UNIVERSIDAD DE LA REPÚBLICA | FACULTAD DE INGENIERÍA | INSTITUTO DE COMPUTACIÓN

Proyecto de Grado
Informe Final

Regionalización de Sistema de Traducción Open Source

Versión 4.0

11/01/2011

Integrantes: Ernesto López

Tutores: Luis Chiruzzo, Dina Wonsever

Departamento de Procesamiento de Lenguaje Natural

Instituto de Computación

Facultad de Ingeniería

Universidad de la República

Noviembre 2011

Resumen

La traducción automática es una de las áreas del procesamiento de lenguaje natural de mayor crecimiento en la última década. El desarrollo de computadores con alta capacidad de procesamiento a bajo costo ha disparado la evolución de sistemas basados en Inteligencia Artificial. La necesidad de obtener traducciones en forma veloz y eficaz en un mundo cada vez más dependiente de la tecnología, donde el consumo de aplicaciones se multiplica año a año, ha impulsado con fuerza la investigación en el área del procesamiento de lenguaje. Sistemas enteros necesitan ser internacionalizados y se deben generar manuales en decenas de lenguas distintas. La búsqueda de una Web independiente del lenguaje es hoy en día una realidad, los navegadores más utilizados en el mundo ya detectan contenido Web en una lengua distinta y permiten traducirlo.

En este contexto globalizador resulta difícil encontrar plataformas de traducción fuertemente regionalizadas, en particular para el español utilizado en el Río de la Plata. Es importante en este sentido contar con herramientas colaborativas, enriquecidas por la comunidad. Tal es el caso de la herramienta Apertium, un sistema Open Source de traducción superficial que cuenta con recursos lingüísticos para el español estándar.

Este proyecto tiene como objetivo investigar las características del español rioplatense para permitir que Apertium pueda ser regionalizado al Río de la Plata. Se analizaron posibles soluciones y aquellas seleccionadas fueron implementadas y evaluadas.

Como resultado se obtuvo un sistema de traducción Open Source que reconoce las características más importantes del Español del Río de la Plata. El mismo mostró importantes mejoras en la traducción de textos propios de esa región.

En este documento se detallan las investigaciones realizadas, las alternativas de solución estudiadas y su implementación, y finalmente se evalúa el sistema construido utilizando métricas reconocidas, comparando además con el traductor de Google, uno de los sistemas de traducción más exitosos.

Palabras clave: Traducción Automática, Open Source, Inteligencia Artificial, Español del Río de la Plata, Procesamiento de Lenguaje Natural.

Contenido

1 Introducción

1.1 Antecedentes

La traducción automática es una de las disciplinas más antiguas de la inteligencia artificial. Aunque desde siempre se ha soñado con la posibilidad de construir una máquina capaz de traducir automáticamente el lenguaje humano, aún se está muy lejos de este objetivo, pero se ha avanzado mucho en los últimos años y se han construido sistemas capaces de lograr traducciones con un margen de error tolerable. Estos sistemas funcionan con diversas lenguas y están construidos utilizando distintas tecnologías, algunas, sumamente costosas.

La eficacia de estos sistemas depende fuertemente de la existencia de recursos lingüísticos suficientes para las lenguas involucradas en la traducción. Estos recursos implican entre otras variables la intervención de lingüistas o la construcción de corpus.

Para el caso del idioma español existen varias herramientas de traducción. Sin embargo, para el español utilizado en el Río de la Plata no existen tantos recursos lingüísticos.

Entre los traductores más populares se encuentra el traductor de Google, una plataforma de traducción cerrada basada en un modelo probabilístico [7]. Cuenta con la gran ventaja de que sus estadísticas son alimentadas constantemente por miles de usuarios que utilizan a diario el traductor y sugieren al sistema correcciones cuando entienden que este presenta un error. Esta retro-alimentación ha provocado que el traductor de Google haya mejorado mucho desde sus inicios. Reconoce además, en forma aceptable el español del Río de la Plata aunque no lo identifica como una variante.

1.2 Definición del Problema

Este proyecto tiene como objetivo obtener una herramienta de traducción para el idioma español adaptada para la variante utilizada en el Río de la Plata, que represente y reconozca las particularidades más importantes de la región. Debe ser una herramienta Open Source, flexible, que respete lo más posible los estándares existentes para los sistemas de procesamiento de lenguaje; con este fin se optó por utilizar la plataforma de traducción Apertium, una herramienta de traducción de código abierto basada en *shallow transfer*[1]. Fue desarrollada por el grupo de investigación *Transducens* del *Departament de Llenguatges i Sistemes Informàtics* de la *Universitat d'Alacant,* originalmente para el par de lenguas Español-Catalán [11]. Tiene un diseño modular que implementa la traducción

[1] El análisis realizado sobre el texto a traducir es superficial, detectando componentes sintácticas sin entrar en detalle en su estructura semántica, la transferencia es prácticamente palabra a palabra realizando correcciones de número y género.

como una cadena de montaje. Apertium es una plataforma cada vez más conocida y en constante evolución debido a la colaboración de la comunidad.

Su velocidad de procesamiento, su facilidad para generar nuevos pares de lenguas y extender sus módulos, y el hecho de tratarse de una plataforma de código abierto, la convierten en una herramienta que es foco de varias investigaciones y base de muchas soluciones comerciales de traducción.

Existen otras herramientas que presentan mejores resultados pero son tecnologías cerradas o que carecen de recursos para el español. Y aunque Apertium no reconoce el Español del Rio de la Plata, tiene soporte para el Español estándar.

1.3 Alcance

Visto el objetivo de estudiar una plataforma Open Source de traducción automática adaptada al Español utilizado en la región del Río de la Plata, se busca:

- Estudiar las técnicas de traducción automática más modernas y ubicar Apertium dentro del conjunto de Sistemas de Traducción.
- Analizar exhaustivamente la herramienta Apertium, estudiando antecedentes y casos de éxito, arquitectura y técnicas utilizadas por los distintos módulos. Evaluar su funcionamiento respecto a otros traductores, identificar ventajas y desventajas y analizar su extensibilidad de cara a la adaptación del sistema al Español del Río de la Plata.
- Investigar las características del Español del Río de la Plata, su origen y región de uso.
- Buscar alternativas para adaptar Apertium al Español del Río de la Plata.
- Implementar las soluciones estudiadas en Apertium. Evaluar las mejoras utilizando métricas conocidas y comparar con las herramientas de traducción más populares.

1.4 Organización del documento

Este documento (“Informe final de proyecto”), tiene como fin introducir los objetivos del proyecto, investigar las herramientas de traducción automática y las particularidades del Español de Río de la plata, analizar la herramienta Apertium y las posibles soluciones para adaptarla al Español Rioplatense, y evaluar las soluciones implementadas. El documento se divide en los siguientes Capítulos:

- **Capítulo 1, Introducción:** Se describe brevemente el objetivo del proyecto y su motivación. Se resume las investigaciones realizadas y sus conclusiones.
- **Capítulo 2, Estado del Arte en traductores:** Con el fin de introducir en el mundo de la traducción automática se repasa la historia y presente de dicha disciplina, su

motivación y objetivo. Se describen las distintas técnicas utilizadas y el estado del arte de los sistemas más modernos de traducción.

- **Capítulo 3, Estudio de Plataforma Apertium:** Este capítulo describe con detalle las características de la plataforma Apertium; arquitectura, módulos, ejecución, casos de éxitos, etc.
- **Capítulo 4, Características del Español del Río de la Plata:** Análisis de la variante del Español que se observa en el Río de la Plata. Se repasan los aspectos históricos, geográficos y sociales que explican esta variante. Se hace especial hincapié en la principal característica del Río de la Plata, el voseo.
- **Capítulo 5, Estudio de soluciones:** En base al estudio del Español del Río de la Plata y de Apertium se buscan y analizan distintas alternativas para regionalizar el Español reconocido por Apertium al Río de la Plata.
- **Capítulo 6, Implementación y evaluación de la solución:** Este capítulo describe la implementación de las soluciones analizadas en el capítulo anterior. Además introduce las métricas a utilizar y se detallan las evaluaciones realizadas sobre el sistema implementado. En todo momento se realiza la comparación con el traductor de Google.
- **Capítulo 7, Conclusiones:** Se elabora un conjunto de conclusiones en base a las investigaciones y experimentos realizados. Se plantean trabajos a futuro.
- **Capítulo 8, Referencias:** Se detallan los documentos (libros, artículos, publicaciones, etc) que se utilizaron como referencia y apoyo a lo largo del proyecto.

2 Estado del Arte

El objetivo de automatizar las tareas de traducción ha existido desde siempre. Con la aparición de la computación, en el siglo XX se dieron los primeros pasos en ese camino. Pues aunque no existe aún una máquina capaz de tomar un texto y generar una traducción perfecta, se han construido sistemas capaces de generar traducciones algo más primitivas desde las que se puede obtener una buena traducción con una mínima intervención humana. Aún está en discusión si es posible construir una máquina capaz de lograr una traducción perfecta [21].

2.1 Motivación

Vale preguntarse entonces qué aplicación tienen en el día de hoy los sistemas de traducción automática si en realidad no es posible obtener una traducción perfecta. ¿Qué es lo que buscamos traducir? La mayoría de los trabajos realizados por traductores profesionales consiste en traducir documentos científicos y tecnológicos, transacciones de negocios y comercios, memorándums, escritos legales, noticias, manuales, etc. [24]. Muchos de estos trabajos son realmente complejos hasta para un traductor humano, sin embargo la gran mayoría implica un trabajo repetitivo el cual requiere de consistencia y precisión. Sería más que interesante entonces poder automatizar lo más posible estas tareas y surge la necesidad de contar con un sistema de traducción automática capaz de generar una traducción aceptable que haga la mayor parte del trabajo.

Además, el brutal crecimiento de recursos web, los cuales abarcan prácticamente todas las lenguas del mundo, demanda la existencia de sistemas de traducción capaces de universalizar el acceso a la información independizándola del lenguaje. Ya se está en este camino pues hoy en día la mayoría de los navegadores cuentan con servicios de traducción que se activan al detectar contenido en una lengua ajena a la del usuario. Esto permite también, dotar a los grandes buscadores web con la capacidad de realizar búsquedas independientes de los lenguajes, tanto del lenguaje de la búsqueda como del contenido.

Pero sin dudas, el objetivo más deseado es el de permitir la comunicación entre seres humanos que hablen distintas lenguas en forma transparente mediante un traductor automático, sin necesidad de un intermediario que conozca las dos lenguas. Basta con imaginar un servicio de chat que permita comunicarse a una persona en China con otra en India, cada uno usando su propia lengua.

2.2 Tipos de Traductores

En forma general los Sistemas de traducción se pueden separar en dos categorías. Traducción automática basada en reglas y traducción automática basada en métodos estadísticos [24]. Estás dos técnicas de traducción automática tienen fundamentos totalmente distintos y hasta contrapuestos. Son los dos extremos respecto a tecnologías de traducción. En la práctica se verá que existen muchos sistemas híbridos que utilizan en mayor o menor medida las dos técnicas [27].

2.2.1 Traducción basada en reglas

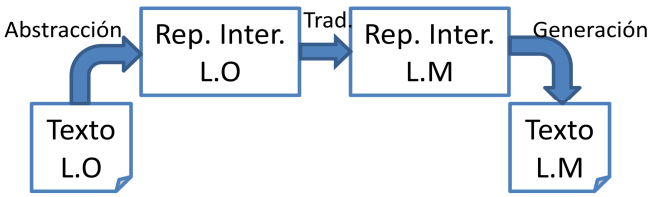
Los traductores basados en reglas fueron los primeros en ser investigados. Tienen como fundamento abstraer las reglas de traducción entre las lenguas. Desde luego esta es una tarea compleja y al día de hoy no existe una representación completa y correcta del lenguaje natural basada en reglas [24].

El método de traducción más simple es el de transferencia directa donde cada lexema es traducido directamente en su representación en la lengua destino, y a no ser por alguna regla de flexión para corregir concordancia de género y número, no existen tratamientos intermedios. Es prácticamente una traducción palabra a palabra. Este tipo de traductores fue descartado rápidamente pues la calidad de las traducciones es mala.

Surgen así los métodos de traducción indirecta basada en reglas. Estos métodos utilizan representaciones intermedias de los lenguajes a fin de abstraer en lo posible el conocimiento de la lengua. Aún está en discusión si es posible construir un sistema computacional capaz de representar la pragmática del lenguaje utilizando reglas. En este sentido se distinguen dos tipos de traducción indirecta: método de transferencia indirecta y método traducción basado en una lengua intermedia.

Método de transferencia indirecta

El método de transferencia indirecta se caracteriza por tener dos representaciones intermedias, una para la lengua origen y otra para la lengua meta. En este caso se puede resumir el proceso de traducción en tres fases: Una fase de análisis del texto en la lengua origen que tiene como resultado una representación intermedia. Una fase de transferencia que tiene como salida la representación intermedia en la lengua meta. Y por último una fase de generación que dada la representación intermedia de la lengua meta genera el texto final. La transferencia puede ser a nivel léxico (palabra a palabra), o a nivel sintáctico, en donde se transforman estructuras sintácticas analizadas para la lengua origen en estructuras sintácticas en la lengua meta. Y en algunos casos también a nivel semántico, en donde hay un análisis de estructuras lógicas, redes semánticas y patrones [24].



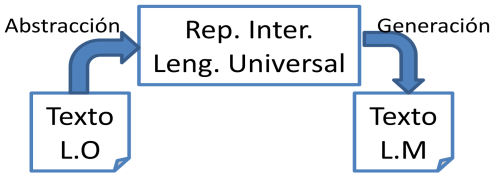
El análisis consta de varias etapas:

- Análisis morfológico: Se analiza lexema a lexema, determinando sus posibles categorías léxicas y sus propiedades (flexión, tipo, etc.). Se identifican fechas, abreviaturas y siglas.
- Des-ambiguación léxica: Para cada lexema se elige una única categoría gramatical de las posibles analizadas en el análisis anterior.
- Análisis sintáctico: Se buscan unidades sintácticas (sintagmas). Se construye uno o varios árboles sintácticos.
- Análisis Semántico: Se analiza el significado del texto. Se buscan estructuras lógicas y redes semánticas.
- Transferencia Léxica: Se realiza la traducción de los lexemas (palabra a palabra) entre la lengua origen y la lengua meta. También se realiza corrección de género y número si corresponde.
- Transferencia Estructural: Se transforman estructuras sintácticas o semánticas (dependiendo de la profundidad del análisis realizado en las primeras etapas) para corregir las diferencias gramaticales entre las dos lenguas. Esta etapa se realiza en conjunto con la transferencia léxica.
- Generación: A partir de la representación intermedia de la lengua meta que se obtiene luego de la transferencia, se genera el texto en dicha lengua.

Los sistemas más conocidos diseñados con este método han sido: METAL, MÉTÉO, SUSY, EUROTRA, LOGOS y GETA (Universidad de Grenoble).

Método de traducción basada en lengua intermedia

En este método, también conocido como interlingua, existe una única representación intermedia común a las lenguas involucradas en la traducción [24]. En este caso el proceso de traducción tiene dos fases: Una fase de análisis en la que se construye la representación intermedia, y una fase de generación en la que se obtiene el texto final en la lengua meta. El fundamento de este método es el de intentar construir un lenguaje que pueda representar el significado universal de todos los lenguajes.



Estos sistemas fueron muy populares en los sesenta, pero hoy en día resulta difícil encontrar traductores de este tipo. La dificultad para definir una representación intermedia común a varias lenguas y para determinar todas las reglas de construcción del lenguaje hacen que sea muy difícil y complejo construir sistemas basados en una lengua intermedia. Sumado esto a la evolución de traductores estadísticos, se ha enlentecido la investigación de este tipo de traductores.

2.2.2 Traductores estadísticos

Los traductores estadísticos tienen un fundamento bien distinto a los sistemas basados en reglas. Se basan en que el conocimiento del lenguaje y de las reglas de traducción esté contenido en un corpus de gran tamaño, el cual se asume es una representación real y completa del lenguaje [17].

Un corpus es un conjunto enorme de ejemplos del uso de la lengua, ya sea en textos o en muestras orales. Existen corpus de distintas naturaleza: Corpus planos donde solo se encuentra el texto tal cual fue tomado de la realidad, corpus anotados en donde se agrega meta-data como ser árboles sintácticos o categorías gramaticales (de uso típico en sistemas de procesamiento lingüístico), corpus paralelos que suelen usarse para alinear textos traducidos entre distintas lenguas, entre otros. Son muy usados por sistemas estadísticos de IA. Uno de los corpus más conocidos es el Corpus Brown, un corpus anotado del inglés [12]. Para el caso del idioma español existe el corpus CREA administrado por la RAE [6].

En base a corpus paralelos los traductores estadísticos calculan probabilidades para determinar la traducción más plausible. Si por ejemplo, un fenómeno sintáctico aparece constantemente en el corpus, es de esperar que ocurra así con un texto cualquiera a traducir. Esta metodología de traducción entra en contraposición con la traducción basada en reglas. Pues en este caso el conocimiento del lenguaje está en el corpus y no en el sistema en sí.

Si bien ya en los años 40’s se empezó a estudiar este tipo de sistemas, recién en los 90’s explotó la investigación de los métodos estadísticos de traducción. El aumento en la velocidad de procesamiento y almacenamiento hizo posible no solo almacenar los gigantescos corpus, sino también ejecutar algoritmos costosos de cálculo estadístico y probabilístico.

Tienen como desventaja que es a veces muy difícil construir corpus. No solo por la calidad de los textos, sino que para el caso de la traducción automática es necesario contar con algún tipo de alineación entre los corpus de las lenguas involucradas. Esto requiere la intervención de lingüistas y traductores que etiqueten el corpus. La habilidad del sistema para traducir depende directamente de la calidad del corpus.

Hoy en día los sistemas de traducción más exitosos son estadísticos.

2.2.3 Sistemas de traducción destacados

SYSTRAN

SYSTRAN es sin duda unos de los sistemas de traducción más populares, en parte debido a que es uno de los más antiguos. Su desarrollo comenzó en los años 70's. En sus inicios fue un traductor basado en reglas. Hoy en día es uno de los sistemas de traducción más prestigiosos con una tecnología de traducción híbrida que intenta conjugar las ventajas de los sistemas basados en reglas y los sistemas estadísticos [10].

Es además motor de traducción de muchos servicios web como Alta Vista y Yahoo, y supo serlo también de Google.

ENGSPAN/SPANAM

Este sistema muy peculiar debido a que tiene un nombre distinto según la dirección de traducción, fue en su origen un proyecto impulsado por la Organización Panamericana de la Salud (OPS) con el fin de contar con una herramienta de traducción de textos técnicos entre las lenguas habladas por los países que conforman la Organización. Así, a mediados de los años 80's surgieron ENGSPAN y SPANAM, el primero un traductor de inglés a castellano, y el segundo de castellano a inglés. Fue uno de los traductores que mostró mejores resultados a fines de los 90's [1].

Se trata de un traductor basado en reglas, muy robusto debido a un muy buen soporte de árboles sintácticos incompletos y a su manejo de palabras con errores ortográficos.

Google Translate

El servicio de traducción de Google es quizás el más utilizado hoy en día. En sus inicios se trataba de un traductor basado en SYSTRAN, pero con el tiempo fue sustituido por un sistema exclusivamente estadístico en un desarrollo propio de la empresa. Liderado por el investigador Franz-Josef Och, uno de los grandes defensores de los sistemas estadísticos.

Es que el éxito de *Google Translate* está en los gigantescos corpus bilingües con los que cuenta, que además reciben la retroalimentación de los usuarios que pueden sugerir

correcciones a las traducciones realizadas por el sistema, al cual pueden acceder en forma gratuita.

También cuenta con una API que puede ser utilizada por otras aplicaciones pero que ya se anunció dejará de ser gratuita debido a la inmensa cantidad de solicitudes que debía atender [23].

InterNOSTRUM

Predecesor de Apertium. Es un traductor superficial entre español y catalán. Su arquitectura, fundamentalmente basada en reglas, es muy similar a Apertium. Consta de un conjunto de módulos que se encargan de generar la representación intermedia de la lengua origen, de realizar la transferencia y de obtener el texto en la lengua meta [4].

Está diseñado para un par de lenguas muy cercanas por lo que se dice que interNOSTRUM realiza una transferencia superficial (*Shallow Transfer*).

AppTek

AppTek es una empresa norteamericana que adquirió a principio de los 90's un software de traducción basado en transferencia directa y lo combinó con un motor de traducción basado en estadísticas. Se trata de un enfoque híbrido. Soporta varios lenguajes como. Inglés, árabe, holandés, persa, portugués, francés, polaco, entre otros [2].

2.2.4 Sistemas de traducción Open Source

Moses

Moses está desarrollado baja la licencia LGPL en C++ y perl. Pueden obtenerse tanto los fuentes como los binarios para Linux o Windows. Se trata de un traductor totalmente basado en estadísticas y puede ser adaptado a cualquier par de lenguas en tanto se cuente con un corpus paralelo para entrenar al sistema [18].

Anusaaraka

Anusaaraka es un traductor del inglés al hindi, desarrollado por la *Chinmaya International Foundation* [29] bajo la licencia GNU. Utiliza algoritmos basados en las gramáticas de Panini [5]. El objetivo es permitir a hindi-hablantes, que no tienen un buen conocimiento de inglés, poder acceder a contenidos en esta lengua a través de *Anusaarka*. De aquí la justificación de que el traductor sea en un solo Sentido.

Apertium

Apertium, como se verá más adelante, es una herramienta de código abierto de transferencia superficial (*Shallow Transfer)*. Desarrollada en sus inicios para pares de lenguas relacionadas [19]. Salvo por el PoS-Tagger, se trata de un traductor basado en reglas. En el siguiente capítulo se detallan a fondo las características de esta herramienta.

3 Estudio de Apertium

Apertium es una herramienta de código abierto para traducción. En principio diseñada para traducir pares de lenguas emparentadas pues se trata de un sistema de traducción superficial.

El sistema está diseñado en módulos que se ejecutan de forma secuencial los cuales van realizando distintas transformaciones sobre el texto en lenguaje fuente hasta producir como salida el texto en el lenguaje objetivo.

Cada módulo va agregando marcas sobre el texto en donde agrega información útil para que el resto de los módulos pueda continuar con la traducción. Los datos lingüísticos son representados usando XML lo que independiza a los módulos de su implementación, por lo que resulta fácil sustituir o agregar módulos al procesamiento. Cada uno puede ser ejecutado por separado. Esto facilita mucho la detección de errores y la integración con otras herramientas de procesamiento de textos. Además el uso de XML como estándar, para el cual existen infinidad de herramientas para su manipulación, facilita enormemente la generación de recursos lingüísticos.

Este diseño hace posible traducir decenas de miles de palabras en unos segundos con una tasa de error aceptable, sobre todo para lenguas emparentadas. Se ha observado que para este tipo de pares de lenguas, una traducción palabra a palabra proporciona resultados con errores que pueden ser solucionados con un análisis simple de la morfología y la estructura sintáctica, y un tratamiento primitivo de la ambigüedad léxica [11]. Estos procesamientos son los que se buscan resolver con Apertium.

La plataforma Apertium está siendo desarrollada por el grupo de investigación *Transducens* del *Departament de Llenguatges i Sistemes Informàtics* de la *Universitat d'Alacant* en colaboración con *Prompsit Language Engineering* [19].

Apertium empieza a ser desarrollado como parte del proyecto OpenTrad (“Traducción automática de código abierto para las lenguas del Estado español”), impulsado y financiado por el Ministerio de Industria y comercio de España desde el año 2004.

Los módulos de procesamiento son: desformateo, análisis morfológico, desambiguación categorial, transferencia estructural superficial, transferencia léxica, generación morfológica y reformateo. Usa transductores de estados finitos para las operaciones de procesamiento léxico (análisis y generación morfológica, transferencia léxica), modelos ocultos de Markov para la desambiguación categorial y chunking multi-etapa basado en estados finitos para la transferencia superficial.

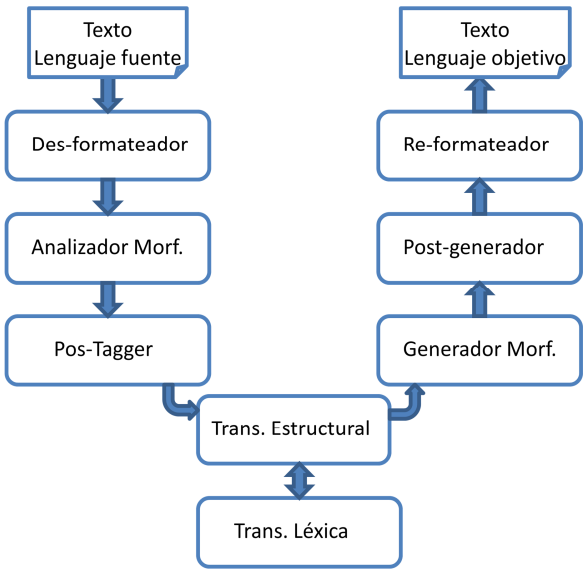
Está desarrollado casi en su totalidad en lenguaje C y compilado para plataformas unix, aunque existen en la red versiones compiladas para Windows y OS/2

En la actualidad existen varias aplicaciones construidas sobre Apertium. Por ejemplo en [31] se exponen los servicios de traducción de Apertium en una capa de Web Service. En [20] se presenta un caso en el que se utiliza Apertium para realizar la localización del software Autodesk y su documentación de español a portugués de Brasil. También se han generado varios pares de lenguas como en [14] y [34], y hasta se han construido o sustituido nuevos módulos como en [28] en donde se utiliza un Pos-tagger basado en trigramas, o en [32] donde se aplican mejoras a la transferencia estructural aplicando métodos estadísticos.

En la actualidad existen recursos Apertium para varios pares de traducción, entre ellos, español-catalán, español-gallego, español-inglés, español-portugués, español-francés, inglés-gallego, portugués-catalán, rumano-español, francés-catalán, entre otros.

Sin duda alguna la gran ventaja de Apertium está en el hecho de que se trata de una herramienta de código abierto y fácilmente extensible. Esto ha llevado a que hoy en día sea una de las herramientas de traducción más utilizadas. La facilidad para generar nuevos pares de lenguas ha llevado a que solo en el último año se hayan generado recursos Apertium para lenguas como checo, alemán, esperanto, islandés, polaco, entre otras.

3.1 Módulos



Apertium consta de 8 módulos de procesamiento. A continuación se describen cada uno de ellos.

Des-formateador: Separa el texto de la información de formato (etiquetas HTML, imágenes, etc.)

Analizador Morfológico: Toma el texto sin formato y genera para cada unidad léxica las posibles formas léxicas. Cada una contiene lema, categoría gramatical, información sobre la flexión morfológica (número, género, persona, tiempo, etc.). Realiza reconocimiento tanto de contracciones como de unidades léxicas de más de una palabra. Es un transductor de estado finito construido a partir de un diccionario morfológico de la lengua origen.

Pos-Tagger: Este módulo se encarga de elegir una única forma léxica de las encontradas por el módulo anterior para cada unidad. Se trata de un desambiguador categorial común basado en un modelo oculto de Markov. El mismo es entrenado en un corpus suficientemente grande de la lengua origen. Permite definir reglas para obligar o prohibir patrones de categorías.

La salida del analizador morfológico es el que se utiliza como entrada para el Pos-Tagger. Sin embargo las etiquetas generadas por este tienen un nivel de detalle excesivo para el Pos-tagger. Se suelen utilizar etiquetas más generales para determinar la categoría gramatical pues no hay que olvidar que el Pos-tagger se construye utilizando información estadística extraída del corpus de la lengua y la utilización de etiquetas demasiado especificas puede llevar a empeorar los resultados del modelo. Es por esto que es necesario agrupar algunos conjuntos de etiquetas del analizador morfológico en una etiqueta de Pos-tagger. También se podrán agregar restricciones así como reglas de preferencia en caso de ambigüedad.

El Pos-tagger puede generarse en modo supervisado, en donde las ambigüedades son resueltas por un humano y también puede generarse en modo sin supervisión en donde las ambigüedades serán resueltas por el sistema.

Módulo de transferencia léxica: Este módulo es invocado por el módulo de transferencia estructural. Entrega para cada forma léxica en el lenguaje origen su correspondiente forma léxica en el lenguaje meta. Por lo tanto no se realiza ningún tratamiento de la polisemia, esto es importante tenerlo en cuenta para la traducción de Entidades con nombre.

Módulo de transferencia estructural: Este módulo se encarga de resolver las diferencias gramaticales entre las lenguas origen y meta (género, número, orden

de palabras, etc.). El módulo se genera a partir de un archivo que contiene reglas que describen la transformación que recibe cada sintagma o patrón.

En este módulo es en donde ocurre efectivamente la traducción. El proceso de transformación se realiza en tres etapas ejecutadas por tres sub-módulos. Una primera etapa en la que se realiza una transferencia superficial en donde se resuelven las diferencias gramaticales entre las lenguas origen y meta (género, número, etc.). Este sub-módulo se construye en base a un archivo en donde se especifican reglas de transformación para una determinada secuencia de categorías léxicas.
En una segunda pasada, que toma como entrada la salida del primer sub-módulo, se realizan transformaciones sintácticas. El objetivo es reorganizar o transformar secuencias de sintagmas. Este proceso es vital para la traducción entre lenguas no relacionadas como por ejemplo inglés y español. Este sub-módulo se construye en base a un archivo muy similar al del primer sub-módulo, solo que en lugar de especificar secuencias de categorías léxicas se especifican secuencias de sintagmas. Estos sintagmas se definen como secuencias de categorías léxicas.

Generador morfológico: Genera a partir de las distintas formas léxicas en la lengua meta las flexiones necesarias para obtener la unidad léxica correspondiente. Se genera a partir de un diccionario morfológico de la lengua meta.

Post-generador: Realiza algunas operaciones ortográficas de la lengua meta tales como contracciones y apostrofaciones. Es generado a partir de un archivo de reglas de transformación con un formato similar al de los diccionarios anteriores.

Re-formateador: Restaura la información de formato.

3.1.1 Formato de textos

El flujo del texto va cambiando de formato desde que ingresa a la cadena de traducción hasta que sale. Los textos entre los módulos tienen tres especificaciones posibles [11].

- Cadena de Texto con formato:
  Es el formato del texto de entrada, HTML, RTF, XML, etc.

- Cadena de Texto sin formato con Super-blancos:
  Al ingresar el texto al módulo Des-Formateador en su formato original, este elimina la información de formato encapsulándola en super-blancos para que sea transparente para los módulos que tratarán el contenido del texto. Luego el módulo de reformateo tomará los super-blancos para restablecer el formato al texto generado como salida.

- Cadena de Texto segmentada:
  Este formato se utiliza para delimitar las unidades léxicas con información generada por los distintos módulos. Luego esta información es eliminada por el último de los módulos. Este formato también permite representar ambigüedades, por ejemplo cuando existe más de un análisis morfológico para alguna unidad léxica.

3.2 Inserción de palabras en los diccionarios

Como ya se mencionó, los diccionarios están definidos mediante archivos XML. Cada vez que se quiera incluir una nueva palabra en diccionario será necesario modificar tres diccionarios. El diccionario monolingüe de la lengua origen, el diccionario monolingüe de la lengua objetivo y también habrá que incluir una entrada en el diccionario bilingüe para indicar la traducción de la palabra entre los dos lenguajes.

El diccionario morfológico consta básicamente de tres secciones: La primera sección que tiene la definición de los símbolos a utilizar, por ejemplo el alfabeto o los atributos que pueden formar parte de las distintas flexiones de la palabra. La segunda sección que define paradigmas, los cuales permiten generalizar las distintas flexiones de las unidades léxicas. De esta forma aquellas palabras que tienen las mismas flexiones, tendrán el mismo paradigma. La tercera sección es el cuerpo del diccionario el cual define las posibles entradas léxicas del lenguaje.

Definición de símbolos:

En esta sección se define el alfabeto a utilizar y los atributos y etiquetas con los que se generan las distintas representaciones lingüísticas. Para el caso del diccionario morfológico se definen todos los atributos que puede definir una unidad léxica [11].

```
:
<sdef n="cnjsub"/>
<sdef n="cnjadv"/>
<sdef n="nt"/>
<sdef n="vbser"/>
<sdef n="vbhaver"/>
<sdef n="vblex"/>
<sdef n="vbmod"/>
<sdef n="inf"/>   <!-- infinitivo -->
<sdef n="ger"/>   <!-- gerundio -->
<sdef n="pp"/>    <!-- participio pasado -->
:
```

Si se quiere agregar un nuevo atributo simplemente hay que definirlo en este conjunto usando la etiqueta <sdef>. Es muy importante tener en cuenta que en los módulos subsiguientes también hay que declarar este nuevo atributo de lo contrario el módulo dará un error.

Paradigmas:

Cada paradigma define una posible flexión, los mismos se indican entre las etiquetas

`<pardef n="<nombreParadigma>" > … </pardef>`

Luego se utiliza la etiqueta <e> para ingresar las distintas terminaciones que identifican cada flexión.

```
<pardef n="s/aber__vblex">
      :
      <e>
        <p>
          <l>abré</l>
          <r>aber<s n="vblex"/><s n="fti"/><s n="p1"/><s n="sg"/></r>
        </p>
      </e>
      <e>
        <l>...
        :
```

Figura 2-3-1

En la figura anterior se muestra parte de la definición que representa las distintas flexiones del verbo saber. En general en Apertium se utilizara la etiqueta <p> para representar pares relacionados, por ejemplo en los diccionarios bilingües relaciona la traducción de cada unidad léxica o sintagma. En el caso del diccionario morfológico monolingüe relaciona la terminación con una serie de atributos que representan la flexión, en el caso de la figura relaciona la terminación ‘abré’ con un verbo, futuro, en primera persona, singular (los atributos existentes son definidos en la primera sección del diccionario). También es posible asociar a la entrada otro paradigma que agregue más información sobre la flexión. En la figura 2-2 se ve cómo para la terminación ‘abiendo’ le agrega atributos para indicar que se trata de un verbo en gerundio y luego adiciona la información de flexión del paradigma “S_cantándo”.

```
:
<e>
  <p>
    <l>abiéndo</l>
    <r>aber<s n="vblex"/><s n="ger"/><j/></r>
  </p>
  <par n="S__cantándo"/>
</e>
:
```

Figura 2-3-2

Existen así paradigmas para las flexiones de adjetivos por ejemplo y detectar así el género y número. También existen paradigmas para los nombres que incluyen entre otros, atributos para indicar si se trata de un nombre propio, de una localidad, etc. Incluso paradigmas para identificar flexiones en unidades léxicas multi-palabra.

La definición de paradigmas es sumamente importante, pues independiza la flexión del lema lo que permite identificar fenómenos regulares de la lengua.

Entradas léxicas:

Finalmente se define el cuerpo del diccionario entre las etiquetas <section>…</section>. Aquí es donde se indican todas las unidades léxicas de la lengua [11]. Por ejemplo para el caso del verbo saber y sus posibles conjugaciones existe la entrada:

```
<e lm="saber"><i>s</i><par n="s/aber__vblex"/>    </e>
```

El atributo lm sirve para identificar la entrada, aunque no es obligatorio. Luego entre las etiquetas … se identifica el lema del verbo (puede ingresarse más de uno) y finalmente se ingresa el paradigma de flexión de la palabra (usando la etiqueta <par>). De esta forma utilizando la entrada de la unidad léxica y la definición del paradigma se puede reconocer el verbo *saber* con sus distintas conjugaciones.

En definitiva utilizando los distintos paradigmas de flexión y las entradas en el cuerpo del diccionario se construye el transductor que realiza el análisis morfológico del texto.

Diccionario bilingüe:

El último paso para la inserción de una palabra es incluir en el diccionario bilingüe la traducción de la misma entre las dos lenguas. Simplemente se ingresa un par que tiene la unidad léxica en cada idioma con su respectiva información de flexión. En la figura 2-3 por ejemplo se traduce el verbo *tener* como el verbo *have* en inglés. Es importante notar que el resto de los atributos de la unidad léxica no se tienen en cuenta para la unificación, solo aquellos los que se incluyen en el par. También es aquí donde se hacen los cambios que

hagan falta sobre la flexión de la palabra en el lenguaje meta. Por ejemplo, cuando se trata de pares de lenguas que tienen género, este no siempre coincide entre ellas, es aquí donde se representan estas diferencias, siempre y cuando no dependa del contexto en cuyo caso se deberá recurrir al archivo de reglas de transferencia estructural.

```
<e>
  <p>
    <l>have <s n="vblex"/></l>
    <r>tener<s n="vblex"/></r>
  </p>
</e>
```

Figura 2-3-3

Esto no pretende ser más que una introducción para entender la estructura de los diccionarios para poder estudiar y entender las soluciones para lograr el objetivo planteado.

3.2.1 Web form para inserción de palabras:

Para facilitar el ingreso manual de palabras existe una aplicación web. Se instala en un servidor Apache (es necesario tener soporte php). La distribución que se puede obtener desde la página de Apertium no soporta el par de lenguajes <Español, Inglés> pero con unas ligeras modificaciones se adaptó para que funcionara. Es necesario indicar la ruta a donde se encuentran los diccionarios iniciales modificando el parámetro *$dicos_path* del archivo *config.php* [11]. El usuario ingresa la palabra en los dos idiomas del par y selecciona la categoría gramatical de la palabra. Luego ingresa la información de la flexión seleccionando para esto un paradigma de flexión para la palabra en ambas lenguas. Finalmente la aplicación valida los datos e inserta la palabra en los diccionarios correspondientes.

3.3 Ejecución de Apertium

Cada módulo puede ser ejecutado por separado y la salida de cada uno puede ser visualizada. Por ejemplo en el siguiente pipe se muestra cómo se invoca a cada módulo en forma independiente utilizando como entrada la salida del módulo anterior (ejecutando desde el directorio donde está el par de lenguas):

```
echo "Quiero que me traduzcas" | apertium-destxt | lt-proc es-
en.automorf.bin |apertium-tagger -g es-en.prob | apertium-
pretransfer | ./es-en.transfer es-en.autobil.bin | ltproc -g es-
en.autogen.bin | ltproc -p es-en.autopgen.bin | apertium-retxt
```

Los módulos utilizados son:

apertium-destxt: Des-formateador

LENG1-LENG2.automorf.bin: Analizador morfológico

LENG1-LENG2.prob: Pos-tagger

apertium-pretransfer: Módulo auxiliar de pre-transferencia, utilizado para generar las unidades léxicas multipalabra.

es-en.autobil.bin: Módulo de transferencia estructural

es-en.autogen.bin: Generador morfológico

es-en.autopgen.bin: Post-generador

apertium-retxt: Re-formateador

Cada vez que se modifican los diccionarios o archivos de reglas es necesario recompilar los módulos afectados.

4 Español del Rio de la plata

Todas las lenguas habladas en el mundo tienen sus variedades. Diferencias que van por ejemplo desde el vocabulario, la conjugación de los verbos, la pronunciación, y hasta en algunos casos diferencias sintácticas.

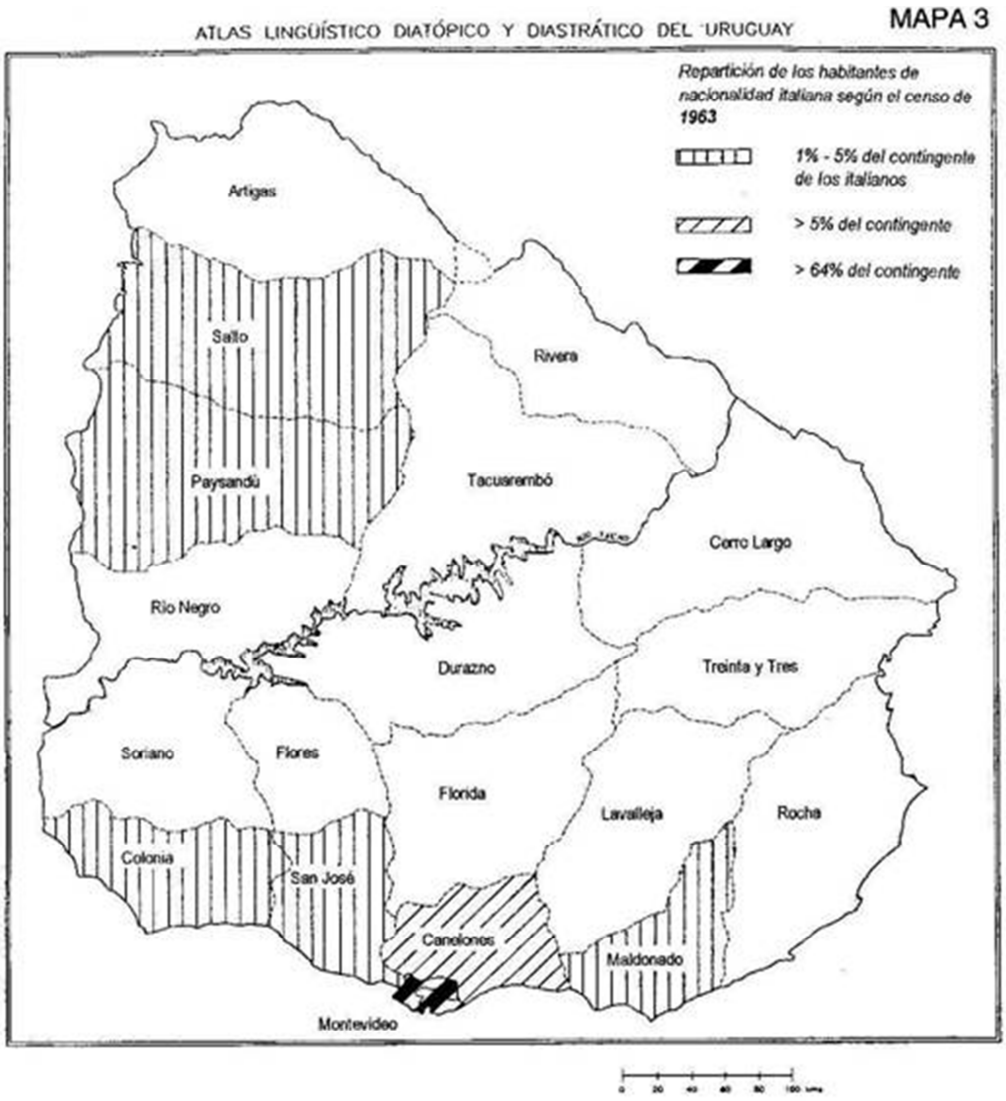
Generalmente estas variedades se agrupan por regiones lo que refleja que surgen como consecuencia de factores culturales y geográficos. Existen numerosos estudios sociológicos que intentan explicar el porqué de éstas variantes. Por ejemplo, el español que se habla en chile, que presenta una pronunciación bastante particular, es producto de la mezcla con la lengua hablada por los indios mapuches y el quechua hablado en el sur. Incluso en las provincias argentinas que limitan con chile se puede encontrar esta variedad del idioma español que tiene mucho en común con el español del Río de la Plata. Sobre todo en aquellas provincias que tienen un lazo cultural (ya sea por motivos económicos o geográficos) más estrecho con Chile [8]. Incluso en Uruguay pueden encontrarse múltiples variedades del español. En las ciudades que limitan con Brasil, en donde la frontera no es más que una calle, se observa una particular mezcla del español con el portugués. Otro ejemplo es el portugués del sur de Brasil, en el cual es aceptado el pronombre 'tu'.

4.1 Origen

El español hablado en el Río de Plata no escapa a estas variantes y es de hecho muy diferente al español de España. Esta variedad se observa principalmente en las ciudades costeras del Río de la plata y del Río Uruguay hasta la desembocadura del Río Negro, pero se extiende también hasta el interior profundo de Uruguay aunque con variantes, observando allí mayor influencia del portugués. Puede encontrarse fusionado también en las provincias del norte argentino y sur de Paraguay [8].

También es conocido como castellano y tiene como principal característica la influencia (en mayor o menor medida según la región) del italiano y el portugués. Este último llega durante la época de la colonia (junto con el Español), al igual que en el resto de América [8]. Y el italiano aparece durante las oleadas migratorias del siglo XIX, sobre todo en Buenos Aires y Montevideo. Desde allí se extiende hacia el interior del país, en mayor medida a aquellos departamentos más enlazados con la capital como Colonia, Canelones, San José, Maldonado, Salto, Paysandú. El italiano no solo influye en modificaciones sobre el vocabulario y algunos modos verbales, sobre todo influye en la fonética, tanto en el acento como en la pronunciación. Es esta influencia italiana la que caracteriza y distingue el castellano del Río de la Plata del resto de las variantes del español. En el litoral del Uruguay sigue apreciándose una influencia más profunda del portugués, mientras que en

Argentina se encuentran variantes bastantes particulares del castellano en las Provincias de Córdoba y Santiago del Estero, sobre todo a nivel fonético.



4.2 Características

Varias son entonces las características que diferencia el castellano del Río de la Plata del resto de las variantes del español, algunas son simplemente fonéticas como el yeísmo, y otras tienen impacto en las conjugaciones de los verbos y el uso de pronombres como el voseo.

4.2.1 Yeísmo

El yeísmo es un cambio fonológico que consiste en la simplificación del sistema consonántico /ĭ/ y /y/, en /y/. Se trata de un proceso fonológico de confusión de dos fonemas originalmente distintos [15].

Se puede observar en toda América Latina, salvo alguna región. Especialmente en Argentina, Uruguay y Paraguay, es típica por su pronunciación rehilada [ʒ]. Tiene su origen en España su aparición se explica por una tendencia a buscar cierta comodidad en el habla y a las oleadas migratorios de las áreas rurales a las urbes que al mezclarse ocasionaban deformaciones a la pronunciación por la cercanía de los dos fonemas palatales.

Por tratarse de un fenómeno estrictamente fonético no se entrará en detalle en él. Pero es importante destacar que se trata de una característica muy propia del Río de la Plata siendo de uso natural en todos los ámbitos.

4.2.2 Queísmo

El queísmo es un fenómeno que consiste en la supresión indebida de una preposición (generalmente *de o en*) delante de la conjunción *que* cuando la preposición viene exigida por alguna otra palabra de la oración [9]. Su uso es común y utilizado en el lenguaje coloquial en España y Latinoamérica. En particular es de uso muy común en el Río de la Plata. Sin embargo su uso es considerado indebido por la Real Academia Española. Algunos ejemplos:

- *"Me alegro QUE haya venido"*
- *"Insistió QUE nos quedáramos a cenar"*
- *"a pesar QUE"*

4.2.3 Voseo

El voseo es una de las peculiaridades del Español del Río de la Plata, aunque no es exclusivo de él y se emplea de distinta forma según la región. En la definición de la RAE se denomina voseo al empleo de la forma pronominal vos para dirigirse al interlocutor [9]. Y se distinguen dos tipos de voseos.

El **Voseo reverencial** que consiste en el uso del pronombre vos para referirse con especial referencia a la segunda persona del plural y singular. Prácticamente en desuso es utilizado en el español antiguo y hoy en día en actos solemnes o recreaciones que reflejan el lenguaje de otra época. Vos es utilizado como forma del sujeto e.j *vos pensáis* y también como termino de proposición e.j *a vos buscáis*. No entraremos en detalle con este tipo de voseo pues no es propio del Río de la Plata.

El **voseo dialectal Americano** es el que captura nuestro interés, se conoce por el uso de formas pronominales o verbales de la segunda persona del plural para dirigirse a un único interlocutor. Es común verlo en distintas variedades del español de América Latina y, a diferencia del voseo reverencial, implica acercamiento y familiaridad pues no es común verlo (al menos en su forma pronominal) en ámbitos de suma formalidad en donde suele utilizarse el ustedeo [16].

Voseo pronominal

El voseo pronominal consiste en el uso de *vos* como pronombre de la segunda persona del singular en sustitución de *tú* o *ti*. Se emplea como:

- Sujeto: *Puede que **vos** tengás razón*
- Vocativo: *¿Por qué la tenés contra Alvaro Arzú, **vos**?*
- Termino de proposición: *Cada vez que sale con **vos**, se enferma*
- Termino de comparación: *Es por lo menos tan actor como **vos***

La RAE también aclara que para el pronombre que se usa con los verbos pronominales y en los complementos sin preposición (pronombre átono), y para el posesivo, se combina con formas de tuteo p.e: "*Vos TE acostaste con el tuerto*", "*Lugar que odio* [...] *como TE odio a vos*"," *No cerrés TUS ojos*". Más adelante entraremos más en detalle en las formas en las que se combinan el voseo y tuteo.

Voseo verbal

El voseo verbal es un poco más complicado que el pronominal. Según la RAE *"el «voseo verbal» consiste en el uso de las desinencias verbales originarias de la segunda persona del plural, más o menos modificadas, para las formas conjugadas de la segunda persona del singular: tú vivís, vos comés o comí"*. Afecta al verbo en distinta forma y en los distintos tiempos según la región. La complejidad del voseo verbal se debe a que su uso varía considerablemente según la región, y no en todos lados es aceptado como norma culta. Se hará especial hincapié en la variedad del Río de la Plata. De hecho solo Argentina, Uruguay y Paraguay reconocen el voseo como norma culta [16]. La Asociación de Letras de Argentina reconoció el voseo recién en 1982 y no en todas sus modalidades como se verá más adelante.

El voseo es, como se vio, utilizado en ámbitos de familiaridad e informalidad y esto tiene mucho que ver con su origen. Pues en sus inicios era rechazado por puristas y tomado por vulgar y denigrante por los gramáticos de la época. Existió un fuerte rechazo a su uso, sobre todo en las clases altas. Lejos de esto, hoy es utilizado por las dos terceras partes de América aunque solo Argentina y Uruguay lo reconocen como norma culta. A mediados del siglo XX toma mayor auge en Argentina, en parte por la influencia de las políticas de izquierda a partir de los años sesenta que por su naturaleza tienden a acortar las distancias sociales y promueven la igualdad; ya en los cuarenta se notan estos cambios durante el peronismo.

Voseo verbal en los tiempos del Presente

Se puede encontrar en el presente indicativo junto a las formas diptongas del plural (habláis), en algunos casos con omisión o perdida de la última ***s*** en la pronunciación (sobre todo en regiones andinas). En el caso del Río de la Plata se observa reducción de diptongo a una vocal abierta (*sabés*) aunque hay documentados reducción a vocales cerradas (*sabís*) Los verbos de la primera conjugación, aquellos cuyo infinitivo termina en *-ar*, nunca presentan en este tiempo formas voseantes en *–ís* [9].

En el presente subjuntivo también se puede ver el voseo en las formas diptongas del plural (habléis), con omisión o pérdida de la última ***s*** en la pronunciación en algunas regiones. En el caso del Río de la Plata se observa reducción de diptongo a una vocal abierta (*subás*) aunque hay documentados reducción a vocales cerradas (*hablís*). En este caso, las formas en -ís solo aparecen en verbos de la primera conjugación.

Voseo verbal en los tiempos del Pasado

El voseo no afecta los tiempos del pasado, al menos no en la región rioplatense. Sin embargo en chile se utilizan variantes del pretérito de indicativo con desinencias de la segunda persona del plural (cantabais, cantarais) con omisión o pérdida de s en la pronunciación [9]: *"¿Dónde andabai que andabai perdido?"* Para el pretérito perfecto simple o pretérito de indicativo, se emplea la segunda persona del plural sin diptongar (volvistes). Sin Embargo esta modalidad es vista como vulgar y no es aceptada como norma culta en la región del Río de la Plata en donde se prefiere en este tiempo el uso de la forma de segunda persona del singular (*volviste*).

Voseo verbal en el imperativo

Las formas voseantes del imperativo surgen a partir de la segunda persona del plural al eliminarse la *d* al final p.e *tomá (tomad), poné (poned)*. Carecen de las irregularidades de la segunda persona del singular que sufre el tuteo por lo que *di, sal, ven, ten,* etc., en su forma voseante resultan *decí, salí, vení, tené*, ext.

Estas formas verbales llevan tilde por tratarse de palabras agudas terminadas en vocal. Cuando van acompañadas de algún pronombre enclítico, siguen las normas generales de acentuación p.e "*Compenetrate en Beethoven, imaginátelo. Imaginate su melena"* [9].

El voseo pronominal y verbal puede combinarse con el tuteo. Se observan así las siguientes modalidades del voseo:

- Plenamente voseante: De mucho uso en el Río de la plata. El sujeto vos va acompañado de formas verbales de voseo p.e "*Vos no podés entregarles los papeles antes de setenta y dos horas"*.

- Voseo exclusivamente verbal: Aquí el sujeto de las formas verbales voseantes es exclusivamente *tú.* Es de uso común en Uruguay, sobre todo en ámbitos de semi-informalidad.
- Voseo exclusivamente pronominal: vos es el sujeto de un verbo en segunda persona del singular: p.e "*Vos tienes la culpa para hacerte tratar mal"*. No es común en el Río de la Plata.

Voseo en el Río de la Plata

Si bien el voseo es aceptado como norma culta en el Río de la Plata, no lo es en todas las acepciones estudiadas. Para la RAE en Argentina, Paraguay y Uruguay la modalidad más generalizada es la que combina el voseo pronominal y el verbal: *vos llegás* [9].

En Montevideo, sin embargo, es más prestigioso el voseo exclusivamente verbal: *tú llegás*. El paradigma verbal de la norma culta está constituido por formas voseantes con reducción del diptongo en el presente de indicativo (*cantás, comés, vivís*), por las formas voseantes propias del imperativo (*cantá, comé, viví)* y por formas tuteantes para el resto de los tiempos verbales. No están asentadas en la norma culta las formas terminadas en -*s* del pretérito perfecto simple o pretérito de indicativo: *cantastes, comistes, vivistes;* ni las formas agudas del presente de subjuntivo: *cantés, comás, vivás*.

5 Estudio de Soluciones

Se planteó el objetivo de regionalizar la herramienta Apertium al Rio de la Plata. Para esto se identificaron las particularidades del castellano rioplatense, y se encontró el voseo como la característica principal de esta variedad del español. Además se analizó el reconocimiento de entidades con nombre realizado por Apertium encontrando dificultades para reconocer nombres propios de la región.

El objetivo de este capítulo es analizar las posibles alternativas para incluir los modos voseantes en la herramienta Apertium (tanto el voseo verbal como el pronominal) y mejorar en lo posible el reconocimiento de entidades con nombre propias del Río de la Plata. Es importante tener en cuenta que se deben buscar soluciones que automaticen lo más posible el proceso de enriquecimiento.

5.1 Voseo

Luego de estudiar las características del voseo surge la pregunta. ¿Cómo representar el voseo en Apertium? ¿Qué recursos serán necesarios para reconocer el voseo con Apertium?

El voseo verbal tiene una connotación claramente morfológica. Está representado por una variación en la flexión de los verbos respecto del español. Y para el caso del voseo pronominal simplemente basta con agregar el pronombre *vos* a los recursos lingüísticos.

Como se vio en el Estudio de Apertium, el tratamiento a nivel morfológico se realiza en el módulo de análisis morfológico.

5.1.1 Módulo de Análisis morfológico

El módulo de Análisis morfológico es un transductor que se construye en base a un archivo con las definiciones necesarias. Como se vio en el Estudio de Apertium este archivo está dividido en varias secciones. Una sección que define los atributos que pueden tomar las etiquetas, otra sección donde se describen los paradigmas de flexión y otra sección con las entradas léxicas. Para el caso del voseo estudiaremos con más detalle las flexiones de los verbos.

Los verbos son ingresados en el cuerpo del diccionario (entradas léxicas) en su forma infinitiva, y se indica con la etiqueta *par* el paradigma de flexión que presenta. Estas flexiones asumen que la raíz del lexema está como prefijo y determinan las características de la flexión en base a la terminación del lexema.

Pongamos como ejemplo el verbo cantar. La entrada en el diccionario para este verbo es la que se ve más abajo. Se ve cómo con el atributo *lm* se indica que el lema es cantar y con la etiqueta *<i>* se indica la raíz del lexema. Esto es muy importante para la construcción del transductor, además limita las flexiones a aquellas que tienen como raíz la indicada con *<i>*. Por último se define el paradigma de flexión con la etiqueta *<par>*. Desde luego el lexema seleccionado tiene que haber sido definido en la sección de paradigmas.

```
<e lm="cantar"><i>cant</i><par n="abandon/ar__vblex"/>     </e>
```

Sin embargo puede llamar la atención el nombre del paradigma verbal asociado a cantar (abandon/ar__vblex). Pero es que el objetivo de definir los paradigmas es el de independizar la flexión del lema del verbo, pues existen muchos verbos que comparten la misma flexión. En el ejemplo mostrado el verbo *cantar* se flexiona en la misma forma que el verbo *abandonar*. Al menos desde el punto de vista de la terminación. En el caso de Apertium para el español, el paradigma *abandon/ar__vblex* es el más utilizado por los verbos.

Los paradigmas definirán entonces la información de la flexión en base a la terminación. Su definición consiste básicamente en: Dada una terminación, se genera una salida con la información.

```
<pardef n="abandon/ar__vblex">
  <e>
    <p>
      <l>aríamos</l>
      <r>ar<s n="vblex"/><s n="cni"/><s n="p1"/><s n="pl"/></r>
    </p>
  </e>
  <e>
    <p>
      <l>aría</l>
      <r>ar<s n="vblex"/><s n="cni"/><s n="p1"/><s n="sg"/></r>
    </p>
  </e>
  <e>
    :
    :
```

En la figura se muestran las primeras dos entradas del paradigma *abandon/ar__vblex* en donde para la terminación *aríamos* se asocia el lexema a una flexión verbal con la etiqueta de nombre “vblex”, y además se incluyen tres atributos más que agregan información de flexión. En este caso, que se trata de un verbo en el condicional del indicativo de la primera persona del plural. La segunda entrada asocia la terminación *aría* a un verbo en condicional indicativo de la primera persona del singular. Para el caso de la entrada *cantar* del diccionario el transductor tomaría este camino para *cantaríamos* y *cantaría*. En la tabla siguiente se muestra el análisis realizado por Apertium para estos dos lexemas. Aquí

se puede apreciar claramente cómo se puede independizar el análisis morfológico del lema.

Lexema	Lema	Análisis morfológico
Cantaríamos	Cant	- Verbo - Condicional indicativo - Primera persona - Plural
Cantaría	Cant	- Verbo - Condicional indicativo - Primera persona - Singular

Es importante notar que esta definición tiene validez en los dos sentidos de traducción. No solo será útil para el análisis morfológico del texto en español, sino que también se utilizará para la generación del texto hacia el español. Desde luego existe un transductor de las mismas características en la otra lengua (en este caso el inglés) como se vio en el estudio de Apertium.

Un elemento más que encontramos en la definición de los paradigmas verbales es el uso de sub-paradigmas. Funcionan en forma similar a los paradigmas de flexión que ya vimos, con la diferencia de que son utilizados dentro de otros paradigmas. También determinan una terminación y asocian a esta un conjunto de atributos. Al invocar un sub-paradigma dentro de un paradigma, el transductor unifica aquellas entradas que tengan como lema el del paradigma, a continuación la terminación de este y tienen como infijo la terminación del sub-paradigma. La imagen siguiente muestra, arriba, un fragmento del paradigma *s/aber__vblex* que define las flexiones del verbo saber, y abajo un fragmento del sub-paradigma *S__cantándo* que se utiliza para reconocer las contracciones pronominales.

```
    :
  <e>
    <p>
      <l>ándo</l>
      <r>ar<s n="vblex"/><s n="ger"/><j/></r>
    </p>
    <par n="s__cantándo"/>
  </e>
    :
```

```
<pardef n="s__cantándo">
      <e>
        <p>
          <l>me</l>
          <r>prpers<s n="prn"/><s n="enc"/><s n="p1"/><s n="mf"/><s n="sg"/></r>
        </p>
      </e>
      <e>
        :
```

La combinación del paradigma y el sub-paradigma de la figura permite reconocer contracciones como *cant-ándo-me* identificando que se trata de un gerundio y el pronombre *–me,* primera persona del singular.

5.1.2 Morfología del Voseo

Ya se estudió en el capítulo 2 las características del voseo verbal en el Río de la Plata. En resumen el voseo verbal que se busca reconocer presenta variantes en la flexión respecto al español estándar en las formas del presente indicativo y del imperativo. Sin embargo estas variantes son de naturaleza regular, aún más que el español estándar. Salvo por contadas excepciones que se verán más adelante.

5.1.3 Variante para el presente indicativo

El modo voseante afecta al presente indicativo provocando reducción de diptongo [9] en la última sílaba. De modo que por ejemplo *Vives* se transforma en *vivís*, *cantas* en *cantás* y *comes* en *comés*.

Respecto del infinitivo se transforma la *r* final en *s* y se acentúa la última silaba. Nótese que se trata de una flexión más regular que para el Español estándar en donde aquellos verbos terminados en *-ir* tienen una flexión más compleja. Por ejemplo el verbo *vivir* se flexiona en *vives*, donde la *–i* se transforma en *–e*. O por ejemplo el verbo *pedir* que se flexiona en *pides*; en este caso no solo se ve modificada la última vocal, sino que también impacta sobre la primera vocal.

5.1.4 Variante para el presente imperativo

El voseo afecta al presente imperativo en forma similar al indicativo. En este caso también hay reducción de diptongo. Por ejemplo *canta* se transforma en *cantá*, *vive* en *viví* y *come* en *comé*.

Respecto del infinitivo, desaparece la *–r* final y se acentúa la última vocal. También en el presente imperativo, el voseo es menos irregular que en el español estándar para aquellos verbos terminados en *-ir*.

5.1.5 Excepción a la regla

Se vio cuáles son las reglas de flexión que aplican al voseo verbal. Estas son regulares para la gran mayoría de los verbos. Sin embargo hay un caso para el cual estas reglas no aplican, el verbo *ser*.

Este verbo presenta de por sí un comportamiento irregular. Desde el punto de vista del voseo se ve modificado su uso respecto del español estándar para la segunda persona del singular, en donde en lugar de utilizar *eres* se utiliza el lexema *sos*, cuyo uso está explicado en el estudio del español del Río de la Plata. Basta con modificar el paradigma */ser_vbser* que define las flexiones del verbo ser e incluir una entrada para la terminación *sos*. En el caso particular del verbo *ser* el lema es nulo.

```
<pardef n="/ser__vbser">
  <e>
    <p>
      <l>sos</l>
      <r>ser<s n="vbser"/><s n="pri"/><s n="p2"/><s n="sg"/></r>
    </p>
  </e>
  :
```

Otra particularidad a tener en cuenta son las contracciones por ejemplo *cómelo* o *vívelo*. Aquí el pronombre se fusiona con el verbo provocando una extensión del lexema respecto a la flexión estudiada más arriba. Simplemente se agrega al final el pronombre. En el caso del voseo las contracciones aparecen para la variante del presente imperativo.

Para solucionar el problema de las contracciones basta con agregar una entrada igual a la utilizada para reconocer el presente imperativo en modo voseante pero invocando a los distintos sub-paradigmas que definen las contracciones pronominales.

5.1.6 Probando en Apertium

Como experimento se probó reconocer el presente indicativo en modo voseante, como prueba se eligió el verbo '*cantar*'. Se intentó reconocer la flexión '*cantás*'. Para eso fue necesario incluir la flexión en el paradigma de este verbo. Se agregó un atributo para identificar que se trata de un verbo en presente indicativo.

```
<e>
  <p>
    <l>ás</l>
    <r>ar<s n="vblex"/><s n="pri"/><s n="p2"/><s n="sg"/></r>
  </p>
</e>
```

En la figura se puede ver cómo se asocia la terminación 'ás' a una flexión verbal (vbflex) del presente indicativo (pri) del singular (sg) de la segunda persona (p2), que corresponde al análisis morfológico estudiado para el presente indicativo en modo voseante.

En el diccionario morfológico todas las formas del presente del verbo '*to sing*' (traducción de '*cantar*' en el diccionario bilingüe) se traducen como '*sing*' por lo tanto no será necesario modificar el diccionario. Se recompila Apertium y al probar la oración '*Conmigo cantás muy bien'* la traducción es '*With me you sing very well'*.

Solo con agregar la flexión para el voseo se logra traducir todos los verbos que utilicen la misma. También es posible agregar un atributo para indicar que se trata de un voseo para, por ejemplo, buscar una traducción más informal en la lengua origen.

Luego se probó incluir '*vos'* como pronombre. El diccionario que provee Apertium tiene un paradigma para todos los pronombres fuertes y una única entrada léxica con lema nulo que utiliza este paradigma. De esta manera solo hace falta agregar *'vos'* dentro de este.

```
<p>
  <l>vos</l>
  <r>prpers<s n="prn"/><s n="tn"/><s n="p2"/><s n="mf"/><s n="sg"/></r>
</p>
```

En la figura se puede ver cómo se asocia la terminación *'vos'* con un pronombre (prn) fuerte (tn) en segunda persona (p2) del singular (sg). Se recompila Apertium y al probar la oración *'Vos aceptás que te maltraten* ', se traduce como '*You accept that they abuse you.*'

En la siguiente figura se puede ver la salida del analizador morfológico para cada una de las oraciones. Resuelve los voseos sin ambigüedad y en forma correcta. Se puede observar también cómo la forma voseante del verbo '*aceptar'* es reconocida correctamente por tener la misma flexión que el verbo '*cantar'*.

```
^Vos/Prpers<prn><tn><p2><mf><sg>$
^aceptás/aceptar<vblex><pri><p2><sg>$
^que/que<cnjcoo>/que<cnjsub>/que<rel><an><mf><sp>$
^te/prpers<prn><pro><p2><mf><sg>$
^maltraten/maltratar<vblex><prs><p3><pl>/maltratar<vblex><imp><p3><pl>$^./.<sent>$^./.<sent>$[][

^Conmigo/Con<pr>+mí<prn><tn><p1><mf><sg>$
^cantas/cantar<vblex><pri><p2><sg>$
^muy/muy<preadv>$
^bien/bien<adv>/bien<preadv>/bien<n><m><sg>$^./.<sent>$^./.<sent>$[][
```

Al traducir desde el inglés al español una oración que contenga algún verbo en presente indicativo, Apertium se encuentra que al generar el texto en español hay dos posibilidades, el presente indicativo clásico y el presente indicativo en modo voseante, puesto que no se eliminó la entrada del español clásico. Apertium toma la opción que aparezca primero en la definición del paradigma. Por lo tanto al traducir desde el inglés, siempre se generará español netamente voseante o netamente tuteante. Esto es una limitante pues el voseo no se utiliza en todos los ámbitos, se observa sobretodo en ámbitos de poca formalidad. Considerando que Apertium realiza un análisis de la pragmática bastante básico, no se cuenta con información para detectar el registro de la oración (formal o informal) y no es posible definir en modo de ejecución si optar por el tuteo o el voseo.

En principio no se eliminarán las entradas para reconocer el tuteo, pues se busca extender el español clásico, y no restringirlo.

Así se logra reconocer voseos utilizando Apertium, En el siguiente capítulo se investiga el método para modificar todos los paradigmas verbales para adaptarlos al voseo.

5.2 Entidades Con nombre

En cuanto a las entidades con nombre, no se hace referencia en la documentación de Apertium acerca de algún tratamiento especial para detectarlas. Sin embargo existe en el diccionario una expresión regular para reconocer nombres propios que está deshabilitada porque enlentece la generación del transductor. El método utilizado para reconocer entidades con nombre es simplemente agregando las mismas como unidades léxicas dentro del diccionario utilizado por el analizador morfológico.

```
<e lm="Uruguay"><i>Uruguay</i><par n="Afganistán__np"/>     </e>
```

Arriba se ve el ejemplo de la entrada del diccionario para Uruguay. Las entidades con nombre también tienen asociado un paradigma, el cual determina atributos como el tipo

de entidad o el género. Por ejemplo, el paradigma *"Afganistán__np"* tiene la siguiente definición:

```
<pardef n="Afganistán__np">
  <e>
    <p>
      <l/>
      <r><s n="np"/><s n="loc"/><s n="m"/><s n="sg"/></r>
    </p>
  </e>
</pardef>
```

Esté paradigma se utiliza para las entidades de tipo localidad, del género masculino, singular. En general las entidades con nombre suelen traducirse como identidad, es decir que su nomenclatura es la misma en las dos lenguas. Pero hay casos en los que sí existe traducción. Esto es muy común para los nombres de países. Por ejemplo *Estados Unidos* se traduce como *United States*. Para el caso de existir una traducción será necesario modificar también el diccionario utilizado por el módulo de transferencia.

El desafío en cuanto a las entidades con nombre es el de la regionalización: Lograr reconocer entidades propias del Río de la Plata que son desconocidas por los recursos lingüísticos con los que cuenta Apertium.

5.2.1 Geonames

Geonames [30] es una base de datos de nombres geográficos que contiene más de 10 millones de nombres. Funciona bajo licenciamiento Common Creative 3.0. Contiene diversos tipos de nombres agrupados por categorías y sub-categorías, como ciudades, accidentes geográficos, plazas etc. Almacena datos como latitud y longitud, población, código postal, elevación, entre otros. Esta inmensa base de datos tiene como fuente entre otros la *National Geospatial-Intelligence Agency's*, la *OrdenanceSurveyOpenData* y el *U.S. Geological Survey Geographic Names Information System*.

Los datos pueden ser descargados en formato texto o accedidos a través de diversos Web Services que permiten acceder a la información de varias formas.

¿Cómo utilizar esta enormidad de información?, ¿Cómo obtener únicamente aquellos nombres propios de la región? Geonames permite obtener los nombres por país, por lo que es posible obtener todos los nombres geográficos de Uruguay. Sin embargo existen más de 50 mil en la lista, pues se puede encontrar variedad de entidades, desde nombres de plazas y comercios, caminos y calles, hasta las ciudades y los pueblos más importantes del Uruguay. No es deseable enlentecer y sobrecargar el transductor con miles de nombres geográficos para los cuales tampoco se conoce la traducción. Se hace difícil

definir alguna regla para delimitar qué nombres se desean reconocer y cuáles no. En definitiva el objetivo es encontrar algún método automático para ingresar entidades con nombre en los recursos de Apertium.

Hay un tipo de entidad para el que sí existe un atributo para medir la relevancia y son los pueblos y ciudades, para los cuales Geonames almacena la población. Utilizando este atributo es posible en forma automática, dado un valor límite de población, extraer de la lista de nombres de Geonames, aquellos pueblos y ciudades cuya población esté por encima del valor límite definido. Luego habrá que ingresar estos nombres en Apertium.

En cuanto a los atributos que podrán tomar estas entidades son:

Tipo de Entidad ->Localidad

Género-> Masculino, Femenino

Número-> Singular, plural

Estos atributos están contenidos por los paradigmas, *Afganistán__np, Estados_Unidos__np, Bahamas__np y Barcelona__np*.

Se asume que la traducción para cada una de las localidades es la identidad. Por ejemplo 'Zapicán' se traducirá como 'Zapicán'. Pero ¿Qué sucede con aquellos nombres compuestos o que pueden tener homónimos? Se distinguen dos casos.

Por ejemplo, para el nombre 'Las Flores' se agregará una entrada para toda la cadena de caracteres (incluyendo el espacio). El transductor del analizador morfológico se quedará con la entrada más larga que unifique el texto. Por lo tanto en lugar de tomar el determinante 'Las' y el nombre 'Flores', dará como salida la localidad 'Las Flores'.

El otro caso es el de los homónimos. Para el caso de 'Salto', existirán varias entradas que unifiquen el texto, una de las cuales será la localidad, por lo que el analizador morfológico dará como salida todas las posibles categorías. Quedará en manos del Pos-Tagger elegir una de ellas. Esto podría parecer un problema dado que el Pos-Tagger no contiene en su corpus las nuevas localidades agregadas al traductor. Sin embargo la categoría asignada a las localidades sí existe en el corpus por lo que si las categorías asignadas a los homónimos tienen probabilidades bajas para el texto de entrada, entonces el Pos-tagger reconoce correctamente la localidad. Oraciones como, 'Salto es un lugar hermoso' o 'Me gustaría ir a Salto' son reconocidas correctamente por el Pos-Tagger.

5.2.2 Traducción de entidades

Existe también una gran base de nombres, la enciclopedia Wikipedia. La gran ventaja que ofrece Wikipedia es la posibilidad de ver un artículo en distintos idiomas. A su vez como cada artículo está identificado por su título se tiene su traducción accediendo al link del artículo traducido. Así es que para el artículo 'República Oriental del Uruguay' tiene como traducción al inglés el artículo 'Uruguay'. Por lo tanto tenemos un par de traducción simplemente utilizando los títulos. Experimentos muestran que en función del idioma entre un 62% y un 92% de los títulos de los artículos son una traducción correcta [35]. En el experimento de Tyers y Pienaar se utilizó un conjunto inicial de palabras extraídas desde el corpus de Apertium en inglés y utilizando el título del artículo traducido de cada palabra se generaron pares de palabras entre los dos lenguajes.

Lo otro interesante es que la mayoría de los artículos hacen referencia a sustantivos o nombres. En [22] se utilizó Wikipedia para alinear a nivel de oraciones un corpus bilingüe basado en el largo de las oraciones y los links interwiki.

Se proponen entonces dos experimentos. Primero recorrer páginas de Wikipedia de artículos de la región y obtener pares de traducción utilizando el título de los artículos como se hizo en [35]. Luego utilizando los links interwiki alinear entidades con nombre dentro del artículo.

Wikipedia es un recurso muy extenso con una gran cantidad de textos tanto en español como en inglés y la existencia de los artículos y sus versiones en otros idiomas puede ser explotado en variedad de formas. Por ejemplo alineando las sentencias de un artículo con traducción sería posible encontrar nuevas reglas de transferencia estructural.

6 Implementación y evaluación de la solución

6.1 Introducción

Con el objetivo de regionalizar se busca que Apertium reconozca los modos voseantes del Río de la Plata, que son la característica más fuerte de esa región, y mejorar el reconocimiento de las entidades con nombre propias del Río de la Plata. En el capítulo anterior se analizaron las siguientes soluciones para lograr este objetivo:

- Modificar todos los paradigmas de flexión verbal en el diccionario del analizador morfológico para reconocer el presente indicativo y el imperativo del modo voseante
- Incluir el pronombre vos en la lista de pronombres del diccionario del analizador morfológico.
- Obtener las ciudades y pueblos más relevantes de Uruguay desde la base de datos de Geonames, utilizando la población como atributo para medir la relevancia.
- Utilizar los nombres de los artículos de Wikipiedia (interwiki) y la existencia de versiones en varios idiomas de cada artículo.

A continuación se describe la solución técnica desarrollada para implementar las soluciones analizadas.

6.2 Implementación

La implementación de la solución se realizó a través de un conjunto de programas *Perl* que reciben como entrada y como salida archivos XML. Se optó por el lenguaje *perl* por tratarse de una plataforma que maneja en forma nativa y cómoda expresiones regulares. Es una tecnología estándar para aplicaciones de procesamiento de lenguaje.

6.2.1 Inclusión de modo voseante

Se vio en el capítulo anterior el método para incluir en Apertium los modos voseantes. Tanto el voseo verbal como el voseo pronominal. La solución estudiada consiste básicamente en modificar el diccionario utilizado por el analizador morfológico del español, y para cada paradigma de flexión verbal incluir las entradas para los modos del voseo verbal. Y además incluir el pronombre *vos* en el analizador morfológico.

El primer paso entonces es identificar todos los paradigmas de flexión, porque no hay que olvidar que los paradigmas se utilizan para determinar el análisis morfológico de todas las categorías gramaticales. Afortunadamente todos los verbos están ingresados juntos en una sección del diccionario. Además los paradigmas de flexión verbal tienen todos el sufijo *–vblex*. Utilizando esta información se desarrolló un programa *perl* que extrae todos los

paradigmas de flexión verbal. Tras ejecutar el programa se extrajeron un total de 161 paradigmas verbales.

Hay que recordar que el impacto del voseo es sobre el presente indicativo y el presente imperativo de la segunda persona del singular. En definitiva lo que se hace es tomar la entrada del español estándar para estas conjugaciones y se modifica la terminación.

La entrada para el presente indicativo es de la forma

```
<e>
  <p>
    <l>$indic</l>
    <r>$infinitivo<s n=\"vblex\"/><s n=\"pri\"/><s n=\"p2\"/><s n=\"sg\"/></r>
  </p>
</e>
```

En la variable *$indic* se indica la terminación para la flexión del presente indicativo según lo estudiado en el capítulo anterior.

En la variable *$infinitivo* se indica la terminación del infinitivo. Como ya se vio se incluyen los tres atributos que representan la flexión, presente indicativo (*pri*), segunda persona (*p2*), singular (*sg*).

La entrada para el presente imperativo es de la forma

```
<e>
  <p>
    <l>$imper</l>
    <r>$infinitivo<s n=\"vblex\"/><s n=\"imp\"/><s n=\"p2\"/><s n=\"sg\"/></r>
  </p>
</e>
```

En la variable *$imper* se indica la terminación para la flexión del presente imperativo según lo estudiado en el capítulo anterior.

En la variable *$infinitivo* se indica la terminación del infinitivo. Como ya se vio se incluyen los tres atributos que representan la flexión, presente imperativo (*imp*), segunda persona (*p2*), singular (*sg*).

Pero tiene que haber también una tercera entrada para el presente imperativo, pues en este caso pueden aparecer contracciones. Como se vio en el capítulo anterior, para resolver esta particularidad se utilizan sub-paradigmas.

```
<e>
  <p>
    <l>$imper</l>
    <r>$infinitivo<s n=\"vblex\"/><s n=\"imp\"/><s n=\"p2\"/><s n=\"sg\"/><j/></r>
  </p>
  <par n="S__cánta"/>
</e>
```

En la variable *$imper* se indica la terminación para la flexión del presente imperativo asumiendo que ocurre una contracción pronominal. Esto es muy importante porque al aparecer la contracción en el presente imperativo del voseo, el lexema pasa de ser una palabra aguda a ser una palabra grave debido a la sílaba que agrega el pronombre y desaparece el acento en la última vocal del presente imperativo del modo voseante. Por ejemplo *cantá* puede contraerse con el pronombre *-lo* en el lexema *cantalo*.

En la variable *$infinitivo* se indica la terminación del infinitivo. Y como ya se vio se incluyen los tres atributos que representan la flexión, presente imperativo (*imp*), segunda persona (*p2*), singular (*sg*).

Pero además hay que agregar el sub-paradigma para identificar el pronombre de la contracción. Para esto se incluye la referencia al sub-paradigma *S__cánta* que reconoce las terminaciones de las contracciones en lexemas como *cant-a-me*.

Como ya se discutió, estas entradas se incluyen al inicio de la definición del paradigma para que tengan prioridad sobre el español estándar. En la figura de abajo se ve la implementación para el paradigma *abandon/ar_vblex*.

```
<pardef n="abandon/ar__vblex">
  <e>
    <p>
      <l>ás</l>
      <r>ar<s n="vblex"/><s n="pri"/><s n="p2"/><s n="sg"/></r>
    </p>
  </e>
  <e>
    <p>
      <l>á</l>
      <r>ar<s n="vblex"/><s n="imp"/><s n="p2"/><s n="sg"/></r>
    </p>
  </e>
  <e>
    <p>
      <l>a</l>
      <r>ar<s n="vblex"/><s n="imp"/><s n="p2"/><s n="sg"/><j/></r>
    </p>
    <par n="S__cánta"/>
  </e>
  :
```

Para agregar estas entradas en el diccionario será necesario previamente determinar para cada uno de los 161 paradigmas verbales las terminaciones de los modos voseantes indicadas más arriba. Esto se hizo en forma manual a través de una interfaz de usuario. Sin embargo no todos los paradigmas verbales extraídos necesitan ser modificados. Ocurre que en el caso del presente imperativo del español estándar, la vocal fuerte se encuentra generalmente en la penúltima sílaba, haciendo del lexema una palabra grave, a diferencia del modo voseante en donde la vocal fuerte es la última. Al utilizar las contracciones aparece una nueva sílaba como sufijo del lexema. En el caso del modo voseante se vio que al aparecer esta nueva sílaba, el lexema pasaba de ser una palabra aguda a ser una palabra grave por lo que desaparece el acento de la última vocal. Esto implica un cambio respecto a la terminación por lo que fue necesario agregar una entrada adicional para el presente imperativo con el fin de reconocer las contracciones como ya se vio. Pero para el español estándar esta sílaba adicional implica que el lexema pase de ser una palabra grave a ser una palabra esdrújula, apareciendo entonces un acento en la antepenúltima sílaba, y en muchos casos esta sílaba es parte de la raíz por lo que no se puede reconocer este lexema desde el paradigma (recordar que el paradigma, es independiente de la raíz).

Para resolver este problema se ingresan dos entradas en el diccionario para los verbos que tienen este problema. Además de ingresar el uso normal del verbo (con la raíz sin el acento), se agrega una entrada con la raíz incluyendo el acento que se genera al aparecer contracciones. A esta nueva entrada se le asocia otro paradigma en el cual sí se invocan a los sub-paradigmas que identifican las contracciones.

En la figura de abajo se ve el ejemplo del verbo olvidar. La primera es la entrada clásica, que tiene asociada el paradigma más común *abandon/ar__vblex*. Se ve como la raíz es *olvid* por lo que el transductor para el lexema *olvídalo*, que corresponde a una contracción del presente imperativo para el español estándar, nunca tomará la primer entrada. Sin embargo la segunda entrada tiene como raíz *olvíd*; Y tiene asociada el paradigma *abandón/ar__vblex* (nótese el acento) cuya definición tiene ya en el inicio la terminación *–a* con la invocación al sub-paradigma *S_cánta* que como ya se vio identifica las contracciones, por lo que el lexema *olvíd-a-lo* ahora es reconocido sin problemas.

```
:
<e lm="olvidar"><i>olvid</i><par n="abandon/ar__vblex"/>     </e>
<e lm="olvidar"><p><l>olvíd</l><r>olvid</r></p><par n="abandón/ar__vblex"/>     </e>
:

:
<pardef n="abandón/ar__vblex">
  <e>
    <p>
      <l>a</l>
      <r>ar<s n="vblex"/><s n="imp"/><s n="p2"/><s n="sg"/><j/></r>
    </p>
    <par n="S__cánta"/>
  </e>
:
```

Estos paradigmas adicionales utilizados para reconocer las flexiones con contracciones para el español estándar no serán modificados pues no se ven afectados por los modos voseantes.

Luego de determinar manualmente las terminaciones del voseo para cada paradigma verbal, se generó un archivo XML con la información ingresada. Este archivo y el diccionario del analizador morfológico del español son tomados como entrada por un programa *Perl* que para cada paradigma del archivo XML, busca su definición en el diccionario y agrega (utilizando las terminaciones incluidas en el XML) las entradas para el modo voseante estudiadas antes.

```
:
<parVoseo paradigma="abat/irse__vblex" indic="ís" imper="í" contr="i"/>
<parVoseo paradigma="dem/ostrar__vblex" indic="ostrás" imper="ostrá" contr="ostra"/>
<parVoseo paradigma="sal/ir__vblex" indic="ís" imper="í" contr="i"/>
<parVoseo paradigma="prohib/ir__vblex" indic="ís" imper="í" contr="i"/>
<parVoseo paradigma="o/ír__vblex" indic="ís" imper="í" contr="i"/>
<parVoseo paradigma="/ser__vbser" indic="sos"/>
:
```

En la imagen de arriba se ve un fragmento del archivo generado. Notar la entrada para el paradigma */ser__vbser* en donde se definen las flexiones del verbo ser. Como ya se vio este verbo tiene un comportamiento irregular respecto al resto y es que en el modo voseante se sustituye en uso del lexema *eres* por el lexema *sos*. Este paradigma queda modificado como se muestra a continuación.

```
<pardef n="/ser__vbser">
  <e>
    <p>
      <l>sos</l>
        <r>ser<s n="vbser"/><s n="pri"/><s n="p2"/><s n="sg"/></r>
    </p>
  </e>
  :
```

El último paso es incluir el voseo pronominal. Apertium tiene un paradigma para todos los pronombres fuertes y una única entrada léxica con lema nulo que utiliza este paradigma.

```
<pardef n="prfuerte__prn">
    <!--Un solo paradigma para los pronombres personales fuertes -->
    <e>
      <p>
        <l>yo</l>
        <r>prpers<s n="prn"/><s n="tn"/><s n="p1"/><s n="mf"/><s n="sg"/></r>
      </p>
    </e>
    <e>
      <p>
        <l>tú</l>
        <r>prpers<s n="prn"/><s n="tn"/><s n="p2"/><s n="mf"/><s n="sg"/></r>
      </p>
    </e>
    :
```

```
 :
 <e lm="pronombres personales fuertes"> <i/> <par n="prfuerte__prn"/> </e>
 :
```

De esta manera solo hace falta agregar *'vos'* dentro de este.

```
<p>
  <l>vos</l>
  <r>prpers<s n="prn"/><s n="tn"/><s n="p2"/><s n="mf"/><s n="sg"/></r>
</p>
```

6.2.2 Inclusión de entidades con nombre

Para el reconocimiento de entidades se habían estudiado dos soluciones. Utilizar la base de nombres de Geonames y extraer de allí ciudades y pueblos del Uruguay, utilizando la

población como atributo para medir la relevancia de cada nombre. Luego utilizar Wikipedia para encontrar una traducción utilizando la propiedad de los links interwiki.

Desde el sitio de Geonames se obtuvo el archivo *UY.txt* que contiene todos los nombres geográficos de Uruguay. Este archivo está estructurado en filas cuyas columnas están delimitadas por tabuladores. En cada fila se almacena un nombre. Los campos de cada fila son:

```
The main 'geoname' table has the following fields :
---------------------------------------------------
geonameid         : integer id of record in geonames database
name              : name of geographical point (utf8) varchar(200)
asciiname         : name of geographical point in plain ascii characters, varchar(200)
alternatenames    : alternatenames, comma separated varchar(5000)
latitude          : latitude in decimal degrees (wgs84)
longitude         : longitude in decimal degrees (wgs84)
feature class     : see http://www.geonames.org/export/codes.html, char(1)
feature code      : see http://www.geonames.org/export/codes.html, varchar(10)
country code      : ISO-3166 2-letter country code, 2 characters
cc2               : alternate country codes, comma separated, ISO-3166 2-letter country
                    code, 60 characters
admin1 code       : fipscode (subject to change to iso code), see exceptions below,
                    see file admin1Codes.txt for display names of this code; varchar(20)
admin2 code       : code for the second administrative division, a county in the US,
                    see file admin2Codes.txt; varchar(80)
admin3 code       : code for third level administrative division, varchar(20)
admin4 code       : code for fourth level administrative division, varchar(20)
population        : bigint (8 byte int)
elevation         : in meters, integer
gtopo30           : average elevation of 30'x30' (ca 900mx900m) area in meters, integer
timezone          : the timezone id (see file timeZone.txt)
modification date : date of last modification in yyyy-MM-dd format
```

En el análisis de soluciones se vio que en este archivo hay toda clase de entidades, pueblos y ciudades, plazas, comercios, entre otros. Se trata de una enorme cantidad de nombres, en su mayoría irrelevantes. Más de cinco mil nombres que de incluirse todos enlentecen innecesariamente la ejecución Apertium (el tiempo de ejecución es una de las grandes ventajas de Apertium) solo para reconocer nombres de muy baja probabilidad. Se decidió entonces utilizar como atributo para medir la relevancia la población y extraer así del archivo *UY.txt* los pueblos y ciudades que tengan una población por encima de un valor límite. Mediante un programa *perl* se extrajeron 120 entidades con nombre a las cuales manualmente mediante una interfaz de usuario se les asigno alguno de los paradigmas utilizados para las entidades con nombre que se estudiaron.

```
:
<e lm="Colonia"><i>Colonia</i><par n="Barcelona__np"/>    </e>
<e lm="Departamento de Colonia"><i>Departamento<b/>de<b/>Colonia</i><par n="Afganistán__np"/>    </e>
<e lm="Chuy"><i>Chuy</i><par n="Afganistán__np"/>    </e>
<e lm="Departamento de Cerro Largo"><i>Departamento<b/>de<b/>Cerro<b/>Largo</i><par n="Afganistán__np"/>    </e>
<e lm="Cerro Colorado"><i>Cerro<b/>Colorado</i><par n="Afganistán__np"/>    </e>
<e lm="Cebollatí"><i>Cebollatí</i><par n="Afganistán__np"/>    </e>
<e lm="Casupá"><i>Casupá</i><par n="Barcelona__np"/>    </e>
<e lm="Castillos"><i>Castillos</i><par n="Afganistán__np"/>    </e>
<e lm="Carmelo"><i>Carmelo</i><par n="Afganistán__np"/>    </e>
:
```

Finalmente se ingresan en el diccionario del analizador morfológico las entidades con nombre extraídas desde Geonames. Notar que en principio se indica mediante el tag *<i>* que la traducción de cada entidad será la misma que para el español. Por ejemplo, *Paso de los Toros* será traducido tal cual al inglés. La gran ventaja es que no es necesario modificar los archivos de transferencia.

6.3 Evaluación

La evaluación de un sistema de traducción es una tarea sumamente compleja. Desde luego el sistema de evaluación más completo es la evaluación hecha por un ser humano. El problema radica en que dada una oración, existen varias traducciones correctas. Algunas varían en el orden de algunas palabras, y otras por ejemplo utilizan palabras distintas. Un humano puede ver dos traducciones y detectar que ambas son correctas. Pero cuando se trata de determinar un método automático de evaluación esto se hace muy difícil porque siempre se va a comparar contra una traducción supuestamente correcta, y solo una. Existen sin embargo algunas métricas de evaluación que intentan mitigar ese problema.

Habiendo definido una métrica, será necesario contar con un corpus bilingüe de testeo contra el cual comparar las traducciones de Apertium. Que además tiene que tener la particularidad de ser representativo del Río de la Plata.

6.3.1 Métricas

Como ya se discutió, la definición de métricas para evaluar sistemas de traducción puede resultar una tarea compleja. Esto se debe a que existe más de una traducción correcta. Sin embargo, las distintas traducciones correctas suelen tener elementos en común. Aparecen lexemas en común, n-gramas en común, y también es de esperar que tengan una longitud similar [26].

En todo momento se asume que existe una traducción de *referencia* contra la cual se va a evaluar el sistema de MT, y siempre la unidad básica de evaluación es la oración. Se busca entonces una métrica de precisión que:

- Evalúe positivamente las traducciones que se adecuen más a la referencia. Aquellas traducciones que compartan más lexemas con la referencia estarán más adecuadas a esta.
- Evalúe positivamente las traducciones que comparten n-gramas más largos respecto a la referencia. Estas traducciones satisfacen la fluidez.

BLEU es una de las métricas de evaluación de sistemas de MT más reconocidas. Busca ponderar las dos características mencionadas (fluidez y adecuación) y además penaliza aquellas traducciones cuya longitud no se condice con la referencia. Esto es muy importante porque de lo contrario pueden aparecen traducciones malas que sin embargo cumplan con la adecuación y fluidez. Y también puede suceder lo contrario, traducciones *buenas* que por ejemplo no cumplen con la fluidez. Perfectamente puede aparecer alguna conjunción en el candidato que no está en la referencia. Esto provoca que un n-grama quede partido, pero esto no quiere decir que la traducción sea mala. Por ejemplo:

Referencia: "*...heed the Party commands...*"

Candidato: "*...heed of the Party commands...*"

Esto se debe a que a veces se debe penalizar aquellas palabras que aparecen en el candidato pero no en la referencia, y a veces no se debe penalizar. Para resolver este problema BLEU penaliza aquellas oraciones de longitud menor a la candidata [26]. Resumiendo, para realizar el cálculo de BLEU se mide:

- La precisión de la traducción contando la cantidad de n-gramas que tienen en común la traducción candidata y la referencia. Solo que al contabilizar los n-gramas en común, se tiene en cuenta el número máximo de veces que aparece cada n-grama en la traducción de referencia. Esto es para evitar que candidatos como "*the the the the*" para la referencia "*the cat is in the box*" salgan favorecido por un conteo tonto de n-gramas.

$$Pn = {CantMax(Ngrmas\ en\ Comun)}\Big/{Cant(NGramas\ Candidato)}$$

- Se penaliza aquellos candidatos cuya longitud no sea similar a la referencia. Esto es porque candidatos como "*on the*" para la referencia "*the cat is on the box*" son muy favorecidos por la medida de precisión y sin embargo no representan una traducción correcta. Para esto se utiliza un coeficiente de penalización por brevedad que se calcula como $PB = \begin{cases} 1\ si\ cand. > ref. \\ e^{(1-\frac{ref}{cand})}\ si\ cand. \leq ref \end{cases}$ donde *cand* y *ref* son las longitudes del candidato y la referencia.

- Finalmente BLEU se calcula combinando la penalización por brevedad con la media geométrica de la precisión para cada longitud de n-gramas
  $BLEU = PB * \exp(\sum_1^N w_n log p_n)$ Donde los pesos Wn deben sumar 1

BLEU es un estándar en evaluación de sistemas de MT. Es ampliamente aceptada y como se muestra en [26] es casi tan efectivo como una evaluación humana. Con este criterio es que se eligió BLEU como métrica para evaluar las modificaciones hechas en Apertium.

Además de utilizar la métrica BLEU se tendrá en cuenta también la métrica NIST, también utilizada para evaluar sistemas de traducción. NIST está basada en BLEU, se diferencia simplemente en que da mejor puntuación a aquellos n-gramas menos comunes y que en definitiva aportan más información al contenido de la oración. Por ejemplo, NIST puntuará mejor el n-grama "cálculos interesantes" que el n-grama "en la". Además difiere en la penalización por brevedad de forma tal que pequeñas variaciones en el largo de la traducción no impactan tanto en el resultado final de la evaluación como en BLEU.

6.3.2 Construcción de Corpus de Prueba

Resulta difícil encontrar un corpus paralelo con textos del Español del Río de la Plata. Porque además, la principal característica de esta variante del español se observa en la segunda persona del plural, la cual se observa generalmente en diálogos.

Pero existe un tipo de documento para el cual existe naturalmente una traducción, y en donde generalmente se observan diálogos y conversaciones. Son los subtítulos de las películas. Para cualquier película existen subtitulaciones en varios idiomas. Estos subtítulos pueden verse como transcripciones de un mismo texto pero en distinto idioma, por lo que de por sí representan un texto bilingüe. Pero los subtítulos cuentan además con una propiedad sumamente valiosa que es la ventana de tiempo en el cual deben aparecer en la pantalla. Esto permite tener información adicional que es de mucha utilidad en el momento de alinear dos subtítulos de una misma película.

Selección de películas para el Corpus

Lo ideal sería contar con las películas que presentan el español rioplatense más representativo. Sin embargo a priori no existe ningún atributo para determinarlo. Se tomaron entonces las películas argentinas y uruguayas mejor puntuadas según IMDB.com con la premisa de que sería más factible encontrar sus respectivos subtítulos en ambos idiomas. Siempre que fue posible se seleccionó la transcripción original extraída desde las versiones de DVD de las películas. Cuando no fue posible se utilizaron subtítulos realizados por usuarios de Internet, intentando siempre seleccionar el mejor puntuado.

Pre-tratamiento de Subtítulos

Previo a la alineación del corpus fue necesario realizar algunos procesamientos sobre los textos de los subtítulos.

- Se codificaron todos los archivos en UTF-8
- Algunos pares de subtítulos no tenían los mismos tiempos pues fueron diseñados para distintas versiones del video. En estos casos se alinearon los tiempos utilizando una herramienta gráfica para manipular subtítulos.
- Se unificó el formato de todos los archivos de subtítulos a Srt (SubRIP).

Se recolectaron finalmente 26 películas argentinas y uruguayas.

Alineando los Corpus

Basado en [35] y [33, 13, 3] se alinearon todos los subtítulos a nivel de oraciones. Utilizando los tiempos de aparición y finalización de las líneas en la pantalla, y el largo de las oraciones se logró alinear con bastante precisión las oraciones de los pares de subtítulos.

Para evaluar la precisión de la heurística se extrajeron 52 oraciones al azar (2 por subtítulo) y se corrigieron manualmente. Un 80.2% de las oraciones fueron alineados correctamente por la heurística. Las sentencias mal alineadas se debían a textos que aparecían en un subtítulo pero en el otro, alineación 1:0 y 0:1.

6.3.3 Evaluación

La evaluación fue una tarea compleja ya que existen condicionantes que hacen difícil medir la mejora que hemos incluido en la plataforma Apertium:

- Los subtítulos son generalmente transcripciones del lenguaje oral por lo que aparecen expresiones propias del lenguaje hablado que redundan en una mala traducción.
- Las modificaciones realizadas fueron exclusivamente a nivel léxico por lo que aquellos sintagmas que eran traducidos incorrectamente por Apertium, seguirán así.
- ¿Cómo comparar dos traducciones? La traducción de Apertium puede no coincidir con la traducción del subtitulo y aún así ser correcta.

Dadas estas condicionantes resulta poco útil comparar exactamente la traducción de Apertium y la del subtítulo para determinar si es correcta o no.

Ya se estudió en este capítulo las métricas utilizadas en el campo de la traducción automática. En particular para este proyecto se tuvieron en cuenta la métrica BLEU y la métrica NIST. Los scripts [25] de evaluación utilizados fueron los desarrollados en la

edición 2008 de la competencia de métodos de evaluación organizada por la NIST (National Institute of Standards and Technology). Están desarrollados en perl y el texto fuente, el texto de referencia y el texto a evaluar son tomados de archivos XML con el formato definido en la competencia. En la imagen se ve un extracto del archivo de evaluación del sistema Apertium, traducción desde el subtítulo de la película *El Baño del Papa.*

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.dtd">
<tstset setid="sample_set" srclang="Spanish" trglang="English" SysID="Apertium">
<doc docid="sample_document_1" genre="nw" SysID="Apertium">
    :
    <seg id="4">it Put this in my dispatch, please.</seg>
    <seg id="5">Good day, Moon.</seg>
    <seg id="6">Care!</seg>
    <seg id="7">Which crapped. We have ~a problem.</seg>
    :
    :
```

El objetivo de la evaluación es medir el impacto de la inclusión del voseo y las entidades con nombre en Apertium. Y además comparar Apertium con otro traductor de uso común, como es el caso de *Google Translator.* Se eligió este sistema en particular por ser uno de los más usados en el mundo, por ser uno de los traductores automáticos que presenta mejores resultados y por ser de naturaleza estadística, en contraposición con Apertium.

En definitiva se evaluaron tres sistemas (Apertium, Apertium Río de la Plata y Google Translator) en los dos sentidos de traducción, Español-Inglés e Inglés-Español. En la siguiente imagen se ve una salida ejemplo del script de evaluación desarrollado:

```
  Evaluation of Spanish-to-English translation using:
    src set "sample_set" (1 docs, 771 segs)
    ref set "sample_set" (1 refs)
    tst set "sample_set" (1 systems)

NIST score = 3.4444  BLEU score = 0.1218 for system "Apertium"

# ------------------------------------------------------------------------

Individual N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.7878   0.5265   0.1048   0.0232   0.0021   0.0001   0.0000   0.0000   0.0000  "Apertium"

 BLEU:  0.4356   0.1816   0.0803   0.0347   0.0163   0.0047   0.0019   0.0007   0.0003  "Apertium"

# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.7878   3.3142   3.4191   3.4422   3.4444   3.4445   3.4445   3.4445   3.4445  "Apertium"

 BLEU:  0.4356   0.2812   0.1852   0.1218   0.0815   0.0507   0.0318   0.0196   0.0122  "Apertium"
```

6.4 Resultados

Tras evaluar los tres sistemas de traducción, se observó como era de esperarse una superioridad importante del Sistema de traducción de Google. La información más interesante está en cómo se vio afectado Apertium con las modificaciones hechas.

6.4.1 Español-Inglés

Para el caso de la traducción de español a inglés, en todos los casos se observaron mejores resultados en Apertium adaptado al Río de la Plata respecto a Apertium Tradicional. En promedio las métricas arrojaron los siguientes resultados:

	BLEU	NIST
Apertium Tradicional	0.118183333	3.414683333
Apertium R.P	0.1246	3.553916667
Google Translator	0.226116667	4.810316667

Es difícil comprobar qué porcentaje del texto a traducir corresponde a sintagmas voseantes, pero se observa un incremento en la performance del Apertium del orden del 5.4% respecto a la versión tradicional del sistema. Recordar que el sistema modificado reconoce los verbos en modo voseantes y sus contracciones, y el pronombre *vos* en todos sus usos.

Se probó el mismo experimento sobre el subtítulo de la película *Mar adentro* de origen Español y las métricas fueron idénticas para Apertium Tradicional, algo esperable dada la escasa o nula existencia del modo voseante en el Español estándar.

Respecto al reconocimiento de entidades con nombre, se etiquetó el texto utilizado en la evaluación con las dos versiones de Apertium (la versión tradicional y la versión mejorada) y se contaron la cantidad de entidades con nombre reconocidas. Ya que la mejora consistió en agregar localidades del Uruguay, se realizaron dos evaluaciones: una teniendo en cuenta únicamente los subtítulos de las películas de origen Uruguayo, y otra sobre el total de los textos. En el anexo se estudia con más detalle el conteo realizado, detallando además los resultados para verbos y pronombre.

Localidades	Apertium R.P	Apertium Trad.	Aper.RP/Aper.Trad
Películas Uruguayas	98	36	2.722
Todas las películas	223	158	1.411

Entidades con Nombre	Apertium R.P	Apertium Trad.	Aper.RP/Aper.Trad
Películas Uruguayas	330	267	1.235

Todas las películas	1458	1396	1.044

En cuanto al reconocimiento de localidades en textos de origen Uruguayo se observa una muy importante mejora de casi el 280% respecto a la versión tradicional de Apertium. Incluso en el total de los textos (subtítulos de películas en su mayoría Argentinas) también se observa una mejora de un 40%. Hay que tener en consideración que solo un 6.5% de las entidades con nombre representan localidades.

6.4.2 Inglés-Español

Al traducir de Inglés al Español hay que tener en cuenta que el traductor Apertium adaptado al Río de la Plata puede funcionar en dos modos al generar Español, pues puede generar en el modo tradicional (exclusivamente tuteante) o en modo exclusivamente voseante. Con el funcionamiento tradicional los resultados obtenidos son idénticos al traductor sin las modificaciones. Aquí se detallan los resultados usando el generador con modos voseantes. En promedio las métricas arrojaron los siguientes resultados:

	BLEU	NIST
Apertium Tradicional	0.112733333	3.35635
Apertium R.P	0.111433333	3.374283333
Google Translator	0.21005	4.597533333

En este caso los resultados variaron según los textos. En algunos tuvo mejores resultados Apertium para Río de la Plata y en otros Apertium Tradicional. Esto se explica debido a que si bien se traducen sintagmas voseantes correctamente, en el texto de testeo no todos los sintagmas son de esta naturaleza. En definitiva traducir exclusivamente en modo voseante no es una representación correcta del Español del Río de la Plata. Se probó el mismo experimento sobre el subtítulo de la película *Mar adentro* de origen Español y las métricas fueron mejores para Apertium Tradicional, algo esperable dada la escasa o nula existencia del modo voseante en el Español estándar.

7 Conclusiones

Con el objetivo de regionalizar al Río de la Plata un sistema de traducción Open Source, se realizó un estudio profundo de las particularidades del castellano del sur de América. Se analizó su origen y se determinó que el uso del voseo es la característica más importante del Español del Río de la Plata.

En base a este estudio, se analizó el funcionamiento del voseo. Se determinaron todas las flexiones aceptadas y se definieron las reglas que las generaban.

Con el fin de incluir el voseo en Apertium se estudió a fondo el funcionamiento de la herramienta y se investigó cómo incluir los modos voseantes en el sistema. Modificando el diccionario del Español de Apertium, se desarrollaron programas para incluir en forma automática las flexiones del voseo. Además utilizando la base de nombres de Geonames se enriqueció el reconocimiento de entidades con nombre de Apertium incluyendo más de cien localidades del Uruguay.

Para evaluar las mejoras incluidas se construyó un corpus paralelo basado en subtítulos de películas del Río de la Plata, las cuales por contener en su mayoría diálogos, presenta un uso bastante frecuente de voseos. Utilizando las métricas BLEU y NIST se evaluó, en base al corpus paralelo, Apertium con las modificaciones realizadas frente a Apertium Tradicional. Como referencia se evaluó también el traductor de Google sobre los mismos textos a fin de conocer la calidad de las traducciones de Apertium frente a un traductor reconocido como es el de Google.

Las evaluaciones fueron hechas con dos métricas estándar, BLEU y NIST, y los resultados arrojaron una mejora de un 5,4% en las traducciones al inglés de subtítulos de películas rioplatenses. En el sentido de traducción inverso (inglés al español), los resultados fueron variados debido a que al generar se debe optar por hacerlo en modo tuteante o voseante exclusivamente.

Además se observó que en textos del Uruguay el reconocimiento de localidades creció un 270% lo que representa una mejora de 23% en el reconocimiento de entidades con nombre en textos de origen uruguayo.

Como es de esperar, en textos que no contengan voseos el funcionamiento de Apertium no se ve modificado y los resultados de las métricas son idénticos para la versión tradicional y la estándar.

Como resultado se logró conseguir una versión de Apertium que reconoce en forma aceptable la principal característica del Río de la Plata, el voseo, en sus dos acepciones, pronominal y verbal. Y reconoce además los pueblos y ciudades del Uruguay. El desarrollo fue incluido en el repositorio de Apertium como un par de lenguas más.

7.1 Trabajo a Futuro

Como trabajo a futuro se plantea la posibilidad de investigar más a fondo el módulo de transferencia estructural de Apertium, encargado de construir los árboles sintácticos, a fin de lograr detectar el registro de la oración (formal, informal, semi-formal, etc.) para poder determinar cuándo utilizar un modo tuteante y cuándo utilizar un modo voseante.

Desarrollar una solución que permita a través de los links interwiki existentes en la Wikipedia, obtener entidades con nombre y su traducción.

8 Referencias

[1] J. G. d. Amores Carredano. Los sistemas de traducción automática engspan tm y spanam tm de la organización panamericana de la salud. *Procesamiento del lenguaje natural. Nº 18 (mayo 1996), pp. 87-101*, 1996.

[2] A. Boretz. Apptek launches hybrid machine translation software. *http://www.speechtechmag.com/Articles/News/News-Feature/AppTek-Launches-Hybrid-Machine-Translation-Software-52871.aspx/*, Accedida en Mayo 2010.

[3] P. F. Brown, J. Lai, and R. Mercer. Aligning sentences in parallel corpora. *ACL*, 1991.

[4] e. a. Canals Marote, Raúl. El sistema de traducción automática castellano-catalán internostrum. *Procesamiento del lenguaje natural. Nº 27 (sept. 2001), pp. 151-156*, 2001.

[5] V. Chaitanya, R. Sangal, and A. B. (Group). *Natural language processing: a Paninian perspective*. Prentice-Hall of India, 1996, 1996.

[6] R. A. E. B. de datos (CORDE) [en línea]. Corpus diacrónico del español. *http://www.rae.es*, Accedida el 04/10/2011.

[7] G. Documentation. Acerca del traductor de google. *http://translate.google.com.uy/about/intl/es_ALL/*, Accedido el 16/08/2011.

[8] A. Elizaincín. Geolingüística, sustrato y contacto lingüístico: español, portugués e italiano en uruguay. *ROSAE – Congresso em Homenagem a Rosa Virgínia Mattos e Silva*, 2009.

[9] R. A. Española. Diccionario panhispánico de dudas, 1era edición 2da tirada. *http://buscon.rae.es/dpdl/SrvltGUIBusDPD?lema=voseo*, Accedida el 5/10/2011.

[10] M. Flanagan and S. Mcclure. Systran and the reinvention of mt. *IDC*, 2002.

[11] M. L. Forcada, B. I. Bonev, S. O. Rojas, J. A. P. Ortiz, G. R. Sánchez, F. S. Martínez, C. Armentano-Oller, M. A. Montava, and F. M. Tyers. Documentation of the open-source shallow-transfer machine translation platform apertium. *Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant*, 2010.

[12] W. N. Francis and H. Kucera. *Brown Corpus Manual*. Brown University, 1979.

[13] W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. *ACL '91 Proceedings of the 29th annual meeting on Association for Computational Linguistics*, 1991.

[14] M. Ginestí-Rosell, G. Ramírez-Sánchez, S. Ortiz-Rojas, F. M. Tyers, and M. L. Forcada. Development of a free basque to spanish machine translation system. *Procesamiento de Lenguaje Natural*, 2009.

[15] R. González. Mi querida elle. *http://www.babab.com/no09/elle.htm*, Accedida el 02/08/2011.

[16] M. Kapovic. Fórmulas de tratamiento en dialectos de español, fenómenos de voseo y ustedeo. *HIERONYMUS I*, 2007.

[17] P. Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[18] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. *Proceedings of the ACL 2007 Demo and Poster Sessions*, page pages 177–180, 2007.

[19] M. L.Forcada. Apertium: free/open-source rule-based machine translation. *Presentation at Fourth Machine Translation Marathon "Open Source Tools for Machine Translation"*, 2010.

[20] F. Masselot, P. Ribiczey, and G. Ramírez-Sánchez. Using the apertium spanish-brazilian portuguese machine translation system for localization. *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, 2010.

[21] A. K. Melby. *The Possibility of Language: A Discussion of the Nature of Language, With Implications for Human and Machine Translation*. The Possibility of Language. A discussion of the nature of language with implications for human and machine translation, 1995.

[22] M. Mohammadi and N. GhasemAghaee. Building bilingual parallel corpora based on wikipedia. *Proceedings of the 2010 Second International Conference on Computer Engineering and Applications*, 2010.

[23] G. News. Google translator api is now avilable as a paid service. *http://code.google.com/intl/es-ES/apis/language/translate/overview.html*, Accedida el 10/09/2011.

[24] S. Nirenburg, H. L. Somers, and Y. A. Wilks. *Readings in Machine Translation*. The MIT Press, 2002.

[25] NIST. Mt08 scoring scripts. *http://www.itl.nist.gov/iad/mig//tests/mt/2008/scoring.html*, Accedida el 23/09/2011.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. *40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.*, 2002.

[27] A. Press. *W.John Hutchins and Harold L. Somers*. An introduction to machine translation, 1992.

[28] Z. Sheikh and F. Sánchez-Martínez. A trigram part-of-speech tagger for the apertium free/open-source machine translation platform. *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, 2009.

[29] A. O. W. Site. *http://anusaaraka.iiit.ac.in*, Accedida el 1/08/2011.

[30] G. W. Site. About geonames. *http://www.geonames.org/about.html*, Accedido el 23/09/2011.

[31] V. M. Sánchez-Cartagena and J. A. Pérez-Ortiz. Scalemt: a free/open source framework for building scalable machine translation web services. *The Prague Bulletin of Mathematical Linguistics 93*, 2010.

[32] F. Sánchez-Martínez, M. Forcada, and A. Way. Hybrid rule-based example-based mt: Feeding apertium with sub-sentential translation units. *Proceedings of the 3rd Workshop on Example-Based Machine Translation*, pages p. 11–18, 2009.

[33] Tiedemann. Improved sentence alignment for movie subtitles. *In Proceedings of RANLP 2007, Borovets, Bulgaria,*, pages 582–588, 2007.

[34] F. M. Tyers. Rule-based breton to french machine translation. *Proceedings of the 14th Annual Conference of EAMT*, 2010.

[35] F. M. Tyers and J. A. Pienaar. Extracting bilingual word pairs from wikipedia. *Proceedings of the SALTMIL Workshop at Language Resources and Evaluation Conference.*, LREC08:19–22, 2008.

9 Anexos

Anexo 1 - Conteo de voseos y entidades

Además de evaluar las modificaciones realizadas sobre Apertium utilizando las métricas BLEU y NIST, se definió una métrica diferente para poder tener una idea más clara del impacto de la mejora sobre el uso del voseo y de las entidades con nombre.

Ya que las modificaciones fueron realizadas en el diccionario del Español, se tomaron todos los textos utilizados en la evaluación y se ejecutaron los dos primeros módulos de Apertium, el analizador morfológico y el Pos-Tagger (etiquetador léxico), a fin de tener el texto etiquetado por Apertium en su versión tradicional y en la versión con las mejoras realizadas. De esta forma se hace un corte transversal en el proceso de traducción para poder entender mejor como impactaron las mejoras realizadas. A través de las etiquetas generadas se pueden contar la cantidad de verbos reconocidos, la cantidad de entidades con nombre y la cantidad de pronombres, siendo estas las categorías afectadas por las modificaciones.

Categorías	Apertium RP	Apertium	Aper.RP/Aper.Trad
Flexiones verbales	22656	20077	1.128
Verbo Ser	3224	3031	1.063
Pronombres fuertes	9874	8564	1.152
Entidades con nombre	1458	1396	1.044
Localidades	223	158	1.411

Como resultado se tiene una mejora importante en la traducción de verbos (aprox. 13%). Aunque analizando con detenimiento se puede observar que se debe a los verbos en modo voseante que no eran soportados por Apertium en su versión original. Sin embargo es de esperar que en otro tipo de corpus la mejora no sea de esta magnitud ya que los subtítulos presentan una gran cantidad de verbos en modo voseante debido a la presencia de diálogos.

Se desglosa además el reconocimiento del verbo ser para medir el impacto de la flexión irregular *sos*. Aunque la mejora en este caso es de un 6.3%

También se ve una mejora en el reconocimiento de pronombres debido a la presencia del pronombre ‘vos’ que también paso a ser traducida por Apertium luego de las modificaciones. Logrando en este caso una mejora de un 15.2%

Sobre las entidades con nombre, se puede observar una leve mejora de 4.4%, sin embargo si se analiza exclusivamente las localidades la mejora es de un 44.1%. El caso más interesante es el de las películas uruguayas en donde el reconocimiento de localidades mejoró un 277% y el global de las entidades con nombre un 23%.

Anexo 2 - Paradigmas modificados

A continuación se detallan los paradigmas verbales modificados con las terminaciones definidas para reconocer el voseo en modo indicativo, imperativo y el imperativo al contraerse en forma pronominal (p.e *cantámelo*).

Paradigma	Indic.	Imper.	imper. Cont.
/oler__vblex	olés	olé	ole
descri/bir__vblex	bís	bí	bi
apr/obar__vblex	obás	obá	oba
d/ar__vblex	as	a	
corr/egir__vblex	egís	egí	egi
v/estirse__vblex	estís	estí	esti
abat/irse__vblex	ís	í	i
dem/ostrar__vblex	ostrás	ostrá	ostra
qu/erer__vbmod	erés	eré	ere
sal/ir__vblex	ís	í	i
prohib/ir__vblex	ís	í	i
o/ír__vblex	ís	í	i
/ser__vbser	sos		
conm/over__vblex	ovés	ové	ove
deb/er__vblex	és	é	e
distíng/uir__vblex	uís	uí	
atribu/ir__vblex	ís	í	i
abandón/ar__vblex	ás	á	
conv/enir__vblex	enís	ení	eni
averg/onzar__vblex	onzás	onzá	onza
atá/car__vblex	cás	cá	
d/ormir__vblex	ormís	ormí	ormi
qu/ebrar__vblex	ebrás	ebrá	ebra
sonr/eír__vblex	eís	eí	ei
t/ener__vblex	enés	ené	ene
dirí/gir__vblex	gís	gí	
and/ar__vblex	ás	á	a
le/er__vblex	és	é	e
emp/ezar__vblex	ezás	ezá	eza
satisf/acer__vblex	acés	acé	ace
averígü/ar__vblex	ás	á	
ret/orcer__vblex	orcés	orcé	orce
vend/er__vblex	és	é	e

Paradigma	Indic.	Imper.	imper. Cont.
p/oblar__vblex	oblás	oblá	obla
ampli/ar__vblex	ás	á	a
abát/ir__vblex	ís	í	
abandon/ar__vblex	ás	á	a
abalan/zar__vblex	zás	zá	za
/erguir__vblex	erguís	erguí	ergui
est/ar__vblex	ás	á	a
parí/r__vblex	ís	í	
atra/er__vblex	és	é	e
anoche/cer__vblex	cés	cé	ce
agradé/cer__vblex	cés	cé	
gob/ernar__vblex	ernás	erná	erna
ll/over__vblex	ovés	ové	ove
p/ensar__vblex	ensás	ensá	ensa
r/odar__vblex	odás	odá	oda
despl/egar__vblex	egás	egá	ega
conf/esar__vblex	esás	esá	esa
manif/estar__vblex	estás	está	esta
agrade/cer__vblex	cés	cé	ce
v/olcar__vblex	olcás	olcá	olca
h/acer__vblex	acés	acé	ace
aco/ger__vblex	gés	gé	ge
evacu/ar__vblex	ás	á	a
frun/cir__vblex	cís	cí	ci
condú/cir__vblex	cís	cí	
m/orir__vblex	orís	orí	ori
com/enzar__vblex	enzás	enzá	enza
contrad/ecir__vblex	ecís	ecí	eci
ró/mper__vblex	mpés	mpé	
equival/er__vblex	és	é	e
ac/ordar__vblex	ordás	ordá	orda
m/order__vblex	ordés	ordé	orde
adv/ertir__vblex	ertís	ertí	erti

v/enir__vblex	enís	ení	eni
v/erter__vblex	ertés	erté	erte
abalán/zar__vblex	zás	zá	
sitú/ar__vblex	ás	á	
r/ogar__vblex	ogás	ogá	oga
conven/cer__vblex	és	é	e
conc/ebir__vblex	ebís	ebí	ebi
cons/eguir__vblex	eguís	eguí	egui
ro/mper__vblex	mpés	mpé	mpe
ac/ostar__vblex	ostás	ostá	osta
alég/ar__vblex	ás	á	
s/ervir__vblex	ervís	erví	ervi
res/onar__vblex	onás	oná	ona
m/entir__vblex	entís	entí	enti
alí/ar__vblex	ás	á	
asc/ender__vblex	endés	endé	ende
dev/olver__vblex	olvés	olvé	olve
p/oner__vblex	onés	oné	one
averigu/ar__vblex	ás	á	a
p/oder__vbmod	odés	odé	ode
pref/erir__vblex	erís	erí	eri
ó/ír__vblex	ís	í	
situ/ar__vblex	ás	á	a
equivál/er__vblex	és	é	
comp/oner__vblex	onés	oné	one
adqui/rir__vblex	rís	rí	ri
ac/ertar__vblex	ertás	ertá	erta
ata/car__vblex	cás	cá	ca
/ir__vblex	vas	andá	anda
/caber__vblex	cabés	cabé	cabe
p/erder__vblex	erdés	erdé	erde
enc/ontrar__vblex	ontrás	ontrá	ontra
as/ir__vblex	ís	í	i
ren/ovar__vblex	ovás	ová	ova
s/aber__vblex	abés	abé	abe
c/eñir__vblex	eñís	eñí	eni
bend/ecir__vblex	ecís	esí	esi
averigü/ar__vblex	ás	á	a
averígu/ar__vblex	ás	á	
atribú/ir__vblex	ís	í	
abat/ir__vblex	ís	í	i

deb/er__vbmod	és	é	e
v/olar__vblex	olás	olá	ola
disting/uir__vblex	uís	uí	ui
abst/ener__vblex	enés	ené	ene
sál/ir__vblex	ís	í	
áb/rir__vblex	rís	rí	
comp/etir__vblex	etís	etí	eti
ju/gar__vblex	gás	gá	ga
par/ir__vblex	ís	í	i
prev/er__vblex	eés	eé	ee
v/estir__vblex	estís	estí	esti
t/emblar__vblex	emblás	emblá	embla
conc/ernir__vblex	ernís	erní	erni
atrá/er__vblex	és	é	
ab/rir__vblex	rís	rí	ri
desp/edir__vblex	edís	edí	edi
acó/ger__vblex	gés	gé	
descrí/bir__vblex	bís	bí	
arr/endar__vblex	endás	endá	enda
apr/etar__vblex	etás	etá	eta
c/ocer__vblex	océs	océ	oce
ás/ir__vblex	ís	í	
diri/gir__vblex	gís	gí	gi
d/ecir__vblex	ecís	ecí	eci
c/ontar__vblex	ontás	ontá	onta
qu/erer__vblex	erés	eré	ere
prohíb/ir__vblex	ís	í	
convén/cer__vblex	cés	cé	
reun/ir__vblex	ís	í	i
rev/entar__vblex	entás	entá	enta
s/oñar__vblex	oñás	oñá	oña
alm/orzar__vblex	orzás	orzá	orza
aleg/ar__vblex	ás	á	a
condu/cir__vblex	cís	cí	ci
vénd/er__vblex	és	é	
frún/cir__vblex	cís	cí	
t/ener__vbmod	enés	ené	ene
c/olgar__vblex	olgás	olgá	olga
ca/er__vblex	és	é	e
c/errar__vblex	errás	erá	erra

Anexo 3 - Entidades incluidas

A continuación se detallan las entidades con nombre extraídas desde la base de Geonames junto con el paradigma definido para cada una de ellas.

Entidad	Paradigma
Young	Afganistán__np
Villa Sara	Barcelona__np
Villa del Carmen	Barcelona__np
Vichadero	Afganistán__np
Vergara	Barcelona__np
Velázquez	Afganistán__np
Veinticinco de Mayo	Afganistán__np
Veinticinco de Agosto	Afganistán__np
Tupambaé	Afganistán__np
Trinidad	Barcelona__np
Departamento de Treinta y Tres	Afganistán__np
Treinta y Tres	Afganistán__np
Tranqueras	Barcelona__np
Tomás Gomensoro	Afganistán__np
Toledo	Afganistán__np
Tarariras	Barcelona__np
Tala	Barcelona__np
Departamento de Tacuarembó	Afganistán__np
Tacuarembó	Afganistán__np
Departamento de Soriano	Afganistán__np
Soriano	Afganistán__np
Solís de Mataojo	Afganistán__np
Soca	Afganistán__np
Sauce	Afganistán__np
Sarandí Grande	Afganistán__np
Sarandí del Yi	Afganistán__np
Santiago Vázquez	Afganistán__np
Santa Rosa	Barcelona__np
Santa Lucía	Barcelona__np
Santa Clara de Olimar	Barcelona__np
Santa Catalina	Barcelona__np

Entidad	Paradigma
Migues	Afganistán__np
Mercedes	Barcelona__np
Melo	Afganistán__np
Mariscala	Barcelona__np
Departamento de Maldonado	Afganistán__np
Maldonado	Afganistán__np
Los Cerrillos	Afganistán__np
Libertad	Barcelona__np
Departamento de Lavalleja	Afganistán__np
Las Toscas	Barcelona__np
Las Piedras	Barcelona__np
Lascano	Afganistán__np
La Paz	Barcelona__np
La Paloma	Barcelona__np
La Floresta	Barcelona__np
Juan L. Lacaze	Afganistán__np
José Pedro Varela	Afganistán__np
José Enrique Rodó	Afganistán__np
José Batlle y Ordóñez	Afganistán__np
Joaquín Suárez	Afganistán__np
Joanicó	Afganistán__np
Isidoro Noblía	Afganistán__np
Guichón	Afganistán__np
Fray Bentos	Afganistán__np
Departamento de Florida	Afganistán__np
Florida	Barcelona__np
Departamento de Flores	Afganistán__np
Florencio Sánchez	Afganistán__np
Empalme Olmos	Afganistán__np
Ecilda Paullier	Barcelona__np
Departamento de	Afganistán__np

Santa Bernardina	Barcelona__np
San Ramón	Afganistán__np
San José de Mayo	Afganistán__np
Departamento de San José	Afganistán__np
San Javier	Afganistán__np
San Jacinto	Afganistán__np
San Félix	Afganistán__np
San Carlos	Afganistán__np
San Bautista	Afganistán__np
San Antonio	Afganistán__np
Departamento de Salto	Afganistán__np
Salto	Afganistán__np
Rosario	Barcelona__np
Rodríguez	Afganistán__np
Departamento de Rocha	Afganistán__np
Rocha	Barcelona__np
Departamento de Rivera	Afganistán__np
Rivera	Afganistán__np
Departamento de Río Negro	Afganistán__np
Río Branco	Afganistán__np
Rafael Perazza	Afganistán__np
Quebracho	Afganistán__np
Punta del Este	Afganistán__np
Progreso	Afganistán__np
Porvenir	Afganistán__np
Piriápolis	Afganistán__np
Piedras Coloradas	Afganistán__np
Departamento de Paysandú	Afganistán__np
Paysandú	Afganistán__np
Paso de los Toros	Afganistán__np
Paso de Carrasco	Afganistán__np
Pando	Afganistán__np
Pan de Azúcar	Afganistán__np
Palmitas	Barcelona__np

Durazno	
Durazno	Afganistán__np
Dolores	Afganistán__np
Dieciocho de Julio	Afganistán__np
Delta del Tigre	Afganistán__np
Curtina	Afganistán__np
Constitución	Afganistán__np
Colonia del Sacramento	Afganistán__np
Colonia	Barcelona__np
Departamento de Colonia	Afganistán__np
Chuy	Afganistán__np
Departamento de Cerro Largo	Afganistán__np
Cerro Colorado	Afganistán__np
Cebollatí	Afganistán__np
Casupá	Barcelona__np
Castillos	Afganistán__np
Carmelo	Afganistán__np
Carlos Reyles	Afganistán__np
Cardona	Barcelona__np
Cardal	Afganistán__np
Departamento de Canelones	Afganistán__np
Canelones	Afganistán__np
Blanquillo	Afganistán__np
Bella Unión	Barcelona__np
Belén	Barcelona__np
Baltasar Brum	Afganistán__np
Atlántida	Barcelona__np
Departamento de Artigas	Afganistán__np
Artigas	Afganistán__np
Aiguá	Barcelona__np
Aguas Corrientes	Barcelona__np
Aceguá	Barcelona__np
Puntas de Valdez	Afganistán__np
Colonia Nicolich	Afganistán__np
Barra de Carrasco	Afganistán__np

Ombúes de Lavalle	Afganistán__np
Nuevo Berlín	Afganistán__np
Nueva Palmira	Afganistán__np
Nueva Helvecia	Afganistán__np
Departamento de Montevideo	Afganistán__np
Montevideo	Afganistán__np
Montes	Afganistán__np
Minas de Corrales	Barcelona__np
Minas	Barcelona__np

Pajas Blancas	Barcelona__np
Punta del Diablo	Afganistán__np

Películas seleccionadas

A continuación se detallan las películas que fueron seleccionadas para evaluar las modificaciones realizadas en Apertium.

Nombre	Año
El baño del Papa	2007
El hijo de la novia	2001
El sueño de Valentín	2002
Esperando la carroza	1985
Felicidades	2000
La historia oficial	1985
La noche de los lapices	1986
La suerte está echada	2005
Lluvia	2008
Nueve Reinas	2000
Un cuento chino	2011
Luna de Avellaneda	2004
Cuestion de principios	2009

Nombre	Año
Mentiras piadosas	2008
Martin (Hache)	1997
Camila	1984
Tierra del Fuego	2000
Viaje hacia el mar	2003
Whisky	2004
25 Watts	2001
Un lugar en el mundo	1992
Plata quemada	2000
Sol de otoño	1996
Crónica de una fuga	2006
Annita	2009
Tiempo de valientes	2005