>>> revista tradumàtica
Traducció i Tecnologies de la informació i la Comunicació
Número 4 : traducció automàtica : desembre 2006 : http://www.fti.uab.cat/tradumatica/revista :
ISSN 1578-7559

La recerca lingüística en la TA

Maite Melero i Nogués
Grup de Lingüística Computacional
Universitat Pompeu Fabra

1. Introducció

La relació de la lingüística amb la Traducció Automàtica ha estat una llarga història d'amor i desamor. Després d'una primera etapa d'ignorància mútua, les dues disciplines van iniciar una etapa de col·laboració intensa, que va germinar en una disciplina nova, la lingüística computacional. Aquesta col·laboració, no exempta de malentesos i dificultats, ha aportat sens dubte beneficis mutus. Si d'una banda, els sistemes de TA han aprofitat les descripcions formals de les llengües i s'han anat fent lingüísticament més complexos, de l'altra, la necessitat de la seva aplicació pràctica ha obligat a les teories lingüístiques a expressar els seus principis de forma completa i coherent.

Ocasionalment els sistemes de TA lingüístics han adoptat una teoria lingüística preexistent, però més freqüentment han establert un nou marc de referència adaptat al problema específic de la traducció, inspirant-se en un conjunt de teories prèvies. La primera estratègia afavoreix la reusabilitat del sistema, així com la transferència del coneixement d'una implementació a una altra, però exigeix de la teoria una major cobertura i robustesa. La segona estratègia, més eclèctica, permet una flexibilitat més gran i és la més habitual.

La majoria dels marcs teòrics distingeixen tres nivells de descripció gramatical: morfològic, sintàctic i semàntic. La morfologia s'ocupa de l'estructura de les paraules; la sintaxi, de l'estructura de les frases; i la semàntica s'ocupa del significat. Els conceptes lingüístics d'aquests tres nivells juguen un paper fonamental a l'hora de definir les representacions abstractes emprades per un determinat sistema de TA. Com veurem, l'elecció de la profunditat de l'anàlisi, i per tant de la naturalesa de les seves representacions, es fonamenta sobretot en la naturalesa i la profunditat de les divergències en la traducció de les llengües implicades.

2. Sistemes pre-lingüístics o de traducció directa

Els primers programes de TA seguien una estratègia de "traducció directa" o paraula a paraula, amb un mínim d'anàlisi lingüística. El sistema de traducció desenvolupat a la Universitat de Georgetown (Kay, 1975: 219-232) que, com la majoria dels d'aquella època, pretenia traduir del rus a l'anglès, va ser un dels que va tenir més èxit. En aquest sistema els aspectes lingüístics i computacionals estaven barrejats i s'anaven modificant de forma ad hoc. El

resultat era un codi de dimensions i complexitat enorme, però amb capacitats d'anàlisi lingüística molt rudimentàries. A la dècada dels 70, després del demolidor informe ALPAC (1966) que desaconsellava la recerca en TA per considerar-la poc útil, de baixa qualitat i exageradament cara, pràcticament l'únic sistema que es va mantenir en desenvolupament va ser SYSTRAN, que va adaptar el sistema de Georgetown a la traducció entre l'anglès i el francès i va ser adoptat per la CEE. SYSTRAN va millorar-ne la modularitat, separant l'aspecte lingüístic del computacional, però no va introduir cap avenç en el processament lingüístic, que es limita als següents aspectes:

- Anàlisi morfològica de la paraula origen, que permet extreure'n el lema i la informació morfològica bàsica, com ara nombre, temps verbal, etc.
- Traducció lèxica a través d'un diccionari bilingüe, normalment format per lemes.
- Flexió de la paraula traduïda i reordenació local d'algunes seqüències de mots, per exemple el parell substantiu-adjectiu, però només en situació d'adjacència estricta, ja que no hi ha informació de caràcter sintagmàtic.

En aquest tipus de sistemes no hi ha una metodologia consistent per abordar els problemes de traducció: hom adopta qualsevol solució capaç de millorar els resultats. Per exemple, per a traduir les preposicions russes a l'anglès, SYSTRAN utilitza un sistema ad hoc de classificació semàntica dels verbs i dels noms que són adjacents a aquestes preposicions. Tanmateix aquesta classificació és diferent d'altres que s'utilitzen per altres tasques i té poc a veure amb la semàntica del rus (Hutchins, 1979: 29-52). Totes aquestes estratègies lingüístiques es van codificant per a cada paraula del diccionari bilingüe, el qual, després de dècades de laboriosa feina manual, acumula molta informació específica, que tanmateix no es pot generalitzar, ni per a tota una classe de paraules, ni per a altres parells de llengües. Malgrat la simplicitat del seu model lingüístic, SYSTRAN s'ha imposat com a un dels principals sistemes de traducció per molts parells de llengües, especialment europees, i el seu nivell de qualitat és homologable al d'altres sistemes conceptualment més complexos.

3. Sistemes lingüístics o basats en regles.

Si fins llavors, la TA i la lingüística s'havien ignorat mútuament, la dècada dels 80 va suposar l'inici d'una relació, no per difícil, menys fructífera. És la dècada d'or dels sistemes de TA inspirats en principis de base lingüística, fortament influïts per la noció de gramàtica formal introduïda per Noam Chomsky (1982) i que va suposar una revolució en la teoria del llenguatge.

Els principals sistemes de TA que apareixen durant aquesta època es basen en l'estratègia de la transferència (o transfer) i fan servir estructures de trets com a eina de representació (p.e. METAL (Alonso, 1990: 189-201) i EUROTRA (Allegranza, 1991: 15-123). La traducció per transferència consta dels següents processos lingüístics, aplicats en successió: anàlisi morfològica, anàlisi sintàctica, transferència, generació sintàctica i generació morfològica. La fase d'anàlisi sintàctica (i la seva contrapartida, la generació) suposa un increment considerable de sofisticació lingüística en els sistemes de TA respecte als de l'etapa anterior, ja que requereix una gramàtica completa de la llengua d'origen, és a dir, un conjunt de regles capaç de descriure adequadament totes les frases possibles d'aquesta llengua. Com apuntàvem més amunt, els analitzadors sintàctics es basen en conceptes afins als models de gramàtica

generativa desenvolupats a partir de la formulació original chomskiana, és a dir: regles recursives lliures de context, trets de subcategorització, teoria de la X amb barra i noció de cas i de valència o de papers temàtics. El resultat d'aquesta anàlisi solen ser arbres de sintagmes -o constituents- que, en els sistemes més sofisticats, es converteixen en estructures progressivament més abstractes, que són les que efectivament es tradueixen (o transfereixen) a la llengua de destí.

En els sistemes anomenats d'interlingua, l'estructura que resulta de l'anàlisi és una representació abstracta de base semàntica, lingüísticament neutral i, per tant, en principi no lligada a cap llengua concreta. A l'hora de la veritat, molts dels sistemes d'interlingua, treballen en realitat amb representacions molt "profundes" o abstractes, però no "universals", i que per tant requereixen un cert grau de transferència. Aquest és el cas de sistemes com el Geta-Ariane, (Boitet, 1989: 54-65), que basa les seves representacions semàntiques en el model de sentit-text del lingüista rus Igor Mel'cuk (1998: 3-20); i el Rosetta de Philips (Eindhoven) (1994).

Els dos tipus principals de representació lingüística d'una frase en TA són, en un primer nivell d'anàlisi, els arbres sintàctics o de constituents, que donen informació de dominància (p.e. la categoria SN domina la categoria N) i de precedència (o sigui, d'ordre superficial de les paraules); i, en un nivell més profund, els arbres de dependència, o d'estructura predicat-argument. En aquestes representacions més profundes, només es mantenen els lemes de les paraules amb contingut lèxic (principalment noms, verbs, adjectius i adverbis), mentre que les paraules funcionals o gramaticals (articles, conjuncions, algunes preposicions) es converteixen en trets binaris o booleans (p.e. +definit / -definit), al igual que la informació morfosintàctica procedent dels sufixes flexius (p.e. plural, present).

Les representacions abstractes també poden expressar-se com una forma canònica que neutralitza diferències sintàctiques superficials, que poden ser variants d'expressió en una mateixa llengua, com ara la veu activa o la veu passiva, o bé divergències estructurals entre dues llengües (Hutchins, 1992:103,138). Les divergències estructurals es manifesten quan les dues llengües implicades utilitzen diferents recursos morfològics i sintàctics per expressar un mateix significat. En teoria, quan més divergents són dues construccions, més abstracta hauria de ser la seva representació.

L'estructura de depèndencia, o predicat-argument s'inspira directament en la gramàtica de Valències (ref). Aquesta teoria resulta particularment adequada als problemes de la traducció perquè encara que un verb i la seva traducció a una altra llengua poden tenir estructures sintàctiques diferents, en canvi, solen coincidir en la seva valència és a dir, en el nombre dels seus arguments. Així, en l'exemple 1, el verb anglès look, amb complement preposicional encapçalat per at, es tradueix pel veb català mirar, que regeix un objecte directe sense preposició.

(Ex. 1) The girl looks at the baby.
La noia mira el nen.

Aquesta representació lingüística també resol de forma convenient una altre problema habitual en la traducció, com és la traducció lèxica de les preposicions

regides per un predicat. Aquestes preposicions rarament poden ser traduïdes de forma literal, tal i com mostren els següents exemples: consistir en (cat.) –> consist of (ang.); comptar amb (cat.) –> count on (ang.); dependre de (cat.) –> depend on (ang.).

La gramàtica de Casos incorpora a la de Valències les funcions semàntiques dels complements, i fins i tot dels adjunts. La naturalesa de la relació d'un argument amb el seu predicat s'expressa en termes de papers temàtics, com ara agent, pacient, instrument, experimentador, etc. Aquest afegit teòric permet resoldre molts casos pràctics de divergències de traducció entre dues llengües, que no queden resolts amb una simple estructura de dependència. Un dels exemples més típics és la traducció del verb anglès like pel català agradar.

(Ex. 2) The girl likes the baby.
A la noia li agrada el nen.

Mentre que en la frase anglesa el subjecte del verb és girl (noia) i l'objecte és baby (nen), en la frase catalana els dos arguments ocupen posicions sintàctiques creuades: el subjecte és nen i l'objecte (en aquest cas, indirecte) és noia. Una representació que utilitzi conceptes temàtics, com experimentador i estímul, permet donar una única representació a les dues frases:

predicat: like (agradar)
experimentador: girl (noia)
estímul: baby (nen)

Un problema important de la teoria de Casos és justament la identificació d'un conjunt adequat de papers temàtics.

Les representacions abstractes tenen doncs com a objectiu apropar les dues llengües eliminant-ne alguns detalls específics i destacant-ne les característiques generals, que poden llavors traduir-se més fàcilment. Sovint, però, les divergències de traducció són tan grans que les representacions poden arribar a fer-se molt complexes. Vegem el següent exemple:

(Ex. 3) He swam across the river.
Va travessar el riu nedant.

En aquest exemple els dos predicats (swim / travessar) no es poden traduir directament. Per resoldre aquest tipus de problemes, Dorr (1993) proposa la descomposició semàntica del predicat tot inspirant-se en la teoria de la Semàntica Conceptual de Jackendoff (1983).

event: go (anar)
cosa: he (ell)
camí: to [across river] (cap a [a través del riu])
manera: swimmingly (natatòriament)

Però si, com dèiem, ja resulta compromès identificar un conjunt adequat de papers temàtics en la teoria de Casos, escollir un conjunt de primitius semàntics encara presenta moltes més dificultats. Així, la manera ad hoc com Dorr tria els seus primitius (e.g. swimmingly) ha estat força criticada (Arnold, 1996: 217-

241; Benett, 2003).

Apart de les relacions del predicat amb els seus arguments, altres àmbits de la lingüística teòrica, com ara la teoria del temps i l'aspecte, han aportat solucions al tractament de les divergències de traducció entre les llengües (Allegranza, 1991:37-68). Altres àmbits de la recerca lingüística, com ara les teories del discurs i l'estructura del text, han estat menys considerades per la TA, amb algunes excepcions (Ramm, 1994: 53-75, 77-115; Eberle, 2003: 15-17). En la immensa majoria dels casos, la TA considera la frase com la unitat més gran de traducció. D'aquesta manera, problemes d'abast més gran, com ara la identificació dels antecedents dels pronoms (resolució de l'anàfora), queden generalment no resolts en els sistemes actuals de TA.

Les representacions lingüístiques, siguin de la profunditat que siguin, solen prendre la forma d'estructures de trets. Un mecanisme habitual per processar aquestes estructures és a través de la unificació. Aquest mecanisme es basa en el principi que dues estructures poden combinar-se en una de sola, sempre i quan els valors dels seus trets siguin compatibles. Aquest principi és comú a algunes de les teories gramaticals més importants dels últims anys, com ara la Gramàtica Lèxico-Funcional (LFG) de Bresnan (1982), molt influent en TA (Kaplan, 1993: 193-202); la Gramàtica d'estructura sintagmàtica regida pel nucli (HPSG) de Pollard i Sag (1994), que ha inspirat diversos sistemes de TA, el més recent dels quals és l'anomenat DELPH-IN (Bond, 2005); i per últim la Gramàtica Categorial, molt utilitzada des dels primers temps de la TA per la simplicitat de les seves regles (Bar-Hillel, 1953: 47-58; Oehrle et al., 1988). A nivell semàntic, la Minimal Recursion Semantics, introduïda per Copestake (2005: 281-332), és utilitzada com a interlingua per diversos sistemes de TA (Stymne, 2006: 9-17; Flickinger, 2005: 165-172).

4. Nous sistemes lingüístics: els sistemes híbrids

És justament durant la dècada dels 90 quan es produeix un gir radical en la recerca a l'àmbit de la TA. A finals dels 80, un grup d'investigadors d'IBM aprofitant la creixent disponibilitat de textos en format digital, construeixen el primer sistema de TA totalment basat en models probabilístics (Brown, 1990: 79-85). A partir d'aquí es produeix un creixement espectacular de les aproximacions inductives a la traducció automàtica, en les quals el sistema aprèn a traduir a partir de corpus bilingües alineats. Paral·lelament, els sistemes lingüísticament rics pateixen un cert desprestigi, ja que malgrat els esforços invertits no han estat capaços de superar un cert llindar de qualitat. Així, mentre que sistemes basats en sòlids principis lingüístics es mostren incapaços d'enfrontar-se a instàncies pràctiques de la llengua de cada dia, els sistemes entrenats en corpus de dimensions cada vegada més grans, són capaços d'aprendre molt ràpidament i d'adaptar-se a la llengua dels textos reals.

Tanmateix, els sistemes purament estadístics aviat troben també el seu sostre. D'alguna manera suposen un retorn als temps de la traducció directa i als vells problemes de la manca de generalitzacions. En llengües molt flexives de seguida es presenta el problema de l'escassetat de les dades, ja que en no haver-hi anàlisi morfològica, cada forma constitueix una paraula diferent. També, com era de preveure, la manca d'informació estructural dificulta el tractament de les dependències no locals (interrogatives, relatives) i dels

sintagmes complexos.

D'aquesta manera veiem com en els últims temps, el pèndul es torna a acostar cap als sistemes basats en coneixement lingüístic, sense abandonar però els avenços aconseguits per la TA basada en corpus. Pràcticament tots els sistemes que es consideren estadístics avui en dia, realitzen com a mínim l'anàlisi morfològica de l'entrada i un reordenament sintàctic de la sortida. D'altra banda, la majoria dels sistemes anomenats lingüístics empren alguna tècnica d'anàlisi de corpus, bé sigui per resoldre ambigüitats lèxiques o adquirir terminologia, com fan fins i tot les implementacions més recents de SYSTRAN, o bé per a construir autèntiques arquitectures multi-motor, com ara el sistema d'interlingua Pangloss (Nirenburg, 1995).

És de preveure, per tant, que el paradigma dominant en els propers anys siguin els sistemes híbrids (Carl, 2002), és a dir una combinació de les capacitats generalitzadores dels sistemes lingüístics o basats en regles i de la flexibilitat i robustesa dels sistemes d'aprenentatge automàtic a partir de corpus. No hi ha encara una metodologia precisa de com es poden combinar els dos tipus de tècniques, és a dir, quins components del sistema són més aptes per a una o altra. Una solució típica consisteix a tenir gramàtiques d'anàlisi i generació basades en regles i un component de transferència entrenat amb exemples (Richardson, 2001: 293-298). Alguns sistemes basats en regles adquireixen part del seu coneixement lingüístic a partir de les dades, emprant tècniques d'aprenentatge automàtic (Melero, 2006). Thurmair (2006: 45-48) creu que el millor model híbrid és aquell que incorpora optimitzacions estadístiques a una base formada per regles lingüístiques.

El gran repte per la TA són encara les diferències estructurals entre les llengües, és a dir, les divergències en la traducció i també és aquí on la recerca lingüística, en combinació amb les aproximacions estadístiques, encara pot fer aportacions importants (Ayan, 2004).

Conclusió

L'evolució històrica dels sistemes de TA ens mostra que la lingüística per ella mateixa no pot resoldre el problema de la TA, però que és imprescindible per al seu progrés. Hem vist com després que l'interès general s'havia desplaçat completament cap als mètodes estadístics, el pèndul tornava a oscil·lar cap a les aproximacions lingüístiques. La recerca lingüística pot encara contribuir molt al desenvolupament de la TA, i al problema fonamental de les divergències en la traducció, amb observacions de fenòmens, amb tècniques i teories, que la recerca en TA pot adoptar i integrar amb aproximacions més empiricistes.

Bibliografia

Allegranza, V., P. Benett, J. Durand, F. van Eynde, L. Humphreys, P. Schmidt and E. Steiner (1991). Linguistics for Machine Translation: The Eurotra Linguistic Specifications, in C. Copeland, J. Durand, S. Krauwer and B. Maegaard (eds), *The Eurotra Linguistic Specifications*, Luxembourg: Office for Official Publications of the European Communities.

Alonso, J.A. (1990). Transfer InterStructure: designing an 'interlingua' for transfer-based MT systems. In *Proceedings of the Third Conference on*

Theoretical and Methodological Issues on Machine Translation of Natural Languages (Austin, TX).

Automatic Language Processing Advisory Committee (ALPAC) (1966). Language and Machines: Computers in *Translation and Linguistics.*

Arnold, D. (1996) Parametrizing Lexical Conceptual Structure for Interlingual Machine translation, In *Machine Translation* 11.

Ayan, N. F., Dorr, B. and N. Habash (2004): Multi-Align: Combining Linguistic and Statistical Techniques To Improve Alignments for Adaptable MT. In *Lecture notes in computer science Springer*, Berlin.

Bar-Hillel, Y. (1953) A quasi-arithmetical notation for syntactic description, In *Language* 29.

Benett, P. (2003) The relevance of linguistics of machine translation. In Somers, H. (ed.) *Computers and Translation. A translator's guide*. Amsterdam: Benjamins.

Boitet, C. (1989) GETA Project. In Nagao (ed.) *Machine Translation Summit,* Ohmsha, Tokyo.

Bond, F., Oepen, S., Siegel, M., Copestake, A. and Flickinger, D. (2005). Open Source Machine Translation with DELPH-IN In: Proc. *Open Source MT workshop at MT Summit X*, Phuket, Thailand.

Bresnan, J. (1982) (ed.) *The mental representation of grammatical relations.* MIT Press, Cambridge, Mass.

Brown, P., J. Cocke, S. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer and P.S. Roossin (1990) A statistical approach to machine translation, *Computational Linguistics* 16.

Carl, M., Way A. and Schäler R. (2002). Toward a Hybrid Integrated Translation Environment. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users,* Springer-Verlag London, UK.

Chomsky, N. (1982) *Lectures on Government and Binding*. Foris, Dordrecht.

Copestake, A., Flickinger, D., Sag, I. and Pollard, C (2005). Minimal Recursion Semantics: An introduction. *Journal of Research on Language and Computation*, 3(2--3).

Dorr, B. J. (1993). *Machine translation: A View from the Lexicon*. Cambridge, Massachusetts: The MIT Press.

Eberle, K. (2003) Coordination, incorporation and dynamic semantic representation in transfer, *Controlled language translation, EAMT-CLAW-03*, Dublin City University.

Flickinger, D., Lønning, J.T, Dyvik, H., Oepen, S. & Bond, F (2005) SEM-I rational MT: enriching deep grammars with a semantic interface for scalable machine translation. In *Proceedings MT Summit X*, Phuket, Thailand.

Hutchins, W. J. (1979). Linguistic models in machine translation. In UEA Papers in *Linguistics* 9.

Hutchins, W. J. and H. L. Somers (1992) *An Introduction to Machine Translation*. London: Academic Press.

Jackendoff, R. (1983) *Semantics and Cognition*. Cambridge, Massachusetts: The MIT Press.

Kaplan, R.M. and Wedekind, J. (1993) Restriction and Correspondence-based Translation. In *Proceedings of Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Manchester.

Kay, M. (1975) Automatic translation of natural language In Haugen, E. & Bloomfield; M. (eds.) *Language as a human problem*. Guilford, Lutterworth

Mel'cuk, I. (1998) The Meaning-Text Approach to the Study of Natural Language and Linguistic Functional Models. In *S. Embleton (ed.): LACUS Forum 24*, Chapel Hill: LACUS.

Melero, M. (2006) "Combining machine-learning and rule-based approaches in Spanish syntactic generation". Tesi doctoral. http://www.mcu.es/roai/en/consulta/registro.cmd?id=37052

Nirenburg, S. (ed.) (1995) *The Pangloss Mark III Machine Translation System.* A Joint Technical Report by NMSU CRL, USC ISI and CMU CMT (CMU-CMT-95-145).

Oehrle, R.T., E. Bach and D. Wheeler (1988) (eds) *Categorial grammars and natural language structures*. Reidel, Dordrecht.

Pollard, C. and Sag, I. (1994). Head-Driven Phrase Structure Grammar. University of Chicago Press, Chicago.

Ramm, W. (ed.) (1994) *Text and Context in Machine Translation: Aspects of Discourse Representation and Discourse Processing*. Luxembourg: Office for Official Publications of the European Communities.

Richardson, S., W. Dolan, A. Menezes, and J. Pinkham. (2001). Achieving commercial-quality translation with example-based methods. In *Proceedings of VIII MT Summit*, Santiago de Compostela, Spain.

Rosetta, M. T. (1994) *Compositional Translation.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

Stymne, S. & Ahrenberg, L. (2006). A bilingual grammar for translation of English-Swedish verb frame divergences. In *Proceedings of 11th Annual*

Conference of the European Association for Machine Translation. Oslo, Norway.

Thurmair, G. (2006). Using corpus information to improve MT quality. In *Fifth International Conference on Language Resources and Evaluation. Third International Workshop on Language Resources for Translation Work, Research & Training (LR4Trans-III),* Genoa, Italy.

Desembre 2006