Capítulo 3

3 CONMAT: La Estructura Experimental

En este capítulo se presenta la estructura bajo la cual se ejecutarán todos los experimentos llamada CONMAT. Aquí se explica porqué CONMAT es un dominio adecuado para experimentar técnicas con habla fuera del vocabulario. Además, se hace una clasificación de los diferentes tipos de habla para los sistemas de reconocimiento automático de habla (ASR) de diálogo dirigido. Al final del capitulo, se propone una metodología de evaluación que permite comparar el desempeño de los experimentos realizados en esta tesis.

3.1 El Corpus

La estructura experimental está basada en el contexto del sistema de conmutador automático instalado en la Universidad de las Américas-Puebla, llamado CONMAT. Este sistema tiene la tarea de atender llamadas telefónicas y transferirlas con base a la palabra clave reconocida. El sistema CONMAT esta constituido por los siguientes dominios: nombres de personas, nombres de departamentos, nombres de lugares dentro de la UDLA, números de extensiones y nombres de aplicaciones. Para simplificar los experimentos en esta tesis, se omiten los números de extensiones.

CONMAT fue elegido por tres razones. Primero, el sistema ya ha sido desarrollado y está funcionando realmente para todo el campus de la UDLA. Segundo, gran parte de las llamadas telefónicas que recibe CONMAT contienen palabras clave combinadas con habla y sonidos fuera del vocabulario, las cuales son de particular interés para esta tesis. Finalmente, existen muchos datos disponibles para entrenamiento y prueba.

El corpus de desarrollo consta de 553 frases, las cuales son respuestas que usuarios reales dijeron a la pregunta "*Por favor diga el nombre de la persona a la que desea llamar*". La población de locutores por género es de 42% de Mujeres y 58% de Hombres, y la población de locutores por origen es de 94% de Mexicanos y 6% de extranjeros. La tabla 3-1 ilustra una clasificación general las frases del corpus. La determinación del tamaño de este corpus fue motivada por la reducción en el tiempo de cada experimento.

FRASE	PORCENTAJE
Palabras clave + Lenguaje natural*	21.2%
Solo palabras clave	78.8%

Tabla 3-1 Tipos generales de frases en el corpus de la estructura experimental.

La siguiente sección describe una clasificación detallada de los tipos de frases que se pueden presentar en los sistemas de diálogo dirigido, considerando tanto un sistema de solo vocabulario y un sistema que incorpora una gramática de lenguaje natural.

3.2 Clasificación de Tipos de Habla

Los sistemas de reconocimiento de habla usan dos tipos de entrada que controlan lo que puede ser reconocido: El vocabulario y la gramática. El vocabulario especifica el conjunto de posibles palabras y la gramática especifica la secuencia de palabras permitidas.

Para hacer un análisis del desempeño de reconocimiento con el enfoque de habla fuera del vocabulario, se propone una clasificación de los diferentes tipos de habla que se pueden presentar en un sistema real, ya que un sistema de reconocimiento puede usar un vocabulario con o sin gramática. La figura 3-1 muestra un árbol que clasifica los tipos de habla para aplicaciones que usan solo vocabulario, que sirve para examinar el desempeño de las aplicaciones. El término *analizable* significa que las frases solo tienen palabras clave y el término *no analizable* significa que las frases pueden tener habla fuera del vocabulario que pueden incluir palabras clave.

*El término lenguaje natural se refiere a que el reconocedor acepta frases con palabras no clave definidas a través de una gramática.

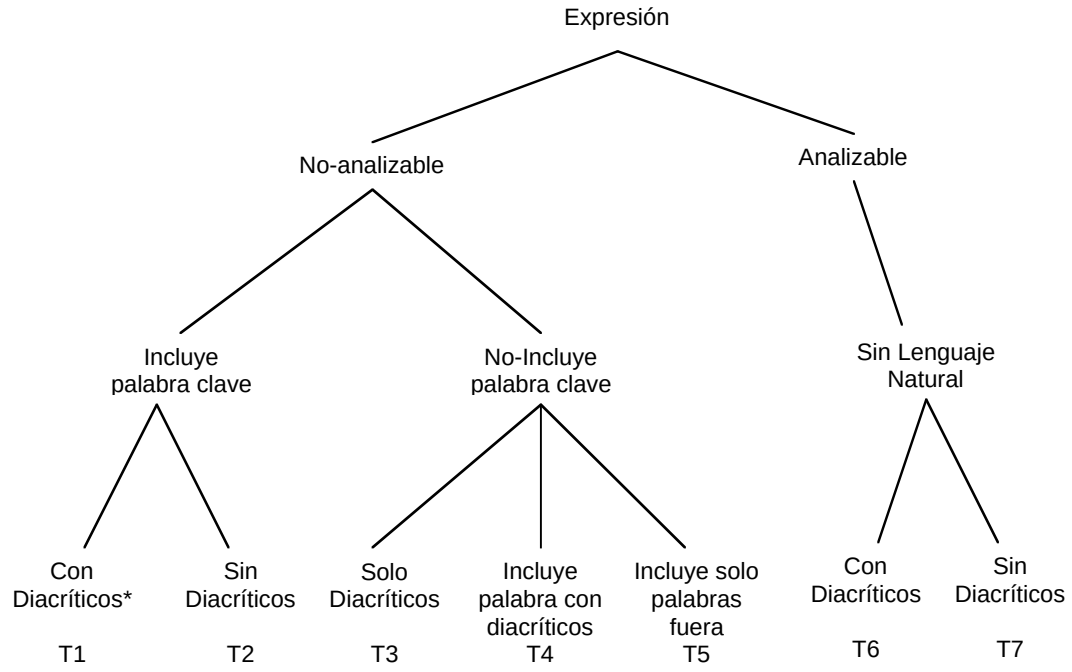

Figura 3-1 Clasificación de tipos de habla para aplicaciones con solo vocabulario (sin gramática).

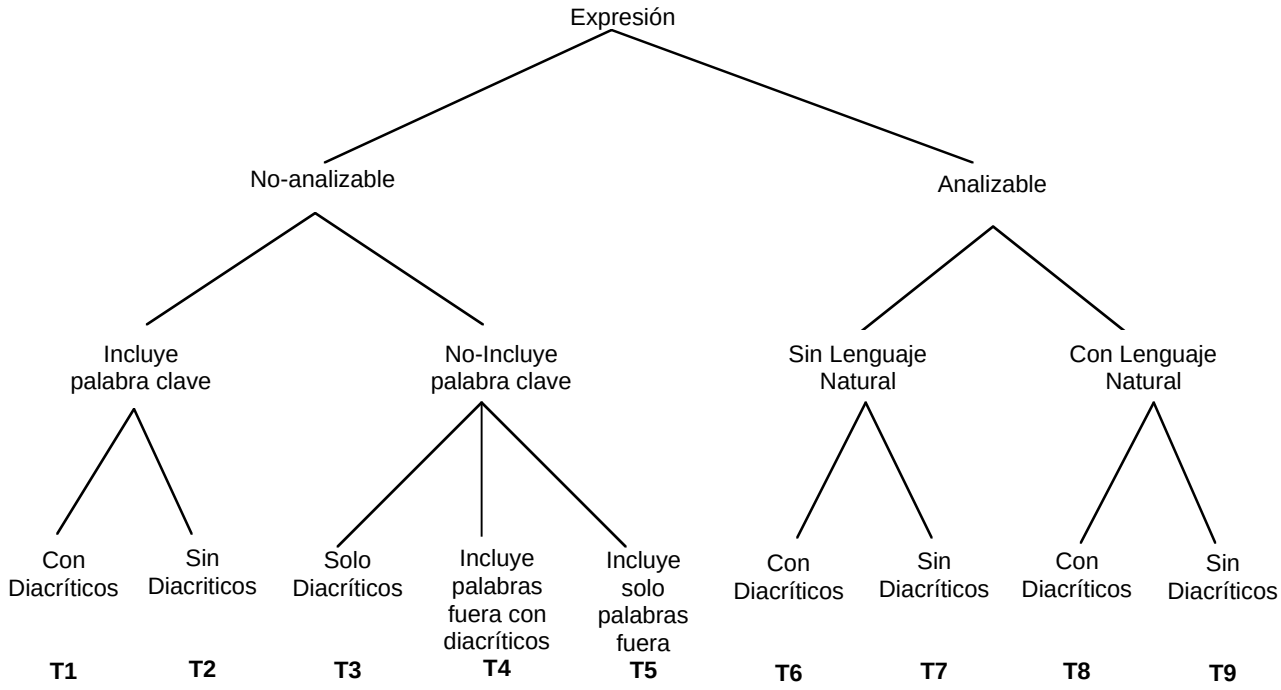

Figura 3-2 Clasificación de tipos de habla para aplicaciones con gramática.

*Los diacríticos representan algún tipo de ruido, ya sea humano o no-humano.

En la figura 3-2 se muestra un árbol que clasifica los tipos de habla para aplicaciones que usan gramática, donde analizable o no analizable depende del parser del reconocedor. Debido a que la problemática de esta tesis se enfoca a tratar el problema del habla fuera del vocabulario, los nodos terminales t_i se presentan con y sin diacríticos (ver tabla 3-2). El número de frases en el corpus de la estructura experimental de acuerdo a la clasificación de los tipos de frases para aplicaciones con y sin gramática, se muestra en la tabla 3-3 (observe que el uso de una gramática permite el análisis de 27 frases en otras categorías).

CÓDIGO	DESCRIPCIÓN
[tc]	Ruidos provocados cuando la persona afina su voz antes de hablar
[c]	Tos
[um, uh, mm]	Expresiones como “mmmh”, “ah”, “oh”, etc.
[l]	Risas
[r]	Ruidos de respiración
[ls]	Ruidos provocados por los labios
[q]	Referente al diálogo, pero no se entiende lo que quiso decir el usuario

Tabla 3-2 Diacríticos considerados en el corpus de la estructura experimental.

SOLO VOCABULARIO		CON GRAMATICA	
NODO TERMINAL	# FRASES	NODO TERMINAL	# FRASES
T1	3	T1	2
T2	37	T2	11
T3	65	T3	65
T4	6	T4	6
T5	33	T5	33
T6	10	T6	10
T7	399	T7	399
		T8	1
		T9	26

Tabla 3-3 Clasificación de corpus de datos con tipos habla para aplicaciones con solo vocabulario y con gramática.

Debido a que en esta tesis de hacen varios experimentos que modelan habla fuera del vocabulario, se requiere de un método que calcule el desempeño de los experimentos y a la vez evalúe el costo computacional que representan. En la siguiente sección se explica un método para evaluar el desempeño de un reconocedor dentro de un dominio específico.

3.3 Metodología de Evaluación

En la sección 2.8 se explicaron las medidas de desempeño del reconocedor dentro de un dominio específico. Cuando se obtienen resultados de varias tareas de reconocimiento, se torna complejo observar el desempeño de todas esas tareas y hacer una comparación entre estas. En esta sección se propone un método que mide el desempeño y costo computacional del reconocedor dentro de un dominio específico, de tal manera que se reduzcan varias medidas en un solo número y cuando se quiera comparar varios experimentos, sea fácil observar su desempeño.

3.3.1 Desempeño de Reconocimiento

En los últimos años se han desarrollado varios criterios para evaluar sistemas de reconocimiento de habla automática [Villarubia y Acero, 1993] y sistemas de entendimiento de lenguaje natural [Polifroni et al, 1998]. En [Villarubia y Acero, 1993], se presenta un criterio de evaluación de un reconocedor de palabras dentro de contexto y sugieren que un criterio de evaluación debería ser dependiente del tipo de aplicación.

La tasa de falsas alarmas es medida por muchos investigadores como falsas alarmas por palabra del vocabulario por hora (*fa/kw/hr*) [Manos, 1996; Lau, 1998]. Esta medida es apropiada para aplicaciones de vigilancia (por ejemplo identificar palabras clave en programas de radio), pero no es la más apropiada para aplicaciones de telecomunicaciones, tales como servicios de automatización de operadoras o sistemas basados en menús de opciones. En este último tipo de aplicaciones, los errores de sustitución y falsas alarmas conducen al sistema a realizar acciones incorrectas.

Para evaluar reconocedores de palabras aisladas, la figura de mérito básica es, sin lugar a dudas, la tasa de error. Tomando como referencia el criterio de evaluación desarrollado por [Villarubia y Acero, 1993] para reconocedores que incorporan mecanismos de rechazo, en esta tesis se propone evaluar el desempeño de reconocimiento a través de una figura de mérito (*FOM*). La figura de mérito representa el complemento de una función de costo *C* que pondera las tasas de error del reconocedor y está definida por la siguiente ecuación:

$$FOM = 1 - C \qquad (3.1)$$

Donde *C* es la función que pondera las tasas de error del reconocedor:

$$C = L_{cc\_in}T_{cc\_in} + L_{fa\_in}T_{fa\_in} + L_{fr\_in}T_{fr\_in} + L_{fc\_in}T_{fc\_in} + L_{fa\_out}T_{fa\_out} + L_{fc\_out}T_{fc\_out} \qquad (3.2)$$

con
$$L_{cc\_in} + L_{fa\_in} + L_{fr\_in} + L_{fc\_in} + L_{fa\_out} + L_{fc\_out} = 1 \qquad (3.3)$$

donde

$$\begin{aligned} L_{cc\_in} &= h\ P_k\ C_{cc\_in} \\ L_{fa\_in} &= h\ P_k\ C_{fa\_in} \\ L_{fr\_in} &= h\ P_k\ C_{fr\_in} \\ L_{fc\_in} &= h\ P_k\ C_{fc\_in} \\ L_{fa\_out} &= h\ (1-P_k)\ C_{fa\_out} \\ L_{fc\_in} &= h\ (1-P_k)\ C_{fc\_in} \end{aligned} \qquad (3.4)$$

En la ecuación 3.2, T_i representa cada medida de error y L_i representa los parámetros que ponderan las tasas de error T_i. Los parámetros L_i deben cumplir con la ecuación 3.3 para normalizar el costo entre 0 y 1. En las ecuaciones 3.4, C_i representa la penalización de la *i-esima* medida de desempeño, dada por el diseñador de la aplicación, donde, P_k es la probabilidad *a priori* de que una palabra este dentro del vocabulario, y h es una constante de normalización definida por la ecuación 3.5 para que se cumpla $L_i = 1$.

$$h = \frac{1}{P_k[C_{cc\_in} + C_{fa\_in} + C_{fr\_in} + C_{fc\_in}] + (1 - P_k)[C_{fa\_out} + C_{fc\_out}]} \qquad (3.5)$$

El valor P_k se obtiene en este trabajo de un corpus con 5835 frases, de las cuales, un subconjunto es el corpus de entrenamiento (553 frases). Los valores de penalización C_i representan una prioridad para cada medida de desempeño y es dada por el diseñador de aplicaciones. Los valores de penalización y el valor P_k definidos para los experimentos de esta tesis son los siguientes:

$P_k = 0.725$

$C_{cc\_in} = 8.3$ $\quad$ $C_{fa\_out} = 95$

$C_{fa\_in} = 41.3$ $\quad$ $C_{fc\_out} = 5$

$C_{fc\_in} = 21.4$

$C_{fr\_in} = 29$

3.3.2 Costo Computacional

Otra medida de desempeño usada para evaluar al reconocedor en un determinado contexto, es el costo computacional por cada segundo de habla, y está definido por la ecuación 3.6.

$$Costo\ \ por\ segundo\ de\ habla = \frac{Tiempo\ total\ de\ reconocimiento}{Tiempo\ total\ de\ habla} \qquad (3.6)$$

Las evaluaciones de desempeño y costo computacional de esta tesis, fueron ejecutadas en una computadora Dell con un procesador Pentium II de 450 Mhz y 128 MB de RAM.

3.4 Resumen

En este capítulo se describe el corpus de desarrollo que será usado por las técnicas a experimentar en esta tesis. Se hace una clasificación de los tipos de habla tanto para aplicaciones con solo vocabulario como para aplicaciones con gramática. Esta clasificación es realizada con el fin de analizar el desempeño de los diferentes tipos de habla. También, en este capítulo se propone una metodología de evaluación que consiste en una medida de desempeño y una medida que representa el costo computacional. La medida de desempeño es una figura de mérito (FOM) que indica el desempeño de reconocimiento, y el costo computacional es calculado para obtener el costo por segundo de habla.

El siguiente capítulo explica la técnica de niveles de confianza y evalúa los experimentos base (solo vocabulario y gramática predefinida)