EL PROYECTO DE TRADUCCION AUTOMATICA 'EUROTRA'

UlrikeTallowitz
CELE - UNAM

Para los que tienen interés en conocer un ejemplo dél uso que la lingüística puede hacer de la computadora, quiero describir este proyecto en que trabajé en los años 1989 y 1990.

Se trata de un proyecto europeo de investigación y desarrollo en el campo de la traducción automática.

Lugar: Universidad de Saarbruecken, RFA.

El motivo para la fundación de este proyecto fue el multilingüismo de la Comunidad Europea, y la necesidad de sus instituciones de traducir una cantidad enorme de documentos a las nueve lenguas de la Comunidad. Estas lenguas son: español, danés, alemán, griego, inglés, francés, italiano, holandés y portugués.

Si se quiere traducir de todas las lenguas a todas las otras, resultan 72 “parejas de lenguas". Esto por supuesto presenta un costo elevado si se hace a base de la traducción humana. (Representa 35% - 65% de los costos personales en las distintas instituciones de la Comunidad Europea).

Para resolver este problema hay dos posibles soluciones:,

1. Reducir el número de lenguas oficiales.
2. Hacer Innovaciones dentro de los servicios de traducción.

Por razones políticas y culturales, sólo la segunda es aceptable.

La meta es de crear un sistema prototípico de traducción automática, para todas las lenguas de la Comunidad, aplicado a un arca de textos limitado: textos oficiales administrativos, en las areas de economía, tecnología y política. Se quiere cubrir un léxico de alrededor de 20000 entradas en cada lengua, como meta final del proyecto.

FUNDAMENTOS TEORICOS LINGÜISTICOS

Se usan teorías lingüísticas que se basan en el mecanismo de "unificación", como son la "Lexical Functional Grammar" (LFG) o la "General Phrase Structure Grammar" (GPSG) para crear un nuevo formalismo para la descripción del proceso de traducción.

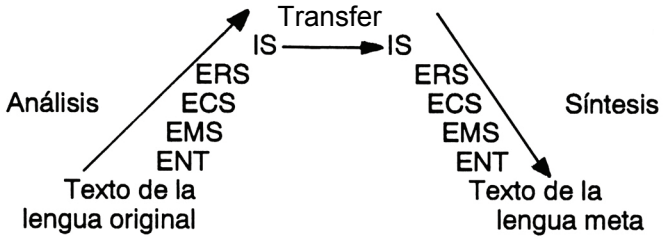
La hipótesis básica es que el proceso de traducción se puede describir

como una sucesión de proyecciones, de tal manera que un texto pasa por un cierto número de representaciones gramaticales. Estas representaciones son determinadas por criterios lingüísticos:

hay un nivel morfológico,
un nivel de sintaxis de superficie,
un nivel de sintaxis profunda, y
un nivel sémantico-sintáctico.

EL PROCESO DE TRADUCCION

Se compone de tres fases:
el Análisis
el Transfer
la Síntesis



ENT= EUROTRA Normalized Text
EMS= EUROTRA Morphological Structure
ECS= EUROTRA Constituent Structure
ERS= EUROTRA Relational Structure
IS = Interface Structure

ENT Y EMS no están implementados todavía. Las traducciones se hacen de ECS a ECS.

EUROTRA es un sistema a base de 'transfer', es decir, en un nivel de gramática profunda (Interface Structure), el transfer de una lengua a la otra se hace estructuralmente. Las estructuras básicas sintácticas de la lengua fuente se transfieren, con reglas complejas, a las estructuras de la lengua meta.

El análisis y las síntesis son monolinguales.

Cada nivel de representación está definido por una GRAMATICA: un generador que consiste en un número finito de reglas.

Los diferentes niveles están relacionados por TRADUCTORES: un número finito de reglas que transforman un nivel en otro.

Un ejemplo:

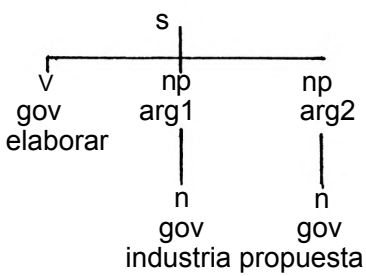
Regla (constructor) en la gramática alemana en el nivel IS, para oraciones con verbos divalentes tal como:

1) Die Industrie arbeitet einen Vorschlag aus.
2) (La industria elabora una propuesta)

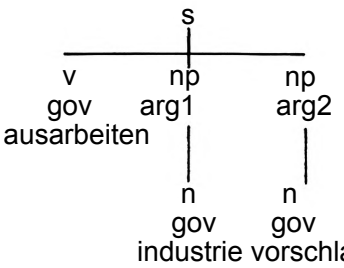
```
bS2          ={cat=s)   [{cat=v,   isframe=arg12, arg1=A, arg2=B,
d_pformarg2=P,
             humarg1=H, humarg2=HH, abstrarg1=AB, abstrarg2=AAB},
            ({role=arg1, cat=np, sr=A, hum=H, abstr=AB};
             {role=arg1, cat=s, sr=A}),
           ^{role=arg2, cat=np, sr=B, d_pform=P, hum=HH, abstr=AAB};
            {role=arg2, cat=s, sr=B, d_pform=P};
            {role=arg2, cat=pp, sr=B, huma=HH, abstr=AAB};
            {role=arg2, cat=ap, sr=B, hum=HH, abstr=AAB);
            {role=arg2, cat=advp, sr=B, hum=HH. abstr=AAB}),
           *{role=mod, cat=pp},
           *{role=mod, cat=advp},
           ^{role=mod, cat=s, stype=subord},
           ^{role=mod, cat=np},
           ^{role=mod, cat=negp}].
```

```
      cat  = category
        s  = sentence
        v  = verb
       np  = nominal phrase
  isframe  = marcodependencial del verbo
      arg  = argument
      mod  = modifier
d_pformarg2 = preposición que toma el argumento2
```

La oración española está representada en el nivel IS de la siguiente forma:



Esta estructura se transforma a la estructura correspondiente alemana, a base de la regla bS2:



Para generar de esta estructura de gramática profunda la representación superficial, se proyecta enseguida al siguiente nivel, en este caso: ERS, y después pasa de ERS a ECS.

Estos procesos se hacen por medio de la 'unificación', es decir si coinciden las características (feature-value-pair) de los objetos lingüísticos con las de las reglas gramaticales y lexicales, se unifican las estructuras para producir una nueva estructura bien formada. Si hay contradicción en los 'features', no se crea la estructura.

De ahí resulta que cuando una regla de la gramática es muy general, o sea muy poco especificada, se producen varios objetos, incluyendo objetos no correctos lingüísticamente. En este caso se habla de 'overgeneration'. Entonces hay que precisar la regla de tal manera queproducesolamente objetos correctos, sin destruir aquellos objetos quetambién sonposibles. Es decir, hay que formular las reglas suficientemente generales para admitir todas las posibles realizaciones de una estructura básica, y al mismo tiempo suficientemente específico para rechazar estructuras falsas.

Esto no es siempre posible. Un ejemplo es el uso del artículo definido en diferentes lenguas:

Español:

3) Trabaja en la industria.
4) Trabaja en el ejercicio.
5) Trabaja en la industria electrónica.
6) Trabaja en la industria europea.
7) La industria europea se ha desarrollado rápidamente desde la guerra.

Inglés:

8) He works in industry.
9) He worksin the military.
10) He works in the electronics industry.

11) He works in European industry.

12) European industry has developed rapidly since the war.

Aquí no es posible formular una regla que dice: el artículo definido en español debe producir un artículo definido en inglés. Pero tampoco es cierto que el artículo en español siempre se traduce en una construcción sin artículo en inglés. Hay que formular muy precisamente los casos en que se debe o no producir el artículo, y esto dentro de un sistema a base de la oración, sin usar el contexto del texto. Este problema no se ha resuelto completamente todavía.

Otro ejemplo:

El participio de pasado se analiza como AP (adjective phrase) en español, y traduce en una oración relativa en alemán:

13) Las medidas tomadas por la industria...

14) Die Massnahmen, dle von der Industrie ergrlffen wurden...
Mientras que

15) Las medidas ya están tomadas,
se debería traducir así:

16) Die Massnahmen sind schon ergrlffen worden. (voz pasiva).

Utilizando la regla que traduce la oración (13), se creó el siguiente objeto:

*Dle Massnahmen sind schon, dle von der Industrie ergrlffen wurden.

Entonces hubo que especificar las reglas para la construcción "estar+participio del pasado" => voz pasiva. (15) => (16) En este caso se necesitaba también un cambio en el análisis del grupo español: ahora se analiza la construcción "estar+participio del pasado" como voz pasiva, lo que hace el transfer más sencillo.

EL TRANSFER COMPLEJO

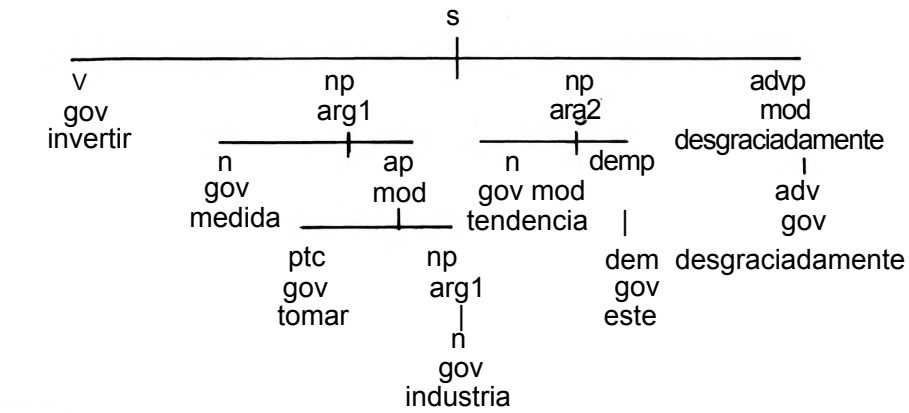
El análisis y la síntesis se hacen independientemente de la lengua meta, o de la lengua fuente, respectivamente. Esto permite a cada grupo nacional de investigación de utilizar los modelos gramaticales que ellos prefieren. La meta es de elaborar estos dos módulos (análisis y síntesis) lo más completo posible para así simplificar el transfer. En el mejor de los casos se trata de un transfer de unidades semánticas solamente, y las diferentes realizaciones sintácticas de cada lengua se crean sólo en los niveles superiores. Sin embargo, esto no es posible en cada caso, y muchas veces hay que recurrir a "transfer complejo".

Ya vimos un ejemplo de 'transfer sencillo' arriba: (1)=> (2). El siguiente ejemplo, sin embargo, requiere 'transfer complejo':

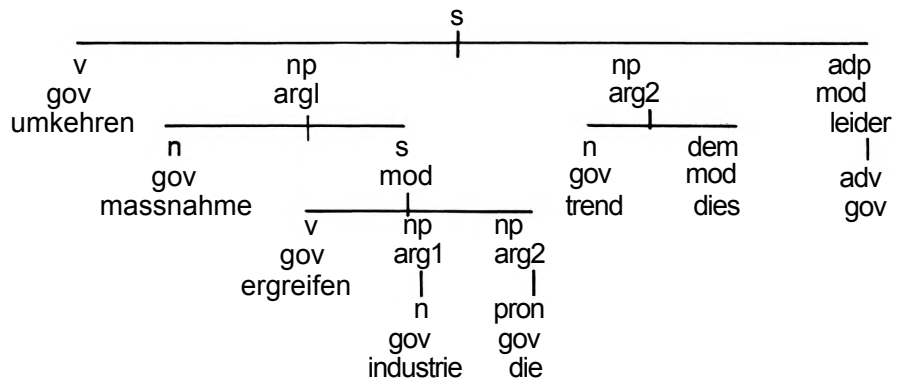
17) Desgraciadamente las medidas tomadas por la Industria no han

invertido esta tendencia.

18) Leider haben die Massnahmen, die von der Industrie ergriffen wurden, diesen Trend nicht umgekehrt.



===>



La regla que hace este transfer se encuentra en ei 'traductor' de la IS española a la IS alemana:

```
tRELPASpor=~: {cat=ap, role=mod} [V: {cat=ptc,nb=NB},
                  NP1: {cat=np, e_pform=por,role=arg1},
                  MOD: *{role=mod}]
=> {cat=s,d_voice=pass,stype=rel,role=R,sASPECT=retrospective,
    sTENSE=simul}
   < {cat=v},NP1, {cat=np,argtype=full,role=arg2,nb=NB,sNEG=no}
   <{cat=n,ntype=pron,prontype=rel,nb=NB}>, MOD > .
```

PROBLEMAS

De los múltiples problemas que todavía quedan en la traducción automática quiero enumerar solamente algunos:

— Anáforas, pronominalización

Como he dicho arriba, el sistema se basa en la unidad lingüística de la oración. No se pueden todavía hacer relaciones entre diferentes oraciones, o tomar en cuenta características del texto.

Consideremos las siguientes oraciones españolas:

19) El consejo vino hoy. Rechazó la propuesta de la Comisión.

No hay forma en este momento de definir cual pronombre en alemán (masculino, femenino o neutro) se debe insertar en la segunda oración.

Se crean tres objetos, con un pronombre diferente en cada uno.

— La selección del artículo definido (ver arriba)

— Disambiguación lexical

En el caso de palabras polisémicas es a veces muy difícil de encontrar el sentido correcto.

Ejemplo:

20) Der Rat verabschiedet das Gesetz.

21) (El consejo decreta la ley)

22) Der General verabschiedet die Soldaten.

23) (El general despide a los soldados).

Como 'verabschleden' en alemán se traduce tanto a 'decretar' como a 'despedir' en español, se necesitan 'lexical semantic features' para disambiguar los significados.

Las entradas en el léxico (en el nivel IS):

```
verabschieden1=...
frame={arg1={sr=agent,sft={abstract=abstr,abstr=phen,phen=soc,
            soc=inst,anim=hum}},
            arg2={sr=affected,sft={info=sem,abstract=abstr,
            anlm=nll}}}).

verabschleden2=...
    frame={arg1 = {sr=agent,sft={anim=hum}},
            arg2={sr=affected,sft={anlm=human,info=nil}}}).
```

En el análisis de la oración alemana la entrada VerabschiedenV está escogida porque el primer argumento 'der fíat' también está codificado con las características phen=soc,soc=inst (institución). La segunda oración está analizada con el verbo 'verbschieden2' que tiene para su primer argumento la característica anim=hum, que coincide, es decir unifica, con el sujeto 'der General'.

En el tansfer al español Verabschiedenl' se traduce a 'decretar' y 'verabschieden2' a 'despedir.

Un grupo de investigadores de diferentes países está tratando de retinar este sistema. Uno de los problemas en este aspecto es en que medida se debe o se puede utilizar "world knowledge" para definir los 'features', o si se debería más bien definirá base de características sintácticas (distributional model).

DATOS TECNICOS

El sistema completo (database tools, menu systems, command interpreter, compilers etc.), es decir, el código binario (binaries) comprende 4,4 megabyte (Mb).

Se agregan 8 - 10 Mb para el código fuente, y 1,5 Mb para los diccionarios, las gramáticas y los traductores (en esta fase de investigación).

El sistema de trabajo es UNIX.

Lenguas de programación: PROLOG, C y UNIX shell scripts.

Se usa la base de datos relacional UNIFY.

Las computadoras que se usan son:
DEC Vax, Microvox y workstations,
HP workstations,
SUN workstations y servers,
con capacidades de
8 - 12 Mb RAM (memoria del acceso al azar) y
800 Mb Disc.

B I B L I O G R A F I A

BRESNAN, J. (1982) *The Mental Representations of Grammatical Relations.* Cambridge, Mass.: MIT-Press.

CHOMSKY, N.(1981) *Lectures on government and binding.* Dordrecht: Foris.

HALLER, J. (1987) *Das EUROTRA-Projekt: Stand 1987 und Ausblick.* Sprache und Datenverarbeitung, Nr. 1, Saarbrücken.

HORN, G. (1983) *Lexical Functional Grammar.* Berlin: Mouton Pub.

JACKENDOFF, R.S. (1977) *X-Bar Syntax: A study of phrase structure.* Cambridge, Mass.: MIT-Press.

---------- (1983) *Semantics and Cognition.* Current Studies in Linguistics Series, Vol.8. Cambridge, Mass.: MIT-Press.

---------- (1987) *Consciousness and the Computational Mind.* Cambridge, Mass.: MIT-Press.

NIRENBURG, S. (ed.) (1987) *Machine Translation: Theoretical and methodological Issues.* Cambridge: Cambridge University Press.

RADFORD, A. (1981) *Transformational syntax.* Cambridge: Cambridge University Press.

SELLS, P. (1985) *Lectures on contemporary syntactic theories.* Stanford: Centre for the study of language and information.

STEINER, E/Schmidt, P./Zelinsky-Wibbelt, C. (eds.) (1988). *From syntax to semantics: insights from machine translation.* London: Frances Pinter and Norwood, N.J.: Ablex.

SCHMIDT, P. (1989). *Papers on Syntax in EUROTRA.* EUROTRA-D Working Papers Nr.13, Vol.1: General Issues. EUROTRA-D Working Papers Nr.14, Vol.2: ID/LP.

WINOGRAD, T. (1983) *Language as a Cognitive Process.* Vol. 1 Syntax. Reading,Mass.: Addison Wesley Publ.Company.

ZELINSKY-WIBBELT, C. (1989) *Machine Translation Based on Cognitive Linguistics:* What Lexical Semantics Contributes to the Semantic Unity of a Sentence. EUROTRA-D Working Papers Nr.16.

ANEXO: Estas publicaciones se pueden pedir de EUROTRA, Saarbrücken (Ulrike Tallowitz):

EUROTRA-D WORKING PAPERS

Nr:	Autor(en) und Titel
WP 2/86[1]	Steiner, Erich, Ursula Eckert, Birgit Weck, Jutta Winter. The Development of the EUROTRA-D System of Semantic Relations, 1986.
WP 3/87	Steiner, Erich (Ed.). Investigating Semantic Relations: Papers on Theory Transfer, and Generation, 1987. E. Steiner, Semantic Relations in LFG and in EUROTRA-D -a comparison. U. Eckert und U. Heid. Methoden zur teilautomatischen Erstellung von Transferwörterbüchern (syntaktische Funktionen nach LFG, semantische Relationen nach STEINER 1986). B. Weck, U. Held, D. Rosner. Zur Generierung deutscher Sätze aus semantischen Repräsentationen auf der Basis der Verbklassifikationen von STEINER: Erfahrungen aus dem EUROTRA-D/SEMSYN-Experiment.
WP 4/87[2]	Zelinsky-Wibbelt, Cornelia. Semantische Merkmale für die automatische Disambiguierung: Ihre Generierung und ihre Verwendung, 1987.
WP 5/87	Steiner, Erich, Jutta Winter, Cornelia Zelinsky-Wibbelt. Aspects of Determination and Focus in a Multilingual MT System, 1987.
WP 6/88	Schütz, Jörg (Ed.). Workshop Semantik und Transfer, 1988. P. Schmidt. Transferprobleme in EUROTRA. U.Heid, K. Netter, J. Wedekind. Zur Funktionsweise des Transfers auf f-Strukturen. Ch. Hauenschild, C. Umbach. Funktor-Argument-Struktur. J. Schütz, R. Sharp. CAT2R - Komplexität eines Formalismus für multilinguale MÜ. E. Steiner, J.Schütz. A first outline of the co-operation between Penman (ISI/USC) and ET-D (IAI/UdS).
WP 7/89	Kapanadze, Oleg. Fragen zur maschinellen Übersetzung zwischen Deutsch und Georgisch, 1989.
WP 8/89	Truar, Matthias. The EUROTRA DataBase Environment, 1989.
WP 9/89	Rochemont, Michael. Implementing Focus in Machine Translation, 1989.

WP 10/89 Schütz, Jorg. Towards a Framework for Knowledge-based Machine Translation, 1989.

WP 11/89 Steiner, Erich (Ed.). Predicate-Argument Structure for Tranfer, 1989.
Erich, Steiner. Argument Structure: Grammatical Issues.
Erich, Steiner, Ursula Reuther. Semantic Relations
Erich, Steiner, Jutta Winter. Focus

WP 12/89 Bateman, John. Upper Modelling for Machine Translation: a level of abstraction for preserving meaning, 1989.

WP 13/89 Schmidt, Paul. Papers on Syntax in EUROTRA. Bd. 1 : General Issues, 1989.

WP 14/89 Schmidt, Paul, Jorg Schütz. Papers on Syntax in EUROTRA. Bd. 2: ID/LP, 1989.

WP 15/89 Abramson, Harvey. Extending a Logic Grammar Formalism with Features and Embedded Rulës, 1989.

WP 16/89 Zelinsky-Wibbelt, Cornelia. Machine Translation Based on Cognitive Linguistics: What Lexical Semantics Contributes to the Semantic Unity of a Sentence, 1989.

WP 17/89 Rothkegel, Annely. Polylexi^alität. Verb-Nomen-Verbindungen und ihre Behandlung in EUROTRA, 1989.

1 Out of print
A revised version of WP 2/86 can be found in Steiner/Eckert/Weck/Winter. 1987. The application of aspects of Systemic Functional Grammar to Machine Translation in EUROTRA-D. Duisburg: LAUD- papers in Linguistics.

or

Steiner E., The Development of the EUROTRA-D System of Semantic Relations, in: Steiner E. & Schmidt, P. & Zelinsky-Wibbelt, C.(eds.) 1988. From syntax to semantics: insights from machine translation. London: Frances Pinter & Norwoordm, N.J.: Ablex

2 Out of print
A revised version of WP 4/87 can be found in Zelinsky-Wibbelt, C. EUROTRA-D Working Papers Nr. 16, 1989. Machine Translation Based on Cognitive Linguistics: What Lexical Semantics Contributes to the Semantic Unity of a Sentence.