http://dx.doi.org/10.12795/PH.1997.v11.i02.05

EL PAPEL DE LA SUBLENGUA EN LA TRADUCCIÓN AUTOMÁTICA

Elena Bárcena

This article considers the role of Sublanguage in Machine Translation (MT). Firstly, the intrinsic linguistic difficulties of the processing and translation of natural languages are discussed. A proposed solution to such difficulties is to limit the type of system input, which, as will be seen, can be done in several ways. A particularly effective way is the restriction of the input to a sub-genre which corresponds to a sublanguage. Secondly, some existing definitions of the term "sublanguage" are presented, along with the main principles of Sublanguage Theory which are relevant to Natural Language Processing, and specifically to MT. Thirdly, a working definition of sublanguage in the context of MT is subsequently proposed, and the advantages and disadvantages of the sublanguage-based approach are presented. Fourthly and finally, some conclusions are drawn about the feasibility of sublanguage-based MT and its future applications.

1. Introducción

Hace varias décadas, muchos investigadores de Traducción Automática (TA) proclamaban la pronta aparición de sistemas de alta calidad para textos arbitrarios y sin intervención humana significativa (Reifler 1958:517). Tales predicciones naturalmente no se cumplieron. Había una insalvable distancia entre la desmesurada ambición de estos investigadores y la simplicidad de sus estrategias de traducción (Weber 1987), por no mencionar las severas limitaciones de la tecnología computacional del momento. Los intensos esfuerzos de investigación lingüística y computacional que se han venido realizando desde entonces no han producido los resultados que cabía esperar (con notables excepciones) (Somers 1993:232). Así se ha llegado a la conclusión, hoy plenamente generalizada, de que la TA de textos no restringidos, sin interacción del usuario y cuyo output sea de calidad comparable a la que producen los traductores profesionales es un objetivo inalcanzable en el futuro previsible.

Las razones de tal inviabilidad son múltiples y se encuentran tanto en las dificultades para alcanzar una formalización del conocimiento lingüístico suficientemente sofisticada, como en las limitaciones de los ordenadores relativas al tamaño de su memoria, capacidad de almacenamiento y velocidad de procesamiento. También se puede hablar de que ha habido cierta falta de compenetración entre el trabajo de los lingüistas teóricos y los requisitos prácticos de la implementación computacional. Pero hay además una serie de razones profundas que se encuentran en la propia naturaleza de las lenguas humanas. Primero, la combinación de los tamaños de los inventarios léxico y gramatical de una

I.S.S.N. 1132-0265

lengua dada produce un número teóricamente infinito de mensajes aceptables -y no aceptables- desde distintas perspectivas lingüísticas. Esto implica que el sistema de TA debe poseer un conocimiento vasto y complejo gramátical, contextual y del mundo real. Segundo, la polisemia de muchas palabras y los múltiples análisis sintácticos de ciertas oraciones resultan en complejos casos de ambigüedad léxica y estructural. Tercero, la equivalencia entre la forma gramatical de las cadenas superficiales y su contenido semántico y lógico es a menudo poco evidente. Las lenguas hacen uso de recursos como la elipsis y la adición de palabras gramaticales. Estos fenómenos deben ser identificados e interpretados por el sistema como exponentes de reglas y principios de la comunicación verbal o restaurados por la correspondiente categoría o construcción más profunda que se supone ocultan. Cuarto, hay importante información pragmática (contextual, intencional) necesaria para la correcta comprensión -y traducción- de un mensaje que queda implícita en éste (Somers et al. 1990:271) y a la que el sistema no suele tener acceso.

Raskin (1987:57) enumera algunos de los cruciales fenómenos translingüísticos que hacen de la TA una empresa particularmente ardua dentro del Procesamiento del Lenguaje Natural:

> "a) no hay una correspondencia exacta entre las formas morfológicas de dos lenguas distintas.
>
> b) las estructuras sintácticas no pueden ser generalmente copiadas de una lengua a otra.
>
> c) debido a diferencias en la articulación semántica, la misma palabra puede ser traducida de modo distinto en dos oraciones.
>
> d) un elemento de signficado puede tener que perderse [o añadirse] al traducir.
>
> e) los cambios significativos en la traducción pueden deberse a la necesidad de controlar la información 'dada-nueva' o 'tópico-foco'.
>
> f) una paráfrasis significativa puede ser necesaria por razones de locución.
>
> g) hay sofisticada información adicional de tipo pragmático, e.g., fórmulas de trato dependiendo del estatus (relativo) de los participantes, que puede determinar el resultado de la traducción." (*mi traducción*).

Por razones así, ya algunos de los primeros críticos como Bar-Hillel (1951:237), a la vez que desistían del tipo de TA que reuniera todas las características de automaticidad, generalidad y calidad a las que se aspiraba al principio, sugirieron continuar la investigación en este campo sacrificando alguna de dichas características en un mismo sistema. Así surgieron, entre otras, dos importantes estrategias: interacción con el usuario y restricción del input. Concentrémonos en la segunda.

2. Restricción del input

La táctica más extendida de restricción de input en la historia de la TA ha sido el 'enfoque general basado en corpus', que limita el objeto de traducción a un sub-campo en sus versiones experimentales e incluso operacionales (e.g., CETA [textos de matemática y física], ALP [textos eclesiásticos mormones], METAL [documentos técnicos] [Slocum 1984:6-9]). Sin embargo, "sus limitaciones [...] se ven como temporales: la expansión a otras áreas temáticas está anticipada" (Hutchins & Somers 1992:323) (*mi traducción*), lo

cual significa que el diseño y la operatividad del sistema no pueden beneficiarse sustancialmente de tales restricciones circunstanciales.

Otra estrategia es la que se ha venido a llamar el 'uso de sintaxis restringida y vocabulario controlado' (Hutchins & Somers 1992:151-152), que consiste en la imposición de restricciones estructurales y léxicas a los autores de los textos que van a ser traducidos, i.e., la expresión de los textos ha de ceñirse a un conjunto predefinido de construcciones y palabras (e.g., el uso de Systran en Rank Xerox y TITUS). A pesar del relativo éxito de estos enfoques, debe admitirse que limitan los escenarios de aplicación del sistema ya que, por ejemplo, se requieren autores entrenados *ex profeso*. Además, la espontaneidad y naturalidad de la lengua quedan inevitablemente afectadas (hasta tal punto que hay autores que hablan, por ejemplo, del "francés-Systran" [Sager 1986:166]).

Veamos las demandas reales de traducción. No hemos de olvidar para qué se investiga la aplicación de los ordenadores a la traducción de textos: principalmente para satisfacer las necesidades urgentes de cuerpos políticos y militares, científicos, tecnólogos, personas de negocio y profesionales que deben acceder a información o comunicarse en idiomas que no conocen (Gross 1988:xi). La lengua de estos textos goza normalmente de tales características que muchos críticos como Kittredge & Lehrberger (1982:3) han llegado pronosticar que "el éxito comercial de la TA en el futuro previsible depende seguramente de la posibilidad de escribir [...] gramáticas para textos en campos particulares" (*mi traducción*). Las lenguas en las que están expresados la mayoría de estos textos se llaman 'sublenguas'.

El 'enfoque de TA basada en sublengua' diverge de los anteriores de dos maneras. Primero, aquí el diseño del sistema está completamente adaptado a las características específicas de la sublengua que va a traducirse, con el fin de aprovechar del mejor modo posible su carácter cerrado, homogéneo y sistemático y excluir el conocimiento lingüístico irrelevante que complicaría el sistema innecesariamente. Segundo, las sublenguas conllevan una serie de usos restringidos y divergentes con respecto a sus correspondientes lenguas estándares que han sido desarrollados por los propios hablantes de acuerdo con sus necesidades comunicativas. A veces algunas normas son dictadas por comités de estandarización. En cualquier caso, su vocabulario y estructuras gramaticales existen antes y al margen de que se considere la creación de un posible sistema de TA para tal sublengua.

3. Principios de Teoría de Sublengua

Una comunidad de especialistas de un sub-campo tecnológico, científico o profesional está normalmente unida por una serie de conocimientos comunes sobre el dominio que va más allá de los conocimientos de los hablantes de la lengua estándar y por tanto comparte usos léxicos, semánticos, sintácticos y pragmáticos. Tras la observación de este hecho surgió una de las primeras definiciones de 'sublengua' como "la lengua utilizada por una comunidad específica de hablantes, esto es, por aquéllos interesados en un tema particular o envueltos en una ocupación especializada" (Bross et al. 1972:1303) (*mi traducción*). Este fenómeno se debe a la capacidad y tendencia de las lenguas humanas a adaptarse a las distintas situaciones comunicativas, manteniendo así un máximo nivel de eficiencia.

Sager (1982:9) dice que "los artículos de investigación de un sub-campo científico dado muestran tales regularidades de co-ocurrencia en comparación a los de la lengua en general que es posible escribir una gramática de la lengua utilizada en el sub-campo" (*mi traducción*). Debe puntualizarse, sin embargo, que la limitación semántica del discurso no es una condición suficiente para la identificación de una sublengua. Ni siquiera todos los textos científicos y técnicos gozan del estatus de sublengua. Además, la comunidad de hablantes de una sublengua no está siempre bien definida. Por ejemplo, hay algunas sublenguas escritas en las que el acceso a los textos es relativamente libre. De acuerdo con un gran número de críticos, este hecho no altera en absoluto su entidad como sublenguas si hacen usos distintivos y homogéneos de formas y fenómenos lingüísticos.

El carácter regular de una sublengua se ve fácilmente en la recurrencia y co-ocurrencia de los elementos léxicos. Las palabras suelen aparecer sólo como una parte de la oración y a veces su uso está incluso más restringido. Hay también una tendencia a emplear la misma palabra para referirse a un concepto específico. Además, las palabras tienen hábitos de aparición junto con otras palabras vecinas. A nivel sintáctico, las sublenguas también se suelen caracterizar por su inflexibilidad para expresar un tipo de mensaje, a diferencia de otras lenguas como las literarias o periodísticas en las que prima y se valora la diversidad o riqueza de expresión. Lo cierto es que cuanto más especializado y estructurado es el contenido del dominio, más regular e inflexible se espera que sea la correspondiente sublengua, i.e., más rígidas las barreras entre lo que se puede y no se puede decir. Esta propiedad es consecuencia del hecho de que las sublenguas reflejan la estricta organización existente en la parte del mundo real que describen, mientras que "la lengua general impone sólo la estructuración más amplia sobre nuestra percepción del mundo" (Harris 1982:235) (*mi traducción*).

Como las sublenguas están limitadas en referencia a un dominio temático específico, utilizan un número relativamente pequeño de palabras y construcciones. Se suele decir, por tanto, que sus inventarios léxico y gramatical están cerrados. Investigaciones sobre esta cuestión han establecido que el número de palabras de una sublengua puede oscilar entre algunas centenas y varios miles. El carácter cerrado no es binario y se debe hacer una distinción entre cierre absoluto y relativo o limitado (Montgomery & Glover 1986:158). Para ello Kittredge (1982:124) propone lo siguiente: "En una sublengua cuyo léxico está especificado hasta un nivel de confianza de 99,99%, esperaríamos encontrar una nueva palabra cada 10.000 palabras de texto nuevo en la sublengua. Para muchas sublenguas, este puede ser un nivel de confianza demasiado alto" (*mi traducción*).

Así pues, otro paso en la caracterización de las sublenguas fue la identificación de notables restricciones en comparación con las lenguas estándares y de aquí la definición de sublengua de Kittredge & Lehrberger (1982:2): "aquellos conjuntos de oraciones cuyas restricciones léxicas y gramaticales reflejan los conjuntos restringidos de objetos y relaciones que se encuentran en un dominio de discurso" (*mi traducción*). El carácter cerrado de las sublenguas, como dice Lehrberger (1986:20), ha hecho que se crea en ocasiones que los diccionarios y gramáticas de las sublenguas derivan de los de las lenguas estándares simplemente por eliminación de los elementos irrelevantes. Sin embargo, para

llegar a la descripción completa de una sublengua, se requiere la supresión de algunas acepciones léxicas y reglas sintácticas en algunas ocasiones, la modificación de otras para que cubran casos particulares y la inserción de algunas nuevas. La aplicación de una regla estándar puede resultar en una oración no gramatical en una sublengua dada y viceversa. Así pues, la relación entre una sublengua y la correspondiente lengua estándar se describe mejor, no como una relación de inclusión (o total independencia), sino como una de semiautonomía o intersección (Lehrberger 1986:23).

Finalmente, es interesante recordar que las sublenguas comparten la mayoría de las propiedades universales de las lenguas estándares, tales como capacidad generativa ilimitada y la completez o capacidad de "describir por sí misma[s] cualquier situación imaginable, cualquier mensaje en el área de la realidad a la que sirve[n]como lengua" (Moskovich 1982:193) (*mi traducción*), en ambos casos, por supuesto, con respecto al submundo que describen.

4. El enfoque de TA basado en sublengua

Una cuestión fundamental al considerar la aplicación del enfoque de sublengua para el diseño y desarrollo de sistemas de TA es hallar una metodología que permita la identificación de una sublengua apropiada para este fin específico. El paradigma de área temática es claramente insuficiente para obtener una sublengua. En los últimos años, varios autores han comenzado a reconocer la existencia de una segunda coordenada. Siguiendo la terminología de Biber (1988:206) la denominamos 'género' y definimos como un criterio de clasificación textual que se basa en elementos externos relacionados con el propósito del autor y la función práctica del texto. En esta línea, un 'subgénero' es la combinación compatible de un género y un área temática dados. Una vez recopilado un corpus multilingüe representativo de un subgénero, el cual va a constituir la fuente de conocimiento del sistema, es necesario realizar un análisis de dicho corpus para distinguir los 'tipos textuales' del subgénero, que son los modelos lingüísticos que siguen sus textos, y clasificar los textos del corpus en base a dichos tipos textuales. El número de tipos textuales de una sublengua para TA ha de ser reducido y los textos que corresponden a cada tipo textual han de compartir una serie de rasgos fundamentales en cuanto a estructura y formato, construcciones gramaticales y usos léxicos, intra- y translingüísticamente.

Veamos ahora de qué modo las propiedades que acabamos de presentar influyen en la construcción y el uso de sistemas de TA basados en sublengua. En primer lugar, debido a las divergencias de las sublenguas entre sí y con respecto a la lengua estándar, hay que decir que un lexicón y una gramática de una sublengua o lengua estándar no proporcionaría una descripción de todo y sólo el contenido léxico y gramatical de otra sublengua, e incluso es posible que algunas entradas y reglas de distintos sistemas lingüísticos entraran en conflicto en el mismo sistema. Esto hace del enfoque de TA basado en sublengua una necesidad, más que una conveniencia, para la traducción de textos expresados en uno de estos sistemas lingüísticos.

4.1. Ventajas

Las ventajas de procesamiento de los sistemas basados en sublengua están en parte relacionadas con restricciones cuantitativas y cualitativas en comparación con los sistemas de aplicación general. Por ejemplo, la capacidad de almacenamiento del ordenador, el tamaño de su memoria y la velocidad de procesamiento que se requiere es mucho menor. También es más fácil en principio reunir un corpus representativo de una sublengua.

En cuanto al análisis sintáctico, a menudo se pueden explotar las características regularidades superficiales de las sublenguas y realizar un análisis sintáctico relativamente superficial. Se espera que los típicos problemas de los analizadores sintácticos como la ambigüedad, la elipsis o la correferencia pronominal (intra- y supra-oracional) se atenuen. Esto es en cierta medida debido a las propias características de las sublenguas: por ejemplo, la información de los textos en la lengua fuente suele estar expresada clara y explícitamente, lo cual reduce el número de posibles interpretaciones y se esperan menos casos de ambigüedad categorial debido a que las palabras tienden a aparecer siempre bajo la misma categoría sintáctica. El otro motivo es que es más fácil captar y formalizar de modo efectivo el conocimiento semántico, pragmático y del mundo real necesario para resolver dichos problemas. Esto puede hacerse de múltiples maneras según las características intra- y translingüísticas de cada sublengua, lo cual se determina tras el análisis de un corpus representativo. El siguiente comentario de Somers & Jones (1991:158) es un ejemplo del rol que desempeña el estatus de sublengua en el sistema EBIGeM para la generación multilingüe de anuncios de empleo:

> "Vemos la sublengua como un factor esencial que auna y define el conocimiento contextual expresado en el modelo intencional, el conocimiento lingüístico en las representaciones y el conocimiento del dominio para el conjunto del sistema." (*mi traducción*).

De particular interés para la TA es el carácter translingüístico de las sublenguas. Cuando el mundo referencial de una sublengua y las intenciones comunicativas de sus hablantes son comunes a través de los distintos idiomas, es probable la sublengua tenga fuertes equivalencias translingüísticas. De hecho, se ha observado que una sublengua a través de distintos idiomas es más homogénea en términos de vocabulario y estructuras oracionales que sublenguas distintas dentro del mismo idioma (Kittredge 1982:109). Esta propiedad es fundamental para simplificar el diseño del sistema, que no requiere un profundo nivel de abstracción en la descripción lingüística ni complejos módulos de transferencia.

Se espera que la simplificación del diseño y mecanismo mejore su robustez y, por tanto, el nivel de confianza del ingeniero y el usuario y que la frecuencia, la duración y el coste necesarios para diseñar, desarrollar, operar y mantener el sistema con sus correspondientes modificaciones y actualizaciones sean considerablemente menores. Finalmente, es interesante mencionar que a pesar de que la diacronía afecta a todos los sistemas lingüísticos naturales y de que los sub-campos de referencia científicos y tecnológicos evolucionan con especial rapidez, las sublenguas pueden considerarse en general sistemas lingüísticos relativamente estables, particularmente aquéllas cuyo medio de expresión es la

escritura. La ventaja de dicha estabilidad es de tipo práctico y está relacionada con la rentabilidad del proceso de construcción y aplicación del sistema.

4.2. Inconvenientes

La sección anterior trataba las ventajas de ajustar el diseño de un sistema para la traducción de una sublengua a las características de dicha sublengua. Hay también dos problemas que pueden surgir precisamente de tan ceñida adaptación: la habilidad del sistema para traducir una ocurrencia no-estándar y su transportabilidad.

En primer lugar, no todas las sublenguas son perfectamente homogéneas. Aunque raramente, como dice Lehrberger (1986:21), "aquello a lo que nos referimos como textos de sublengua contienen normalmente algún material que no pertenece a la sublengua propiamente dicha" (*mi traducción*). Este material extra, 'hapax legamona' (i.e., tipos léxicos que ocurren una vez [Sebba 1989:33-35) y términos infrecuentes, puede ocurrir como palabras, cláusulas u oraciones enteras intercaladas entre las de la sublengua y puede pertenecer a la lengua estándar o a otra sublengua cualquiera. El problema es que no puede ser analizado por la específica gramática del sistema ya que, además, su presencia y forma son impredecibles.

El segundo problema de la TA basada en sublengua es su transportabilidad para cubrir otras sublenguas (Raskin 1987:55), sin que las modificaciones afecten al diseño básico del sistema. Se puede, sin embargo, hacer una distinción entre sublenguas relacionadas entre sí y las que no lo están. Hay algunas sublenguas que comparten la misma sintaxis pero diferente vocabulario. En este caso se puede, por ejemplo, pensar en un diseño de diccionario que conlleve cierta modularización (Lehrberger & Bourbeau 1988:217), e.g., un diccionario común y una serie de diccionarios especializados de cada sub-campo. Cuando lo que difiere entre las sublenguas es el léxico y la sintaxis, se requieren cambios más fundamentales. La transportabilidad del sistema es entonces complicada y modificarlo para que cubra otras sublenguas puede afectar al buen funcionamiento de reglas existentes. Los escenarios de aplicación y las posibilidades de transportabilidad de cada sistema basado en sublengua son muy limitados, así que se ha de averiguar antes de su construcción si el uso que va a tener el sistema (e.g., el volumen de traducción) compensa el esfuerzo y el coste que conlleva su construcción.

5. Conclusión

El problema de la TA puede verse desde la perspectiva de los esfuerzos necesarios para escribir descripciones formales de lenguas humanas. Confinar el sistema a la traducción de una sublengua es una interesante opción que merece ser explorada y profundizada en distintas direcciones. Además, hoy en día hay un acuerdo prácticamente total entre los críticos de que, dados el estado de evolución actual de la Lingüística Computacional y las demandas reales del mercado, el trabajo de construcción de sistemas prácticos de traducción totalmente automatizada debería concentrarse en las sublenguas ya que, como hemos visto, hay razones para esperar que los mejores resultados prácticos se alcancen con textos que estén constreñidos en forma, contenido y función (e.g., Boitet 1990:130; Hutchins 1986:325; Lehrberger 1982:82; Lehrberger & Bourbeau 1988:128).

BIBLIOGRAFÍA.

Y. Bar-Hillel, "The state of machine translation in 1951", *American Documentation 2* (1951) 153-165.

D. Biber, *Variation across speech and writing* (Cambridge 1988).

C. Boitet, "Towards personal MT: .general design, dialogue structure, potential role of speech", *COLING-90* 3 (Helsinki 1990) 30-35.

I. D. J. Bross, P. A. Shapiro & B. B. Anderson, "How information is carried in scientific sub-languages", *Science* 176 (1972) 1303-1307.

R. Grishman & R. Kittredge (eds.), *Analyzing Language in Restricted Domains: Sublanguage Description and Processing* (Hillsdale 1986).

M. Gross, "Preface", en Lehrberger & Bourbeau xi-xiii.

Z. Harris, "Discourse and Sublanguage", en Kittredge & Lehrberger 231-236.

W. J. Hutchins, *Machine Translation: Past, Present, Future* (Chichester 1986).

W. J. Hutchins & H. L. Somers, *An Introduction to Machine Translation* (Cambridge 1992).

R. Kittredge & J. Lehrberger (eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains* (Berlin 1982).

R. Kittredge, "Variation and Homogeneity of Sublanguages", en Kittredge & Lehrberger 107-137.

J. Lehrberger, "Sublanguage Analysis", en Grishman & Kittredge 18-38.

J. Lehrberger & L. Bourbeau, "Machine Translation. Linguistic characteristics of MT systems and general methodology of evaluation", *Linguisticae Investigationes Supplementa* 15 (Amsterdam 1988).

C. A. Montgomery & B. C. Glover, "A Sublanguage for Repoting and Analysis of Space Events", en Grishman & Kittredge 129-161.

W. Moskovich, "What is a sublanguage? The notion of sublanguage in modern Soviet linguistics", en Kittredge & Lehrberger 191-205.

S. Nireburg (ed.), *Machine translation. Theoretical and methodological issues* (Cambridge 1987).

V. Raskin, "Linguistics & Natural Language Processing", en Nirenburg 42-58.

E. Reifler," The Machine Translation Project at the University of Washington, Seattle", *Proceedings of the 8th International Congress of Linguists* (Oslo1958) 514-518.

J. C. Sager, "Conclusions", *World Systran Conference* (1986) 161-166.

N. Sager, "Syntactic Formatting of Science Information", en Kittredge & Lehrberger 9-26.

M. Sebba, *The Adequacy of Corpora* (Manchester 1989).

J. Slocum, *Practical Issues. Tutorial* (Austin 1984).

H. L. Somers, "Current Research in Machine Translation", *Machine Translation* 7 (1993) 231-246.

H. L. Somers & D. Jones, "Machine Translation seen as interactive multilingual text generation", *Translating and the Computer. The Theory & Practice of Machine Translation a Marriage of Convenience?* 13 (London, 1991) 153-165.

H. L. Somers, J. Tsujii & D. Jones, "Machine Translation without a Source Text", *COLING-90* 3 (Helsinki 1990) 271-276.

H. J. Weber, "Converging Approaches in Machine Translation: Domain Knowledge and Dicours [sic] Knowledge", *Linguistic Agency* B 164 (Duisburg 1987).