Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 343-350 recibido 28-04-2005; aceptado 01-06-2005

Sistema de Traducción Oral para el Castellano, Catalán e Inglés

Victoria Arranz
ELDA – Evaluation and Language Resources Distribution Agency
55-57, rue Brillat Savarin
75013 París, FRANCE
arranz@elda.org

Elisabet Comelles
Centre de Recerca TALP
Universitat Politècnica de Catalunya
C/ Jordi Girona 1-3
08034 Barcelona
comelles@lsi.upc.edu

David Farwell
Institució Catalana de Recerca i Estudis Avançats
Centre de Recerca TALP
Universitat Politècnica de Catalunya
C/ Jordi Girona 1-3
08034 Barcelona
farwell@lsi.upc.edu

Resumen: En este artículo se describe el Sistema de Traducción del Habla FAME, basado en una interlingua y desarrollado para el castellano, el catalán y el inglés. Este sistema es una extensión del sistema NESPOLE!, que traduce el alemán, el francés, el inglés y el italiano, pero cuyos módulos se han integrado en una Arquitectura de Agentes Abierta. El artículo describe la arquitectura general del sistema y el formalismo interlingua utilizado, llamado Interchange Format (IF). A continuación se describen los componentes del módulo de traducción donde se incluyen el reconocedor de voz, la cadena de análisis, la cadena de generación y el sintetizador de voz. También se muestran los resultados de una evaluación del sistema de traducción.
Palabras clave: Traducción Automática, Traducción del Habla, Interlingua, Evaluación, Tecnologías del Habla, Catalán, Castellano, Inglés.

Abstract: This paper describes the FAME Interlingua-based Speech-to-Speech Translation System for Catalan, English and Spanish. This is an extension of the already existing NESPOLE! that translates between English, French, German and Italian, but all modules have now been integrated in an Open Agen Architecture. This article describes the system architecture and the interlingua formalism used, called Interchange Format (IF). In what follows we describe the components of the translation module including the speech recognizer, the analysis chain, the generation chain and the speech synthesizer. We also show the results obtained from an evaluation of the system.
Keywords: Machine Translation, Speech-to-Speech Translation, Interlingua, Evaluation, Speech Technologies, Catalan, English, Spanish.

1 Introducción

En este artículo se describe el sistema de traducción del habla FAME, basado en una interlingua y desarrollado para el castellano, catalán e inglés. Este sistema se ha desarrollado en la Universitat Politècnica de Catalunya, Barcelona, como parte del proyecto FAME financiado por la Unión Europea (para más detalles véase http://isl.ira.uka.de/fame/).

Este sistema de traducción es una extensión del sistema NESPOLE! (Metze et al., 2002) al catalán y castellano y dentro de un dominio restringido, el de reservas de hotel. Es decir, el escenario que tenemos es dos hablantes monolingües de lenguas diferentes, donde uno es el cliente que llama para reservar un hotel y el otro el agente de viajes. Aunque la arquitectura del sistema se basaba inicialmente en la plataforma NESPOLE!, la arquitectura general que integra todos los módulos en la actualidad se basa en una Arquitectura de Agentes Abierta (Holzapfel et al., 2003).

La arquitectura del sistema está constituida por los siguientes componentes: un reconocedor de voz (RV) que convierte la señal de audio de la lengua origen en texto; la cadena de análisis, dónde un analizador de la lengua origen transforma este texto en una representación interlingua; la cadena de generación, donde se transforma esta representación interlingua en texto en la lengua de destino; y por último, un

ISSN: 1135-5948

sintetizador de voz (SV) que convierte el texto en señal de audio en la lengua de destino.

Se ha adoptado un enfoque interlingua debido a la ventaja que esto representa a la hora de incorporar nuevas lenguas al sistema: sólo se necesita desarrollar la parte de análisis y de generación para la nueva lengua, sin tener que modificar los componentes ya existentes. Al contrario que los sistemas de transferencia, en los que hay que desarrollar el módulo de transferencia para cada par de lenguas que queremos cubrir, los sistemas basados en una interlingua convierten la lengua de origen en una representación conceptual y a partir de esta se genera la lengua de destino. De esta forma, los dos únicos módulos que hay que desarrollar son el de análisis y el de generación de cada lengua. Esto permite que ambos módulos puedan ser desarrollados por separado y por un desarrollador monolingüe en la lengua que analiza o genera.

La interlingua utilizada en nuestro sistema se llama Interchange Format (IF) (Levin et al., 2002), utilizada en el consorcio C-STAR (véase http://www.c-star.org) y adaptada a nuestras necesidades. El IF se centra en la identificación de actos de habla y en los diferentes tipos de preguntas y respuestas típicas de un dominio específico, es decir, más que fijarse en las distinciones semánticas y estilísticas, intenta representar la intención del hablante. Aún así, al convertir la lengua de origen en la representación IF hay que tener en cuenta una serie de propiedades, tanto léxicas como estructurales, del castellano y del catalán.

En cuanto a la evaluación se ha realizado partiendo de grabaciones reales donde el cliente es siempre angloparlante y el agente castellano o catalanoparlante. Se han hecho dos tipos de evaluación: una manual, y otra usando métodos estadísticos. Los resultados obtenidos de ambas evaluaciones se han comparado y estudiado.

A continuación describiremos la arquitectura general del sistema, después hablaremos del formalismo interlingua con el que hemos trabajado y continuaremos con una descripción del módulo de traducción para acabar con la evaluación y las conclusiones.

2 *Arquitectura general del sistema*

La arquitectura usada para integrar todos los módulos está basada en una Arquitectura de Agentes Abierta (Figura 1). Este tipo de arquitectura permite integrar fácilmente el resto de módulos que forman parte de la arquitectura general del proyecto (p. ej. la recuperación de información, el procesamiento de imagen, la detección de tópicos, etc.), así como facilitar la comunicación y el intercambio entre ellos. En el caso particular de nuestro sistema de traducción representó una agilización del proceso considerable, efectuando la comunicación entre módulos de un modo mucho más rápido que con la plataforma de NESPOLE!

Como se puede obeservar en la figura 1, hay un gestor de diálogo que acepta la salida de uno de los reconocedores de voz, identifica el idioma de origen y de destino y lo manda al traductor. Este a su vez lo envía al componente de análisis del idioma de origen del servidor HLT, que produce un IF. Este IF pasa entonces por el componente de generación del servidor HLT del idioma de destino. Este lo manda de nuevo al gestor de diálogo que lo envía al sintetizador de voz (SV) del idioma de destino, que produce una versión oral de la traducción.

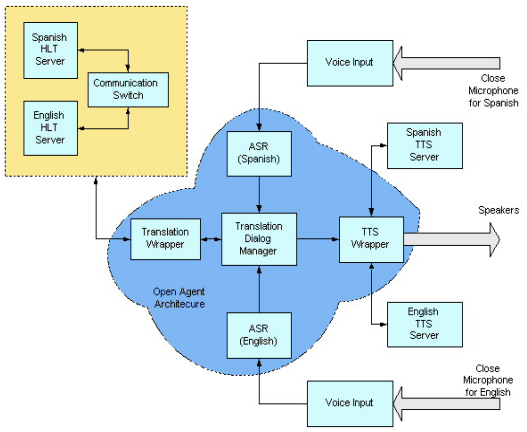

Figura 1: Arquitectura de Agentes Abierta del sistema de traducción

3 *Formalismo Interlingua*

El Interchange Format (IF), la interlingua usada para desarrollar este sistema de traducción automática del habla, está basado en la Teoría de los Actos de Habla de Searle (Searle, 1969) e intenta representar la intención del hablante más que el significado de la oración. Dentro del dominio de reservas de hotel hay diferentes actos de habla como: dar información sobre un precio, verificar una reserva, pedir información sobre el tipo de habitaciones, etc.

En nuestro sistema, estos actos de habla se llaman Actos de Dominio (ADs) y representan el tipo de acciones que se expresan mediante las oraciones. Dentro del IF, estos ADs están

compuestos por un número de elementos que desglosan toda la información semántica que precisa ser comunicada y que se traduce en pequeños elementos que se pueden combinar. Una representación IF está formada por los siguientes elementos:

Etiqueta del Hablante+AD+Argumentos

A continuación pasamos a desglosar y explicar estos elementos:

- La etiqueta del hablante puede ser:
  - *c:* para las intervenciones del cliente
  - *a:* para las intervenciones del agente
- El AD está compuesto por un Acto de Habla obligatorio seguido de uno o más conceptos opcionales:

Acto de Habla + (Actitudes) + (Predicación Principal) + (Participantes de la Predicación)

  - El Acto de Habla es el primer elemento de un AD y es obligatorio. Puede formar un AD por sí mismo o puede ir acompañado de otros elementos que ocupen una posición posterior. A continuación se muestran algunos actos de habla: *give-information, request-information, affirm, apologize*.
  - Las Actitudes son el elemento que representa la actitud del hablante. Este elemento es opcional y puede haber uno o más de uno. Cuando hay más de una actitud, la más general se situa a la izquierda. Algunos ejemplos de actitudes son: *+obligation*, *+disposition*.
  - La Predicación Principal representa el objeto principal del que se habla. Es un elemento obligatorio, a no ser que el AD esté formado tan solo por un Acto de Habla. A continuación se especifican algunas Predicaciones Principales: *+contain*, *+reservation*, *+concept*.
  - Los Participantes de la Predicación representan los objetos de los que se predica. Se trata de un elemento opcional y puede haber uno o más. Algunos Participantes de la Predicación son: *+accommodation*, *+transportation*.
- Los Argumentos. Los elementos del AD se traducen en una serie de parejas argumento-valor que se posicionan en una lista después del AD. Estas parejas se escriben en una lista separadas por ",". Cada argumento contiene subargumentos y/o valores correspondientes.

A continuación ofrecemos un ejemplo (1) junto con su representación interlingua que contiene todos los elementos de los que hemos hablado en esta sección:

(1) ¿Quiere que le reserve una habitación?

a: request-information+disposition+ reservation+room (disposition=(who=you, desire), for-whom=you, who=i, room-spec=(quantity=1, room))

En este caso, el AD está formado por un Acto de Habla *request-information*, que nos indica que se trata de una demanda de información, y una Actitud, *+disposition*, que indica que hay una disposición por parte de un participante. A continuación encontramos una Predicación Principal, *+reservation*, que representa la predicación principal de la oración "hacer una reserva" y un Participante del Predicado, *+room*, que indica el objeto del que se quiere hacer una reserva, "una habitación".

A nivel de argumentos tenemos el argumento correspondiente a *+disposition* que es *disposition=(who=you, desire)* y que expresa que el oyente es una segunda persona del singular y que desea alguna cosa. Los argumentos correspondientes a *+reservation* son *who=i* indicando que quien hace la reserva es una primera persona del singular y *for-whom=you*, que representa que la reserva se hace para una segunda persona del singular. Por último, el argumento que representa el concepto *+room* es *room-spec=(quantity=1, room)*, que nos indica que el número de habitaciones a reservar es una.

Una vez descrita la interlingua utilizada pasamos a describir el módulo de traducción.

4 Módulo de traducción

4.1 El reconocimiento de voz y la cadena de análisis

Para el reconocimiento de voz tanto en castellano como en catalán se ha utilizado el *JANUS Recognition Toolkit (JRTk)* desarrollado en UKA y CMU (Woszczyna et al., 1993). Los

modelos del lenguaje (ML) tanto del castellano como del catalán son modelos de trigramas. Ambos han sido entrenados a partir de textos del mismo dominio pertenecientes a los corpus C-STAR-II[1] y LC-STAR[2], utilizados a su vez para desarrollar el sistema de traducción texto-a-texto. Los modelos acústicos del castellano y del catalán se han entrenado a partir de una base de datos de habla de 30 horas cada uno.

Una vez obtenida la salida del reconocedor, se lleva a cabo un preprocesado muy sencillo que prepara el texto para su posterior análisis con el analizador SOUP (Gavaldà, 2000) y las gramáticas de análisis. Analizar habla espontánea es considerablemente más complejo ya que se deben tratar disfluencias, fragmentos de habla, oraciones agramaticales, etc. que no se dan en la lengua escrita. Por esta razón el módulo de análisis tiene que ser capaz de producir un análisis razonable de todos los tipos de entrada hablada que se dan en el dominio que tratamos. El SOUP es un analizador descendente basado en *charts* diseñado específicamente para analizar habla espontánea. En principio, no se esperaría que un solo árbol de análisis cubriera un turno de diálogo completo (TU) en habla espontánea, ya que este tipo de intervenciones suelen contener múltiples segmentos de significado. Estos segmentos se llaman Unidades Semánticas de Diálogo (USDs) y corresponden a un único AD. Una de las características más relevantes del SOUP es la capacidad de poder reproducir en un solo análisis una secuencia de árboles de análisis para cada turno, segmentando el texto de entrada en USDs con éxito.

Las gramáticas de análisis utilizadas por el SOUP son gramáticas libres de contexto y, para el castellano y el catalán, se han desarrollado de forma manual a partir de un corpus de diálogos obtenido de la base de datos del C-STAR-II[1]. Inicialmente se desarrolló la gramática del castellano y a partir de ésta la del catalán. Las gramáticas de análisis son tres:

- La Gramática *Full-DA*: se utiliza para escribir las reglas de nivel superior de las ADs, que son más complejas y están formadas por diversos argumentos. Contiene 714 líneas de código[3] y 450 reglas.
- La Gramática *Cross-Domain*: contiene reglas de nivel superior de las ADs más frecuentes que no forman parte de un dominio restringido (p. ej. *negate*) y que tienen pocos o ningún argumento en su definición. Esta gramática está formada por 141 líneas de código y 81 reglas.
- La Gramática *Shared*: contiene 36.829 líneas de código y es la más importante de todas ya que las otras gramáticas van a esta para ver la definición de los argumentos, subargumentos y valores.

Durante el desarrollo de las gramáticas hemos tenido que solventar algunas de las diferencias más importantes que hay entre el inglés y el castellano y catalán. A continuación se detallan algunas de ellas:

- El castellano y el catalán son lenguas muy flexionadas.
- El orden del Sintagma Nominal. En inglés, en general, los adjetivos tienden a preceder al nombre mientras que en castellano y catalán acostumbran a seguirlo.
- El orden de los constituyentes dentro de la oración. El inglés presenta un orden de constituyentes mucho más fijo que el castellano y el catalán, que tienen un orden de constituyentes mucho más flexible.

Otras líneas en las que se ha trabajado durante el desarrollo de las gramáticas de castellano y catalán tienen que ver con la mejora de la flexibilidad y la robustez del sistema con la finalidad de hacerlo más atractivo para el usuario. Se han intentado cubrir las diferentes formas que puede utilizar un usuario para expresarse. Esto se ha solucionado mediante la expansión de las reglas de nivel superior y la ampliación del léxico. Así, el usuario se siente más cómodo ya que el sistema es capaz de analizar una amplia variedad de estructuras y en consecuencia él/ella puede expresarse de manera más libre, sin tener que ceñirse a determinadas estructuras.

En el último eslabón de la cadena de análisis, encontramos un conversor de análisis escrito en Perl, que convierte la salida del analizador en la representación IF canónica usada en el proceso de generación. En general, el conversor sólo extrae información de conceptos y también añade o elimina paréntesis y comas, para conseguir la representación IF

[1] http://www.is.cs.cmu.edu/nespole/db/current/cstar-examples.db

[2] http://www.lc-star.com

[3] Se refiere al número total de líneas que ocupan las reglas e información léxica.

esperada antes de ser enviada al módulo de generación.

4.2 La cadena de generación y la síntesis de voz

La cadena de generación está formada por el conversor de generación de NESPOLE! y el generador GenKit (Tomita, 1988), un sistema de generación basado en pseudounificación. El conversor de generación reformatea la sintaxis de la representación IF canónica en una estructura de rasgos que será utilizada por GenKit.

Para generar una oración a partir de una estructura de rasgos IF el generador utiliza gramáticas sintáctico-semánticas híbridas. Una gramática para GenKit se puede escribir generalizando las reglas semánticas para poder ser utilizadas con muchos actos de habla y diferentes combinaciones de conceptos, o especificando las reglas para un dominio específico. Durante el desarrollo de las gramáticas se actuó en ambos sentidos: por una parte, se utilizó un número pequeño de reglas generales para cubrir un conjunto más amplio de posibles actos de dominio, en algunos casos sacrificando el estilo; por otra parte, se escribieron reglas más específicas para cubrir conceptos IF más frecuentes y así asegurar una generación más fluida y natural. La combinación de estos dos sentidos intenta llegar a un equilibrio entre una amplia cobertura y una alta calidad estilística.

El conocimiento utilizado por el GenKit consiste en conocimiento gramatical, léxico y morfológico. Las palabras están asociadas con los conceptos semánticos y valores del IF mediante las entradas léxicas. Estas entradas léxicas no sólo contienen los lemas de estas palabras sino que también están enriquecidas, por ejemplo, con información léxica pertinente a la generación morfológica (por ejemplo la información de género en el caso de los nombres) y, en el caso de los verbos los requisitos de subcategorización. La generación de la forma morfológica correcta se hace mediante reglas gramaticales de flexión que utilizan información adicional almacenada en las entradas léxicas. Para los casos más complejos de la morfología verbal, la forma correcta se genera a partir de una tabla de información morfológica. Antes de ser tratada por el sintetizador de voz, la salida de generación pasa por un posprocesado.

Para el castellano y el catalán se han tratado diversos fenómenos específicos que no se habían implementado anteriormente en las gramáticas utilizadas por el GenKit. Algunos de estos fenómenos específicos de cada lengua son la construcción *que+subjuntivo*, los clíticos, los pronombres en las construcciones ditransitivas, los verbos pronominales, etc. Algunos de estos fenómenos se reflejan en los léxicos que se han desarrollado y que están enriquecidos con una gran cantidad de información vital para la generación.

En cuanto a la síntesis de voz, tanto para el castellano como para el catalán, hemos utilizado el sistema Text-to-Speech (TTS) (Bonafonte et al., 1988) desarrollado en la UPC. Se trata de un sistema de síntesis de voz concatenativo basado en la selección de unidades, un enfoque que aporta altos niveles de inteligibilidad y naturalidad al habla de salida cuando las bases de datos para la selección son suficientemente grandes.

5 Evaluación del módulo de traducción

Se ha hecho una evaluación del sistema de traducción del habla con usuarios externos y en una situación que ha intentado reproducir una situación real. Cabe decir que no sólo se evaluó el sistema interlingua sino también un sistema estadístico desarrollado dentro del proyecto, aunque en este artículo tan sólo presentamos el sistema interlingua. Los objetivos de esta evaluación eran los siguientes:

- Examinar la actuación del sistema en una situación tan real como sea posible, como si fuera usado por un hablante de inglés que quisiera hacer una reserva de hotel en Barcelona.
- Estudiar la influencia del uso del reconocedor de voz en la traducción.

Para conseguir estos objetivos, se diseñaron diferentes tests en los que se utilizan diferentes módulos del sistema. Por esta razón, esta evaluación se dividió en un número de tareas que se enfocaban a diferentes objetivos. Estas tareas eran las siguientes:

- Evaluación de los componentes de traducción de texto-a-texto.
- Evaluación del sistema de traducción del habla. Para cuantificar la diferencia de resultados obtenidos mediante la traducción

con reconocimiento de voz y los obtenidos sin el reconocedor, sólo de texto-a-texto.
- Comparación de los métodos de evaluación para la traducción. Investigar la validez de métodos de evaluación frecuentemente usados como BLEU y mWER para enfoques de traducción basados en la semántica como el interlingua. La aplicación de estas métricas se compara con el uso de un método de evaluación orientado a la tarea.
- Comparación del sistema de traducción basado en reglas y el sistema de traducción estadístico.

En este artículo presentamos los resultados obtenidos en la evaluación del sistema completo (con RV). Para ello mostramos y comparamos los resultados obtenidos utilizando las métricas BLEU y mWER y aquellos obtenidos en la evaluación con el método orientado a la tarea.

5.1 Preparación de la evaluación, grabación y preparación de datos

Antes de la evaluación, se llevaron a cabo una serie de tareas para obtener y preparar los datos usados para evaluar: grabación de diálogos y datos durante el uso del sistema, reclutamiento de voluntarios que hicieran de usuarios, diseño de los diferentes escenarios para los usuarios, transcripción de los datos grabados, etc.

Las conversaciones tuvieron lugar entre un cliente de habla inglesa y un agente de viajes que hablaba castellano o catalán. Se grabaron 20 diálogos con un total de 12 hablantes. De estos 12, 10 eran totalmente ajenos a la tarea y no estaban familiarizados con el sistema, mientras que 2 sí lo estaban. Cada uno de los 10 hablantes participó en 2 diálogos, mientras que los dos restantes participaron en 10 cada uno. De los 10 hablantes no familiarizados con el sistema 5 hicieron de agente y 5 de cliente. Cabe destacar que los hablantes de inglés reclutados no eran ingleses nativos y eso debe tenerse en cuenta al considerar los resultados provinientes de reconocimiento de voz.

Se diseñaron 5 escenarios diferentes por hablante, disponibles en todas las lenguas relevantes (en catalán y castellano los escenarios del agente y en inglés los del cliente). Antes de empezar las grabaciones se explicó a los voluntarios el funcionamiento básico del sistema (dónde pulsar para empezar o detener la grabación, dónde tenían la información de los escenarios, etc.). La pantalla del ordenador sólo les mostraba sus respectivos escenarios y la interfaz del sistema. La interfaz tan sólo les mostraba su salida de RV y la traducción, tanto del traductor estadístico como del traductor basado en reglas, de la intervención del otro usuario. El hecho de poder ver la salida de RV permitía que los voluntarios pudieran comprobar si se había reconocido bien su intervención y en consecuencia permitir la traducción, o en caso contrario, repetir su intervención antes de producirse un fallo. La razón para ver la salida de los dos traductores en la pantalla es que el sintetizador tan sólo reproducía una de ellas.

La grabación de los diálogos se hizo en una habitación preparada con este fin. Los hablantes se situaron por separado con sus respectivos ordenadores de manera que sólo pudieran ver su propia pantalla. Una vez acabada la grabación, se prepararon los datos de evaluación:

- Se transcribieron todos los ficheros de audio y se concatenaron en TUs, agrupándolos por diálogo.
- Se mantuvieron y marcaron las intervenciones fallidas (errores en el manejo del sistema con RV de ruidos y clics), pero no se consideraron para la evaluación de las tecnologías.
- Se crearon las referencias para cada hablante y diálogo, para poder llevar a cabo la evaluación con las métricas BLEU y mWER.

5.2 Evaluación

Como se ha mencionado anteriormente, se planearon diferentes tipos de evaluación: la primera es una metodología de evaluación orientada a la tarea que se centra en juicios de fidelidad y naturalidad de las traducciones obtenidas, la segunda es la aplicación de métricas como BLEU y mWER.

Las métricas usadas en la evaluación orientada a la tarea se explican a continuación. La evaluación se hizo para TUs y se evaluó de manera separada la *forma* y el *contenido* (la semántica). Para evaluar la forma se dio solamente la salida de generación a los evaluadores y para evaluar el contenido se les dio tanto la entrada como la salida. De esta manera, las métricas varían según si lo que se evalúa es la forma o el contenido:

- **Bien**: si la salida está bien formada (forma) o si se ha transmitido toda la información del hablante (contenido).
- **Ok+**/**Ok**/**Ok-**: si la salida es aceptable, con una gradación que puede ir desde un pequeño error en la forma (como por ejemplo la falta de un determinante) o una pequeña pérdida de información (Ok+), hasta un error más importante de forma o una pérdida más grande de información (Ok-).
- **Mal**: si la salida es inaceptable, ya sea una salida ininteligible o que no esté relacionada con la entrada.

5.2.1 Resultados de la evaluación

Los resultados obtenidos durante la evaluación del sistema completo utilizando la metodología orientada a la tarea se muestran en las Tablas 1, 2, 3 y 4. Una vez estudiados los resultados podemos decir que muchos de los errores obtenidos se deben al componente RV. Como se puede observar, el sistema obtiene un elevado porcentaje de acierto para la dirección castellano-inglés. Esto es debido a que la lengua más desarrollada tanto a nivel de reconocimiento como de gramáticas de análisis y de generación ha sido el castellano y en menor proporción el catalán. Además, para el RV de inglés se utilizaron los modelos del lenguaje proporcionados con el JRTk (cf. Sección 4.1). De todos modos, cabe destacar que el peor resultado obtenido en cuanto a contenido es un 62.4% de traducciones aceptables para el par de lenguas inglés-castellano. Este resultado es comparable al de evaluaciones como la del sistema NESPOLE! (Lavie et al., 2002) que obtuvo unos resultados ligeramente inferiores y fue hecha a partir de USDs y no de TUs, simplificando la tarea de traducción. Por ello, estamos satisfechos con la actuación de nuestro sistema, aunque debemos mejorar el componente de RV.

MÉTRICAS	FORMA	CONTENIDO
BIEN	70.59%	31.93%
OK+	5.04%	15.12%
OK	6.72%	9.25%
OK-	9.25%	16.80%
MAL	8.40%	26.90%

Tabla 1: Evaluación de las traducciones usando el sistema completo (con RV) para el par catalán-inglés. Evaluación basada en 119 TUs

MÉTRICAS	FORMA	CONTENIDO
BIEN	92.85%	71.42%
OK+	4.77%	11.90%
OK	1.19%	7.14%
OK-	0%	5.96%
MAL	1.19%	3.58%

Tabla 2: Evaluación de las traducciones usando el sistema completo (con RV) para el par castellano-inglés. Evaluación basada en 84 TUs

MÉTRICAS	FORMA	CONTENIDO
BIEN	64.96%	34.19%
OK+	15.39%	11.97%
OK	8.54%	14.52%
OK-	5.12%	12.82%
MAL	5.99%	26.50%

Tabla 3: Evaluación de las traducciones usando el sistema completo (con RV) para el par inglés-catalán. Evaluación basada en 117 TUs

MÉTRICAS	FORMA	CONTENIDO
BIEN	64.8%	17.6%
OK+	4.8%	10.4%
OK	12%	18.4%
OK-	8.8%	16%
MAL	9.6%	37.6%

Tabla 4: Evaluación de las traducciones usando el sistema completo (con RV) para el par inglés-castellano. Evaluación basada en 125 TUs

Pares de lenguas	# oraciones	mWER	BLEU
CAT-ENG	119	78.98	0.1456
ENG-CAT	117	81.19	0.2036
SPA-ENG	84	60.93	0.3462
ENG-SPA	125	86.71	0.1214

Tabla 5: Evaluación del sistema completo usando las métricas mWER y BLEU

Los resultados obtenidos utilizando las métricas BLEU y mWER se presentan en la Tabla 5. Como se puede observar en las diferentes tablas, los resultados obtenidos mediante las métricas mWER y BLEU son considerablemente inferiores en comparación con aquellos obtenidos en la evaluación manual. Esto puede ser debido a los siguientes factores:

- La traducción de salida se compara en los métodos estadísticos con una única referencia, que no cubre la flexibilidad y variedad del lenguaje. A esto debe añadirse

la dificultad extra adquirida al trabajar con habla espontánea.

- Los métodos estadísticos penalizan todo aquello que difiere de la referencia, que pueden ser errores mínimos de forma que no afectan el resultado final.
- En general, los resultados bajan más cuando la lengua de origen es el inglés. Esto se debe a que: a) el trabajo de desarrollo se centró mayormente en el castellano y catalán; b) los hablantes de inglés voluntarios no eran ingleses nativos y esto dificulta la tarea del reconocedor.

6 Conclusiones y trabajo futuro

En este artículo hemos presentado un sistema de traducción del habla basado en una interlingua. Hemos explicado la arquitectura general y los diferentes módulos que componen el sistema de traducción. A su vez, hemos presentado la interlingua utilizada, y también hemos mostrado y comparado los resultados obtenidos mediante dos metodologías de evaluación diferentes, una manual y otra basada en métodos estadísticos.

Una vez llegado a este nivel de desarrollo, el próximo paso es resolver los problemas técnicos acaecidos y extender el sistema tanto para el dominio de reservas como para otros dominios. En cuanto a las dificultades, una de las más importantes a solventar es la recopilación de más datos específicos para desarrollar modelos del lenguaje más trabajados para el reconocedor de voz. También hay que mejorar la capacidad del sistema para trabajar con transcripciones degradadas. Una opción es incorporar dentro del modelo de diálogo estrategias para que los mismos interlocutores puedan pedir repeticiones y reformulaciones. Otro es entrenar un sistema, reconocer y corregir fallos en la salida del reconocedor automáticamente.

Agradecimientos

Esta investigación ha sido financiada parcialmente por los proyectos FAME (IST-2000-28323) y ALIADO (TIC2002-04447-C02). Nuestro más sincero agradecimiento a Climent Nadeu y a Jaume Padrell por su esfuerzo y su ayuda en los diversos aspectos del proyecto.

Bibliografia

Bonafonte, A., I. Esquerra, A. Febrer, J. A. R. Fonollosa y F. Vallverdú. 1998. The UPC Text-to-Speech System for Spanish and Catalan. En *Proceedings of 5th International Conference on Spoken Language Processing*, Sydney, Australia.

Gavaldà, M. 2000. Soup: A Parser for Real-World Spontaneous Speech. En *Proceedings of the 6th International Workshop on Parsing Technologies*, Trento, Italia.

Holzapfel, H., I. Rogina, M. Wölfel y T. Kluge. 2003. XXX Deliverable D3.1: Testbed Software, Middleware and Communication Architecture.

Lavie, A., F. Metze, R. Cattoni y E. Costantini. 2002. "A Multi-Perspective Evaluation of the NESPOLE! Speech-to-Speech Translation System". En *Proceedings of ACL-2002 Workshop on Speech-to-Speech Translation: Algorithms and Systems*, Philadelphia, PA, E.E.U.U.

Levin, L., D. Gates, D. Wallace, K. Peterson, A. Lavie, F. Pianesi, E. Pianta, R. Cattoni y N. Mana. 2002. Balancing Expressiveness and Simplicity in an Interlingua for Task based Dialogue. En *Proceedings of ACL-2002 workshop on Speech-to-Speech Translation: Algorithms and Systems*, Philadelphia, PA, E.E.U.U..

Metze, F., J. McDonough, J. Soltau, C. Langley, A. Lavie, L. Levin, T. Schultz, A. Waibel, L. Cattoni, G. Lazzari, N. Mana, F. Pianesi y E. Pianta. 2002. The NESPOLE! Speech-to-Speech Translation System. En *Proceedings of HLT-2002*, San Diego, California, E.E.U.U.

Searle, J. 1969. *Actos de habla*, Madrid, Cátedra 1980.

Tomita, M. 1988. *Generation Kit and Transformation Kit Version 3.2. User's Manual*. Carnegie Mellon University, E.E.U.U.

Woszczyna, M., N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki-Waibel, A. Waibel y W. Ward. 1993. Recent Advances in JANUS: A Speech Translation System. En *Proceedings of Eurospeech-1993*, páginas 1295-1298.