instanţe ale acestora
  - EVENT: arată situaţii care se întâmplă, stări, circumstanţe în care ceva se întâmplă sau se menţine adevărat: *Female pilots were held up until now by the lack of piloting opportunities for them in the military.*
  - MAKEINSTANCE: marchează câte instanţieri sau realizări are un eveniment; marcajul include şi categoriile de timp şi aspect ale verbului: *But they still have catching up to do two hundred and thirty four Americans have flown in space, only twenty six of them women.*
- pentru expresii şi semnale temporale:
  - TIMEX3: indică momente ale zilei, date şi durate: *25 October; nowadays*.
  - SIGNAL: marchează modul în care se leagă obiectele temporale: *after, until*.
- pentru legături între evenimente sau între evenimente şi expresii temporale:
  - TLINK: indică 13 tipuri de relaţii temporale (corespunzătoare celor definite de Allen)
  - ALINK: indică o legatură aspectuală între un verb aspectual şi argumentul acestuia.
  - SLINK: indică o legatură de subordonare între două evenimente.

De provenienţă lingvistică, teoriile temporale au fost studiate şi formalizate cu predilecţie pentru limba engleză. Lucrarea prezintă cercetările efectuate pentru obţinerea corpusului paralel TimeBank, englez-român, care să fie folosit, printr-un import automat, la validarea acestor teorii pentru limba română. Secţiunea a doua prezintă corpusul TimeBank, realizarea versiunii româneşti a corpusului şi prelucrările corpusului paralel (alinierea versiunilor în română şi engleză, şi corectarea acesteia). Importul adnotărilor temporale şi al celor adiţionale din TimeBank 1.2 şi o analiză preliminară a acestui import sunt prezentate în capitolul 3. Ultimul capitol prezintă concluziile şi obiectivele viitoare ale cercetărilor.

2. Corpusul paralel TimeBank - englez-român

2.1. Corpusul TimeBank

Realizarea corpusului TimeBank a început în 2002 în cadrul proiectului şi atelierului TERQAS [1]. Corpusul conţine în versiunea actuală 183 de fişiere de rapoarte de ştiri în limba engleză, adnotate conform cu TimeML v.1.2. (Pustejovsky et al., 2006). Corpusul este de dimensiuni mici, întrucât nu a fost creat ca un scop în sine, ci pentru a testa validitatea şi a demonstra aplicabilitatea limbajului TimeML. Documentele au fost selectate astfel încât să conţină multă informaţie temporală; ele conţin şi alte marcaje XML, inclusiv formatul documentelor, informaţie structurală, nume de entităţi, marcaje de propoziţie.

TimeBank 1.2. este versiunea actuală a corpusului, conformă cu specificaţiile TimeML 1.2.1. Structura şi adnotarea corpusului sunt, în esenţă, aceleaşi cu cele din prima versiune a corpusului. TimeBank 1.2 este distribuit prin LDC [2] (Pustejovsky et al., 2006). Statisticile extrase automat din TimeBank 1.2 sunt ilustrate în Tabelul 1.

Adnotarea documentelor a început cu o fază de preprocesare, când unele articole lexicale de tip evenimente şi semnale au fost marcate cu unele clase, timpuri sau aspecte ale acestora. După această etapă 5 adnotatori umani au completat adnotarea,

verificând şi paşii de preprocesare şi corectitudinea în conformitate cu noua specificaţie TimeML 1.2.1.

Tabelul 1: Statistici asupra corpusului TimeBank 1.2

TimeML tags	#
events	7935
Instances	7940
timexes	1414
signals	688
alinks	265
slinks	2932
tlinks	6418
TOTAL	27592

General	#
propoziţii	4715
unităţi lexicale	61042
unităţi lexicale unice	10586

Adnotarea temporală iniţială a corpusului este considerată „preliminară" întrucât s-a arătat (Boguraev, Ando, 2006) că apar greşeli sistematice datorate dimensiunii relativ reduse a corpusului şi inconsistenţelor în adnotare: legături temporale sau de subordonare inconsistente sau incomplete, clasificarea evenimentelor - în perfectare, adnotare incompletă a timpului şi aspectului unor evenimente.

2.2. Corpusul TimeBank - versiunea în limba română

Textul englez, fără marcajele XML, a fost iniţial distribuit la două masterande la Lingvistică Computaţională, Facultatea de Informatică Iaşi. Un set minimal de recomandări de traducere a avut ca scop obţinerea unei traduceri unitare între traducători, o exprimare cursivă în limba română şi mai ales facilitarea unei alinieri lexicale optime între engleză şi română.

- Propoziţiile în română sunt traduse într-o corespondenţă 1:1, ori de câte ori limba o permite. Alinierea la propoziţii se obţine direct prin traducere.
- Se folosesc echivalenţi de traducere cu aceeaşi parte de vorbire, cuvintele româneşti fiind cât mai "apropiate" de corespondentele lor englezeşti: atunci când cuvântul englez are un cognate în română, acesta din urmă va fi folosit (*sporadic -> sporadic* şi nu *mai rar*), evitându-se parafrazările.
- Se traduc toate cuvintele şi nu se introduc în traducere, din motive stilistice, cuvinte sau expresii fără corespondent în engleză.
- Se foloseşte scrierea cu diacritice, conformă cu normele în vigoare ale limbii.
- Timpurile verbelor se păstrează pe cât posibil, modificările fiind acceptate doar pe temeiuri lingvistice şi nu stilistice.
- Se păstrează formatul folosit în engleză pentru date, momente ale zilei şi numere.

Varianta actuală în limba română a fost verificată manual, mai ales din cauza unor inconsistenţe şi lipsuri în traduceri, care nu ar fi permis o aliniere a unor elemente temporale esenţiale.

În cele 4715 propoziţii în corpusul în limba română sunt 65375 unităţi lexicale (inclusiv semne de punctuaţie), din care 12640 sunt unice.

2.3. Adnotări ale corpusurilor TimeBank englez şi român

În vederea alinierii lexicale, s-a folosit o preadnotare unitară a textelor paralele care să fie utilă aliniatorului lexical YAWA (Tufiş et al., 2006). Această procesare preliminară se referă la segmentarea la nivel de cuvânt, adnotarea cu etichete morfosintactice şi lematizarea textelor în engleză şi română. Modulul TTL (Ion, 2007) oferă aceste adnotări şi în plus, asigură o reprezentare uniformă a textelor adnotate în termenii codificării corpusului paralel într-un format XML similar cu formatul XCES (Ide et al., 2000).

Segmentarea la nivel de cuvânt consideră că spaţiul nu este singurul delimitator de cuvinte şi nici nu este întotdeauna delimitator de cuvinte. Atât în engleză cât şi în română există expresii idiomatice care vor trebui considerate ca unităţi lexicale în procesul de aliniere

Adnotarea morfosintactică se face cu ajutorul unui adnotator probabilistic care implementează adnotatorul TnT (Brants, 2000) bazat pe Modele Markov Ascunse; varianta folosită a TnT permite identificarea părţii de vorbire a cuvintelor necunoscute, mai exact cuvintele din clasele deschise. Setul de etichete morfosintactice este compatibil cu specificaţiile MULTEXT-East [3] fapt care permite reprezentarea uniformă a informaţiei morfosintactice în engleză şi în română.

Lematizorul implementat în TTL este de asemenea unul probabilistic. O lemă candidată se generează pe baza unei mulţimi de reguli extrase automat dintr-un lexicon care conţine pentru fiecare formă ocurenţă a unui cuvânt, lema şi eticheta morfosintactică a acesteia. Lema unei noi forme ocurenţe a unui cuvânt de o etichetă morfosintactică dată este lema cea mai probabilă dintre toate lemele candidate după Modelul Markov al tuturor lemelor de aceeaşi etichetă din lexicon (Ion, 2007).

Tot ca o cerinţă a alinierii lexicale, în corespondenţă *n : m*, s-au recunoscut, folosind expresii regulate peste secvenţe de etichete morfosintactice, grupuri nominale şi prepoziţionale nerecursive, complecşi verbali, adjectivali şi adverbiali.

2.4. Alinierea lexicală a corpusurilor

Alinierea lexicală a corpusului paralel a fost realizată cu YAWA (Tufiş et al., 2006) pe ieşirea modulului TTL. Corpusul paralel TimeBank 1.2 a fost aliniat la nivel de unitate lexicală din română în engleză urmându-se patru faze specifice acestui aliniator. În fiecare dintre primele trei etape YAWA adaugă noi legături la cele create în etapa anterioară, fără a şterge nimic:

1. alinierea cuvintelor conţinut (din clasele deschise) folosind un dicţionar de echivalenţi de traducere extras automat (Tufiş, Barbu, 2002); după această etapă precizia este mare însă recall-ul este îmbunătăţit în continuare: P = 94.08%, R = 34.99%, F = 51.00%.

2. aliniere la nivel de grup sintactic: se aliniază cuvintele aflate în acelaşi grup sintactic cu cuvintele deja aliniate la pasul anterior, utilizându-se reguli de aliniere. De exemplu, un substantiv românesc aliniat la unul englezesc care este precedat de un determinator, aliniază determinatorul englezesc la substantivul românesc (*fata - the girl*). Evaluarea alinierii după această etapă: P = 89.90%, R = 53.90%, F = 67.40%.

3. aliniere în secvenţe continue de cuvinte nealiniate: pe scheletul alinierii anterioare, se aliniază toate blocurile de indecşi consecutivi rămaşi nealiniaţi (Tufiş et al., 2006);

4. pasul de corecţie: corectează, chiar prin ştergerea unor legături, alinierea de la pasul 3.

La încheierea alinierii evaluarea YAWA este: P = 88.80%, R = 74.83%, F = 81.22%

Fazele 2 şi 4 sunt evident dependente de perechea de limbi aliniate dar regulile de aliniere şi cele de corecţie nu sunt integrate în corpul aliniatorului astfel încât să poată fi schimbate atunci când se doreşte alinierea altei perechi de limbi.

În urma alinierii automate română-engleză a 181 de fişiere dintre cele 183 ale bitextului iniţial, au rezultat 91714 de corespondenţe, din care 25346 sunt alinieri la NULL (Forăscu, Ion, 2006). Neincluderea în aliniere a două fişiere a avut la bază lipsurile şi greşelile de traducere.

Pentru a obţine un import optim al marcajelor temporale din corpusul original englezesc, alinierile au fost verificate manual. Majoritatea greşelilor s-au datorat unei delimitări incorecte a unor cuvinte, în special a numeralelor şi valorilor numerice, unei marcări morfo-sintactice perfectibile în cazul pronumelor posesive, a negaţiilor şi a unor adjective din engleză.

3. Marcarea corpusului TimeBank românesc

3.1. Importul automat

Deoarece traducerea din engleză a corpusului a păstrat structura XML propoziţională, a fost posibilă parsarea corpusului englezesc şi pentru fiecare marcaj XML de propoziţie, conţinutul acestuia a fost înlocuit cu traducerea lui în română. Deşi traducerile au folosit un set de reguli suficient de restrictive, în cadrul unei propoziţii traduse în română ordinea cuvintelor se va schimba faţă de cea în engleză. Din acest motiv a fost folosită alinierea din română în engleză pentru a importa marcajul XML în sens invers.

Importul marcajelor se face pentru orice pereche (S_{ro}; S_{en}) din corpusul paralel TimeBank în care T_{en} este propoziţia echivalentă (T_{en} are acelaşi text brut ca şi S_{en}, însă T_{en} are structura XML pe care urmărim să o importăm (Forăscu & all, 2007):

Figura 1 Transfer al marcajului XML din română în engleză

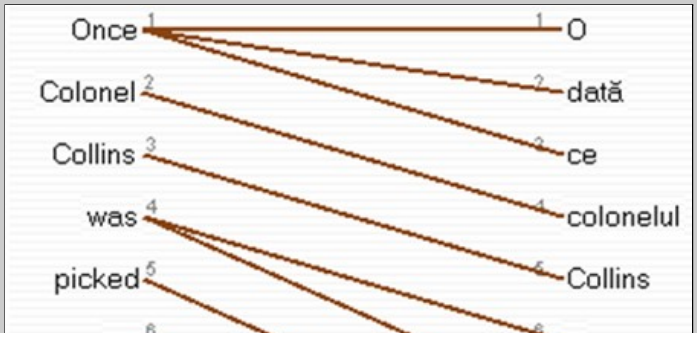
română	engleză
<s> <SIGNAL sid="s72">O dată ce</SIGNAL> <ENAMEX TYPE="PERSON">colonelul Collins</ENAMEX> a fost <EVENT class="I_ACTION" eid="e22">aleasă</EVENT> ca <EVENT class="STATE" eid="e52">astronaut</EVENT> <ENAMEX TYPE="ORGANIZATION"> NASA</ENAMEX> </s>	<s> <SIGNAL sid="s72">Once</SIGNAL> <ENAMEX TYPE="PERSON">Colonel Collins</ENAMEX> was <EVENT eid="e22" class="I_ACTION">picked</EVENT> as a <ENAMEX TYPE="ORGANIZATION">NASA</ENAMEX> <EVENT eid="e52" class="STATE">astronaut</EVENT> </s>

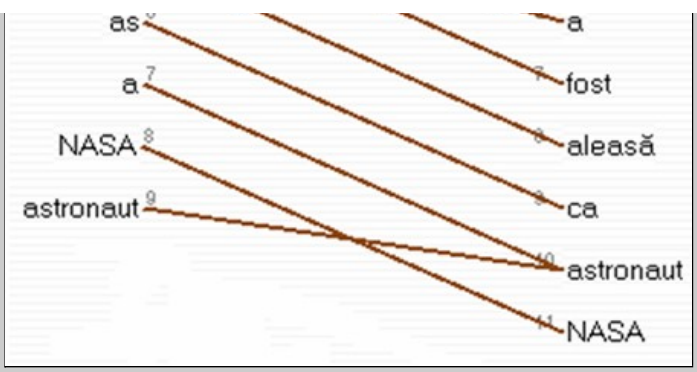
- se construieşte lista **E** a perechilor formate din fragmente de text englezesc şi şirul de indecşi ai fragmentelor din S_{en}, în T_{en}. Deoarece tokenizarea în **S**en este uneori diferită de cea din T_{en}, se foloseşte această listă pentru a mapa fragmente englezeşti din T_{en} pe secvenţe de indecşi din S_{en}, astfel încât să fie folosită alinierea lexicală dinspre română a acestor indecşi.
  De exemplu - pentru textul ilustrat în figurile 1 şi 2 lista **E** = *{<"Once";1>, <"Colonel Collins";2,3>, <"was";4>, <"picked"; 5>, <"as";6>, <"a";7>, <"NASA";8>, <"astronaut";9>}*.

- se adaugă la fiecare element al listei **E** contextul XML în care apare.
  Folosind exemplul anterior, lista devine **E** = *{<"Once";1; s,SIGNAL>, <"Colonel Collins";2,3; s,ENAMEX>, <"was";4; s>, <"picked"; 5; s,EVENT>, <"as";6; s>, <"a";7; s>, <"NASA";8; s,ENAMEX>, <"astronaut";9; s,EVENT>}*.
  Simultan cu marcajul XML, se păstrează şi atributele acestuia (când există).

- se construieşte lista **RW** de cuvinte corespondente româneşti şi marcaje XML transferate, folosind alinierea lexicală dintre S_{ro} şi S_{en}. Dacă un cuvânt din S_{ro} nu este aliniat (aliniat cu cuvântul de index 0), contextul maximal (s) este folosit. Lista devine **RW** = *{<"O dată ce";1,2,3; s,SIGNAL>, <"Colonel Collins";4,5; s,ENAMEX>, <"a fost";6,7; s>, <"aleasă"; 8; s,EVENT>, <"ca";9; s>, <"astronaut";10; s,EVENT>, <"NASA";11; s,ENAMEX> }*.

Figura 2 Alinierea lexicală a secvenţei din figura 1





- se construieşte lista finală **R** de fragmente de text româneşti din **RW** folosind elementele adiacente din **RW** care apar în acelaşi context XML. Lista este întoarsă în format XML - figura 1.

Un document ce include marcaje TimeML este privit ca fiind compus din trei părţi: partea de început a documentului (header), textul propriu-zis (ce conţine marcajele inline) şi partea finală ce conţine marcajele offline, cele de instanţiere a evenimentelor şi de legătură dintre elementele temporale ale textului. Procedura de import automat descrisă mai sus are în vedere doar marcajele inline - cele din header şi text. În cazul importului marcajelor offline s-a ţinut cont că acestea folosesc atribute ale marcajelor inline. Astfel, s-au importat automat doar acele marcaje offline pentru care atributele aparţin unor elemente deja importate în română. Tabelul 2 conţine o statistică asupra importului automat astfel realizat.

Tabelul 2 Statistici la importul marcajelor în corpusul TimeBank 1.2. român

Marcaje TimeML	#	% transfer
EVENT	7703	97.07
MAKEINSTANCE	7706	97.05
TIMEX3	1356	95.89
SIGNAL	668	97.09
ALINK	249	93.96
SLINK	2831	96.55
TLINK	6122	95.38
TOTAL	**26635**	**96.53**

3.2. Analiza marcajelor importate

Rezultatele obţinute prin importul automat sunt promiţătoare. Însă pentru a atinge unul dintre obiectivele iniţiale - crearea unui corpus paralel englez- român adnotat temporal în ambele limbi - am început validarea manuală, simultan cu evaluarea importului temporal pe corpus. Am avut în vedere analiza distinctă a patru situaţii de transfer al marcajelor:

1. Transfer perfect.

2. Transfer cu modificări ulterioare - atunci când în cele două limbi construcţiile temporale nu sunt identice, însă transferul se poate realiza prin reguli specifice.

3. Transfer cu modificări specifice fenomenelor lingvistice române, ca de exemplu cliticele, intercalarea conjuncţiilor şi adverbelor între verbele unui acelaşi grup verbal.

4. Transfer imposibil.

Simultan cu validarea manuală a adnotărilor importate în română din TimeBank, am avut în vedere şi urmărirea tuturor elementelor temporale din corpusul englezesc. Aşa cum se va vedea, unele construcţii temporale nemarcate în varianta engleză au fost identificate şi propuse spre validare creatorilor TimeBank.

Analiza preliminară a folosit zece fişiere ale corpusului, reprezentând aproximativ 10% din acesta. Fişierele marcate au fost urmărite în paralel (pentru română şi engleză), fără a interveni şi la marcajele offline (MAKEINSTANCE, ALINK, TLINK şi SLINK), întrucât acestea au fost importate automat doar dacă elementele pe care le identifică sau leagă erau deja marcate ca atare în textul românesc: o legătura temporală TLINK

este importată dacă marcajele elementele ai căror identificatori îi foloseşte sunt deja importate.

Tabelul 3 rezumă toate cele patru situaţii întâlnite în analiza importului automat

Tabelul 3. Situaţiile obţinute la importul automat

Marcaje / Transfer	EVENT	TIMEX3	SIGNAL
Perfect	785	33	29
Cu modificări	37	3	-
Bazat pe fenomene specifice limbii	3	-	-
Imposibil	4	-	4
TOTAL (engleză)	829	36	33

Modificările necesare la importul automat în cazul marcajului EVENT (37 cazuri - 4.5%) se datorează regulii TimeML care impune ca în cazul grupurilor sintactice, marcajul să se pună doar pe elementul principal (head-ul) al construcţiei. Acesta este cazul verbelor reflexive (în unele cazuri marcajul EVENT a inclus automat şi pronumele reflexiv), verbelor compuse din română şi locuţiunilor verbale.

Fenomenele specifice limbii care trebuie luate în considerare la import sunt: intercalarea adverbelor şi conjuncţiilor între verbele ce formează un verb compus: ***also*** *said* - *au* ***mai*** *spus*; *(he)* ***also*** *criticised* - *a* ***şi*** *criticat*, situaţii în care marcajul EVENT s-a pus separat pe verbul auxiliar, creând astfel două „evenimente" identice ca identificatori, dar diferite ca lexicalizare; fenomenul de cliticizare a fost corect tratat la import.

Situaţiile de transfer imposibil se datorează lipsei traducerilor: *forces that harbor ill intentions - forţe străine cu intenţii rele* - situaţie în care lipsa traducerii nu schimbă înţelesul; în cazul alinierii ***Poliţia a descoperit corpul (dezmembrat al) unui bărbat răpit -*** *Police discover dismembered body of man kidnapped* lipsa traducerii modifică înţelesul. Cazuri imposibile de transfer sunt şi cele în care traducerea este corectă, însă în română evenimentele din engleză nu sunt lexicalizate. Din cele 40 de evenimente neimportate automat doar patru au rămas nemarcate în urma validării manuale, celelalte fiind corectate.

În ceea ce priveşte marcajul TIMEX3 - s-a întâlnit o situaţie de lipsă a alinierii: expresia temporală *some time* nu a inclus în alinierea cu limba română *un timp mai lung* grupul adjectival *mai lung*. Celelalte modificări de luat în considerare au în vedere demarcarea prepoziţiilor ca părţi din TIMEX3, ca de exemplu în: *eight years (war) - (războiul)* ***de*** *opt ani*. Pentru cele patru marcaje TIMEX3 nemarcate automat s-a făcut corecţia manual.

Singura situaţie întâlnită de imposibilitate de transfer a marcajului SIGNAL are la bază nnelexicalizarea în limba română a prepoziţiei din engleză, în exprimări de genul: *on Tuesday - marţi*, unde prepoziţia *on* este marcată ca SIGNAL în engleză.

Activitatea de validare şi corectare a permis identificarea unor elemente temporale (încă) nemarcate în corpusul TimeBank. Pentru marcajul EVENT am identificat 104 elemente noi: 5 din clasa REPORTING (*say, said*), 21 de tip stare (STATE) (*belongs, look, ceiling, staying, war, policies*), 70 din clasa OCCURENCE (substantive: *missions, training, fight, (mediation) effort, demarcation, move*, dar şi verbe: *supervising, leading, include*), 1 de tip I_ACTION (*include*) şi 7 din clasa I_STATE (*like, think, (have the) power*).

Cauza pentru care am propus aceste noi evenimente este, cel mai des, faptul că orice propoziţie exprimă un eveniment, în sensul definiţiei din TimeML, chiar dacă uneori aceste evenimente nu sunt uşor sau chiar deloc ancorabile temporal.

Cele două noi expresii temporale identificate pot primi amândouă valoarea PAST_REF, ceea ce înseamnă că valoarea nu este exactă, dar poate fi normalizată, conform cu standardul ISO 8601 extins, la o valoare non-specifică din trecut: *once, not that long ago*.

Cele 19 elemente de legătura de tip SIGNAL nemarcate se datorează, cel mai probabil, greşelilor inevitabile de adnotare: *several, when, meanwhile, time and again, after, on.* Aşa cum am arătat, unele dintre aceste elemente nu sunt lexicalizate în limba română. Absenţa unor semnale (SIGNAL) nu a permis identificarea unor instanţieri multiple ale unor evenimente.

Observaţiile de mai sus sunt consistente cu creatorii corpusului TimeBank [4]: corpusul încă necesită îmbunătăţiri şi recorectări mai ales în ceea ce priveşte clasele evenimentelor, marcajele şi legăturile temporale incomplete, legături de subordonare incomplete.

4. Concluzii şi etape viitoare

Analiza arată că importul automat al marcajelor temporale din engleză reprezintă o soluţie la problema identificării informaţiei temporale, cu condiţia existenţei unui corpus paralel. Rata de succes a importului automat conduce spre concluzia că adnotarea manuală nu reprezintă o soluţie, deoarece este demonstrat că adnotarea temporală este foarte consumatoare de timp, deci scumpă (Forăscu, Solomon, 2004, pentru un studiu pentru limba română). Validarea manuală a marcării temporale automate în limba română (prin import) va fi încheiată. Corpusul astfel creat va fi folosit, în combinaţie cu metode specifice de învăţare automată, pentru crearea şi antrenarea unui adnotator temporal pentru limba română. Pentru evaluarea adnotatorului se vor avea în vedere şi alte domenii pe lângă cel de ştiri, precum beletristică, legislaţie etc.

Adnotările temporale combinate cu cele de discurs vor fi aplicate pentru determinarea structurii temporale a discursului, rezumarea multi-document şi folosirea ontologiilor temporale pentru a obţine inferenţe despre evenimente în timp.

Bibliografie

Boguraev, B., Ando, R. (2006). Analysis of TimeBank as a Resource for TimeML Parsing. In *Proceedings of LREC 2006*, Genoa, Italy, pp. 71-76.

Brants, T. (2000). TnT - a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference*, ANLP-2000, Seattle, WA, pp. 224-231.

Cristea, D., Forăscu, C. (2006). Linguistic Resources and Technologies for Romanian Language. In *Journal of Computer Science of Moldova*, Academy of Science of Moldova, vol. 14, nr. 1(40), pp. 34-73, ISSN 1561-4042.

Forăscu, C., Solomon, D. (2004). Towards a Time Tagger for Romanian. In *Proceedings of the ESSLLI Student Session*, August 2004, Nancy, France.

Forăscu, C., Ion, R. (2006). TimeBank 1.2: O versiune adnotată în limba română. In C. Forăscu, D. Tufis, D. Cristea (eds.) *Lucrările Atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române*, Iaşi, noiembrie 2006. Editura Universităţii Al.I. Cuza, ISBN 978-973-703-208-9.

Forăscu, C., Ion, R., Tufiş, D. (2007). Semi-automatic Annotation of the Romanian TimeBank 1.2. In Orasan, C., Kubler, S.(eds.) *Proceedings of the RANLP 2007 CALP (Computer-Aided Language Processing) workshop*, september 2007, Borovets, Bulgaria.

Ion, R. (2007). *Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română*. Teză de doctorat susţinută la Academia Română.

Ide, N., Bonhomme, P., Romary, L. (2000). XCES: An XML-based Encoding Standard for Linguistic Corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference*., pp. 825-830.

Mani, I., Pustejovsky, J., Gaizauskas, R. (eds.) (2005). *The Language of Time: A Reader*. Oxford University Press, ISBN-13: 978-0-19-926853-5, May 2005.

Pustejovsky, J., Verhagen, M., Sauri, R., Littman, J., Gaizauskas, R., Katz, G., Mani, I., Knippen, B., Setzer, A. (2006). *TimeBank 1.2*. Linguistic Data Consortium, Philadelphia, ISBN: 1-58563-386-0.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (EACL2006), Trento, Italy, 3-7 April, 2006, pp. 153-160.

Tufiş, D., Barbu, A.M. (2002). Revealing translators knowledge: statistical methods in constructing practical translation lexicons for language and speech processing. In *International Journal of Speech Technology*. Kluwer Academic Publishers, no.5, pp.199-209, 2002, ISSN 1381-2416.

[1] *Temporal and Event Recognition for Question Answering Systems*, http://www.timeml.org/site/terqas/index.html

[2]http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T08

[3] http://nl.ijs.si/ME

[4] http://www.timeml.org/site/timebank/timebank.html

Sponsor de aur / Partenaire Or

Sponsor de argint / Partenaire Argent

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / Partenaire Bronze

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latină/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



RESURSE LEXICALE ELABORATE LA FACULTATEA DE LITERE DIN CLUJ-NAPOCA

REZUMAT

Sanda Cherata

În lucrare sunt prezentate resursele lexicale elaborate la Facultatea de litere din Cluj-Napoca, precum şi aplicațiile informatice cu care aceste resurse au fost create.
Începând cu anul 1995, la Facultatea de litere din Cluj-Napoca au existat preocupări pentru producerea unor resurse lexicale pentru limba română. Au fost create, astfel:

- *concordanțe* ale textelor poetice ale lui B. Fundoianu şi George Bacovia;
- *baze de date terminologice multilingve* în diferite domenii: mediu, medicamente brevetate, turism, alimentație publică.

Pentru realizarea acestor produse au fost elaborate:

– un sistem informatic pentru crearea concordanțelor textelor poetice româneşti - **CONCORD**, care încorporează şi un analizor lexical pentru limba română;
– diferite sisteme pentru realizarea bazelor de date terminologice.

Lucrarea prezintă aceste aplicații informatice, precum şi aspecte ale realizării concrete a produselor respective.

Sanda Cherata
Universitatea Babeş-Bolyai
Facultatea de Litere
Catedra de Limbi Moderne Aplicate
scherata@lett.ubbcluj.ro, sanda_cherata@yahoo.com



Baza de date terminologice a Centrului Național de Terminologie din Chișină

DRUȚĂ Inga, Centrul Național de Terminologie din cadrul Institutului de Filologie al Academiei de Științe a Moldovei

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Preocupările terminologice au fost determinate de progresele apărute în noile domenii ale științei - generatoare de neologisme -, cât și de tendința societății de a se globaliza, de a ajunge la o societate europeană multilingvă, de a-și armoniza rapid și eficient terminologiile.

Crearea bazelor de date terminologice (BDT) reprezintă una dintre prioritățile teoriei și practicii terminologice. BDT favorizează comunicarea prin oferirea unor standarde de limbă validate, aplicabile în industrie și tehnologie, iar ca o reacție inversă, aceste modele servesc la afirmarea unei limbi.

Elaborarea bazelor de date terminologice are multiple implicații sociale, politice, științifice, culturale, economice. Crearea la centrele de cercetare din Republica Moldova a unor bănci terminologice moderne care funcționează în regim interactiv contribuie la instituționalizarea limbii române ca limbă oficială.

Din 1989 (anul înființării), Centrul Național de Terminologie și-a propus următoarele obiective majore:

- crearea unui cadru favorabil cercetărilor de terminologie;
- elaborarea de resurse terminologice (dicționare, lexicoane, glosare) multilingve în vederea armonizării terminologiei românești cu cea din Uniunea Europeană;
- crearea și actualizarea unei baze de date terminologice;
- realizarea de instrumente moderne de lucru destinate traducătorilor;
- adoptarea și unificarea standardelor naționale;
- formarea de traducători-terminologi profesioniști etc.

Colectarea informațiilor terminologice din diverse surse, în special din dicționare/glosare specializate, standarde de terminologie, manuale, instrucțiuni, texte oficiale etc. a devenit o activitate constantă de terminografie în cadrul CNT, din care au rezultat, în primă etapă, glosare bilingve ce au constituit seria „Dicționarul profesiei" (în care au apărut *Dicționarul funcționarului public*, 1996, 1997; *Denumiri de profesii și ocupații*, 2000; *Dicționarul silvicultorului*, 2000; *Mic dicționar automobilistic*, 2002; *Dicționarul textilistului*, 2003; *Dicționarul feroviarului*, 2004; *Denumiri de mărfuri nealimentare*, 2005; *Ghidul taximetristului*, 2005 ș.a.), iar în etapele următoare, BDT și proiectul LOGOTEH (dicționar electronic de echivalente dificile).

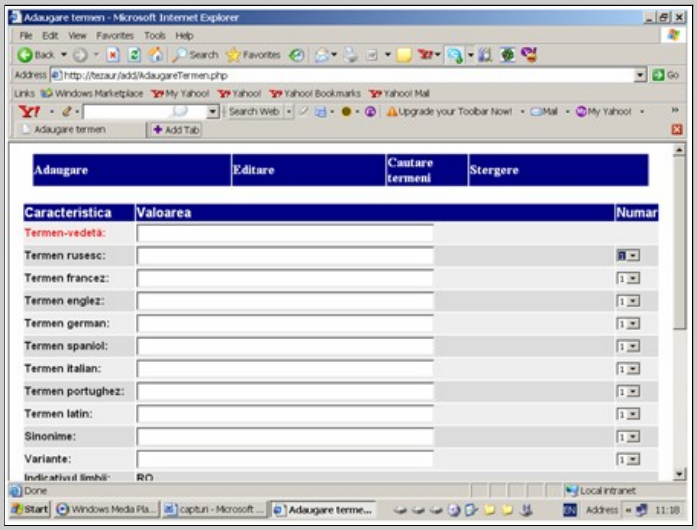
Eforturile Centrului Național de Terminologie din Chișinău în vederea creării unei baze de date terminologice au început în anul 2000, iar finalizarea proiectului se preconizează pentru 2008 (sub rezerva revizuirilor/actualizărilor ulterioare).

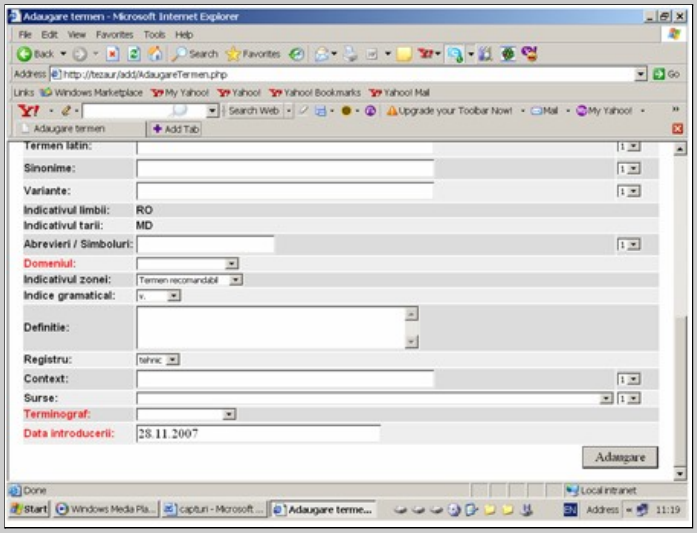
În cadrul BDT, conceptele sunt reprezentate prin termeni încadrabili în categoriile:

- termeni **recomandabili** (descriptori), care reprezintă cu exactitate conceptele cărora le sunt asociați și a căror utilizare este recomandată de anumite surse;
- termeni **standardizați**, a căror utilizare este reglementată de standarde;

- termeni **favoriţi**, utilizaţi mai frecvent decât descriptorii;
- termeni **nerecomandabili**, a căror utilizare trebuie evitată.

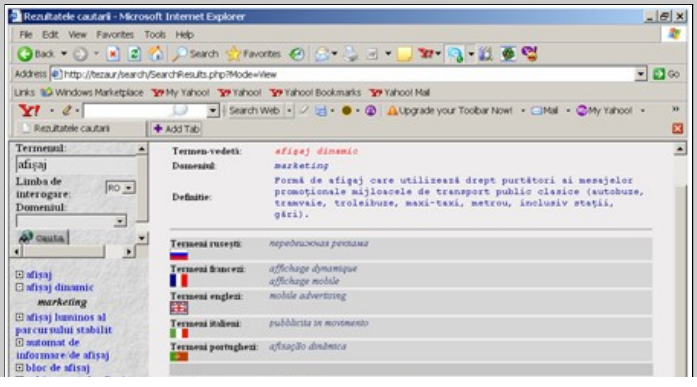
Fişa terminologică a BDT a CNT cuprinde 20 de câmpuri, la care se adaugă data înregistrării, care se ataşează automat fişei:

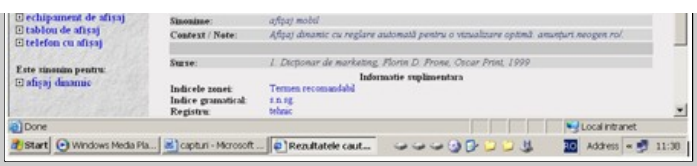




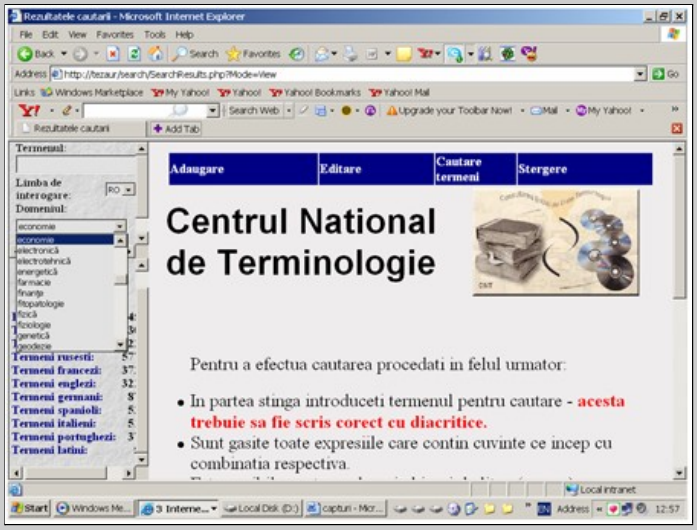
O înregistrare terminografică din baza de date consemnează pentru un concept: eticheta lingvistică (termenul prin care este desemnat conceptul), domeniul, definiţia în limba română, sursa definiţiei, contextul de utilizare, echivalentele în limbile rusă, franceză, engleză, germană, italiană, spaniolă, portugheză (pentru plante, specii, termeni din medicină etc. - şi în limba latină), sinonimele acceptate sau nerecomandabile, simbolul, abrevierea, statutul înregistrării (validată, propunere). Ca metadate, pentru fiecare concept se consemnează caracteristicile gramaticale şi numele terminografului. Unele înregistrări sunt însoţite de note privind statutul sau utilizarea termenilor stocaţi.

Consultarea BDT constă în alegerea limbii, urmată de tastarea termenului căutat în caseta de căutare, indicându-se sau nu domeniul de cunoaştere:





Actualizarea BDT se face prin operaţiile clasice de adăugare, modificare, ştergere:



În momentul de faţă, baza de date terminologice a CNT conţine 54.454 de fişe terminologice vizând peste 60 de domenii În etapa dată, specialiştii de la Centrul Naţional de Terminologie actualizează şi completează înregistrările terminologice din BDT, precizează domeniile de cunoaştere şi depun eforturi în vederea validării înregistrărilor.

Cu toate că proiectul nu este finalizat, Baza de date terminologice a CNT este vizibilă pe internet la adresa: http://cnt.dnt.md/romint/index.php şi poate fi consultată de cei interesaţi fără restricţii, constituind un instrument viabil, eficient de promovare a terminologiilor naţionale.

Bibliografie selectivă

Bălănescu, Olga, *Limbaje de specialitate*, Bucureşti, Editura Universităţii din Bucureşti, 2000.

Bidu-Vrânceanu, Angela, *Lexic comun, lexic specializat*, Bucureşti, Editura Universităţii din Bucureşti, 2000.

Ciobanu, Georgeta, *Elemente de terminologie*, Timişoara, 1998.

Pavel, Eugeniu şi Rucăreanu, Costin, *Introducere în terminologie*, Bucureşti, 2001.

Terminologia în România şi în Republica Moldova, Editura Clusium, 2000.

Vintilă-Rădulescu, Ioana, *Terminologia şi problemele ei actuale*, Bucureşti

Sponsor de aur / Partenaire Or

Sponsor de argint / Partenaire Argent

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / Partenaire Bronze

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latina/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



Baze de date lexicale: dicționar morfologic, silabisitor, dicționar de valențe verbal

BARBU Ana-Maria, Institutul de Lingvistică „Iorgu Iordan - Al. Rosetti

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Rezumat

Lucrarea de față prezintă trei baze de date lexicale pentru limba română: un dicționar morfologic, un dicționar de cuvinte despărțite în silabe (silabisitor) și un dicționar de valențe verbale. Dicționarul morfologic conține toate cuvintele aparținând părților de vorbire flexionare în paradigme complete și a părților de vorbire neflexionare. O intrare de dicționar (corespunzătoare unei forme flexionare) conține, la rândul ei, informații despre cuvântul-titlu, caracteristicile gramaticale, clasa de flexiune și accentul formei respective. Dicționarul este disponibil în două variante: una în format text și alta în format XML. Silabisitorul conține, descrierea în format XML a formelor flexionare, din dicționarului morfologic, despărțite în silabe. Cea de-a treia bază de date prezentată, dicționarul de valențe verbale, conține informații sintactice și semantice referitoare la verb și la argumentele sale. O intrare de dicționar reprezintă câte un verb, căruia îi sunt asociate structurile argumentale specifice, fiecărei asemenea structuri fiindu-i subordonat unul sau mai multe sensuri. Realizarea acestor baze de date se înscrie în preocupările generale de construire a unor resurse lingvistice în format electronic pentru prelucrarea automată a limbii române.

1. Introducere

Subiectul propus in această prezentare se înscrie în preocupările generale de construire a unor resurse lingvistice în format electronic pentru prelucrarea automată a limbii române. Mai precis, vom prezenta rezultatele a două proiecte de cercetare finalizate: un dicționar morfologic [1] și un silabisitor [2] , precum și rezultatele parțiale ale unui proiect în curs de desfășurare: un dicționar de valențe verbale [3].

Lucrarea de față prezintă, prin urmare, două baze de date: una de paradigme flexionare ale cuvintelor (numită "dicționar morfologic"), cealaltă de forme flexionare despărțite în silabe, (numită "dicționar de silabe"). Fiecare bază este disponibilă în două formate Unicode, text și XML. În dicționarul morfologic, principala informație a unei intrări constă în forma flexionară, lemă și descrierea morfologică. Dacă un cuvânt nu are forme flexionare poziția pentru forma flexionară este ocupată de lemă. În dicționarul de silabe, o intrare conține o formă flexionară, despărțirea ei în silabe și un câmp pentru observații. Procedurile de construire ale acestor componente ale dicționarelor, precum și rezultatele obținute formează conținutul următoarelor trei secțiuni ale prezentului articol. În secțiunea 5 sunt descrise aspecte de reprezentare a intrărilor unui dicționar de valențe, și a modul lui de construire. Articolul se încheie cu câteva concluzii și perspective de continuare a cercetărilor.

2. Utilitatea unor astfel de dicționare pentru prelucrarea automată a limbii naturale (PALN).

Un dicționar morfologic poate fi folosit oriunde e nevoie de un lematizator [4]. A preciza la ce este bun un lematizator ar fi o pierdere de vreme, dată fiind natura sa de instrument de bază în PALN.

Pentru limba română există mai curând preocupări de a crea analizoare morfologice decât astfel de dicționare de dimensiuni mari. Menționăm în acest sens cel mai recent articol, de care avem știință, cel al lui Bârlădeanu și Burciu (2006). O altă contribuție

este un instrument complex, numit RoLingva, care cuprinde forme flexionare, leme segmentate în silabe, informaţii de accentuare şi un analizor morfologic. RoLingva este însă un produs comercial autonom care nu poate fi folosit în aplicaţii PALN. Premisle de construire a unui dicţionar morfologic sunt prospectate şi în Ionescu (2003).

Credem că un lematizor (care analizează sau generează) are două mari dezavantaje. Pe de-o parte, el nu e adecvat unei limbi cu un bogat sistem de alternanţe fonetice şi de forme neregulate, precum româna. Pe de altă parte, funcţionarea sa este în timp real prea lentă. Avantajul totuşi constă în capacitatea sa de a trata cuvintele necunoscute. În schimb, un dicţionar morfologic are un grad înalt de acurateţe şi funcţionează mult mai rapid. Punctul său slab îl reprezintă tocmai cuvintele noi.

În ceea ce priveşte dicţionarul de silabe, el este util în mod clar în cercetările asupra vorbirii. Pentru cercetări în domeniul segmentării silabice, a se vedea Dinu (2006).

În privinţa dicţionarului de valenţe, acesta este util în multe şi importante domenii ale PALN, amintim doar analizele de adâncime, folosite în Traducerea Automată şi sistemele Întrebare-răspuns, analizele de suprafaţă folosite, de pildă, în Recuperarea şi Extragerea de Informaţii şi chiar în dezambiguizarea părţilor de vorbire utilizată în adnotarea la nivel de cuvânt a corpusurilor şi în recunoaşterea vorbirii. În proiectul nostru orientăm dicţionarul de valenţe în special pentru folosirea lui în Dezambiguizarea Sensurilor Cuvintelor, care este o operaţie orientată semantic şi în care legătura dintre sensuri şi structuri de valenţe este pe deplin exploatată.

3. Proceduri de construire

3.1. Dicţionarul morfologic: RoMorfoDict

Dicţionarul morfologic al limbii române, numit RoMorfoDict, se bazează pe un dicţionar tipărit (DOOM 1989) care indică scrierea corectă, pronunţarea şi câteva forme flexionare ale cuvintelor româneşti. El conţine aproximativ 65000 de intrări lexicale din vocabularul limbii române contemporane şi acoperă toate părţile de vorbire. Dicţionarul cuprinde şi combinaţii de cuvinte care prezintă dificultăţi de scriere, însă aceste dificultăţi au fost ignorate în demersul nostru. De fapt, în construirea automată a dicţionarului, am avut la dispoziţie o copie electronică a DOOM şi un inventar explicit de paradigme ale numelor şi verbelor. Paradigmele sunt explicite în sensul că am considerat a avea de a face cu două paradigme distincte dacă ele se deosebesc prin cel puţin o formă. Fiecărei paradigme îi sunt astfel specificate toate desinenţele corespunzătoare.

O intrare din DOOM are următoarea structură de bază, în care PV înseamnă partea de vorbire, DMS - descrierea morfosintactică, iar FLEX - forma flexionară.

lemma	POS	MSD1	INF1	MSD2	INF2
↓	↓	↓	↓	↓	↓
abandoná	vb.,	ind. prez. 1 sg.	abandonéz,	3 sg. şi pl.	abandoneázã

Intrarea de mai sus descrie verbul *a abandona* care are forma *abandonez* pentru indicativ prezent, persoana întâi singular şi forma *abandonează* pentru indicativ prezent persoana a treia singular şi plural. De observat că intrarea consemnează şi vocala accentuată chiar dacă acest lucru nu se practică în româna scrisă.

Numărul de perechi DMS-FLEX merge de la zero (pentru cuvintele neflexibile) până la valori care depind fie de numărul de forme neregulate, fie de alegerea autorilor lingvişti.

Pentru generarea automată a formelor flexionare au fost scrise două gramatici care să analizeze intrările verbale şi respectiv pe cele nominale (adică substantiv şi adjectiv). Pronumele a fost tratat ca o clasă închisă şi descris manual.

În scrierea gramaticilor, am întâmpinat dificultăţi cauzate de contradicţiile de descriere prezente în dicţionarul tipărit., de lipsa unor informaţii explicite, dar şi de greşeli existente în copia electronică.

După interpretarea intrării din DOOM, pasul următor a fost să se adauge restul de forme flexionare Aceasta s-a făcut prin parcurgerea următoarelor etape:

- S-a identificat paradigma care furniza desinenţele tuturor formelor specificate în DOOM. A apărut posibilitatea să se obţină mai multe paradigme care să se potrivească cu desinenţele furnizate. In această

situaţie, s-a creat o listă de paradigme, iar prima paradigmă identificată a fost luată în mod convenţional în consideraţie în vederea pasului următor. Pe de altă parte, dacă nu s-a găsit nicio paradigmă, inventarul a fost îmbogăţit cu paradigma ilustrată de respectiva intrare.

- S-au îndepărtat desinenţele uneia sau a mai multe forme din DOOM cu scopul de a se obţine astfel rădăcina potrivită unor moduri sau timpuri diferite.

- S-au reconstruit formele flexionare rămase (reprezentate prin căsuţele goale în tabloul de mai jos) din rădăcini şi din desinenţele corespunzătoare ale paradigmei.

Nr. paradigmă	Indicativ prezent					
	1 sg	2 sg	3 sg	1 pl	2.pl	3 pl
Terminaţiile paradigmei	-éz	-ézi	-eáz**ă**	-ăm	-áţi	-eáz**ă**
Formele din DOOM	abandon-éz		abandoneáz**ă**			abandoneáz**ă**

Rezultatul acestor operaţiuni a fost apoi verificat de studenţii implicaţi în grupul de cercetare. Prin intermediul unui program special verificarea s-a concentrat asupra următoarelor greşeli posibile:

- ***Ambiguitatea*** - când există o listă de paradigme posibile. Ambiguitatea provine din faptul că uneori DOOM nu specifică toate formele distincte. De pildă, dacă pentru două verbe cu aceeaşi conjugare, mai puţin formele de imperativ, DOOM nu specifică în fiecare caz forma de imperativ, rezultatul este o ambiguitate de paradigmă. Într-un astfel de caz, trebuia ca studenţi i să dezambiguizeze paradigmele.

- ***Accentul***. Sunt cazuri când în cursul conjugării accentul unui verb oscilează de pe rădăcină pe desinenţe. Programul de reconstrucţie a formelor flexionare a omis uneori să semnaleze acest fel de oscilaţie. În acest caz s-a urmărit manual ca fiecare formă să aibă un singur accent.

- ***Rădăcina***. Cuvintele româneşti, după cum se ştie, prezintă un bogat inventar de alternanţe fonetice. De exemplu, formele flexionare ale unui verb pot avea până la cinci rădăcini fonetic diferite. Cazurile de acest fel puteau induce erori în procesul de flexiune automată.

- ***Interpretarea erorilor***. După cum e cunoscut, descrierile din dicţionarele tipărite nu sunt suficient de explicite pentru scopurile procesării automate. Era de aceea posibil ca gramaticile care analizau intrările să producă erori de interpretare.

În acest moment, RoMorphoDict are 775.969 intrări pentru aproximativ 65.000 leme.

3.2. Silabisitorul: RoSyllabiDict

Construirea dicţionarului de silabe pentru cuvintele limbii române a reprezentat o continuare a dicţionarului morfologic prin aceea că formele flexionare au servit acum ca bază pentru segmentarea silabică. Am utilizat următoarele resurse:

- Un program de implementare a regulilor de segmentare silabică.

- Informaţia cuprinsă în DOOM privitoare la segmentarea silabică.

- Un inventar al diftongilor şi triftongilor româneşti.

Punctul critic al segmentării silabice în română îl reprezintă secvenţele vocalice care pot fi pronunţate ca diftongi/triftongi sau în hiat. În multe cazuri tipul de pronunţare nu poate fi dedus din context (a se vedea Dinu 2003). Pentru unele intrări, DOOM specifică vocalele în hiat, ca în exemplul următor: adáug**ă** (sil. -da-u-). Uneori această informaţie e dată numai pentru lemă, alteori numai pentru o formă flexionară. Sunt multe situaţii de hiat care totuşi nu sunt specificate în DOOM. Această inconstanţă în descriere a reprezentat o sursă de erori în procesarea automată.

Este de menţionat că noi am aplicat procedura pe forme nemarcate accentual deoarece aşa erau formele furnizate de resursele noastre de segmentare silabică amintite mai sus, chiar dacă dicţionarul morfologic supus segmentării silabice conţine accentuarea formelor flexionare. Merită totuşi spus că s-ar putea obţine rezultate mai bune dacă procedurile de segmentare silabică ar putea lua în consideraţie şi informaţia de accentuare, deoarece aceasta reduce numărul de ambiguităţi de tipul diftong/triftong/hiat. De exemplu, secvenţa -ei- poate fi analizată ca un diftong sau un hiat deşi -eí- este de fapt totdeauna un hiat. Am făcut, în continuare, unele îmbunătăţiri post prelucrare legate de accent, ca şi unele verificări parţiale.

Dicţionarul are acum 525.530 de intrări al căror format este dat în secţiunea următoare.

4. Formatul intrărilor

4.1. Dicţionarul morfologic.

RoMorphoDict este disponibil în format text pe trei coloane şi în format XML

4.1.1 Formatul cu trei coloane

Intrările pe trei coloane au structura următoare: FLEX<tab> LEM<tab> DMS, unde FLEX înseamnă formă flexionară, LEM -lemă, iar DMS -descriere morfosintactică.

Exemplul (1) conţine diferite părţi de vorbire. Unele omonime precum *a abandona*, trebuie explicit descrise, altele pot fi contrase pe un singur rând dacă DMS rezultată nu produce ambiguităţi sau erori de interpretare. De exemplu, ultima linie în exemplul 'japonéze japonéz s/adj.f/f-n.pl.n-a.neart.' stă pentru 'japonéze japonéz s.f.pl.n-a.neart.' şi 'japonéze japonéz adj.f-n.pl.n-a.neart.'.

(1)
```
abandoná    abandoná    v.inf.
abandoná    abandoná    v.ind.imperf.3sg.
abandoná    abandoná    v.imper.neg.2sg.

únde    únde    adv/conjct.
zup     zup     interj.
doi     doi     num.
oricé   oricé   pr/det.m-f-n.sg.n-a.
sub     sub     prep.
japonéze    japonéz    s/adj.f/f-n.pl.n-a.neart.
```

4.1.2. Formatul XML

Varianta XML a RoMorphoDict este mai informativă decât precedenta. În afară de descrierea morfosintactică (ca atribute ale elementelor entity şi form), ea furnizează informaţie despre numărul paradigmei (elementul parad), despre rădăcini (elementul rad) şi despre corespondenţele dintre rădăcini şi forme flexionare (prin intermediul atributului culoare), aşa cum se observă din următorul exemplu.

(2)
```
<entity type="verb">
<parad>a-10</parad>
<form mod="inf" color="0">usc&#225;</form>
<form mod="ind" timp="prez." pers="1" nr="sg." color="1">us&#250;c</form>
…….
<rad color="1">usuc</rad>
<rad color="0">usc</rad>
</entity>
```

4.2. Silabisitorul

RoSyllabiDict este în format XML. O intrare este descrisă prin intermediul elementului form . Ea oferă informaţie despre cuvântul a cărui segmentare silabică este dată (atributul **w**, a se vedea exemplul 3a) şi situaţia în care se aplică segmentarea respectivă (atributul obs), dacă este necesar (3b, c).

DOOM defineşte două tipuri de segmentare silabică: una preferată datorită pronunţării şi alta conformă cu structura internă a cuvântului. Cea de-a doua, denumită segmentare structurală, înseamnă despărţirea la limita afixelor pe care cuvântul le

conţine, ca în (3b)

Omonimele sunt diferenţiate numai dacă ele au segmentare silabică diferită, ca în cuvântul *aburi*, care ca verb este segmentat a-bu-ri, iar ca nume este segmentat a-buri.

(3) **a.** <form w="abandona" obs=""> a-ban-do-ná</form>

b. <form w="ignorant" obs=""> ig-no-ránt</form>
<form w="ignorant" obs="structural syllabification"> i-gno-ránt</form>

c. <form w="aburi" obs="v.inf./v.perf.s."> a-bu-rí</form>
<form w="aburi" obs="s."> á-buri</form>

Să observăm că numai formele segmentate conţin informaţie de accent, Unicode, codificată ca vocală cu marcă diacritică de accent, ca în (1).

4.3. Dicţionarul de valenţe

Dicţionarul de valenţe propune descrierea valenţelor verbelor într-o manieră care se depărtează întrucâtva de descrierile anterioare propuse în Ionescu şi Steriu (1999) şi în Drăghicescu (2002) prin principiile adoptate în stabilirea valenţelor şi a legăturilor dintre valenţe şi sensuri, precum şi prin convenţiile de reprezentare care se doresc mult mai explicite şi riguroase.

Principiile adoptate le urmăresc în parte pe cele descrise în Şerbănescu (1994). Unele dintre ele se referă la ceea ce autoarea numeşte 'reguli de redundanţă' potrivit cărora nu sunt prinse în descrierile de valenţă fenomenele regulate din limba română, cum ar fi:

- posibilitatea selecţiei alternative între membrii clasei de substituţie.
  Ex. complementul direct poate fi infinitiv sau conjunctiv;
- realizarea unei poziţii sintactice printr-o parte de propoziţie sau o propoziţie;
- fenomene de dublare a comp. direct sau a comp. indirect;
- fenomene de neexprimare a subiectului;
- realizarea complementul direct cu prepoziţia-morfem *pe*;
- selecţia liberă a prepoziţiilor semantice;
- promovarea posesorului ca dativ posesiv;
- transformările de diateză pasivă şi impersonală.

O intrare în dicţionarul nostru are structura reprezentată în exemplul (4).

(4) ***a trăi*** *'to live'*
Argument structures:
1. NP[*nom*, +animate]
Senses:
- to live: *Victima trăieşte.* 'The victim lives'.

2. NP[*nom*, +animate]
NP[*ac*, +period]
(AdvP[manner] or PP[*la/cu*, -])
Senses:
- to spend: *Ion îşi trăieşte tinereţea (intens / la maxim)*.
  'John lives his youth (intensively / at maximum)'.
- to feel intensively: Spectatorii au trăit momentul (cu entuziasm).
  'The public lived the moment (enthusiastically)'.

3. NP[*nom*, +animate]
PP[*pentru*, +goal]
Senses:
- to devote his/her life: Femeia trăieşte pentru răzbunarea soţului.
  'The woman lives for avenging his husband'.

După cum reiese din exemplul de mai sus descrierea valenţelor precizează următoarele tipuri de informaţii referitoare la verb sau la argumentele sale.

- Restricţii morfo-sintactice:
  GN[nom],
  GP[*la*], GP[loc]
  GV[*că*], GV[*să*], GV[3sg] etc.

- Restricţii pe cuvântul-titlu:
  V[*îşi*], V[se] -reflexive
  V[3sg] -impersonale

- Restricţii semantice: animat, scop, perioadă etc.

- Alternanţe libere: **/**

- Complemente opţionale: **( )**

- Sensurile subordonate valenţelor

Descrierea din (4) este destinată uzului uman. Din aceasta va fi dedusă, intenţionăm prin mijloace automate, o descriere în format XML destinată aplicaţiilor PALN care va avea forma exemplificată în (5), corespunzătoare structurii argumentale 2 din exemplul (4) de mai sus.

(5)
```
<entry>
    <verb msr="">trăi</verb>
    <str-arg. nr="2">
        <arg nr="1" caz="nom", restr-sem="+animat">GN</arg>
        <arg nr="2" caz="ac", restr-sem="+perioadă">GN</arg>
        <arg nr="3">
            <arg nr="3.1" tip="mod" restr-sem="">GAdv</arg>
            <arg nr="3.2" tip="" prep="la, cu" restr-sem="">GP</arg>
        </arg>
        <sensGrup>
            <sens nr.1>
              <sin>a petrece</sin>
              <eg> Ion îşi trăieşte tinereţea (intens / la maxim);</eg>
            </sens>
            <sens nr.2>
              <par>a resimţi intens</par>
              <eg> Spectatorii au trăit momentul (cu entuziasm)</eg>
            </sens>
        </sensGrup>
    </entry>
```

Până în prezent s-a realizat descrierea în format text a circa 1500 de verbe, alese din lexiconul de bază al limbii române.

Concluzii şi perspective

Articolul prezentat aici este destinat să umple un gol în câmpul resurselor electronice pentru limba română. Resursele vor fi făcute disponibile pe web, cel mai probabil pe site-ul Institutului de Lingvistică.

Dicţionarele vor fi suplimentate cu intrări noi , corespunzătoare recentei ediţii DOOM (2005). Nu intenţionăm să introducem cuvinte noi din corpusuri, deoarece nu toate cuvintele din corpusuri rămân în vocabular, iar noi intenţionăm să păstrăm dicţionarele cât se poate de aproape de norma literară. Cuvintele din corpusuri pot fi înregistrate separat în dicţionare speciale.

În privinţa dicţionarului de valenţe, a doua etapă principală este transpunerea din format text în format XML a valenţelor verbale descrise în dicţionar.

Referinţe

Bîrlădeanu, A., N, Burciu, 2006, *Crearea unui generator morfologic pentru verbele din limba română* in C. Forescu, D. Tufis, D. Cristea (eds.) "Lucrarile atelierului Resurse Lingvistice si Instrumente pentru Prelucrarea Limbii Romane Iasi -noiembrie 2006, Editura Universitatii "Al. Ioan Cuza", Iaşi, p. 119-122.

DOOM - *Dicţionarul ortografic, ortoepic şi morfologic al limbii române*, Ed. Academiei, 1989.

Dinu, L. P. 2003. *An approach to syllables via some extensions of Marcus-contextual grammars*. Grammars 6(1), , 2003, pp 1-12

Dinu, L.P., 2006. *On the quantitative and formal aspects of the Romanian syllables*, Revue Roumaine de Linguistique, LI (3-4), 2006, p. 477-498.

Drăghicescu, J. (coord.), 2002. *Dicţionar de Construcţii Verbale Român-Francez-Italian-Englez*, Editura Universitaria, Craiova.

Ionescu, E., 2003 *Premiseale unui dicţionar morfologic electronic al limbii române* în Florentina Hristea şi Marius Popescu (eds.) "Building Awarness in Language Technology", Editura Universităţii din Bucureşti, 2003, p. 461-468.

Ionescu, A., M. Steriu. 1999. *Verbul Românesc. Dicţionar Sintactic*, Editura Universităţii din Bucureşti.

Şerbănescu, A., 1994, *Pentru un dicţionar sintactic al verbelor româneşti*, în SCL, *XLV*, nr. 3-4, Bucureşti, p. 133-150.

[1]Proiect finanţat de CNCSIS prin grantul nr. 33549/18A/2002.

[2] Proiect "Silabisitor", finanţat de Institutul Limbii Române prin contractul nr.8 / 2005.

[3] Proiect finanţat de CNCSIS prin grantul nr.1156/2005

[4] Lematizor este un instrument software care face corespomdenţa dintre lema unui cuvânt (i.e. cuvântul său titlu) şi o formă flexionară a acestuia. întâlnită în text.

Sponsor de aur / **Partenaire Or**

Sponsor de argint / **Partenaire Argent**

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / **Partenaire Bronze**

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latina/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org

Sisteme de întrebare-răspuns în limbaj natural pentru spații de căutare deschise

Seminar internațional „Instrumente pentru asistarea traducerii"

Dan Tufiş
Institutul de Cercetări pentru Inteligență Artificială
Academia Română

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Întrebări şi răspunsuri (I)

- **Cunoaşterea în societatea actuală**

Seminar internațional „Instrumente pentru asistarea traducerii"

– volumul cunoştinţelor creşte într-un ritm ce depăşeşte capacitatea de acumulare a oricărei fiinţe umane; în fiecare zi apar, în practic orice domeniu al cunoaşterii, mai multe lucrări decât poate citi un om într-un an de lectură neîntreruptă;

– cunoaşterea enciclopedică devine imposibilă, lărgimea spațiului de cunoaştere fiind, inerent, invers proporţională cu adâncimea/profunzimea cunoaşterii;

– pe lângă mecanismele rezolutive fundamentale ale procesului de cunoaştere, a învăța CE, CUM şi UNDE să cauți cunoştinţele necesare rezolvării unei situații/probleme particulare devine imperativul metodologic al oricărui proces educațional modern: *o întrebare bine formulată, poate conține în ea sâmburele soluției căutate.*

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Întrebări şi răspunsuri (II)

Seminar internațional „Instrumente pentru asistarea traducerii"

- WEB-ul actual (şi mai ales cel al viitorului) este considerat cel mai reprezentativ depozitar al cunoaşterii umane (*locul unde dacă ştii cum să cauți găseşti informații despre oricine şi orice; orice întrebare rațională îşi găseşte pe WEB cel puțin un răspuns; etc)*
- dimensiuni (estimări) ale părții vizibile (indexate de principalele motoare de căutare):
  - 19,2 miliarde pagini indexate de motorul de căutare Yahoo (august, 2005)
  - 8,6 miliarde de pagini indexate de motorul de căutare Google (august, 2005)
  - 29,7 miliarde de pagini indexate in februarie 2007 (cf. **www.boutell.com**)
  - >25,76 miliarde de pagini (**www.worldwidewebsize.com**) -31 decembrie 2007

Cunoaştere potenţială vs. Cunoaştere efectivă (I)

Seminar internațional „Instrumente pentru asistarea traducerii"

- Considerând numai partea vizibilă (indexată) a WEB-ului se poate vorbi de un imens volum **potențial** de cunoştinţe.
  - De pildă, pentru cererea "size of the WEB“ motorul de căutare Google întoarce 1.100.000 documente, Yahoo întoarce 91.800 documente, MSN Live Search 48.600 documente, ASK.com întoarce 7.340 documente
- Informația utilă, existentă într-un noian de documente necesită inspectarea unui număr mare de texte, aproape niciodată exhaustivă.
- Depinzând de ordinea furnizată de motorul de căutare, documente conținând informație extrem de relevantă rămân “îngropate” în cunoaşterea potențială

Cunoaştere potenţială vs. Cunoaştere efectivă (II)

Seminar internațional „Instrumente pentru asistarea traducerii"

- Economia "captării atenției" se bazează în mare măsură pe utilizarea unor artificii tehnice de promovare a paginilor web între primele documente întoarse de motoarele de căutare pentru cererile formulate de utilizatori

- Tehnicile folosite de motoarele de căutare actuale reflectă doar în mică măsură intenția cognitivă a utilizatorului, fie din cauza insuficienței expresive a limbajului de interogare (cuvinte cheie, neordonate) fie din incapacitatea sistemelor suport de a efectua în timp real prelucrări morfologice, sintactice si/sau semantice în cele peste 150 de limbi sau dialecte cu reprezentare semnificativă pe WEB (Google oferă indexare şi căutare pentru 117 de limbi si dialecte).

- Trecerea de la potențialitate la efectivitate a cunoaşterii de pe WEB pe mai multe căi. Una din ele o constituie sistemele de întrebare-răspuns (QA)

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

O clasificare a sistemelor de întrebare-răspuns (QA)

Seminar internațional „Instrumente pentru asistarea traducerii"

Tipuri de sisteme de QA:

1. cu domeniu de discurs precizat (sisteme de QA circumscrise)
2. Cu domeniu de discurs neprecizat (sisteme de QA deschise)

Distincția 1 vs. 2 este puțin sensibilă la volumul datelor şi se referă la posibilitatea modelării predictive a semanticii unei întrebări arbitrare, dar relevantă în raport cu informațiile existente în spațiul de căutare, precum şi la reprezentarea informației factuale ce ar putea reprezenta răspunsul la o întrebare.

– Sistemele de tip 1 se bazează pe surse structurate de cunoaştere (baze şi bănci de date, rețele semantice factuale) care pot conține volume foarte mari de date

– Sistemele de tip 2 utilizează de regulă cunoştințe lexicale (adeseori ontologii lexicale), euristici şi colecții foarte generale de pattern-uri; cunoaşterea factuală este ne-explicitată şi este "ascunsă" în colecții de documente textuale sau multi-media. Pentru un sistem de tip 2, în cazul cel mai general, spațiul de căutare al răspunsului/răspunsurilor la o întrebare este întregul WEB (însă chiar dacă spațiul de căutare este o colecție de 1000 de documente selectate aleatoriu, un sistem capabil să răspundă corect la o întrebare al cărui răspuns se găseşte in cele 1000 de documente, tot sistem de tip 2 rămâne)

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Un caz particular de sistem Q&A deschis

- Spațiul de căutare: o colecție mare de documente (textuale sau multimedia) constituite în arhive tematice:
  - PubMed Central (http://www.pubmedcentral.nih.gov/): arhiva digitală publică de articole în domeniul bio-medical şi al ştiinţelor vieţii
  - Wikipedia (http://en.wikipedia.org/wiki/Main_Page), The Oxford Text Archive (OTA- http://ota.ahds.ac.uk/) The Arts and Humanities Hub (Humbul Humanities Hub http://www.humbul.ac.uk/), diferite arhive digitale naționale (The Japan Digital National Archives - http://www.digital.archives.go.jp/index_e.html; The SCOTSMAN Digital Archive-http://archive.scotsman.com/; The NATION Digital archive-http://www.thenation.com/archive şi foarte multe altele)
  - Arhive ale unor organizaţii şi organisme internaţionale (UE, ONU, UNESCO, NATO, etc) conţinând cel mai adesea documente paralele (Acquis Communautaire, Europarl)

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Observaţii statistice şi psiho-lingvistice în interacţiunea om-calculator

Seminar internaţional „Instrumente pentru asistarea traducerii"

Numeroase studii experimentale (Long, 1994, Slator et al., 1986, Brennan and Ohaeri, 1994, Veronis 2001, Wahlster, 2001 etc.) asupra proprietăţilor structurale şi stilistice ale limbajului natural folosit în comunicarea om-maşină (interface language) au relevat particularităţi extrem de utile pentru proiectanţii de sisteme QA:

- Pentru o colecţie suficient de mare de documente tematice, chiar dacă colecţia este dinamică (cum ar fi Wikipedia, PubMed sau Acquis Communautaire) ***curba de creştere a vocabularului este aplatizată*** deja (adică prin introducerea la momentul T_{i+1} de texte noi in colecţie, numărul de cuvinte noi faţă de cele întâlnite până la momentul T_i va fi neglijabil);
- de asemenea este raţional a presupune că, pentru cuvintele polisemantice din fondul lexical al colecţiei de texte, dintre toate sensurile posibile (listate de un dicţionar explicativ) ***sensurile relevante pentru limbajul reflectat in colecţia dată sunt doar cele folosite în textele colecţiei.***
- În cazul unui utilizator interesat de informaţia căutată, ***complexitatea lexicală, sintactică şi semantică a întrebărilor sale este scăzută***, de regulă sub nivelul competenţei sistemului de QA

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

O ipoteză de lucru pentru sisteme QA deschise cu spații de căutare tematice

Seminar internațional „Instrumente pentru asistarea traducerii"

În interpretarea semantică a unei întrebări, un cuvânt nou (cu excepția entităților numite - named entities), neîntâlnit anterior în colecția tematică, este fie nesemnificativ (şi deci ignorat), fie reductibil la unul dintre cuvintele văzute cel mai des în contexte sintactice sau semantice similare;

Pentru o astfel de abordare, sunt necesare o serie de pre-procesări elementare asupra textelor din colecția tematică: segmentare lexicală, etichetare morfo-sintactică (POS tagging), recunoaşterea grupurilor frazale (chunking), determinarea dependențelor între cuvinte sau grupuri frazale (dependency linking).

– clasificarea tematică a documentelor din spațiul web este o premisă a abordării; documentele pot avea o clasificare implicită (foarte generală) prin apartenența la o colecție gestionată de o anumită instituție publică sau privată (e.g. Wikipedia, PubMed Public, Acquis Communautaire)
– instrumentele moderne de clasificare a documentelor (document categorizing) pot ajuta la constituirea dinamică de colecții tematice foarte mari de documente în raport cu criterii de selecții arbitrare. Relevanța documentelor selectate în colecția de documente ce va constitui spațiul de căutare pentru un sistem de QA depinde de granularitatea criteriilor de clasificare: cu cât criteriile sunt mai generale, cu atât precizia selecției este mai mare.

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

ACADEMIA ROMÂNĂ

Sisteme timpurii de întrebare-răspuns circumscrise

- Sistemele timpurii (anii '60- '80) de QA circumscrise au folosit ipoteze simplificatoare pentru a compensa resursele lingvistice şi de calcul limitate;
  - Interpretarea întrebărilor bazate pe:
    - cuvinte cheie şi şabloane prestabilite:
      - BASEBALL – Green 1963, ELIZA – Weizembaum, 1966, SIR – Raphael 1968, STUDENT – Bobrow 1968,
    - gramatici semantice/conceptuale:
      - LUNAR – Woods 1973,, LIFER – Hendrix 1977, IURES – Tufiş&Cristea 1983
    - Formalisme acţionale (*înţelegerea unui mesaj de către calculator înseamnă o acţiune conformă cu aşteptările utilizatorului)* combinând elemente declarative şi procedurale la nivelul lexical, semantic şi pragmatic în definirea contextelor cognitive:
      - SHRDLU – Winograd, 1972, MARGIE, SAM –Schank 1973, PAM – Wilensky 1978
  - spaţiile de căutare a răspunsurilor codificate explicit în baze de date sau în reţele semantice;

Seminar internaţional „Instrumente pentru asistarea traducerii"

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Principalele probleme ale sistemelor timpurii de QA (circumscrise):

Seminar internațional „Instrumente pentru asistarea traducerii"

- Scalabilitatea
  - Toate soluțiile întrebărilor plauzibile erau codificate explicit sau erau deductibile din cunoştinţele stocate, în baza unui raționament controlabil;
  - Universurile de discurs: micro-lumi statice cu limbaj controlat, cu expectații bine definite
  - Codificarea explicită a cunoaşterii: extrem de costisitoare, incapabilă să țină pasul cu cunoştinţele domeniului de discurs (indiferent de îngustimea sa)
  - Trecerea de la un univers de discurs la altul: un nou sistem
- Timpul de răspuns
  - Rezonabil pentru spații de căutare reduse, dar complet inacceptabil pentru colecții factuale de mari dimensiuni
- Dependența structurilor de reprezentare a cunoştinţelor lingvistice şi de domeniu față de limba folosită la nivelul interfeței: de cele mai multe ori structurile conceptuale de reprezentare ale domeniului de discurs şi lexicul (sintaxa) limbajului de interogare erau puse în corespondență directă

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

O schimbare de paradigmă: QA deschise

- Internetul şi web-ul au adus o modificare majoră de paradigmă în proiectarea sistemelor de QA.
  - Leit-motivul: "strangularea" procesului de achiziţie a cunoştinţelor (**KA bottleneck**)
    - cea mai naturală şi mai completă formă de codificare a cunoaşterii umane este limbajul natural (scris sau vorbit)
    - textele au devenit disponibile într-un volum atât de mare încât explicitarea manuală a cunoaşterii conţinute desfide orice încercare raţională
    - dinamica cunoştinţelor factuale este necontrolabilă
  - Consecinţa:
    - un sistem de QA de utilitate publică sau specializat trebuie să fie capabil să prelucreze fluxuri mari de informaţie primară şi să răspundă în timp util
    - multilingvismul a devenit un factor esenţial de acceptabilitate
      - informaţii căutate pot exista în documente scrise sau vorbite în limbi diferite; utilizatorul poate dori informaţii doar în limba sa ori într-un număr limitat de limbi
      - traducerea automată (cel puţin rudimente) trebuie să poată fi uşor incorporată in sistemele de QA
    - Graniţa dintre sistemele de QA circumscrise şi cele deschise, va deveni din ce în ce mai difuză atunci când explicitarea cunoaşterii din spaţiul web se va realiza automat (unul din obiectivele Semantic-Web). Până atunci capabilităţile inferenţiale ale unui sistem de QA deschis rămân inerent inferioare celor ale unui sistem QA circumscris

Seminar internaţional „Instrumente pentru asistarea traducerii"

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Structura unui QA deschis şi monolingual (1)

Seminar internațional „Instrumente pentru asistarea traducerii"

1. Sistemul de indexare şi căutare a colecției de documente ce reprezintă spațiul de căutare (pentru întregul web, indexarea este realizată de firme ca Google, Yahoo, Altavista etc, iar căutarea este furnizată prin intermediul maşinilor de căutare ale companiilor respective)
   - în funcție de granularitatea indexării, la momentul căutării, întoarce o listă ordonată de documente sau fragmente textuale (paragrafe, şiruri de lungime predeterminată conținând cuvintele folosite drept criterii de căutare-snippets);
   - ordonarea se face în raport cu o funcție de relevanță, proprie motorului de căutare

2. Interpretorul de întrebări:
   - analiza lingvistică a întrebării
   - clasificarea întrebării în raport cu o tipologie predefinită
   - determinarea tipului de răspuns aşteptat
   - generarea unei cereri de căutare în limbajul de specificare al motorului de căutare utilizat (motor propriu, local sau motor public)

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Structura unui QA deschis şi monolingual (2)

3. Extractorul şi evaluatorul celui mai probabil răspuns (componenta cea mai sensibilă):

   – filtrarea fragmentelor textuale întoarse de motorul de căutare în conformitate cu structura răspunsului aşteptat

   – analiza lingvistică a candidaţilor rămaşi şi reordonarea răspunsurilor probabile; criteriile de reordonare pot fi (şi de regulă sunt) mult mai precise decât ale motorului de căutare;

   – în funcţie de protocolul sesiunii de întrebare-răspuns se selectează un singur fragment (cel mai frecvent caz) sau cele mai bine cotate K fragmente conţinând răspunsurile posibile;

   – extracţia fragmentului minimal de text (un cuvânt, un grup frazal, o clauză sau o frază întreagă) conţinând răspunsul complet la întrebare (1 sau K răspunsuri); sistemele de QA cele mai avansate pot intra într-un dialog de clarificare cu utilizatorul pentru a selecta din cele K răspunsuri posibile pe cel mai adecvat

Structura unui QA deschis şi croslingual

Seminar internațional „Instrumente pentru asistarea traducerii"

Întrebarea în limba L_S dar documentele din spațiul de căutare în limba L_T

În mare măsură, structura rămâne aceeaşi dar

⇒ interpretorul de întrebări trebuie să includă un sub-modul de traducere a cererii de căutare;

Extractorul răspunsului trebuie să fie capabil să efectueze analiza lingvistică pentru limba L_T.

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Tipuri de întrebări şi răspunsuri aşteptate

Seminar internațional „Instrumente pentru asistarea traducerii"

Factoid: Răspunsul este o entitate de tip:

- **Thing** (Ce album a lansat trupa Queen în anul 1991 la casa de discuri EMI?)
- **Organization** (Ce organizație internațională supervizează acordurile ce definesc regulile comerciale dintre statele membre ale sale?)
- **Location** (Care este lacul cu suprafața cea mai mare din lume)
- **Person** (Cine a fost întemeiatorul Imperiului Carolingian?)
- **Measure** (Ce procent de apă conține castravetele?)
- **Time** (În ce an a fost nominalizat la Oscar filmul The Towering Inferno?)
- **Time interval (**Când a domnit Alexandru Ioan Cuza?)

Lists: Răspunsul este o listă de elemente faptice ca mai sus

(Numiți şase partenere cu care a dansat Fred Astaire in filme)

Definition: or Explanation: Răspunsul este un text narativ

(Ce este Cerbul de Aur?)

Temporally restricted: Răspunsul este tip factuale, dar cu respectarea restricțiilor

(Cine a fost preşedintele SUA între 1993 şi 2001?)

NIL: Răspunsul nu există în colecția tematică

(Cine a fost soția lui Abu al-Fida?)

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Întrebări de la competiția CLEF2007

Modalități de evaluare a unui QA deschis (1)

Seminar internațional „Instrumente pentru asistarea traducerii"

- Un răspuns este considerat corect nu numai prin conținutul său ci şi prin identificarea corectă a textului ce susține răspunsul
- Evaluări ale răspunsurilor:
  - ***Corect*** (răspunsul este corect, textul din care a fost extras răspunsul este corect)
  - ***Greşit*** (răspunsul este greşit, indiferent dacă textul din care s-a făcut extracția sa este corect sau nu)
  - ***Inexact*** (răspunsul este parțial corect, fie prea scurt fie prea lung, dar textul din care s-a extras răspunsul este corect)
  - ***Neargumentat*** (răspunsul este corect, dar textul din care a fost extras răspunsul este incorect)

Modalități de evaluare a unui QA deschis (2)

- Dacă protocolul de întrebare-răspuns permite un număr mai mare de alternative, ordinea în care sunt furnizate răspunsurile contează în scorul răspunsului corect:
  - Prima alternativă este corectă -> 1 punct
  - Alternativa M este corectă -> 1/M puncte
- Evaluări ale răspunsurilor:
  - De regulă efectuată de subiecți umani (pe baza unei grile de răspuns) pentru a evita evaluarea mecanică, insensibilă la sinonimie, polisemie, alternanțe structurale (sintactice)

Seminar internațional „Instrumente pentru asistarea traducerii"

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

ACADEMIA ROMÂNĂ

Modalități de evaluare a unui QA deschis (3)

- Şi totuşi…
  - Ce este Selena?
    1. Satelitul natural al Pământului
    2. Rolul binecunoscutei şi regretatei Selena în filmul cu acelaşi nume, interpretat de Jennifer Lopez în 1997
    3. Zeița care conduce Luna
    4. Cântăreață internațională

Răspunsurile 1 şi 2 sunt corecte, 3 şi 4 putând fi respinse în baza tipului de răspuns factual aşteptat (Thing – indus de pronumele interogativ “ce”), dar în contextul unui dialog coerent doar unul este rațional

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Sisteme de QA deschise pentru limba română prezente la CLEF

Seminar internațional „Instrumente pentru asistarea traducerii"

Limba română a fost inclusă în competiția CLEF în anul 2006

- QA-RACAI 2006 (cros-lingual RO-EN)
- QA-UAIC 2006 (cros-lingual RO-EN)
- QA-WHM 2006 (cross-lingual RO-EN)
- QA-RACAI 2007 (monolingual RO)
- QA-UAIC2007 (monolingual RO)
- QA-WHM 2007 (cross-lingual RO-EN)

Câteva particularități ale sistemelor QA-RACAI (1)

- Pentru varianta croslinguală, traducerea cererii este realizată pe baza ontologiei lexicale RO-WordNet aliniată la nivel de serie sinonimică cu ontologia pentru limba engleză WordNet
- Pentru indexare şi regăsire este folosită platforma open-source LUCENE
- Indexarea colecției tematice de documente se realizează multicriterial pe toate cuvintele "conținut" (categorie gramaticală, formă ocurență, lemă, poziție în document, termeni multi-cuvânt)
- Precizia şi completitudinea fazei de regăsire documentară bazată pe indexarea LUCENE este superioară unui motor general de căutare (P=68%, R=80%, F=73,51). O îmbunătățire a performanței acestei faze se reflectă substanțial în perfoemanta sistemului QA

Câteva particularități ale sistemelor QA-RACAI (2)

- Clasificarea întrebărilor se realizează cu un sistem bazat pe modelul maximizării entropiei, antrenabil în raport cu o tipologie dată. Sistemul de clasificare este extrem de precis (la CLEF2007, din cele 200 de întrebări, 199 au fost corect clasificate în una din cele 8 categorii definite de organizatori - temporal (TMP), time interval (ITMP), definition (DEF), measure (MES), list (LST), location (LOC), names (N) and explanation (WHY).

- Prelucrările lingvistice (segmentare lexicală, adnotare morfo-sintactică, lematizare, analiză sintactică parțială (chunking & dependency linking) este folosită platforma de servicii web dezvoltată în institutul nostru.

- Extragerea celui mai probabil răspuns la o întrebare se realizează cu ajutorul unui algoritm recursiv de potrivire structurală între arborele de dependențe al întrebării şi arborele de dependențe al candidatului de răspuns (metoda este originală şi în mare măsură responsabilă de performanța foarte bună a sistemului)

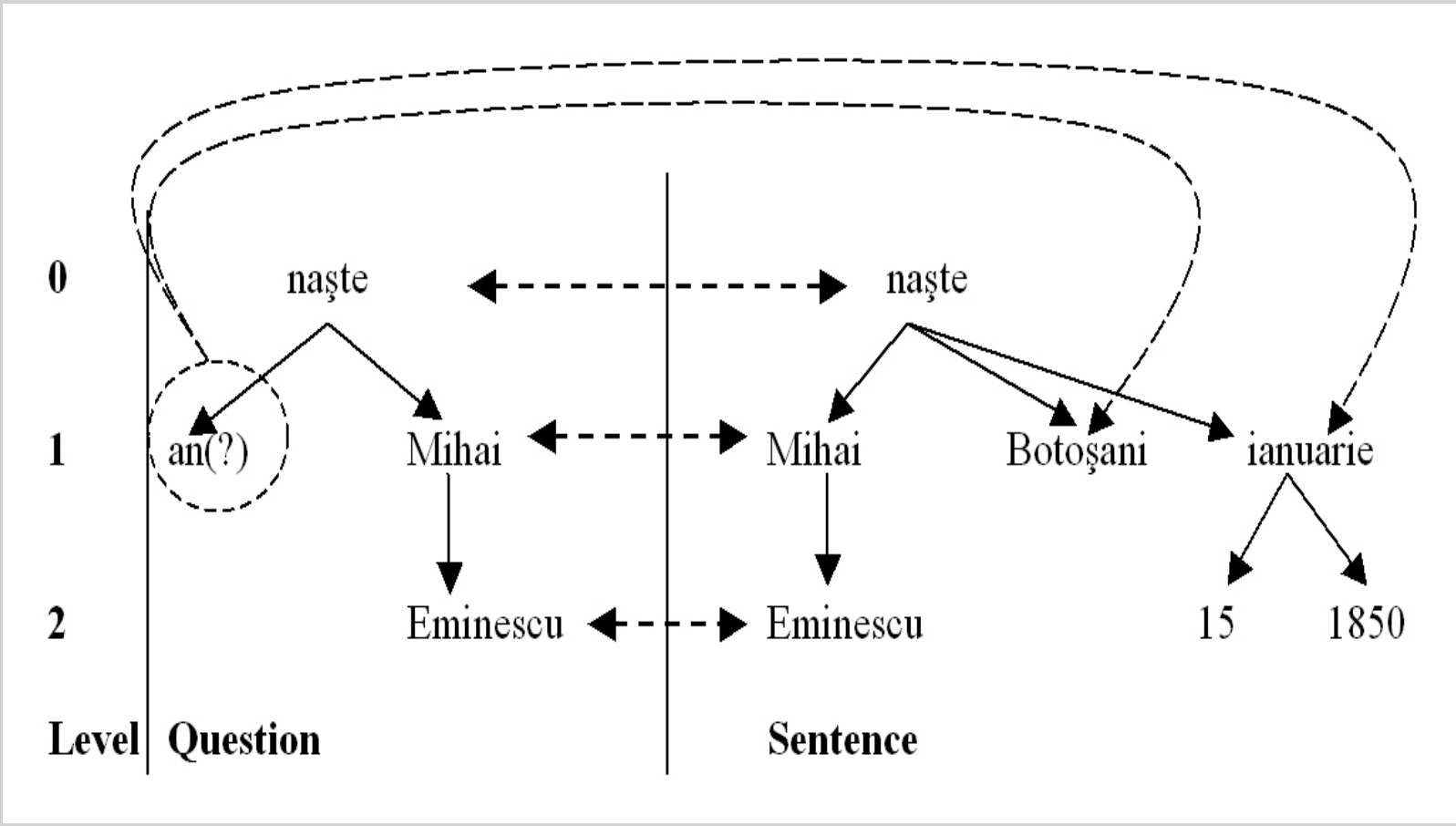
Seminar internațional „Instrumente pentru asistarea traducerii"

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Potrivire structurală între arborele de dependențe al întrebării şi arborele de dependențe al răspunsului

Seminar internațional „Instrumente pentru asistarea traducerii"

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ



Obiective

- Competițiile de tip CLEF, TREC, NTCIR etc., sunt extrem de benefice pentru evaluarea avansului tehnologic în domeniul QA, pentru identificarea punctelor slabe şi direcționarea cercetării în direcțiile cele mai promițătoare.
- Experiența colectivelor din ICIA (RACAI) si UAIC prin participarea la CLEF a condus la idei şi soluții noi, la o focalizare mai precisă a eforturilor noastre de cercetare (întrebări cu răspuns de tip listă, întrebări fără răspuns)
- Performanțele din 2007 (cel mai bun scor 30%) mult mai bune decât cele din 2006 (cel mai bun scor 14,64%); evaluarea s-a făcut considerând un singur răspuns
- O analiză a rezultatelor sistemelor QA-RACAI 2007 şi QA-UAIC 2007 a arătat că un combinator ideal al răspunsurilor ar fi obținut un scor de 40,5%
- Prin proiectul PNII *SIR-RESDEC* (parteneri RACAI, UAIC şi ICI) ținta este un sistem QA deschis cu o acuratețe de cel puțin 50%

Seminar internațional „Instrumente pentru asistarea traducerii"

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Referințe bibliografice

- Brennan, S. and Ohaeri, J. (1994). "Effects of Message Style on Users' Attributions toward Agents." In *CHI '94: Human Factors in Computing Systems, Conference Companion.* ACM. 281-282.
- Gaizauskas, R., Humphreys, K. (2000). A Combined IR/NLP Approach to Question Answering Against Large Text Collections. In: 6th Content-Based Multimedia Information Access Conference (RIAO-2000), pp. 1288–1304. Paris, France)
- Harabagiu, S., Moldovan, D., Clark, C., Bowden, M., Hickl, A., Wang, P. (2005). : Employing Two Question Answering Systems in TREC-2005. In: Text Retrieval Conference (TREC-14). Gaithersburg, Maryland
- Harabagiu, S., Moldovan, D., Paşca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., Gîrju, R., Rus, V., Morărescu, P. (2000): FALCON: Boosting Knowledge for Answer Engines. In: Text Retrieval Conference (TREC-9), pp. 479—489. Gaithersburg, Maryland
- Iftene, A., Pistol, I., C., Forăscu, C., Trandabăț, D., Balahur-Dobrescu, A., Cotelea, D., Drăghici, I.(2007): Construirea unui sistem de tip Întrebare-Răspuns pentru limba română, în Lucrările atelierului de lucru *Resurse lingvistice şi instrumente pentru prelucrarea limbii române*, 14-15 decembrie 2007, Univ. A.I. Cuza, Iaşi
- Ion, R. (2007) : Word Sense Disambiguation Methods Applied to English and Romanian. PhD thesis, Romanian Academy, Bucharest
- Ion, R., Barbu Mititelu, V. (2006): Constrained Lexical Attraction Models. In: Nineteenth International Florida Artificial Intelligence Research Society Conference, pp. 297–302. AAAI Press, Menlo Park, Calif., USA
- Long, B. (1994): Natural Language as an Interface Style http://www.dgp.toronto.edu/ people/byron/papers/nli.html
- Moldovan, D., Bowden, M., Tatu, M. (2006): A Temporally-Enhanced PowerAnswer in TREC 2006. In: Text Retrieval Conference (TREC-15). Gaithersburg, Maryland
- Moldovan, D.I., Clark, C., Harabagiu, S.M., Hodges, D. (2007): COGEX: A semantically and contextually enriched logic prover for question answering. J. Applied Logic 5(1): 49-69
- Moldovan, D., Harabagiu, S., Paşca, M., Mihalcea, R., Goodrum, R., Gîrju, R., Rus, V. (1999) : Lasso: A Tool for Surfing the Answer Net. In: Text Retrieval Conference (TREC-8), pp. 175—184. Gaithersburg, Maryland
- Ratnaparkhi, A. (1998): Maximum Entropy Models for Natural Language Ambiguity Resolution. PhD thesis, University of Pennsylvania, Philadelphia, PA
- Puscaşu, G., Iftene, A., Pistol, I., Bălaşa, D., Tufiş, D., Ceauşu, A., Ştefănescu, D., Orăşan, C., Dornescu, I., Ion, R., Cristea, D. (2007): A Cross-Lingual Romanian to English Question Answering System. In Peters, C.; Clough, P.; Gey, F.C.; Karlgren, J.; Magnini, B.; Oard, D.W.; de Rijke, M.; Stempfhuber, M. (Eds.) LNCS 4730 Lecture Notes in Computer Science, Springer-Verlag, ISBN: 978-3-540-74998-1, pp. 385-394
- Puscaşu, G., Orăşan, C (2007): University of Wolverhampton at CLEF 2007, CLEF2007 Workshop, 10 pp., September, 2007, Budapest, Hungary
- Slator, B. M. (1985) Pygmalion at the interface: Impatient users and foreign-speak (Memoranda in computer and cognitive science), Research Report Computing Research Laboratory, New Mexico State University.
- Slator, B.M., Anderson, M.P., Conley, W. (1986). Pygmalion at the interface. *Communications of the ACM, 29, 7,* 599-604.
- Tufiş D., Cristea D. (1985) "IURES: A Human Engineering Approach To Natural Language Question Answering", in W. Bibel, B.Petkoff (eds.) Artificial Intelligence: Systems, Applications, Methodology, North Holland, 1985
- Tufiş D. (1989) "A Functional Implementation of Some Pragmatic Question-Answering Principles", in International Journal on Studies and Researches in Computers & Informatics, Bucharest, 1989
- Tufiş, D., Ştefănescu, D., Ion, R., Ceauşu, A. (2007): RACAI's Question Answering System at QA@CLEF 2007, CLEF2007 Workshop, 15 pp., September, 2007, Budapest, Hungary
- Véronis, J. (1991). "Error in natural language dialogue between man and machine." *International Journal of Man-Machine Studies* 35, 187-217.
- Voorhees, E.M.(2005): Overview of the TREC 2005 Question Answering Track. In: Text Retrieval Conference (TREC-14). Gaithersburg, Maryland
- Wahlster, W. (1991). "User and Discourse Models for Multimodal Communication" In J. Sullivan and S, Tyler (Eds.) *Intelligent User Interfaces* . Addison-Wesley. 45-67.
- Waldinger, R.J., Appelt, D.E., Dungan, J.L., Fry, J., Hobbs, J.R., Israel, D.J., Jarvis, P., Martin, D., Riehemann, S., Stickel, M.E., Tyson, M. (2004) : Deductive Question Answering from Multiple Resources. In: New Directions in Question Answering 2004, pp. 253—262.
- Winograd, T. (1972). Understanding Natural Language. Academic Press.
- Yuret, D. (1998) : Discovery of linguistic relations using lexical attraction. PhD thesis, MIT, Cambridge, Massachusetts

Seminar internațional „Instrumente pentru asistarea traducerii"

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ



INTELIGENTA ARTIFICIALA – O TEHNOLOGIE FIABILA PENTRU O INFRASTRUCTURA LINGVISTICA

REZUMAT

Felicia Albescu

Traducerea automata se înscrie în sfera de interes a Inteligentei Artificiale nu atât ca scop în sine – desi unul foarte onorabil si foarte greu de atins – cât mai degraba ca etapa intermediara în îndelungatul proces de construire de sisteme artificiale cu comportament cât mai apropiat de comportamentul uman. Majoritatea sistemelor artificiale sunt construite pe baza unor descrieri facute într-un limbaj natural – descriere ce se constituie în modelul verbal al realitatii. Acuratetea captarii semnificatiei textului ce descrie realitatea este definitorie pentru calitatea sistemului artificial.
Întelegerea semnificatiei unui text se datoreaza existentei unor modele conceptuale construite în creierul nostru pe masura acumularii de stimuli veniti pe diverse canale senzoriale si de legaturi între aceste reprezentari ale realitatii, atât reprezentarile cât si legaturile dintre ele fiind asociate ulterior cu stimuli sonori ce se vor constitui într-un sistem conventional de comunicare – limbajul natural.

Modelele conceptuale construite în creierul nostru organizeaza cunostintele despre realitate si fac posibila procesarea acestora – procesare denumita în mod curent rationament. Modelele conceptuale reprezinta realitatea – limbajul este doar un set de conventii utilizate pentru reprezentarea elementelor modelelor conceptuale. În efortul de realizare a unor sisteme artificiale care sa reproduca cât mai exact structura si comportamentul sistemelor reale, modelele conceptuale au un rol fundamental. O data construite, ele pot fi descrise mot-à- mot în orice limbaj (în particular într-un limbaj artificial care permite implementarea unui sistem informatic). Plasate între modelul verbal care descrie sistemul real si modelul verbal care descrie sistemul artificial, modelele conceptuale sunt piatra de temelie a întelegerii semnificatiei realitatii si, în consecinta, indispensabile procesului de traducere automata.

Asigurând independenta de limbaj, modelele conceptuale reprezinta în consecinta o infrastructura capabila sa transporte semnificatia textului dintr-un limbaj în altul, un suport pentru traducerea automata în mai multe limbi simultan. Infrastructura lingvistica implementata intr-un limbaj virtual si asociata cu tehnologia comunicatiilor prin Internet poate deschide calea spre o reala comunicare fara bariere lingvistice

Prof. univ. dr. Albescu Felicia
Director Institutul National de dezvoltare Economica - Academia de Studii Economice Bucuresti
Tel. 0040723581942



Realizarea inferențelor textuale pe limba română

Iftene Adrian, Universitatea "Al.I.Cuza", Facultatea de Informatică, Iaşi - România

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Rezumat

Informaţiile dintr-un corpus pot fi reprezentate într-o varietate de forme. Sistemele de tip întrebare-răspuns (ÎR) trebuie să rezolve variabilitatea semantică, şi să identifice într-o anumită colecţie de date răspunsul la o anumită întrebare. O soluţie potrivită la această problemă constă în folosirea unui sistem de inferenţe textuale (SIT) care să implementeze paşii descrişi în (Bar-Haim et al., 2006). Anul acesta în cadrul competiţiei QA@CLEF [1], ne-am confruntat cu problema variabilităţii semantice şi ne-am decis să includem în sistemul nostru de ÎR un modul care să se ocupe de rezolvarea inferenţelor textuale. Rezultatele au fost încurajatoare ducând la o creştere semnificativă a preciziei. Prin urmare, am decis să construim un SIT pentru limba română care să fie la rândul lui parte componentă a unui sistem de ÎR pe limba română.

1. Introducere

Recunoaşterea inferenţelor textuale (En: *Textual entailment recognition*) RTE [2] (Dagan et al, 2005) este o competiţie, în care, fiind date două fragmente de text se cere precizarea dacă înţelesul unuia din texte poate fi dedus din celălalt text. Scopul acestei competiţii este de a crea o platformă independentă de aplicaţie, care să fie capabilă să identifice inferenţele semantice atât de folositoare în aplicaţiile din lingvistica computaţională.

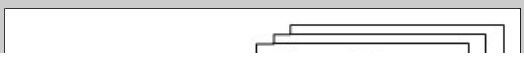
Formal, **inferenţa textuală** - IT - (En: *Textual Entailment*) definită în (Dagan et al, 2005) este o relaţie unidirecţională între două fragmente de text, denumite *T* - textul, şi *H* - ipoteza. Se spune că din *T* se poate infera *H* dacă, un om care citeşte *T* poate infera faptul că *H* este de regulă adevărată. Această definiţie se bazează pe faptul (şi presupune) cunoaşterea umană a limbajului şi cunoaşterea suplimentară despre lume.

Sistemele de IT participă în fiecare an în competiţia RTE, organizată de PASCAL [3] (Pattern Analysis, Statistical Modelling and Computational Learning), o comisie Europeană de excelenţă. Acest an, în cadrul competiţiei RTE3, am participat pentru prima dată în această competiţie cu un sistem realizat pe limba engleză.

2. Sistemul de inferenţe textuale Românesc

În **SIT-ul englezesc** construit de noi anul acesta pentru competiţia RTE3 [4], ideea principală este de a transforma ipoteza folosind cunoaşterea semantică din resurse precum DIRT (Lin and Pantel, 2001), WordNet, Wikipedia, o baza de date de acronime. În plus, am construit un sistem capabil să achiziţioneze cunoaştere suplimentară din Wikipedia englezească. De asemenea, rularea sistemului necesită o parte de pre-procesare realizată cu MINIPAR (care construieşte arborii de dependenţă asociaţi textului şi ipotezei) (Lin, 1998) şi cu LingPipe [5] (care identifică entităţile de tip nume din text şi ipoteză), urmată de încercarea găsirii distanţei minime dintre arborii asociaţi (Kouylekov, Magnini, 2005).

După terminarea competiţiei am construit un **SIT românesc** care să poată fi inclus într-un sistem de ÎR. Pentru a putea face acest lucru am înlocuit majoritatea componentelor din sistemul englezesc cu variante ale acestora care funcţionează pe limba română. În figura următoare putem observa arhitectura acestui sistem.



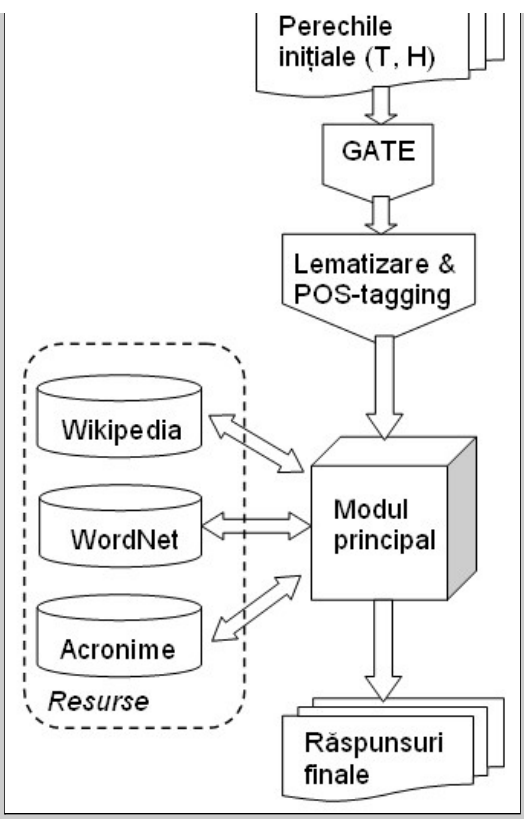


Figura 1: SIT-ul românesc

În continuare vom prezenta resursele şi uneltele folosite de noi în construcţia SIT-ului românesc.

2.1. GATE

Am pornit cu identificarea entităţilor de tip nume, unde am folosit **GATE** (Cunningham et al., 2002) **setat pe limba română**, obţinând o lista cu entităţi de tip nume specifice limbii române. Deoarece, în sistemul englezesc am avut o regulă care se ocupa cu identificarea numelor de entităţi, numere şi date, şi a cărei folosire a dus la o îmbunătăţire a preciziei sistemului cu 16 %, am acordat o atenţie deosebită acestui modul.

În (Hamza et al., 2002), un sistem de identificare a entităţilor de tip nume a fost dezvoltat pe limba română folosind ANNIE, componenta centrală a sistemului de entităţi de tip nume pe limba engleză construit în cadrul arhitecturii GATE, şi prezentată în (Pastra et al., 2002). Sistemul de identificare a entităţilor de tip nume pe limba română foloseşte împărţirea în cuvinte, un dicţionar geografic şi un modul cu reguli gramaticale din ANNIE.

Însă, deoarece multe din întrebările din cadrul competiţiei QA@CLEF au inclus nume de scriitori ai literaturii universale sau personalităţi universale, am rulat şi cu GATE setat pe limba engleză, iar în final am considerat ambele mulţimi de entităţi.

2.2. Lematizare şi POS-tagging

Pentru lematizare am utilizat un lematizator implementat în Perl ce utilizează un dicţionar de forme flexionate pentru limba română realizat de grupul de la Chişinău şi completat de noi în cadrul competiţiei QA@CLEF.

La POS-tagging am utilizat un model de limbă dezvoltat la RACAI de colectivul prof. Dr. Dan Tufiş.

2.3. Acronime

Baza de date cu acronime ne ajută să găsim relaţii între acronim şi semnificaţia lui: **"***UE - Uniunea Europeană***"**. Pentru a găsi **acronimele pentru limba română** am extras automat o listă de acronime dintr-o colecţie de ziare românești cu articole din economie şi politică folosind un algoritm asemănător celui prezentat în (Shinyama et al., 2002). De asemenea, am folosit o listă de acronime românești de pe Internet [6].

2.4. Cunoaşterea suplimentară

Cunoaşterea suplimentară pentru entităţile de tip nume şi pentru numere a fost construită semi-automat pentru entităţile din ipoteză fără corespondent în text. Pentru acestea, am folosit un modul asemănător celui din (Iftene, Balahur, 2007) care extrage într-un fişier fragmente de text din Wikipedia [7], dar care are setată limba pe română şi care foloseşte Wikipedia [8] românească.

În continuare am folosit fişierul cu fragmentele extras mai sus şi şabloane cu relaţii între entităţile de tip nume, cu scopul de a identifica relaţii cunoscute între entitatea cu probleme şi altă entitate. Aceste şabloane construite pentru limba română sunt asemănătoare şabloanelor construite pentru limba engleză, dar am adăugat în plus şabloane specifice limbii române. Aceste reguli vin în mare parte din regulile de extragere a contextelor de definiţii româneşti descrise în (Iftene et al., 2007).

După cum se poate observa în tabela 1, şabloanele noastre identifică două tipuri de relaţii între cuvinte:

- "is", când modulul extrage informaţii folosind exact regulile de identificare a contextelor de definiţii româneşti;
- "in", când în plus informaţiile extrase conţin cuvinte specifice precum: ***în, din, inclus, regiune*** etc.

Tabela 1: Cunoaşterea suplimentară

Bucureşti [in] România
American [in] America
America [is] Statele Unite ale Americii
II [is] Februarie
Chinez [in] China

Rezultatele pe limba română sunt incomplete, întrucât numărul articolelor Wikipedia în limba română este încă redus. De aceea, atunci când nu avem rezultate satisfăcătoare pe limba română folosim cunoaşterea suplimentară obţinută pe limba engleză.

2.5. WordNet

WordNet-ul românesc (Tufiş et al., 2004) a fost folosit pentru a găsi sinseturile cuvintelor din ipoteză fără corespondent în text, urmând ca mai apoi să încercăm să găsim corespondent pentru acestea în text.

Datorită temerilor generale ale unor anumiţi lexicografi, conform cărora simpla traducere a sinseturilor din WordNet-ul Princeton (Fellbaum, 1999) nu va avea ca rezultat un dicţionar reprezentativ pentru limba vizată, în (Tufiş, 1999) a fost adoptată o metodă centrată pe limbă (în contrast cu o metodă mai simplă bazată pe traducerea cuvintelor din Princeton WordNet), bazată pe resurse lexicografice de referinţă: Dicţionarul Explicativ al Limbii Române, Dicţionarul de Sinonime, şi de asemenea un dicţionar propriu român-englez.(Tufiş et al., 1999)

2.6. Regulile de variabilitate semantică: negaţii şi termeni contextuali

Regulile de variabilitate semantică pentru sistemul IT pentru limba engleză au inclus reguli de negaţie pentru termeni specifici ca "*no*", "*never*", "*don't*" etc., utilizarea verbelor modale în formă condiţională sau folosirea verbelor la forma infinitivă. Alte reguli au ca scop surprinderea influenţei pe care o au cuvintele ***pozitive*** asupra contextului - accentuarea înţelesului unui verb şi a influenţei cuvintelor ***negative*** micşorarea gradului de probabilitate a acţiunii reprezentate de verb şi introducerea incertitudinii. În varianta sistemului pentru limba română, am identificat reguli de negaţie şi cuvinte care influenţează contextul şi am introdus reguli similare.

Pentru **regulile de variabilitate semantică** am considerat negaţia cu următoarele cuvinte "*nu*", "*poate*" (care reprezintă forma pură de negaţie). De asemenea, subjunctivele au fost identificate prin faptul că sunt precedate de particula "*să*". În

acest caz, dacă subjunctivul este precedat de un cuvânt precum "*permite, impune, indica, propune*" sau sinonimele lor, de adjective ca "*necesar*", "*obligatoriu*", "*liber*" sau sinonimele lor, sau substantive precum "*încercare*", "*posibilitate*", "*opțiune*" și sinonimele lor, înțelesul devine pozitiv. Pentru cazul cuvintelor care influențează contextul, am construit, ca și în cazul limbii engleze, două liste, una conținând cuvinte precum "*sigur*", "*absolut*", "*categoric*", "*cert*", "*precis*", "*inevitabil*", "*infailibil*" care accentuează certitudinea contextului și "*probabil*", "*posibil*", "*fezabil*", "*realizabil*", "*practicabil*" - care micșorează certitudinea contextului.

2.7. Calcularea potrivirii globale

Ideea pentru calculul acestei valori constă în verificarea potrivirilor dintre cuvintele din ipoteză cu toate cuvinte din text, urmată de calcularea unei valori globale care reprezintă valoarea normalizată a sumei tuturor valorilor cuvintelor. Toate aceste calcule se fac după folosirea tuturor resurselor prezentate anterior: WordNet, baza de date de acronime, și cunoașterea suplimentară.

Ideea principală din cadrul abordării constă în determinarea cuvintelor cheie din ipoteză care se găsesc de asemenea în text și marcarea pozițiilor în care se află. Cuvintele cheie reprezintă termenii din propoziție în afara stop word-urilor. Primul pas după eliminarea stop word-urilor îl reprezintă expandarea termenilor cheie din ipoteză, utilizând următoarele resurse: WordNet, baza de date de acronime și cunoașterea suplimentară.

De exemplu, pentru ipoteza:

H: *Ernest Hemingway, faimos romancier, nuvelist, realizator de povestiri American, a trăit între anii 1899 și 1961.*

După eliminarea stop word-urilor, obținem următoarea listă de termini cheie, ce conține lemele cuvintelor din ipoteză:

{*Ernest Hemingway, faimos, romancier, nuvelist, realizator, povestire, American, trăi, an, 1899, 1961*}

Această listă este mai apoi expandată folosind WordNet-ul Românesc, iar rezultatul este următorul:

{*Ernest Hemingway, {faimos, celebru, excelent}, {romancier, scriitor}, nuvelist, {realizator, producător, creator, participant}, {povestire, mit, parabolă, narațiune}, American, {trăi, exista, viețui}, an, 1899, 1961*}

În faza următoare, lista expandată este completată utilizând cunoașterea suplimentară. În această colecție, găsim *American [in] America* și înlocuim *American* cu lista {*American, America*}.

În final, folosind colecția de acronime, expandăm încă o dată lista de termini pentru America cu *US* și *USA*.

Lista completă rezultată este: {*Ernest Hemingway, {faimos, celebru, excelent}, {romancier, scriitor}, nuvelist, {realizator, producător, creator, participant, autor}, {povestire, mit, parabolă, narațiune}, {American, America, US, USA}, {trăi, exista, viețui}, an, 1899, 1961*}.

Rolul "textului" pentru sistemul de inferențe textuale este jucat de fragmentul de text următor rezultat în urma interogării cu Lucene: "*Ernest Hemingway (n.21 iulie 1899 - d.2 iulie 1961), faimos romancier, nuvelist, realizator de povestiri (short stories în limba engleză), reporter de război, laureat al Premiului Pulitzer în 1953, laureat al Premiului Nobel pentru Literatură în 1954, probabil cel mai cunoscut autor american în întreaga lume.*"

Folosind această listă, construim o matrice care conține aparițiile cuvintelor din ipoteză în textul fără stop word-uri:

Tabela 2: Maparea ipotezei pe text

Nr.	Cuvânt	Poziții în Text
1	*Ernest Hemingway*	**1**
2	*faimos, celebru, excelent*	**8, 30**
3	*romancier, scriitor*	**9**

4	*nuvelist*	**10**
5	*realizator, producător, creator, participant, autor*	**11, 31**
6	*povestire, mit, parabolă, narațiune*	**12**
7	*American, America, US, USA*	**32**
8	*trăi, exista, viețui*	**-**
9	*an*	**-**
10	*1899*	**4**
11	*1961*	**7**

Formula pentru calcularea **potrivirii globale** este următoarea:

$$GF = \frac{\sum_{i}^{\max} \frac{1}{abs(PositionInText_i - PositionInText_{i-1})}}{NumberOfWords}$$

Pentru cazul considerat, rezultatul aplicării formulei este:

$$GF = \frac{1+\frac{1}{7}+1+1+1+1+\frac{1}{20}+0+0+\frac{1}{28}+\frac{1}{3}}{11} = \frac{5.56}{11} = 0.51$$

Pragul dintre perechile pentru care relaţia de inferenţă textuală este adevărată şi cele pentru care este falsă a fost stabilită utilizând perechile de antrenament de la competiţia RTE3. Valoarea sa a fost stabilită la 0.42. În cazul de faţă, deoarece 0.51 este mai mare decât 0.42, decidem că relaţia de inferenţă textuală este adevărată pentru această pereche.

3. Rezultate

Sistemul de inferenţe textuale are un nivel de acurateţe de 69.13% şi pentru limba română, rezultatele actuale au o valoare a acurateţii de aproximativ 57%. Cea mai importantă problemă cu care suntem confruntaţi în construirea sistemului de inferenţe textuale pentru limba română este reprezentată încă de lipsa de resurse suficiente, ce constituie motivul principal pentru diferenţa dintre rezultate.

Prin adăugarea modulului de inferenţe textuale la sistemul de tip întrebare răspuns, pentru fragmentele de text ce nu reprezintă texte coerente, sistemul de inferenţe textuale este inutil; însă, pentru fragmente de text complexe, care exprimă aceeaşi idee, dar cu actori şi contexte diferite, diferenţa pentru alegerea răspunsului corect este obţinută clar, cu un grad mai mare de certitudine, utilizând modulul de inferenţe textuale. În prezent, utilizăm sistemul de inferenţe textuale numai în găsirea răspunsului corect pentru întrebări ce au răspunsuri de tip PERSOANĂ şi LOCALITATE. În cazul acestora, rezultatele demonstrează o creştere a acurateţii de până la 5%.

Pentru viitor, dorim să continuăm dezvoltarea sistemului pentru a putea fi capabil de a procesa întrebări cu răspunsuri de tip DATĂ şi ORGANIZAŢIE. De asemenea, vom utiliza un modul de traducere român-englez, pentru a putea utiliza resursele consistente ce există pentru limba engleză.

Mulţumiri. Autorul mulţumeşte membrilor grupului de lingvistică computaţională de la Facultatea de Informatică Iaşi pentru ajutorul şi sprijinul acordat la diferite stagii ale dezvoltării sistemului.

Lucrul din cadrul acestui proiect este parţial finanţat de Siemens VDO Iaşi şi de proiectul CEEX Rotel numărul 29.

Referinţe bibliografice

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B. and Szpektor, I. (2006) The Second PASCAL Recognising Textual Entailment Challenge. *In Proc. of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment*. Venice. Italy.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (ACL'02). Philadelphia, July (2002)

Dagan, I., Glickman, O. and Magninim, B. (2006). The PASCAL Recognising Textual Entailment Challenge. *In Quiñonero-Candela et al., editors, MLCW 2005*, LNAI Volume 3944, pages 177-190. Springer-Verlag.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Mass.

Kouylekov, M. and Magnini, B. (2005) Recognizing Textual Entailment with Tree Edit Distance Algorithms. *In Proceedings of the First Challenge Workshop Recognising Textual Entailment*, Pages 17-20, 25-28 April, 2005, Southampton, U.K.

Hamza, O., Tablan, V., Maynard, D., Ursu, C., Cunningham, H. and Wilks, Y. (2002). Name entity recognition in Romanian. *Technical report, Department of Computer Science, University of Sheffield*. Forthcoming.

Iftene, A., Balahur-Dobrescu, A. (2007). Hypothesis Transformation and Semantic Variability Rules Used in Recognizing Textual Entailment. *In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Pp.125-130. 28-29 June, Prague, Czech Republic.

Iftene, A., Trandabăţ, D. and Pistol, I. (2007). Grammar-based Automatic Extraction of Definitions and Applications for Romanian. *In Proceedings of RANLP workshop "Natural Language Processing and Knowledge Representation for eLearning environments"*. September 26, Borovets, Bulgaria.

Lin, D. (1998). Dependency-based Evaluation of MINIPAR. *In Workshop on the Evaluation of Parsing Systems*, Granada, Spain, May.

Lin, D. and Pantel, P. (2001). DIRT - Discovery of Inference Rules from Text. *In Proceedings of ACM Conference on Knowledge Discovery and Data Mining* (KDD-01). pp. 323-328. San Francisco, CA.

Pastra, K., Maynard, D., Hamza, O., Cunningham, H. and Wilks, Y. (2002). How feasible is the reuse of grammars for Named Entity Recognition? *In Proceedings of 3rd Language Resources and Evaluation Conference*.

Shinyama, Y., Sekine, S., Sudo, K. and Grishman, R. (2002). Automatic Paraphrase Acquisition from News Articles. *Proceedings of Human Language Technology Conference*, San Diego, USA.

Tufiş, D., Barbu, E., Barbu Mititelu, V., Ion, R., Bozianu, L. (2004) The Romanian Wordnet. *Romanian Journal of Information Science and Technology*, Volume 7, Numbers 1-2, pp. 107-124.

Tufiş, D. (1999). Blurring the distinction between machine readable dictionaries and lexical databases, *Research Report*, RACAI-RR56.

Tufiş, D., Rotariu, G., Barbu, A.M. (1999). TEI-Encoding of a Core Explanatory Dictionary of Romanian, *In Papers in Computational Lexicography*, 219-228, Kiefer, F., Pajzs J. (Eds.), Hungarian Academy of Sciences.

---

[1] CLEF: http://clef-qa.itc.it/

[2] RTE: http://www.pascal-network.org/Challenges/RTE/

[3] PASCAL: http://www.pascal-network.org/

[4] Competiţia RTE3: http://www.pascal-network.org/Challenges/RTE3

[5] Lingpipe: http://www.alias-i.com/lingpipe/

[6] Acronime: http://www.abbreviations.com/acronyms/ROMANIAN

[7] Wikipedia englezească: http://en.wikipedia.org

[8] Wikipedia românească: http://ro.wikipedia.org

Sponsor de aur / Partenaire Or

Sponsor de argint / Partenaire Argent

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / Partenaire Bronze

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latină/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



Requêtes linguistiques sur alignements multilingues

PAUMIER Sébastien, Université Paris-Est, Institut Gaspard-Monge
DUMITRIU Marina, Université Paris-Est, Institut Gaspard-Monge

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Résumé
Les aligneurs de textes multilingues sont des outils très appréciés des traducteurs, car ils permettent de localiser plus rapidement les traductions d'expressions. Toutefois, ils ne permettent généralement que des opérations très simples comme la recherche d'un mot, voire d'une expression régulière sur des caractères. Ces possibilités restent malheureusement très limitées. Pour pallier à ce manque de puissance, nous avons combiné l'aligneur XAlign avec le logiciel de traitement de corpus Unitex. Nous présenterons dans cet article un dictionnaire électronique du roumain, et nous montrerons comment l'on peut exploiter la richesse d'une telle ressource linguistique pour tirer parti de textes alignés.

1) Introduction

De tous les outils utilisés pour l'aide à la traduction, les aligneurs de textes sont, avec les mémoires de traduction, sans doute les plus précieux. Le principe de l'alignement est simple: il s'agit de mettre en correspondance deux textes (ou plus), supposés être la traduction l'un de l'autre. La correspondance se fait généralement au niveau des phrases, le mot restant une échelle inaccessible et à vrai dire, probablement non pertinente. Il est ensuite possible de rechercher des expressions sur un des textes, et de repérer alors leurs traductions dans les phrases alignées avec celles contenant les occurrences, ce qui permet d'exploiter le bi-texte obtenu. Depuis les premiers travaux (Gale&Church, 1991), de nombreuses études ont contribué à améliorer la technique. Le lecteur pourra consulter, entre autres (Simard *et al.*, 1992), (Ma, 2006), (Véronis, 2000), (Dan Melamed, 1996) ou (Macklovitch&Langlais, 2004).

Malheureusement, les logiciels créant ou manipulant des alignements souffrent de deux défauts. Tout d'abord, il est rarement possible de pouvoir éditer l'alignement obtenu pour corriger les inévitables erreurs. Ensuite, la recherche d'expressions dans des bi-textes se limite souvent à une recherche littérale, voire par expressions régulières. Ainsi, la moindre recherche de toutes les formes conjuguées d'un verbe se traduira par une énumération fastidieuse, et cela empirera lorsque l'on voudra effectuer des requêtes plus complexes.

Nous avons donc combiné la puissance de l'aligneur XAlign (Loria, 2006) avec les capacités du logiciel de traitement de corpus Unitex (Paumier, 2006), afin de pallier à ces deux inconvénients. Nous commencerons par présenter XAlign et l'intégration qui en a été faite dans Unitex, puis nous donnerons un bref aperçu des possibilités offertes par ce dernier logiciel, notamment en ce qui concerne les requêtes linguistiques. Nous illustrerons les applications de cette association à travers un exemple simple d'étude linguistique sur un bi-texte français-roumain. Pour cela, nous présenterons un dictionnaire électronique du roumain, et nous montrerons comment en tirer parti sur notre bi-texte.

2) XAlign

Ce logiciel d'alignement de textes a été développé au Loria par Patrice Bonhomme, Thi Minh Huyen NGuyen et Sean O'Rourke. Il travaille sur des textes encodés en TEI (http://www.tei-c.org) et suppose que ceux-ci sont découpés en phrases et paragraphes. Son fonctionnement entièrement statistique lui permet de tourner sur n'importe quelles langues. De plus, il utilise la notion de *cognates*, parfois appelés îlots

de certitude, qui correspondent à des portions de texte non ambiguës sur lesquelles on peut ancrer l'alignement de façon fiable. Il s'agit généralement de noms propres ou d'expressions numériques. La particularité de XAlign est qu'il est réentrant, c'est-à-dire que l'on peut lui fournir en entrée un bi-texte partiellement aligné, qu'il utilisera comme ensemble de cognates. Ainsi, on peut utiliser un traitement quelconque pour repérer des cognates et utiliser ensuite XAlign pour construire un alignement plus fiable que s'il avait dû construire un alignement en partant de zéro.

Ses performances sont comparables à celles de ses concurrents, c'est-à-dire qu'elles dépendent directement de la qualité de la segmentation en phrases et paragraphes des textes d'entrée. Nous effectuons cette segmentation à l'aide de grammaires de découpage complexes (Friburger *et al.*, 2000), permettant de gérer des cas particuliers de façon très fine. Une fois les textes découpés, un programme se charge de les convertir au format TEI. Il ne reste alors plus qu'à lancer XAlign et à visualiser les résultats.

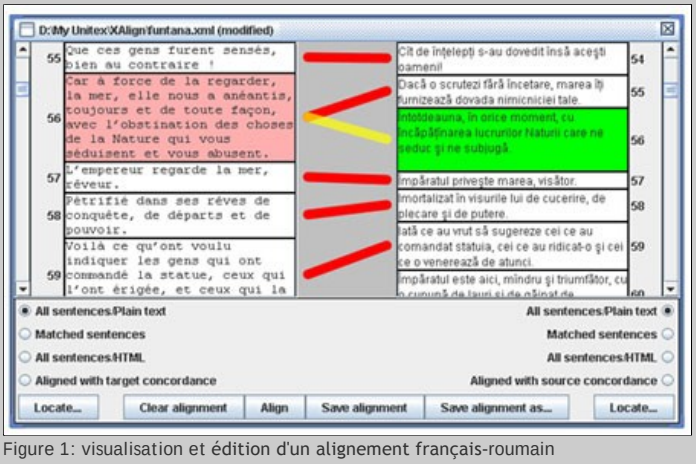


Figure 1: visualisation et édition d'un alignement français-roumain

Comme le montre la figure 1, il est possible d'éditer très simplement l'alignement obtenu par simples clics. Nous verrons dans la section suivante comment on agir sur cet alignement au moyen de requêtes linguistiques.

3) Unitex

Ce logiciel, développé à l'IGM, permet d'effectuer de puissantes requêtes linguistiques sur des textes. En effet, Unitex est fondé sur l'utilisation de larges ressources linguistiques, notamment des dictionnaires électroniques permettant d'identifier pour chaque mot simple ou composé, quelles sont ses interprétations possibles. Grâce à ces ressources, il est possible de formuler des requêtes linguistiques avancées, comme par exemple, obtenir toutes les formes d'un verbe à un temps donné. Ces requêtes peuvent se présenter sous la forme d'expressions régulières. Ainsi, l'expression:

<avoir.V>.(<E>+<ADV>).<V:K>

permet de reconnaître le verbe *avoir*, suivi par un adverbe facultatif, suivi par un participe passé, c'est-à-dire des formes au passé composé avec insertion éventuelle d'un adverbe. Il est même possible de formuler des requêtes encore plus complexes en recourant à des graphes décrivant des combinaisons bien plus riches que de simples expressions régulières:

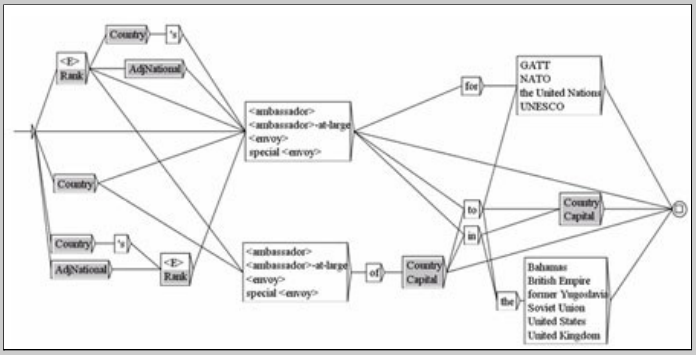


Figure 2: exemple de grammaire pour l'anglais

Notre module d'alignement permet de tirer parti de telles requêtes linguistiques à des fins de traduction. Il est en effet possible d'effectuer une requête sur un texte, d'en obtenir la concordance et de visualiser la liste des phrases dans l'autre langue qui sont alignées avec des phrases de cette concordance. Ainsi, en utilisant l'expression ci-

dessus, on pourrait aisément localiser les phrases contenant les traductions de participes passés français.

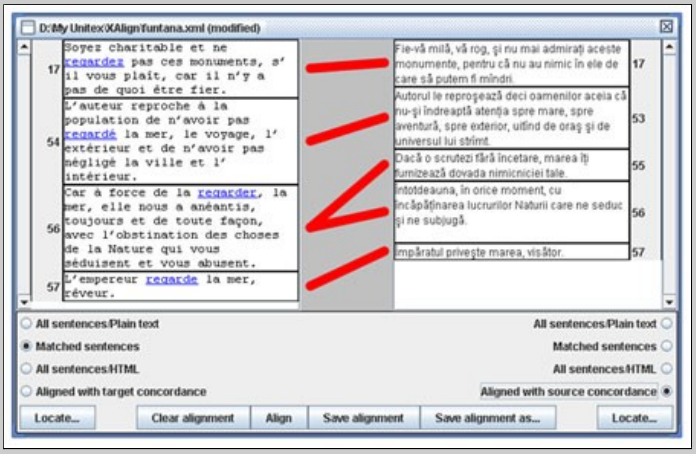


Figure 3: phrases roumaines alignées aux phrases françaises contenant le verbe *regarder*

La figure 3 montre ce que l'on obtient en recherchant toutes les occurrences, conjuguées ou non, du verbe français *regarder*, dans le roman *Les glycines d'Altea* (Thiers, 1992) et leurs correspondants dans la traduction roumaine de ce roman. En comparant les résultats, nous constatons un nombre de divergences qui se prêtent à des commentaires d'ordre linguistique et traductologique. Ainsi constatons-nous que le verbe français apparaît dans les fragments illustrés par la fig. 3 quatre fois, chaque fois à une autre forme et que les hétéronymes roumains – dans la plupart des cas, des équivalents contextuels – ne se trouvent pas toujours à la même forme flexionnelle.

Verbe français	Forme verbale	Hétéronyme roumain	Forme verbale
ne regardez pas	impératif	*nu admiraţi* (« n'admirez pas »)	impératif
n'avoir pas regardé	infinitif passé	*nu-şi îndreaptă atenţia* (« ne dirige pas son attention »	indicatif présent
regarder	infinitif	*scrutezi* (« examines attentivement)	indicatif présent
regarde	indicatif présent	*priveşte* (« regarde »)	indicatif présent

4) Une expérience linguistique

Afin d'illustrer les possibilités de cet outil, nous avons étudié la correspondance des expressions au vocatif dans les deux textes. Pour cela, nous sommes partis du fait que le vocatif est morphologiquement marqué en roumain, ce qui rend son repérage relativement aisé. Nous avons donc décrit au moyen des graphes de la figure 4 une approximation de groupe nominal au vocatif. L'écriture de ces graphes est rendue très simple par l'utilisation sous-jacente du dictionnaire du roumain que nous présenterons plus loin, et grâce auquel, un simple code comme <*N:V*> permet de reconnaître n'importe quel nom au vocatif.

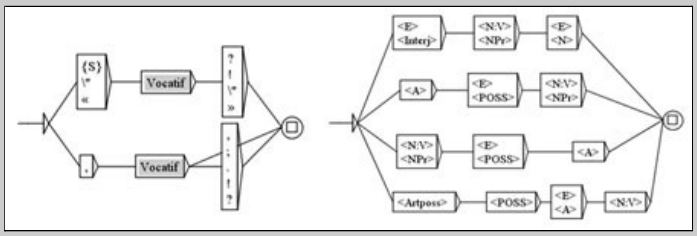
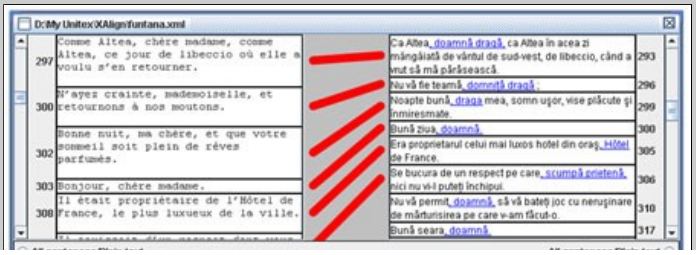


Figure 4: Graphes reconnaissant des groupes nominaux au vocatif



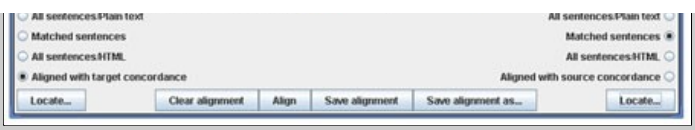


Figure 5: alignement groupes nominaux roumain au vocatif

En appliquant la grammaire de la figure 4 au texte roumain, on obtient l'alignement de la figure 5. En examinant tous les occurrences repérées, on en déduit que le vocatif est représenté en français par les expressions suivantes: *chère madame, madame, mademoiselle, ma chère, chère amie, chère Altea, Altea chérie, Tante, Pancrace, mon ami, cafard, les enfants,* et *monsieur le maire*. En triant selon les traductions françaises, on obtient le tableau suivant:

chère madame	∅, *doamnă, dragă doamnă, stimată doamnă, doamnă dragă*
madame	∅, *doamnă*
mademoiselle	∅, *domniță dragă, domnișoară*
ma chère	∅, *draga mea, doamna mea, scumpă doamnă*
chère amie	∅, *dragă prietenă, scumpă prietenă, doamnă dragă*
chère Altea	∅, *dragă Altea, scumpă Altea*
Altea chérie	∅, *dragă Altea, scumpă Altea*
Tante	*Tanti*
Pancrace	*Pancrațiu*
mon ami	*prietene*
cafard	*fățarnicule, fățarnic împuțit*
les enfants	*copii*
monsieur le maire	*domnule primar*

Ce tableau met en évidence le fait que la traduction n'est pas constante, ni dans un sens, ni dans l'autre. Ainsi, *chère madame* est traduit par 4 séquences différentes, voire par la séquence vide, et l'une de ces traductions, *doamnă dragă*, correspond dans le texte français tantôt à *chère madame*, tantôt à *chère amie*. En revanche, il ne semble pas y avoir de nuance dans le texte roumain pour distinguer *chère Altea* et *Altea chérie*.

Naturellement, ce procédé est soumis aux diverses ambiguïtés du texte. Ainsi, les formes vocatives des noms qui sont homographes avec d'autres cas ne peuvent être distinguées par le programme de recherche. C'est donc au niveau de la grammaire qu'il faut travailler pour donner un contexte permettant de lever le doute. Dans notre grammaire de la figure 4, on utilise des signes de ponctuation comme contexte. Il est cependant toujours possible de rencontrer des erreurs. Par exemple, on voit sur la figure 5, que le mot *Hôtel* a été considéré comme un vocatif par la grammaire. Il s'agit ici d'une erreur due à l'interprétation erronée de *Hôtel* comme étant un nom propre, à cause de la majuscule.

5) Un dictionnaire électronique du roumain

Le dictionnaire électronique que nous utilisons est un dictionnaire DELA que nous avons élaboré à l'Institut Gaspard-Monge. Pour la réalisation de ce dictionnaire électronique du roumain nous avons utilisé plusieurs types de ressources lexicographiques: dictionnaires explicatifs (DEX'98, NODEX et la version numérique dexonline: http://dexonline.ro/), dictionnaires orthographiques (DOOM 2 et *Dicționar ortografic*), dictionnaire morphologique (DMLR). Par ailleurs nous avons consulté plusieurs travaux de grammaires, dès plus généraux (GALR, GLR, Avram 1986, Lombard 1974) jusqu'aux plus ponctuels (Diaconescu 1962, Graur 1988), ce qui nous a permis de réfléchir à un formalisme plus unitaire de certains formes flexionnelles, i.e.: les formes de vocatif (Dumitriu & Guta, *à paraître*).

Le principe des dictionnaires DELA est d'associer à chaque lemme qui constitue l'entrée d'un dictionnaire de formes canoniques, un code flexionnel renvoyant à une classe de flexion et à une grammaire de flexion spécifique représentée par un graphe permettant la génération en temps réel des paradigmes flexionnels qui constituent un dictionnaire des formes fléchies. Les grammaires de flexion tiennent compte aussi bien de la flexion externe (désinences) que de la flexion interne (alternances vocaliques et/ou consonantiques, position de l'alternance etc). La figure 6 montre un exemple de graphe de flexion correspondant à une certaine classe de noms masculins

à double alternance (vocalique et consonantique).

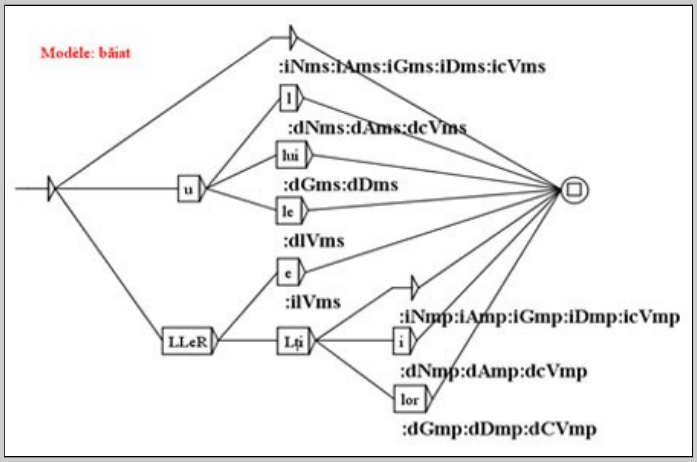


Figure 6: example de graphe de flexion de noms

Le nombre des classes de flexion (et de grammaires associées) obtenu dépasse de loin le nombre de classes traditionnelles (Dumitriu, 2006a, 2006b). Nous donnons quelques chiffres à titre d'exemple: 296 classes et sous classes de flexion nominale (dont: 75 pour les noms masculins, 179 pour les noms féminins et 42 pour les noms neutres); 74 classes des adjectifs qualificatifs auxquelles s'ajoutent les classes des adjectifs pronominaux, etc.

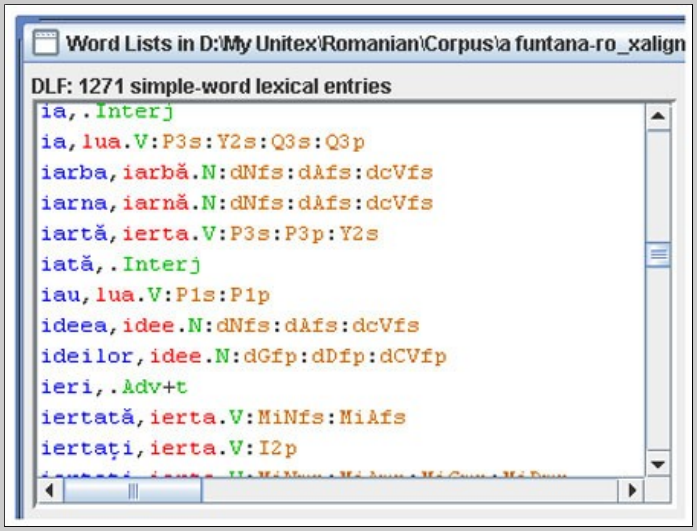


Figure 7: extrait du dictionnaire roumain

La figure 7 montre un extrait du dictionnaire roumain. Comme on peut le voir, chaque entrée est constituée d'une forme fléchie (en bleu), d'une forme canonique (en rouge), d'un code grammatical (*N*, *V*, *Interj*, etc.) éventuellement complété par des codes sémantiques (*t*), ainsi que d'une série optionnelle de codes flexionnels (*dNfs*, *MiAfs*, etc). Pour plus d'information sur ces dictionnaires électroniques et la méthode de flexion automatique utilisée pour les construire, le lecteur pourra se reporter à (Paumier, 2006).

Conclusion

Le logiciel Unitex, librement accessible sur le site de l'IGM, est désormais doté de l'aligneur XAlign. La possibilité d'éditer à la souris les bi-textes obtenus et la puissance des requêtes linguistiques qu'il est possible de faire sur ces bi-textes en font un outil très puissant pour l'exploitation de données textuelles alignées. De plus, nous avons vu que les capacités d'exploration des données augmentaient nettement si l'on disposait de dictionnaires électroniques grâce auxquels on peut identifier simplement des classes entières de mots. Grâce aux différents types de codes prévus par les dictionnaires, les requêtes peuvent être d'ordre morpho-syntaxique, sémantique, stylistique, etc. Le dictionnaire roumain que nous avons présenté constitue donc une ressource linguistique très intéressante, susceptible d'être utilisée dans d'autres applications.

Remerciements

Nous remercions vivement le Loria pour avoir accepté de mettre XAlign sous licence libre, ainsi que Patrick Watrin qui a contribué à l'intégration de cet aligneur dans Unitex. Nous remercions également Jacques Thiers, l'auteur du roman utilisé dans cet article qui a bien voulu nous laisser utiliser la forme électronique de son texte.

Références

Avram, M., 1986, Gramatica pentru toţi, Ed. Academiei RSR, Bucureşti.

Dan Melamed, I., 1996. A Geometric Approach to Mapping Bitext Correspondence, IRCS Technical Report #96-22, a revised version of the paper presented at the First Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, PA, May.

Diaconescu, P., 1962, Vocativul în -(u)le, dans SCL XIX, 5, p. 469 - 472.

Dicţionarul explicativ al limbii române [DEX '98], Academia Română, Institutul de Lingvistică "Iorgu Iordan", Ed. Univers Enciclopedic, 1998.

Dicţionar ortografic al limbii române [Dicţionar ortografic], 2002, Ed. Litera Internaţional.

Dicţionar ortografic, ortoepic şi morfologic al limbii române [DOOM 2], 2005, IIe éd, Ed. Univers Enciclopedic.

Dumitriu, D.-M. & A. Guta. Le vocatif des noms communs et des adjectifs du roumain, Ed. Aius, coll. Infolingua, 7, Craiova (à paraître).

Dumitriu, D.-M., 2006a, Grammaires de flexion des noms roumains par automates finis, Ed. Aius, coll. Infolingua, 2, Craiova.

Dumitriu, D.-M., 2006b, Grammaires de flexion des adjectifs roumains par automates finis, Ed. Aius, coll. Infolingua, 4, Craiova.

Friburger N., Dister A. et D. Maure, 2000, Améliorer le découpage des phrases sous Intex, Revue Informatique et Statistique dans les Sciences Humaines, vol. 36, n°1-4, p. 181-200.

Gale, W. & K. W. Church, 1991, "A Program for Aligning Sentences in Bilingual Corpora" Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA.

Gramatica Limbii Române [GALR] 2005, Editura Academiei Române, Bucureşti.

Graur, A., 1988, Puţină gramatică, Editura Academiei RSR, Bucureşti (les articles sur le vocatif.

Irimia, D., 1997, Gramatica Limbii Române [GLR], Ed. Polirom, Iaşi.

Lombard, A., 1974, La langue roumaine. Une présentation, Ed. Klincksieck, Paris.

Lombard, A & C. Gâdei, 1981, Dictionnaire morphologique de la langue roumaine [DMLR], Ed. Academiei, Bucureşti.

Loria, 2006, XAlign(Alignement multilingue).
http://led.loria.fr/outils/ALIGN/align.html

Ma, X., 2006, "Champollion: A robust parallel text sentence aligner". Proceedings of the fifth international conference on Language Resources and Evaluation (LREC), Genoa, Italy.

Macklovitch, E. et P. Langlais, 2004, Le bitexte et ses applications tutoriel donnée à la conférence TALN , 19-22 avril, Fès, Maroc.

Noul dicţionar explicativ al limbii române [NODEX], 2002, Ed. Litera Internaţional.

Paumier, S., 2006, Unitex 1.2 User Manual,
http://igm.univ-mlv.fr/~unitex/UnitexManual.pdf.

Simard, M., Foster, G. and Isabelle, 1992, "Using Cognates to Align Sentences in Bilingual Corpora", in Proceedings of the Fourth International Conference on Theoretical and Methodogical Issues in Machine translation (TMI92), (Montreal), 67-81.

Thiers, J., 1992, Les glycines d'Altea, Ed. Albiana, Levie.

Thiers, J., 1996, Parfum de glicine, Ed. Macedonski, Craiova (trad : Dana-Marina Dumitriu).

Véronis, J., 2000, Alignement de corpus multilingues. In J.-M. Pierrel (Ed.), Ingénierie des langues (pp. 151-171). Paris: Editions Hermès.

Sponsor de aur / **Partenaire Or**

Sponsor de argint / **Partenaire Argent**

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / **Partenaire Bronze**

Comitetul Electrotehnic Român

TermRom Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latina/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



Construirea unui sistem de tip întrebare răspuns pentru limba română

Iftene Adrian, Universitatea "Al.I.Cuza", Facultatea de Informatică, Iaşi

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanţi / Exposants

Rezumat

Începând cu anul 2007, în cadrul competiţie QA@CLEF a fost introdusă pentru prima dată posibilitatea de a căuta răspunsuri la întrebări puse în limbaj natural, într-o colecţie de date în română. Această colecţie de date reprezintă o variantă îngheţată a Wikipediei româneşti din luna decembrie anul 2006.

Articolul prezintă eforturile depuse de echipa noastră în încercarea construirii unui sistem de tip întrebare-răspuns care funcţionează în totalitate pe limba română. Se va putea observa faptul că, spre deosebire de competiţia similară din 2006 accentul a căzut pe o nouă caracteristică introdusă anul acesta în competiţie: gruparea întrebărilor pe topici, a cărei rezolvare a depins de rezoluţia anaforei. De asemenea o componentă importantă a sistemului o constituie sistemul de inferenţe textuale care ne-a ajutat în ordonarea răspunsurilor.

1. Introducere

Primul sistem de Întrebare-răspuns [1] (ÎR) românesc a fost dezvoltat în anii '80 (Tufiş şi Cristea 1985) şi era reprezentat de o interfaţă ce facilita comunicarea cu o reţea semantică (care codifica cunoaşterea). Astăzi sistemele de ÎR folosesc documente text ca bază de cunoaştere şi integrează tehnici de prelucrare a limbajului natural (PLN) pentru a găsi (într-o colecţie dată de documente sau prin căutare pe web) răspunsul la o întrebare pusă în limbaj natural.

România a participat pentru prima dată la o competiţie CLEF în 2006, în cadrul secţiunii QA@CLEF [2] cu un sistem dezvoltat de UAIC [3] şi RACAI [4] (Puşcaşu, 2006) cu un sistem pe română-engleză. În anul 2007 limba română a fost inclusă pentru a doua oară în competiţie, organizatorii dând posibilitatea participanţilor să aleagă între exerciţiile RO-RO, EN-RO, RO-EN (prima limbă reprezentând limba sursă, cea în care se pun întrebările şi cea de a doua limba ţintă, cea a documentelor în care se caută răspunsurile). Anul acesta pentru limba română colecţia de documente în care s-a căutat răspunsul a fost formată dintr-o variantă îngheţată a Wikipediei româneşti din luna decembrie 2006.

De regulă, sistemele de ÎR folosesc o arhitectură generală de tip *pipe-line*, în care prelucrarea parcurge trei etape principale: analiza întrebării, căutarea documentară şi extragerea răspunsului (Harabagiu, Moldovan, 2003). Sistemul creat este o variantă a arhitecturii generale, cu particularizări specifice legate de reprezentare şi procesare pentru fiecare din componentele amintite mai sus.

De la an la an în competiţia QA@CLEF se adaugă noi caracteristici, cu scopul de a încerca să se simuleze cât mai bine situaţii concrete din lumea reală. Anul acesta caracteristica nou introdusă a fost gruparea întrebărilor pe topici. În cadrul unei topici, toate întrebările se referă la o anumită temă, precizată în prima întrebare sau în răspunsul la aceasta, şi există posibilitatea folosirii legăturilor anaforice între întrebările din aceeaşi topică.

Arhitectura generală a sistemului este ilustrată în figura de mai jos, iar modulele mai importante vor fi prezentate în capitolele următoare, insistând pe componentele nou introduse. În ultima parte vom prezenta rezultatele obţinute la competiţia de anul acesta, concluziile, precum şi direcţiile de lucru viitoare.

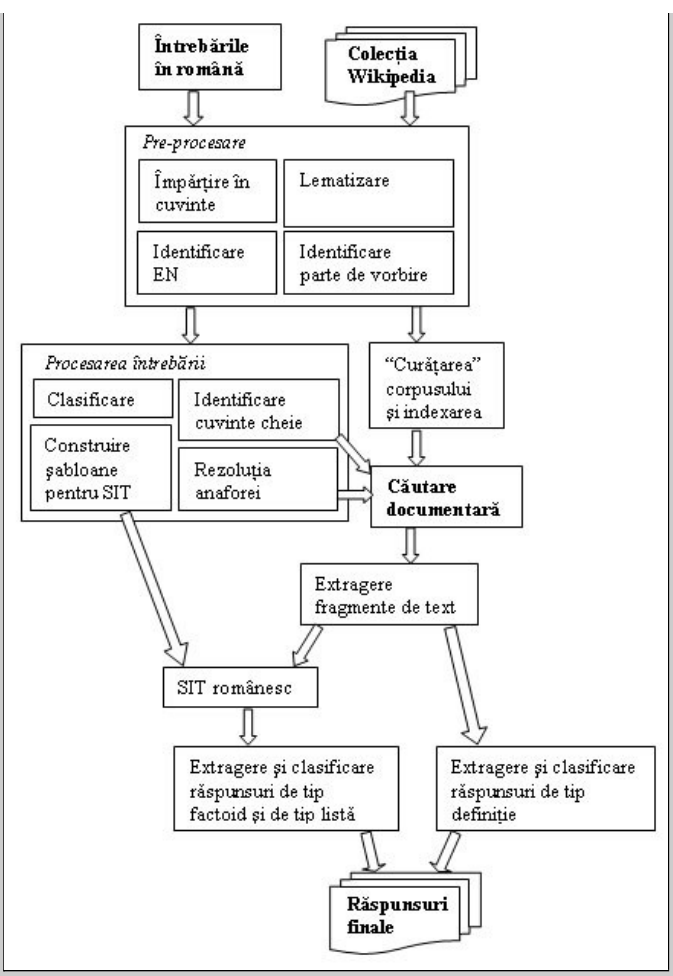


Figura 1: Arhitectura generală a sistemului de întrebare-răspuns românesc

2. Etapele parcurse de Sistemul ÎR

2.1. Pre - procesări asupra corpusului

Întrebările au fost pre-procesate secvențial cu module care au realizat tokenizare, POS-tagging și lematizare. Entitățile de tip nume au fost identificate și clasificate folosind GATE (Cunningham et al., 2002) în: Persoană, Locație, Măsură și Dată.

Similar această parte de pre-procesare a fost aplicată și colecției de documente din Wikipedia.

2.2. Analiza întrebării

Această etapă are ca scop principal identificarea tipului răspunsului așteptat. În plus, se identifică focusul întrebării, tipul întrebării și mulțimea cuvintelor cheie relevante pentru întrebare. Pentru atingerea acestor obiective, s-au parcurs următorii pași:

a) Identificarea grupurilor nominale. Extragerea entităților de tip nume

Folosind mulțimea de întrebări ca intrare, pe baza adnotărilor morfo-sintactice oferite de uneltele noastre, s-au implementat reguli de identificare a grupurilor nominale. Același instrument folosit în faza de pre-procesare a permis identificarea entităților de tip nume din întrebare.

b) Focusul întrebării

Focusul întrebării este cuvântul sau secvența de cuvinte care precizează ce anume căutăm ca răspuns sau despre ce anume este vorba. Pentru aceasta am considerat fie primul substantiv din întrebare (ca în *Ce țară*) sau prima componentă a primului grup substantival atunci când acesta apare după verbul principal al întrebării sau dacă apare după verbul "*a fi*".

c) Tipul răspunsului

Sistemul nostru este capabil să identifice următoarele tipuri de răspuns: PERSOANĂ, LOCAȚIE, ORGANIZAȚIE, TEMPORALĂ, NUMERICĂ, DEFINIȚIE și ALTA. Ataşarea unuia din aceste tipuri unei întrebări analizate s-a efectuat folosind şabloane specifice pentru fiecare tip în parte. În cazul întrebărilor ambigue (de exemplu cele care încep cu Ce), am folosit focusul întrebării (de exemplu, în cazul întrebării *Ce oraş este identificat cu Troia Homerică?*, care are focusul oraş, folosim tipul asociat focusului care este LOCAȚIE).

d) Tipul întrebării

Tipul întrebării poate fi unul din următoarele factoid, definiţie sau listă. Pentru identificarea acestuia am folosit două reguli simple: dacă tipul răspunsului este DEFINIȚIE, atunci evident tipul întrebării este definiţie; dacă focusul întrebării este un substantiv la plural, atunci tipul întrebării este listă, în celelalte cazuri fiind factoid.

e) Rezoluţia anaforei

Nouă caracteristică introdusă anul acesta în competiţia QA@CLEF, gruparea întrebărilor pe topici, a dus la mărirea gradului de dificultate, prin faptul că a fost necesară introducerea unui modul special în arhitectura generală a sistemului, care să se ocupe de rezoluţia anaforei. De exemplu pentru primul grup din setul de întrebări avem:

Tabela 1: Primul grup de întrebări

```
<Group id=1>
    <Question id=1> Ce faimos romancier, nuvelist şi realizator american de
povestiri a trăit între anii 1899 şi 1961?
    </Question>
    <Question id=2> Pentru ce premiu a fost laureat în anul 1954?
    </Question>
    <Question id=3> În ce an a fost laureat al Premiului Pulitzer?
    </Question>
</Group>
```

Se observă cum în întrebările cu id-urile 2 şi 3 trebuie înlocuit pronumele "*el*" cu răspunsul de la prima întrebare.

Pentru a rezolva această problemă am adoptat două metode de rezoluţie a anaforei prezentate în cele ce urmează:

1. *Forţa Brută*

   În această situaţie toate cuvintele cheie de la prima întrebare din grup sunt adăugate la cuvintele cheie ale următoarelor întrebări din grup. Iată ce ar însemna pentru exemplul prezentat mai sus acest lucru:

   Cuvintele cheie pentru prima întrebare din grup sunt {*faimos, romancier, nuvelist, realizator, american, a trăi, 1899, 1961*}, şi doar acestea vor fi folosite în procesul de căutare a răspunsului.

   Cuvintele cheie pentru a doua întrebare sunt {*premiu, laureat, 1954*}, dar în procesul de căutare a răspunsului pentru această întrebare, vom folosi în virtutea regulii de mai sus şi cuvintele cheie de la prima întrebare: {*faimos, romancier, nuvelist, realizator, american, a trăi, 1899, 1961, premiu, laureat, 1954*}.

   2. Folosirea răspunsurilor

   În acest caz folosim răspunsul de la prima întrebare din grup care se adaugă la lista curentă de cuvinte cheie. Pentru exemplul nostru ar însemna că la cuvintele cheie de la întrebarea a doua din grup s-ar adăuga răspunsul de la prima, care este "Ernest Hemingway" şi s-ar obţine mulţimea: {*premiu, laureat, 1954, Ernest Hemingway*}.

Desigur a doua metodă este mai bună, dar ea depinde de capacitatea sistemului nostru de a identifica răspunsul corect. Deoarece, în multe cazuri nu am reuşit să extragem răspunsul corect al primei întrebări din grup, am preferat să folosim o combinaţie a celor două metode.

f) Generarea cuvintelor cheie

Iniţial se porneşte cu o mulţime formată din: focus, entităţile de tip nume, celelalte substantive din întrebare, şi toate verbele care nu sunt auxiliare în întrebare. Pentru aceasta se construieşte mai apoi o mulţime mai mare care conţine pentru fiecare cuvânt şi lista sinonimelor acestuia.

2.3. Crearea indexului şi căutarea documentară

În rulările noastre indexarea şi căutarea s-a realizat cu motorul Lucene [5]. Colecţia de documente a fost indexată atât la nivel de document cât şi la nivel de paragraf folosind lema cuvintelor conţinute. Când nu este găsit nici un paragraf pentru o anumită interogare, se folosesc două strategii: fie se măresc segmentele de la paragrafe la documente, fie se reformulează interogarea folosind pentru anumite cuvinte alte variante posibile de traducere.

2.4. Extragerea Răspunsului

Folosind tipul răspunsului aşteptat, focusul întrebării, mulţimea de cuvinte cheie, părţile de text obţinute în urma căutării pe partea de vorbire, lemă şi informaţii de tip entităţi de tip nume şi indicatorul de relevanţă al paragrafelor determinat de Lucene. Procesul de extragere depinde de tipul aşteptat al răspunsului: Când răspunsul are un anumit tip de entitate de tip nume, modulul de extragere a răspunsului identifică în fiecare propoziţie întoarsă de Lucene aceste entităţi. Când tipul răspunsului nu este un nume de entitate, procesul de extragere se bazează în principal pe recunoaşterea focusului, în acest caz şabloanele sintactice de găsire a răspunsului bazate pe focus fiind foarte importante.

3. Analiza Rezultatelor

Rezultatele evaluării oficiale pentru sistemul nostru sunt prezentate în tabelul de mai jos.

Tabela 2: Rezultatele obţinute în competiţia din 2007

Evaluarea Rezultatelor		
Z	NECUNOSCUT	0
R	CORECT	24
U	NEJUSTIFICAT	1
W	INCORECT	171
X	INEXACT	4
TOTAL		**200**

Fiecare răspuns a fost evaluat ca fiind NECUNOSCUT (răspunsurile neevaluate), CORECT (răspunsurile corecte), NEJUSTIFICAT (răspunsuri care nu puteau fi găsite în bucăţile de text justificatoare), INCORECT (răspunsurile greşite) sau INEXACT (răspunsuri incomplete). După cum se observă rezultatele, precizia sistemului a fost de 12%, valoare care este foarte apropiată de valoarea obţinută anul trecut.

4. Concluzii

În configuraţia actuală, realizarea implementează cele trei niveluri esenţiale unui astfel de sistem. Evaluarea arată o precizie de aproximativ 12%, care, deşi încă insuficientă, indică o plasare în rând cu alte sisteme similare din străinătate. Experienţa câştigată va fi folosită la îmbunătăţirea sistemului pentru participări la ediţii viitoare ale QA@CLEF cât şi în cadrul proiectului SIR-RESDEC, aprobat recent spre finanţare de CMNP.

Din testele pe care le-am realizat (Iftene şi Balahur, 2007), am remarcat că utilizarea unul modul de inferenţe textuale are ca rezultat o îmbunătăţire a clasificării răspunsurilor posibile. Acest lucru este posibil deoarece sistemul de inferenţe textuale face o analiză semantică a contextului întrebării şi nu doar calculează distanţe lexicale între cuvinte. Sperăm ca pe viitor utilitate acestei componente să se dovedească a fi benefică pentru rezultatele pe care le vom obţine.

Mulţumiri. Autorul mulţumeşte tuturor membrilor echipei de la Facultatea de Informatică, care au fost implicaţi în acest proiect. Lucrul la acest proiect este parţial suportat de proiectele STREP FP6-2004-IST-4, şi a proiectului naţional CEEX Rotel 29/2006.

Referinţe bibliografice

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July

Harabagiu, S., Moldovan, D. 2003. *Question Answering*. In: The Oxford Handbook of Computational Linguistics. Oxford; New York: Oxford University Press.

Iftene, A., Balahur-Dobrescu, A. 2007. Improving a QA System for Romanian Using Textual Entailment. *In Proceedings of RANLP workshop "A Common Natural Language Processing Paradigm For Balkan Languages"*. Pp. 7-14, September 26, 2007, Borovets, Bulgaria.

Puscaşu, G., Iftene, A., Pistol, I., Bălaşa, D., Tufiş, D., Ceauşu, A., Ştefănescu, D., Orăsan, C., Dornescu, I., Ion, R., Cristea, D. 2006. *Developing a Question Answering System for the Romanian-English Track at CLEF 2006, in Proceedings of CLEF2006, Alicante, Spain.*

Tufiş, D., Cristea, D. 1985. *IURES: A Human Engineering Approach to Natural Language Question Answering*, in W. Bibel, B.Petkoff (eds), Artificial Intelligence: Systems, Applications, Methodology, North Holland.

---

[1] Question Answering (QA) - rom.: Întrebare-Răspuns (ÎR)

[2] Multilingual Question Answering at CLEF: http://clef-qa.itc.it/

[3] Universitatea "Al.I.Cuza" Iaşi: http://www.uaic.ro/

[4] Romanian Academy Center for Artificial Intelligence: http://www.racai.ro/

[5] http://lucene.apache.org/

Sponsor de aur / **Partenaire Or**

Sponsor de argint / **Partenaire Argent**

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / **Partenaire Bronze**

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org



PORTALUL CONSILR – MIJLOC DE INFORMARE ASUPRA ACTIVITĂȚILOR DE INFORMATIZARE A LIMBII ROMÂNE

REZUMAT

Corina Forăscu, Dan Cristea, Ionuț Pistol

Explozia de informații pe Internet, necesitatea tot mai mare de a accede la acestea prin mijloace inteligente utilizînd limba nativă fac domeniile Tehnologiilor Limbajului Uman şi cel al Prelucrării Limbajului Natural ca unele prioritare la nivel mondial. Necesitatea de a ridica limba română în privinta informațiilor, resurselor şi a tipurilor de prelucrări automate la nivelul altor limbi europene este din ce în ce mai stringentă.

Constituită în Academia Română în 2001, Comisia pentru Informatizarea Limbii Române militează pentru colaborarea deschisă a tuturor specialiştilor, accesul neîngrădit la resurse lingvistice, utilizarea tehnologiilor lingvistice moderne, lansarea de proiecte prioritare interdisciplinare, formarea de noi specialişti. Consorțiul pentru Informatizarea Limbii Române este forul executiv al Comisiei omonime şi are ca principale activități organizarea de evenimente dedicate resurselor şi instrumentelor limbii române şi întreținerea portalului ConsILR[1].

Portalul ConsILR încearcă să țină la curent comunitatea mondială interesată de problematica limbii române din perspectiva tehnologiei informatice, asupra progreselor efectuate şi a posibilităților existente în acest domeniu.

Portalul este bilingv (română şi engleză) şi conține pagini dedicate pentru următoarele:

- Resurse lingvistice şi fonetice (corpusuri scrise şi vorbite, dicționare electronice, gramatici, documentații etc.); membrii ConsILR pot adăuga şi utiliza liber resursele depuse pe portal.
- Instrumente şi tehnologii de prelucrare a limbajului, dedicate limbii române; unele sunt disponibile pentru descărcare şi utilizare off-line, iar o parte dintre instrumente sunt disponibile ca servicii web sau sunt menționate paginile web care găzduiesc asemenea servicii. Între instrumentele ce pot fi accesate din portal este şi un serviciu de traducere automată din română în engleză.
- Universități şi Institute de Lingvistică Computațională şi Prelucrare a Limbajului Natural.
- Evenimente naționale (inclusiv cele organizate de Consorțiu şi membrii acestuia) şi internaționale, importante pentru comunitatea de cercetători lingvişti şi informaticieni interesați de limba română.
- Proiecte de cercetare încheiate sau în derulare, în cadrul cărora s-au dezvoltat resurse sau tehnologii dedicate limbii române.
- Posibilități de finanțare a cercetării în domeniile de interes.
- Articole, publicații, lucrări de licență şi dizertație din domeniul Lingvisticii Computaționale sau domenii înrudite.

[1] http://consilr.info.uaic.ro/

În cadrul comunicării vor fi prezentate mai detaliat activitățile Comisiei şi Consorțiului de Informatizare pentru Limba Română precumş şi situl ConsILR, opiniile şi sugestiile participanților urmând a fi folosite pentru îmbunătățirea portalului.

Corina Forăscu[1,2], Dan Cristea[1,3], Ionuț Pistol[1]
[1]*Facultatea de Informatică, Universitatea "Al. I. Cuza", Iaşi*
[2] *Institutul de Cercetări pentru Inteligență Artificială, Academia Română, Bucureşti*
[3] *Institutul de Informatică Teoretică, Academia Română, Filiala Iaşi*
corinfor@info.uaic.ro, mdcristea@info.uaic.ro

.



Extraction de collocations monolingues et bilingues : application à la traduction

STEFANESCU Dan, ICIA, Academia Română
CEAUSU Alexandru, ICIA, Academia Română
ION Radu, ICIA, Academia Română
TODIRASCU Amalia, LILPA, Université Marc Bloch Strasbourg
HEID Ulrich, IMS Stuttgart, Universität Stuttgart
GLEDHILL Christopher, LILPA, Université Marc Bloch Strasbourg
ROUSSELOT François, INSA Strasbourg

Avant-propos
Program / Programme
Comunicări / Communications
Alte comunicări / Autres communications
Sponsorizare / Partenariat
Expozanți / Exposants

Introduction

Le projet « Collocations en contexte : extraction et analyse contrastive» (financé par l'*Agence Universitaire pour la Francophonie*) a comme objectif le développement d'un système d'extraction semi-automatique de collocations, qui exploitent des corpus alignés. Les collocations posent des problèmes pour les traducteurs, ainsi que pour les systèmes d'aide à la traduction, en ce qui concerne leur utilisation dans le bon contexte et sur le plan des choix des composants (*prendre une décision* se traduit en roumain par *a lua o decizie*, mais en anglais la traduction est réalisée avec le verbe 'make' : *to make a decision ; to make good damages* se traduit par *dédommager* en français, mais par *'a compensa daunele'* en roumain).

Pour ces raisons, les collocations ont fait l'objet de nombreuses études qui se concentrent sur la définition de la notion de collocation, l'identification de leurs propriétés syntaxiques, sémantiques et pragmatiques (Grossmann, Tutin, 2003), la création de dictionnaires électroniques (Blumenthal, 2007), (Mel'cuk et. al, 1984) et le développement d'outils d'extraction automatique des collocations. Malgré l'existence de quelques dictionnaires électroniques qui contiennent des descriptions de collocations en plusieurs langues (BLF (Verlinde et al, 2003)), peu de ressources qui proposent une description complète des propriétés morpho-syntaxiques des collocations et la plupart des dictionnaires offrent une vue limitée sur les données du dictionnaire. Pour pallier ce manque de ressources, des outils d'extraction de collocations à partir des corpus utilisent des méthodes statistiques (Evert, 2005) ou syntaxiques (Tutin, 2004, Seretan et al, 2004). Si les méthodes statistiques identifient beaucoup de candidats invalides (la précision est alors faible), les méthodes syntaxiques nécessitent beaucoup de ressources linguistiques (la précision est meilleure, mais les ressources nécessaires sont très complexes). Pour trouver un compromis entre la précision et le rappel, des méthodes hybrides (Daille, 1996) (appliquées pour identifier des termes dans un domaine limité) ont été développés (Smadja et al, 1990). Notre outil d'extraction combine méthodes statistiques et informations morpho-syntaxiques sur le comportement linguistique des collocations. Nous privilégions une approche avec peu de connaissances linguistiques, qui exploitent des corpus étiquetés et alignés.

Nous présentons la méthodologie adoptée pour ce projet et les corpus utilisés. Nous allons présenter l'analyse linguistique qui serve de base pour définir les filtres linguistiques. Les outils d'extraction monolingues et bilingues sont présentés en détail, ainsi qu'une série de données extraites à partir des corpus anglais et français.

Le projet

Pour atteindre cet objectif, nous avons adopté d'une part la définition traditionnelle de Hausmann (2004), qui considère la collocation comme une relation syntaxique entre une base et son collocatif, et d'autre part la notion morpho-syntaxique des collocations proposée pour l'allemand par (Heid, Ritz 2005), (Ritz, Heid 2006) qui prennent en compte les informations contextuelles autour des collocations (le

complément direct ou indirect, le nombre, genre de la base ou du collocatif etc.). Selon cette approche, le contexte permet de définir des propriétés qui identifient les classes de collocations. La méthodologie que nous avons adoptée ici comporte donc plusieurs étapes de travail, et combine des méthodes statistiques et linguistiques:

- l'étude des données contextuelles d'une classe particulière des collocations (les constructions verbo-nominales (VN)) pour le roumain, le français et l'anglais, à partir des corpus monolingues et multilingues.
- l'identification des classes de collocations pour les langues étudiées, à partir des propriétés morpho-syntaxiques contextuelles.
- l'extraction des candidats collocationnels pour chaque langue, en s'appuyant sur des méthodes statistiques.
- l'application, sur l'ensemble des candidats extraits pour les trois langues, des critères de classification syntaxique pour le filtrage des candidats (à base de patrons definis au 2)).
- l'identification des équivalences de traduction pour un ensemble de corpus similaires alignés (roumain, français, allemande, anglais).
- la constitution d'une base de données linguistiques des collocations multilingue ;

Nous présentons ici une méthode d'extraction qui s'appuie sur une analyse linguistique concernant une classe spécifique des collocations, de type V-N, réalisée sur plusieurs corpus.

Les corpus

Nous avons utilisé des corpus monolingues étiquetés qui permettront d'explorer les contextes des collocations dans chaque langue, et des corpus alignés multilingues qui seront utilisables pour une évaluation des résultats des trois outils. Nous utilisons un corpus multilingue aligné (*AcquisCommunautaire-ACQ,* (Steinberger et al, 2006)), en français, roumain, anglais et allemand. Il s'agit d'un corpus très spécialisé, contenant la législation européenne publiée depuis 1950 et caractérisé par un style juridico-administratif impersonnel et riche en expressions figées. Nous avons sélectionné un échantillon de documents comparables dans les trois langues (environ 15 millions de mots pour chaque langue du projet). Nous avons également utilisé des corpus monolingues contenant des journaux en français (*Le Monde 2004, le Monde diplomatique 1980-1998* : environ 44 millions de mots), en allemand (une collection de journaux, 60 millions de mots) et en roumain (des textes journalistiques, des romans et des textes médicaux - 10 millions de mots). Pour l'anglais, nous avons utilisé le corpus BNC Baby. Ces corpus ont été étiquetés à l'aide du TreeTagger (Schmid, 1994) pour l'allemand, le français et l'anglais (*ACQ* et les corpus monolingues pour l'allemand et le français) et le TTL pour le roumain (Ion, 2007). Le corpus *ACQ* est aligné au niveau de propositions et des mots. Les outils d'extraction, ainsi que les outils d'alignement seront présentés dans les sections suivantes.

Les constructions Verbe-Nom

Vu la durée du projet, nous nous sommes concentrés sur l'étude d'une classe de collocations spécifique, les collocations de type Verbe-Nom. Ainsi, nous considérons que les collocations sont plus que des co-occurences (Sinclair, 1992), mais plutôt des constructions (une relation lexico-grammaticale qui est établie entre les deux éléments) (Golberg, …), (Hausmann, 2004).

Nous avons analysé les combinaisons Verbe-Nom dans un modèle inspiré par (Halliday, 1985), sur trois niveaux (Gledhill, 2007) : la structure lexicale, la fonction syntaxique (Sujet, Prédicat, Complément), le type de procès exprimé par la construction Verbe-Nom. Du point de vue structurel, (Glehill, 2007) identifie deux classes de constructions VN à identifier: les prédicateurs complexes (le verbe et le nom forment un bloc et jouent le rôle de prédicat ensemble : *avoir peur, a ține seama/tenir compte, take account*) et les structures prédicat+complément, ou le complément est plus indépendant (*lua decizii/'prendre des decisions', make decisions*). Pour identifier ces classes d'une manière automatique, nous avons identifié un ensemble de propriétés morpho-syntaxiques, caractéristiques aux verbes et aux noms. Ainsi, les propriétés des constructions VN propres aux verbes sont : être équivalent à un seul verbe (*make decisions = to decide*, mais non *a face obiectul=*a obiectal 'to make objections'*), la valence et le passif (*Comisia face un raport președintelui/'La Commission fait un*

rapport au président' et *Raportul a fost făcut de către comisie/'Le rapport a été fait par la commission'*, mais *Comisia face față situației/'La comission fait face à la situation'* mais il n'est pas possible d'utiliser la forme passive: **Față a fost făcută*). En ce qui concerne les propriétés caractéristiques au nom, nous étudions le déterminant (l'absence ou la présence systématique de l'article), la modification par une clause relative (qui n'est pas possible pour les prédicateurs complexes) ou la conversion vers un groupe nominal (*comisia a luat o decizie = luarea deciziei*, mais pas *Nous avons fait l'hypothèse=le fait de l'hypothèse*).

Aucune de ces propriétés n'est pas suffisante pour identifier d'une manière unique la classe de constructions, seul le rôle sémantique joué par le nom (qui précise le procès exprimé par le verbe), permets de faire la distinction entre la classe prédicateur complexe et une structure classique prédicat+complément. Si le nom remplit la fonction de complément de portée (qui spécifie le procès exprimé par le prédicat), alors il s'agit d'un prédicateur complexe.

Cette dernière analyse peut être validée seulement manuellement, mais les propriétés morpho-syntaxiques sont identifiés automatiquement dans un corpus étiqueté. Les **prédicateurs complexes** sont caractérisés par une invariabilité morpho-syntaxique prononcée (absence du déterminant, pas de passivation possible etc.), alors que les structures **prédicat+complément** sont plus flexibles et le complément est plus indépendant sur un plan syntagmatique.

L'outil d'extraction monolingue

Nous avons développé un outil d'extraction monolingue des collocations, et nous utilisons la technologie développée par RACAI pour aligner des textes en deux langues et pour dériver des équivalences au niveau des lexèmes. Ce même système permet également d'identifier des équivalences pour des couples de mots, non seulement pour des mots contigües, mais pour des paires qui ont les deux propriétés suivantes:

a) la distance entre les deux mots est relativement stable à l'intérieur des textes analysés (la déviation standard calculée est inférieure à un seuil de 1,5)

b) la cooccurrence des deux éléments est statistiquement significative, les deux mots apparaissent ensemble plus souvent que par hasard (LogLikelihood - LL).

Les résultats obtenus suite à la combinaison des deux propriétés indiquent que l'utilisation du score LL calculé pour des paires de mots qui remplissent des critères morpho-syntaxiques ou bien de la moyenne de la distance entre les mots constitue une approche efficace pour l'extraction des collocations du type verbe-nom (V-N/N-V).

Une fenêtre de 11 mots (qui est le contexte où on cherche les cooccurrences) parcourt chaque phrase du texte lemmatisé et étiqueté, et chaque mot devienne, à un moment donné, le centre de la fenêtre [1]. Les mots qui sont introduits dans la fenêtre sont des noms ou des verbes; les autres parties des unités morphologiques ne sont pas pris en compte. Une base de données stocke toutes les paires de mots (sous forme de lemme) formées entre le centre de la fenêtre et les autres mots de cette dernière, et la distance [2] entre les mots qui composent ces paires. A titre de paires intéressantes, nous avons gardé les paires auxquelles correspond une déviation standard inférieure à un seuil de 1,5 (Manning & Schutze, 1999). Pour les paires intéressantes extraites, on calcule le score LL (qui doit être supérieur au seuil de 9).

Employant la méthode ci-dessus présentée sur le corpus ACC roumain et français, nous avons obtenu une liste de collocations ordonnées en fonction du score LL. Nous retrouvons un ensemble de collocations commun aux deux langues (*take account / ține cont; make decisions / lua masuri ; give rise / face obiectul*). Certaines, comme *enter into force/entrer en vigueur, a fi adoptat la Bruxelles/done in Brussells* sont des expressions spécifiques du langage juridique du corpus ACC.

Mot1	Mot 2	LL	Art	Nb	Comp.	Classe
aduce/ 'apporter'	Atingere/ 'ateinte'	51567.34864	-	Sg	Datif	Prédicateur complexe
Înlocui/ 'remplacer'	Text/ 'texte'	43992.3067	Déf	Sg, pl	Acc	Prédicat+ complément
Intra/ 'entrer'	Vigoare/ 'vigueur'	42527.03736	-	Sg	Acc (în/'en')	Prédicateur complexe
Avea/ 'avoir'	Tratat/ 'tratat'	32050.11219	Déf	Sg, pl	Acc	Non
Face/	Obiectul/	30729.47663	Déf	Sg	Datif	Prédicateur

'faire'	'obiectul'					complexe
Modifica/ 'modifier'	Regulamentul/ 'reglement-le'	29141.39454	déf, -	Sg, pl	Acc (la/'à', din/'de')	Prédicat+ complément
Lua/ 'prendre'	Considerare/ 'considération'	27062.0349	-	Sg	Nom	Prédicateur complexe
Ține/ 'tenir'	Cont/ 'compte'	26635.12649	-	Sg	Acc (de/'de')	Prédicateur complexe

Fig. 2. Les paires V-N les plus fréquentes, identifiées par un ensemble de propriétés morpho-syntaxiques plus ou moins invariables (nombre, déterminant, cas du complément indirect).

Nous pouvons constater la présence des propriétés morpho-syntaxiques pertinents comme l'absence du déterminant (-), la préférence pour le nombre singulier et pour un cas spécifique pour les prédicateurs complexes dans les deux langues. Nous pouvons observer que les co-occurrences proposées pour l'anglais contiennent moins de prédicateurs complexes:

Mot1	Mot2	LL	Art	Nombre	Prép.	Classe
Have/avoir	Regard/ 'vue'	139337.613681525	-	Sg	to	Non
Do/ 'faireé	Brussels/ Bruxelles	58421.7707215154	-	Sg	in	Non
Treaty/ traité	Establish/ etablir	55994.7655599668	déf	Sg, pl	-	Non
Regulation/ règle	Bind/ lier	30228.6672675113	-			Non
Replace/ remplacer	Text/ texte	28022.7117842671	Déf,	Sg, pl	With, at	Préd+ complément
Day/ jour	Follow/ suivre	27337.5653023155	Déf	Sg, pl		Non
Take/ prendre	Account/ compte	26833.7653197018	-	Sg	of	Prédicateur complexe
Give/ donner	Rise/ croissance	16367,174	-	Sg	to	Prédicateur complexe

Les candidats collocationnels identifiés par l'outil d'extraction statistique monolingue, seront filtrés à l'aide des propriétés morpho-syntaxiques identifiées sur les corpus : la présence ou l'absence d'un article, la préférence pour une classe de prépositions etc.). Voilà quelques exemples de patrons qui permettent l'identification des prédicateurs complexes en roumain et en anglais :

Exemples

1) face *{1} NSRY *{1,5} NxOY

> NSRY est un nom défini, nombre singulier, en cas accusatif ;NxOY est un nom défini, en cas datif ; *{x,y} - entre x et y mots peuvent apparaître entre le nom et le complément indirect.
> Candidats sélectionnés parmi la liste des candidats les plus fréquents: *face obiectul unui contract/faire l'objet d'un contrat, face față unei situații/faire face à une situation…*

2) take *{1, 2} NNx

> NNx - nom (singulier ou pluriel), *{1, 2} - un ou 2 mot entre le verbe et le nom
> Les candidat sélectionnés parmi la liste des candidats les plus fréquents: *take into account, take measures*

L'outil d'extraction bilingue

L'alignement

L'alignement automatique des textes est une étape nécessaire pour plusieurs applications TAL utilisées dans un contexte multilingue. L'alignement de deux corpus comparables offre des informations supplémentaires entre les relations établies entre un texte et sa traduction. Ces relations sont établies entre divers unités textuelles, des paragraphes, des phrases ou des mots. Dans notre cas, l'alignement lexical (au niveau des mots et des unités lexicales) sont utilisés pour trouver des constructions VN équivalentes (des collocations) dans les quatre langues du projet (anglais, français, allemand, roumain). Les alignements sont initialement réalisées sur les paires anglais-français, anglais-allemand, anglais-roumain. A partir de ces alignements, nous pouvons générer les paires français-allemand, français-roumain et allemand-roumain. Avant l'alignement lexical, il est nécessaire de passer par une étape d'alignement au niveau de la proposition ou de la phrase (qui n'est pas dépendante des langues source et cible). Une des méthodes d'alignement au niveau de la phrase (Gale, Church, 1991) utilise la longueur moyenne des phrases (le nombre de mots ou de caractères). Chen (1993) a développé une méthode qui s'appuie sur l'optimisation des probabilités de traduction des mots, une méthode qui est plus efficace que la méthode de Church, mais qui nécessite des ressources complexes. Le système de (Melamed, 1996) s'appuie sur un mappage géométrique et fait l'hypothèse qu'une proposition relative (calculée en fonction de la longueur du texte) va s'aligner avec une proposition située sur la même position relative. Une des méthodes d'alignement au niveau de la proposition qui est très répandue est celle de Moore (2002). Cette méthode hybride fonctionne en trois étapes : pendant la 1ere étape l'application utilise des méthodes qui s'appuient sur la longueur de la proposition et l'alignement géométrique. La deuxième étape filtre les candidats pour retrouver les alignements qui sont sûrs et, sur cette base, l'algorithme construit une liste d'équivalents de traduction. La troisième étape utilise les listes d'équivalents de traduction pour corriger l'alignement.

Le système de Moore, ainsi qu'une méthode hybride développé par RACAI (Ceaușu et al, 2006) vont constituer les outils pour aligner les corpus ACQ au niveau des propositions pour les 4 langues étudiées. Le système d'alignement de RACAI utilise un classifieur de type vecteur support et il n'est pas dépendent de la langue. Son entraiment doit se faire à l'aide d'une partie restreinte du corpus (200 propositions, alignées manuellement). L'application utilise la distribution LIBSVM (Fan et al., 2005), avec les valeurs implicites pour les paramètres d'entraînement.

L'alignement lexical a été réalisé initialement manuellement, et à l'aide de Uplug (Tiedemann, 2003). Pour l'alignement des corpus anglais-roumain, nous avons utilisé un outil d'alignement développé par RACAI (Tufis et al, 2005) qui utilise un algorithme iteratif. Pour chaque itération, l'alignement lexical va aligner divers catégories de mots comme les entités nommées, les nombres les dates, les mots contenu, les mots grammaticaux et les signes de ponctuation. L'alignement utilise plusieurs valeurs seuils pour identifier les alignements pertinents, à chaque itération. Un alignement entre deux unités lexicales (mots ou expressions) est caractérisé par une serie de propriétés dont les valeurs font partie de l'intervalle [0,1]. L'alignement lexical est indépendent de corpus et des unités lexicales qui le composent et peut être exploité dans un autre contexte. Ainsi, l'alignement lexical peut servir pour identifier les équivalents de traduction pour les collocations de type VN.

Un exemple d'extraction

Les candidats confirmés par l'outil d'extraction bilingue, sont, pour la plupart, des collocations pertinentes, où chaque élément de la collocation est traduit de la même façon qu'en dehors de cette combinaison. Les candidats collocationnels non confirmés sont également intéressants pour les traducteurs : la majorité de ces cas sont non-compositionnels ; c'est-à-dire leurs traductions impliquent souvent des traductions non-standard (mots uniques ou expressions). Nous présentons quelques données extraites à partir des corpus anglais et roumain (ACQ). Les équivalents de traduction communs peuvent être intégrés dans le dictionnaire multilingue de collocations, puisque nous retrouvons des spécificités pour chaque langue (le verbe utilisé qui n'est pas toujours la simple traduction de l'autre verbe, certains prépositions *-into, on-* sont nécessaires et sont intégrés dans le groupe verbal):

To ask support= a cere sprijin/ demander de l'aide
To take into account= a ține cont/tenir compte
To find out the truth=a afla adevarul/apprendre la vérité
To carry on the activities= a exercita activitățile
To make decisions=a lua decizii/prendre des décisions

Quelques cas spécifiques ont été identifiés dans le corpus aligné et qui n'ont pas d'équivalent de traduction, la traduction est un verbe. Les candidats extraits du corpus anglais sont caractérisés par une invariance des propriétés morpho-syntaxiques et sont des prédicateurs complexes L'alignement lexical permet alors d'identifier les traductions correctes :

Turned a blind eye/lit. 'tourner un oeil aveugle'= a ignorat/ignorer
Meets the eye/lit. 'rencontrer l'oeil'= a se vedea/voir
Give a hard time to/lit. 'donner un temps difficile'=a priponi/attacher

Conclusion

Nous avons présenté un outil d'extraction de collocations monolingue, qui combine méthodes statistiques et filtrage linguistique, mais également un outil d'extraction de collocations multilingue, qui nécessite des corpus alignés. La méthode d'alignement (propositionnel et lexical) est présentée en détail. Les candidats extraits pour chaque langue sont confrontés à une liste d'équivalents de traduction établie à partir des corpus alignés et sont utilisés pour alimenter un dictionnaire multilingue de collocations.

Remerciements éventuels

Ce projet est financé par le réseau « Lexicologie, Terminologie, Traduction » de l'Agence Universitaire de la Francophonie. Nous remercions : D. Bourigault pour le corpus ACC français annoté syntaxiquement, R. Mihalcea (Université de Texas) pour le corpus NAACL, Dan Tufis (Académie Roumaine) pour le corpus RoCo et Dan Cristea (Université de Iasi, Roumaine) pour le corpus L4TE.

Références

Blumenthal, P., (2007) A Usage-based French Dictionary of Collocations, in: Y. Kawaguchi/T. Takagaki/N. Tomimori/Y. Tsuruga (éds.): *Corpus-Based Perspectives in Linguistics*, Amsterdam u.a.: Benjamins (Usage-Based Linguistic Informatics 6), 67-83.

Ceauşu, A., Ştefănescu, D., Tufiş, D. (2006). Acquis Communautaire Sentence Alignment using Support Vector Machines, LREC 2006, Genoa, Italy

Chen, S.F. (1993). Aligning Sentences in Bilingual Corpora Using Lexical Information. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, 9-16

Daille, B. (1996) Study and Implementation of Combined Techniques for Automatic Extraction of Terminology , in Resnik, P. (ed.) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, MIT Press, page 49—66.

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart

Fan, R.-E., Chen, P.-H. and Lin, C.-J. (2005). Working set selection using the second order information for training SVM. Technical report, Department of Computer Science, National Taiwan University

Gale, W.A., Church, K.W. (1991). A program for Aligning Sentences in Bilingual Corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California, 177-184

Gledhill, C., (2007). La portée : seul dénominateur commun dans les constructions verbo-nominales, in Frath, P., Pauchard, J. & Gledhill, C. (éds) *Actes du 1er colloque Res per nomen,* Reims 24-36 mai 2007, Université de Reims, Champagne, 113-124.

Gledhill C., Heid U., Mihăilă C., Rousselot F., Ştefănescu D., Todiraşcu A., Tufiş D. & Weller M. 2007. *Collocations en contexte: extraction et analyse contrastive*, Project Report for the *Agence Universitaire pour la Francophonie* 'Réseau Lexicologie, Terminologie, Traduction', Paris :1-38.

Grossmann, F., Tutin, A.(eds.) (2003). Les collocations: analyse et traitement, Numéro special: « *Travaux et Recherches en Linguistique Appliquée* », ISSN: 1572-042X

Hausmann, F.J. (2004). Was sind eigentlich Kollokationnen?, en K.Steyer (eds.) *Wortverbindungen - mehr oder weniger fest*, 309-334

Halliday, M., (1985). *An Introduction to Functional Grammar*. London, Arnold.

Heid, U., Ritz, J. (2005) Extracting collocations and their contexts from corpora, *Actes de COMPLEX-2005*, *Conference on Computational Lexicography and Text Research*, Budapest

Ion, R. (2007). TTL: A portable framework for to-kenization, tagging and lemmatization of large corpora. Research Institute for Artificial Intelligence, Romanian Academy, Bucharest (in Romanian), 22p.

Manning, C. D., Schütze, H. (1999) Foundations of statistical natural language processing, MIT Press.

Mel'čuck I. & all (1984) Dictionnaire explicatif et combinatoire du français contemporain, Recherches Lexico-Sémantiques, Montrèal, Presses Universitaires de Montréal.

Melamed, I. D. (1996). A Geometric Approach to Mapping Bitext Correspondence. IRCS Technical Report 96-22, University of Pennsylvania

Moore, R. C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In Machine Translation: From Research to Real Users (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, pp. 135-244

Ritz, J., Heid, U. (2006) Extraction tools for collocations and their morphosyntactic specificities, in: *Proceedings of the Linguistic Resources and Evaluation Conference, LREC-2006*, Genova, Italia, 2006.

Seretan, V., Nerima, L., Wehrli, E. (2004). A tool for multi-word collocation extraction and visualization in multilingual corpora, *Proceedings of EURALEX'2004*, Lorient, France, Vol2, pp.755-766

Sinclair, J., (1991). *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

Smadja, F. A., McKeown, K. R. (1990), Automatically extracting and representing collocations for language generation, Proceedings of the 28th annual meeting on Association for Computational Linguistics, 252-259, Pittsburgh, Pennsylvania.

Schmid, D. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C. Erjavec, T., Tufiş, D., Varga, D. (2006), The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, *Proceedings of the 5th LREC Conference*, pp.2142-2147.

Tiedemann, J (2003). Combining clues for word alignment. In Proceedings of the 10th EACL, Budapest, Hungary: 339-346

Todirascu, A, Gledhill, C, Stefănescu, D. (2007), Extracting Collocations in Context: the case of Romanian VN constructions, in *Proceedings of RANLP'2007*, Bulgaria

Tutin, A (2004). Pour une modélisation dynamique des collocations dans les textes, Actes du congrès EURALEX'2004, Lorient, France, 2004, Vol. 1, 207-221

Tufiş, D., Ion, R., Ceauşu, A., Stefănescu D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments, *EACL2006,* Italy.

Tufiş, D., Ion, R., Ceauşu, A., Stefănescu D. (2005c). Combined Aligners. In Proceeding of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond", Ann Arbor, Michigan, Association for Computational Linguistics, pp. 107-110

Verlinde, S., Selva, T., Binon, J. (2003) Les collocations dans les dictionnaires d'apprentissage: repérage, présentation et accès, en Grosman F., Tutin, A. (eds.). *Les collocations: analyse et traitement* / - Amsterdam: De Werelt, 2003. - p. 105-115

[1] A ce stade, nous appliquons la méthode de Smajda. Ce dernier nous permet d'identifier des paires intéressantes de mots non nécessairement adjacents.

[2] La distance est négative si la paire est composée du mot-cible et d'un mot placé avant

ce dernier.

Sponsor de aur / Partenaire Or

Sponsor de argint / Partenaire Argent

Ambassade de France en Roumanie

Association Européenne de Terminologie

Sponsor de bronz / Partenaire Bronze

Comitetul Electrotehnic Român

Asociaţia Română de Terminologie

Seminar Internaţional „Instrumente pentru asistarea traducerii" - Acte / Séminaire international « Les outils d'aide à la traduction » - Actes
Editat de / Edité par : Uniunea latină/Union Latine
ISBN: 978-9-291220-37-3
2008

Uniunea Latină
Direcţia Terminologie şi Inginerie Lingvistică - DTIL
131, rue du Bac - F-75007 Paris
T: (33) 1 45 49 60 62 / F: (33) 1 45 49 67 39
dtil@unilat.org