TEXTLYNX PLUS

Ampliación de funciones para un analizador avanzado de textos

AUTOR: Víctor Annier Barrios Cañizares
TUTOR: Lic. Manuel Llanes Abeijón

Problema científico

- ¿Cómo aumentar la funcionalidad de ***TextLynx*** para obtener los vínculos entre textos mediante la comparación de corpora, el agrupamiento (clasificación de los textos) y la clasificación del léxico por sus propiedades distributivas en el corpus?

Objetivo General

- Aumentar la funcionalidad de ***TextLynx*** a través de la implementación de herramientas que permitan un análisis exhaustivo del texto

Tareas específicas

- Revisar la bibliografía actualizada relacionada con la lingüística de corpus, la minería de textos y el procesamiento del lenguaje natural.
- Estudiar el código de los programas ***TextLynx*** y ***CorpusMiner***.
- Definir las funciones que debe desarrollar cada herramienta.
- Probar e implementar cada herramienta

Referencia tecnológica

- Rational Suite 2003 Enterprise Edition.
- En cuanto al lenguaje se decidió usar el Object Pascal.
- Como herramienta de programación se utilizó el Delphi 6.

Herramientas diseñadas para satisfacer el objetivo general

- Corpus Classifier
- Corpus Reducer
- Corpus Codec
- Corpus Transformations

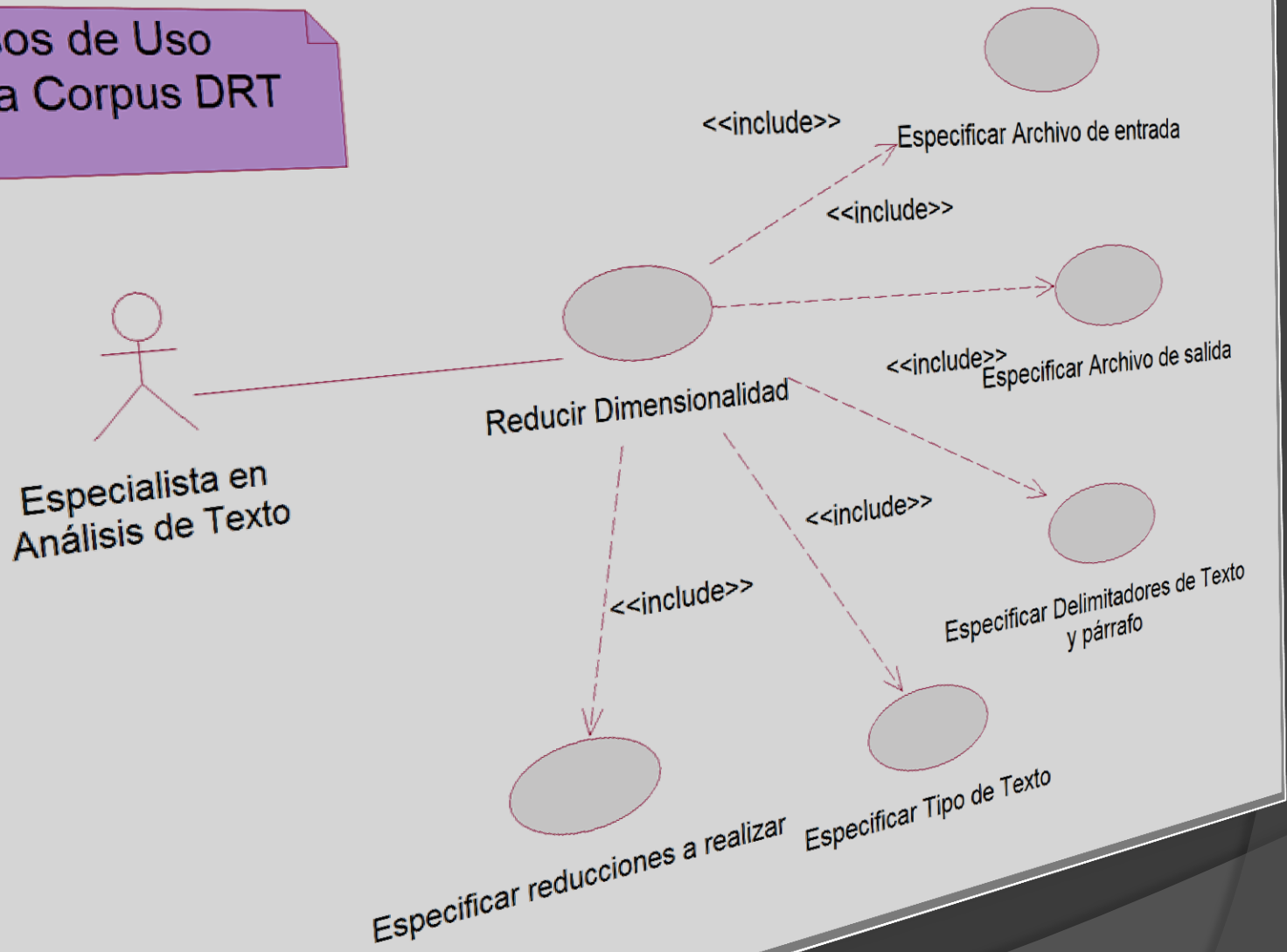
Corpus Classifier

- Clasifica el léxico del corpus
  - Frecuencia en el corpus
  - Rango en el corpus
  - Frecuencia clave en el corpus
  - Rango clave en el corpus
  - Posición textual
  - Posición oracional
  - Proximidad de los tókenes
  - Tamaño de la oración
- Compara dos corpus según la clasificación realizada.

Corpus Reducer

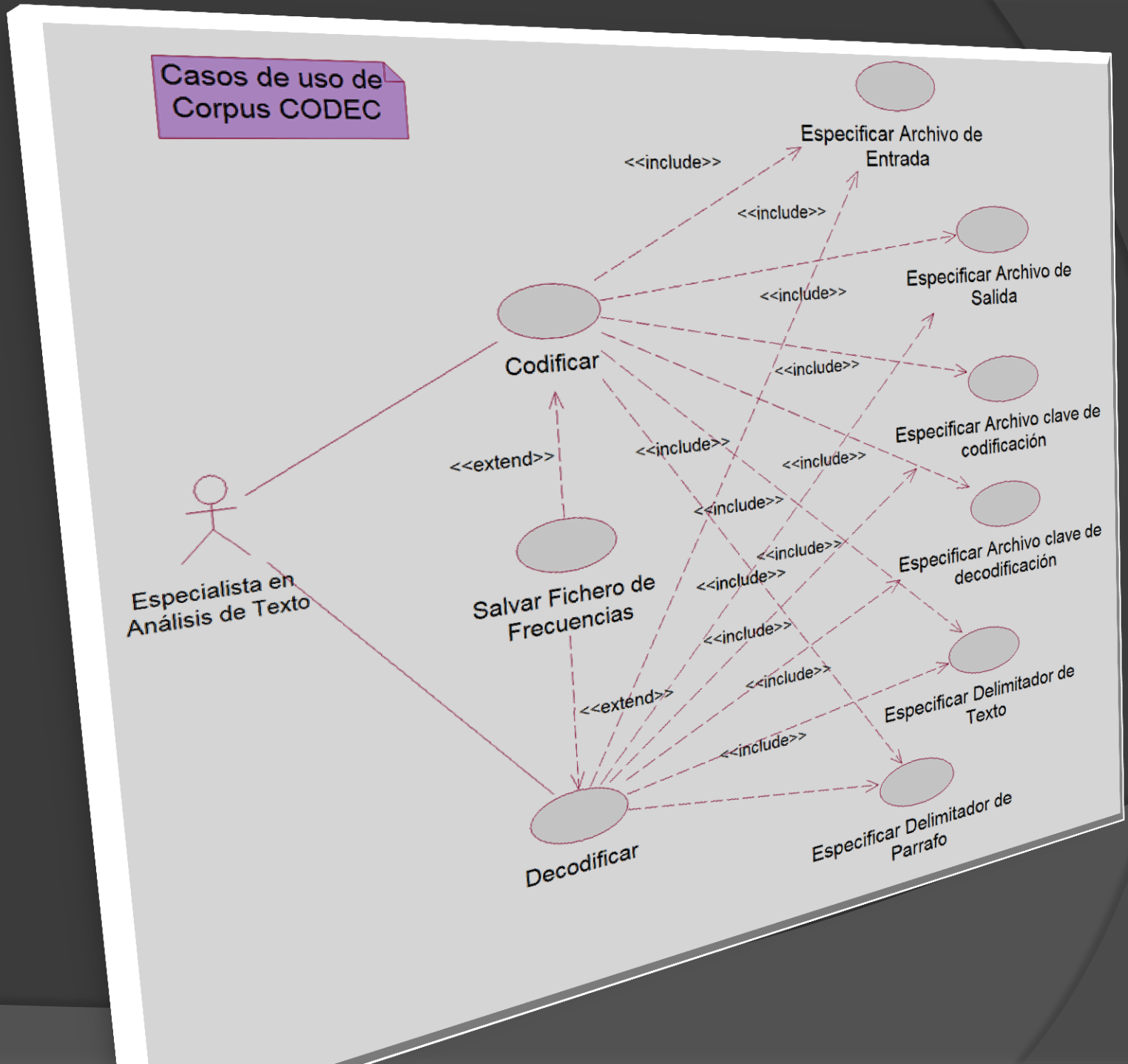
- Reduce las dimensiones de un corpus formado por lexico común en base a las siguientes tecnicas:
  - Degramatización
  - Reducción de sinónimos
  - Reducción de contracciones
  - Reducción por familia de palabras
  - Eliminación de palabras cuyo rango de frecuencia no se incluye en uno especificado.

Casos de Uso para Corpus DRT



Corpus Codec

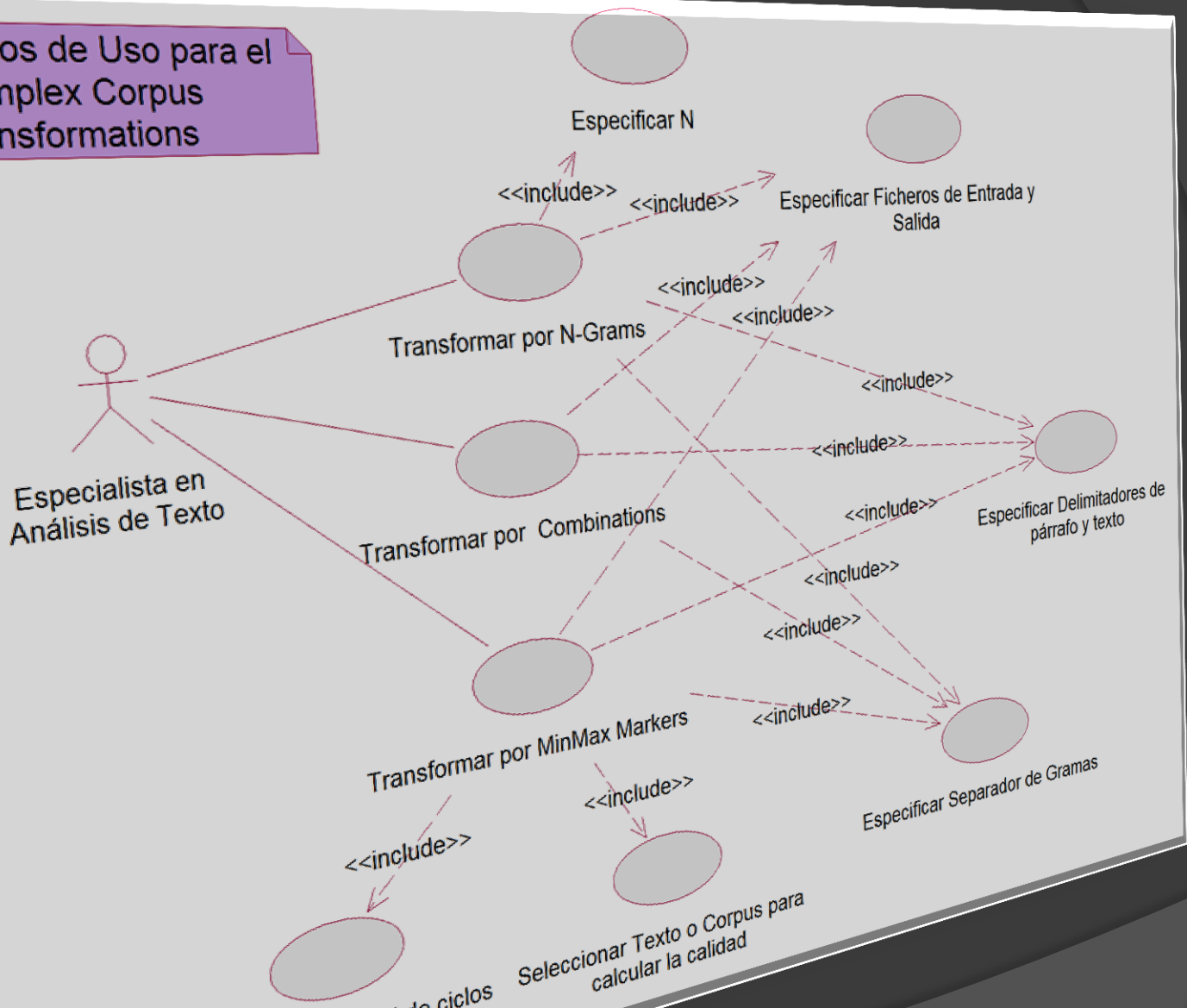
- Codifica un corpus para reducir su dimensión física y facilitar su procesamiento más rápido
- Decodifica el resultado obtenido después de aplicar un procesamiento


Casos de uso de Corpus CODEC
Especificar Archivo de Entrada
Especificar Archivo de Salida
Especificar Archivo clave de codificación
Especificar Archivo clave de decodificación
Especificar Delimitador de Texto
Especificar Delimitador de Parrafo
Codificar
Decodificar
Salvar Fichero de Frecuencias
Especialista en Análisis de Texto
<<include>>
<<extend>>

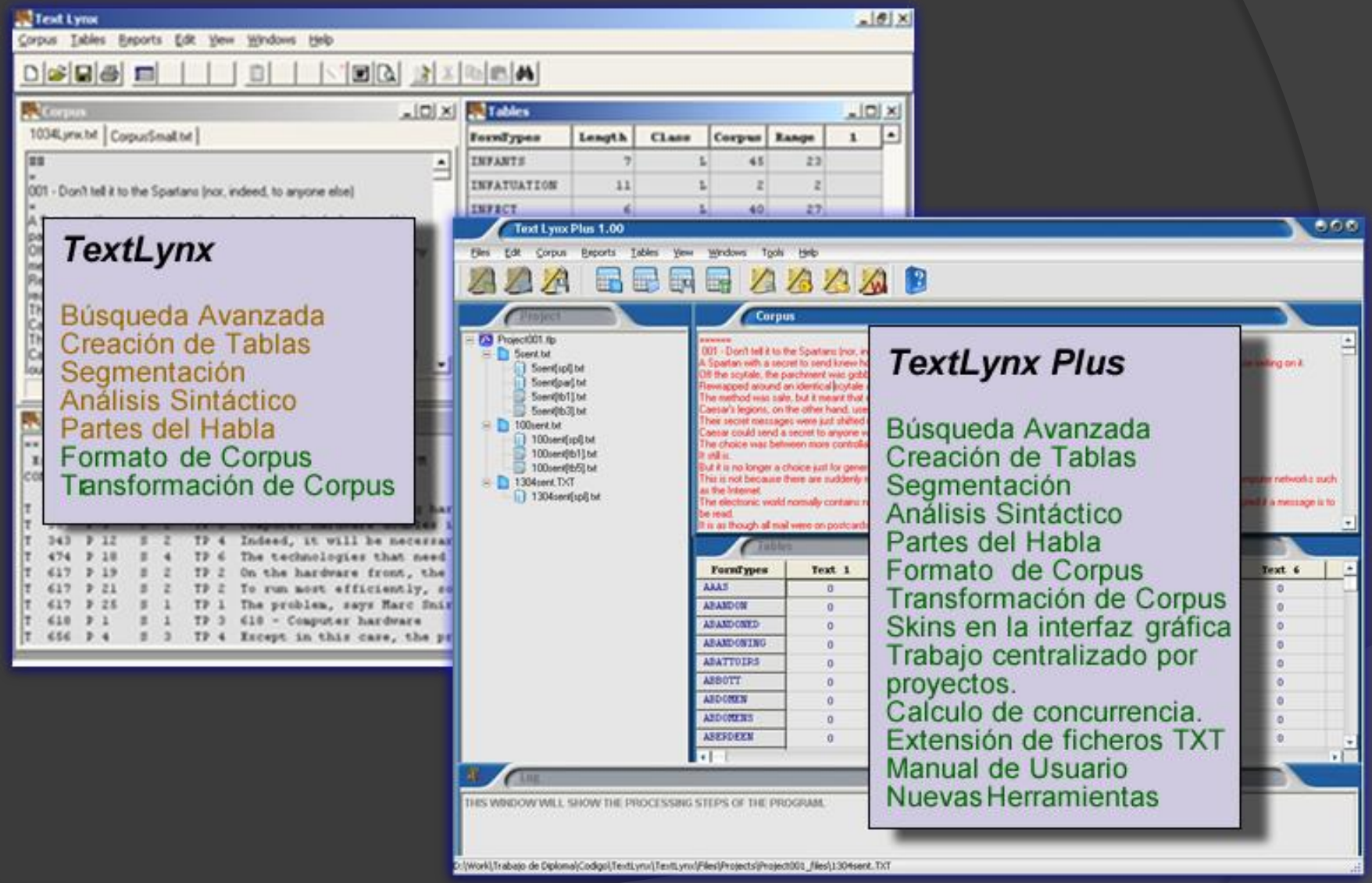
Corpus Transformations

- Transforma el léxico de un Corpus de forma tal que el léxico resultante incluya información léxico-combinatoria. Los algoritmos de transformación son:
  - NGrams
  - Combinations
  - MinMax Markers (Para texto y para Corpus)

Casos de Uso para el Complex Corpus Transformations



Resumen del trabajo realizado



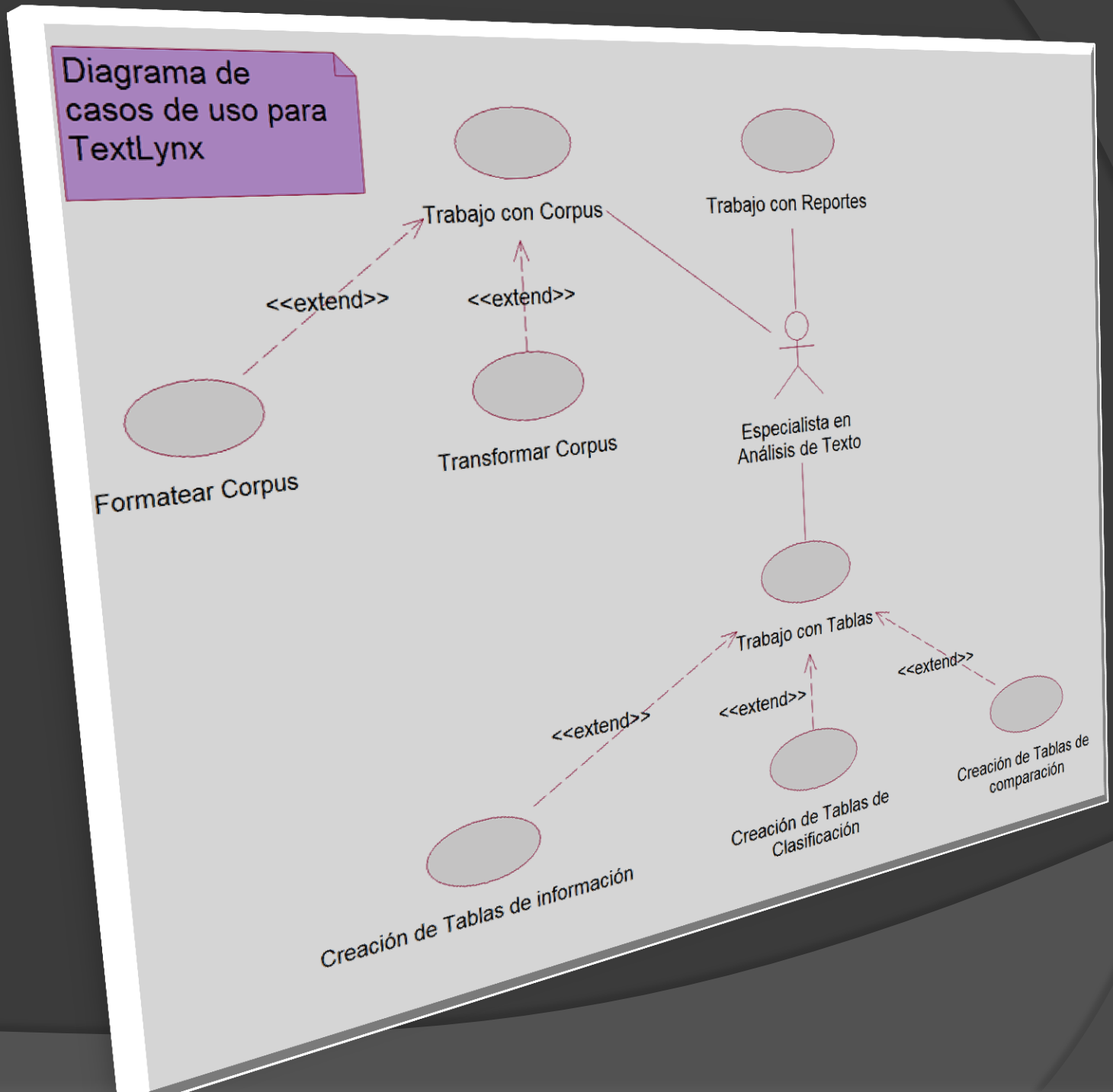
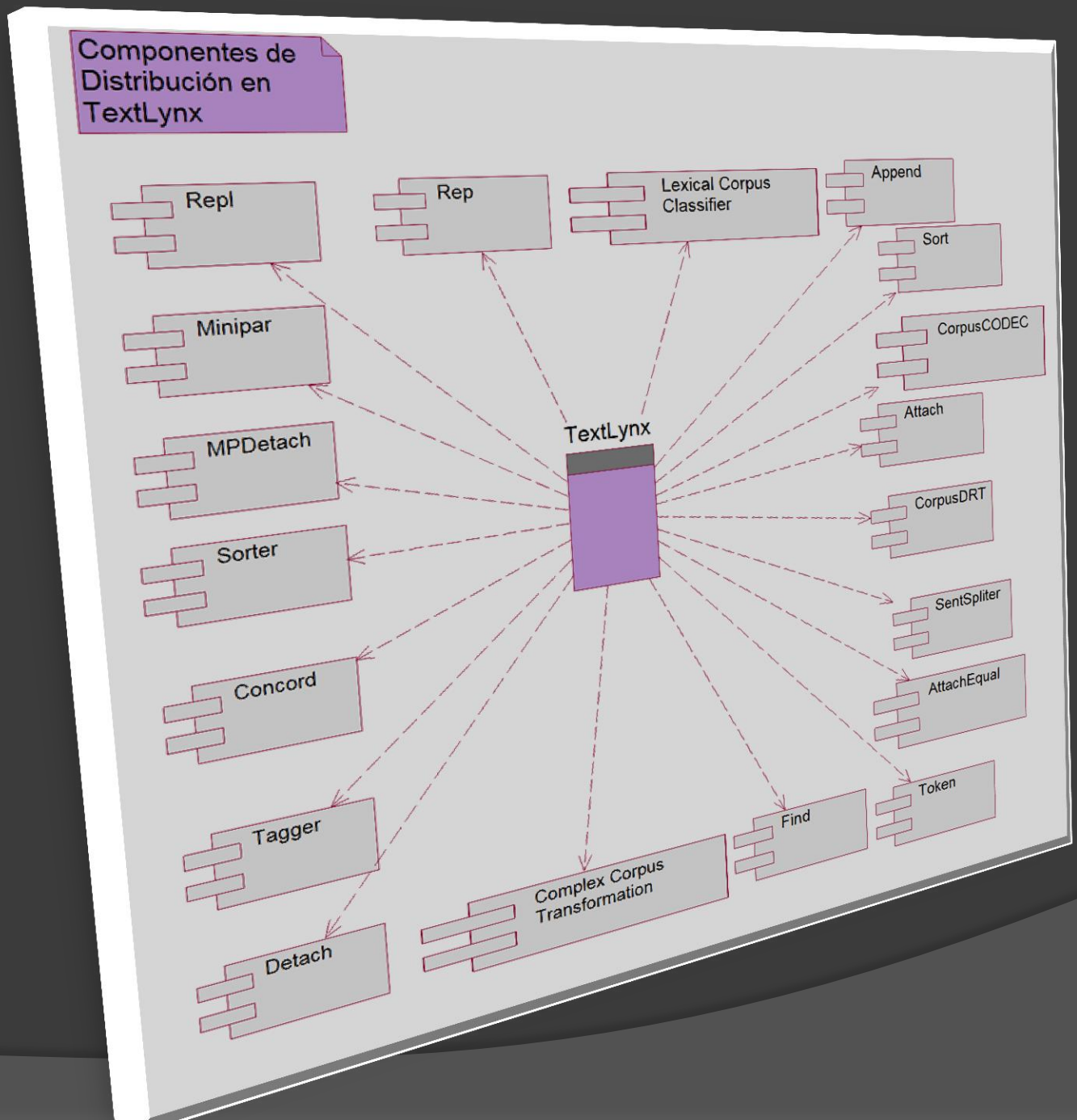
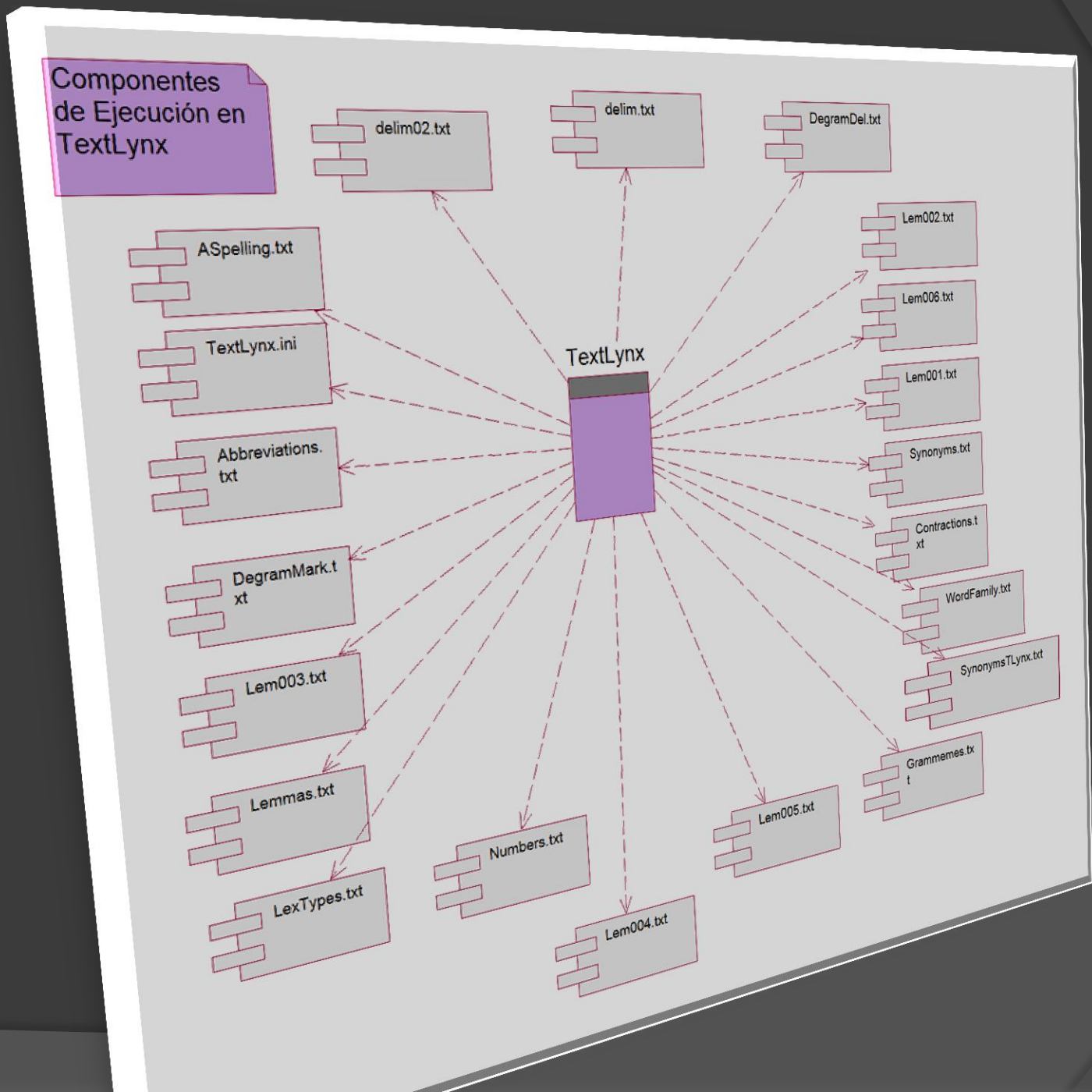


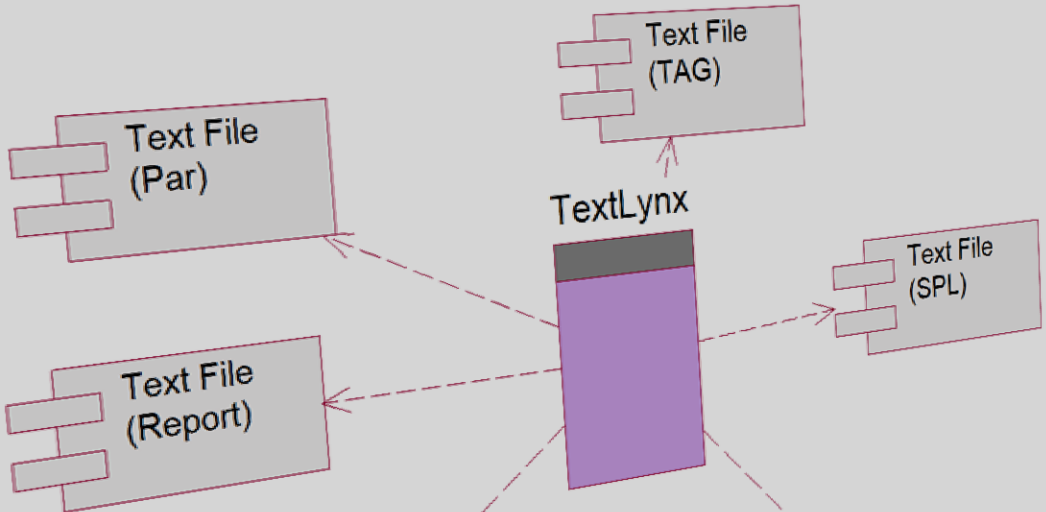
Diagrama de casos de uso para TextLynx
Trabajo con Corpus
Trabajo con Reportes
<<extend>>
<<extend>>
Formatear Corpus
Transformar Corpus
Especialista en Análisis de Texto
Trabajo con Tablas
<<extend>>
<<extend>>
<<extend>>
Creación de Tablas de información
Creación de Tablas de Clasificación
Creación de Tablas de comparación


Componentes de Distribución en TextLynx
Repl
Rep
Lexical Corpus Classifier
Append
Sort
Minipar
CorpusCODEC
TextLynx
Attach
MPDetach
CorpusDRT
Sorter
SentSpliter
Concord
AttachEqual
Token
Tagger
Find
Complex Corpus Transformation
Detach


Componentes de Ejecución en TextLynx
delim02.txt
delim.txt
DegramDel.txt
ASpelling.txt
Lem002.txt
Lem006.txt
TextLynx.ini
TextLynx
Lem001.txt
Abbreviations.txt
Synonyms.txt
Contractions.txt
DegramMark.txt
WordFamily.txt
SynonymsTLynx.txt
Lem003.txt
Grammemes.txt
Lemmas.txt
Lem005.txt
Numbers.txt
LexTypes.txt
Lem004.txt

Componentes para Trabajar en TextLynx

Nota: Solo se tomaron en cuenta aquellos componetes que no tienen un uso temporal.



Este componente indica todos los archivos de tablas que se crean

Resultados

- Corpus Classifier
- Corpus Codec
- Corpus Reducer
- Corpus Transformations

Conclusiones

- Se corrigió todo el funcionamiento de ***TextLynx*** incluyendo las operaciones que supuestamente el software realizaba en su primera versión.
- Se ha aumentado la funcionalidad de ***TextLynx*** diseñando y programando las herramientas: Complex Corpus Transformations, Corpus CODEC, Corpus Reducer, y Corpus Classiffier.
- Se ha duplicado el número de tablas generadas por ***TextLynx*** entre las que se incluye Text-Text comparison. Esta tabla añade al software comparaciones entre textos de un corpus y funciona como primer eslabón para la obtención de novedosos resultados.

Recomendaciones

- Agregar gráficos. Los gráficos pueden ser una manera muy efectiva de devolver los resultados del análisis de texto realizado por ***TextLynx Plus*** y además permitirá dar una mayor claridad a los resultados.

Preguntas del oponente

1. Varias veces mencionas el uso de un “arbol de caracteres” en tu implementación como una representación de un corpus que utiliza un mínimo de memoria ¿En qué te basas para decir esto?
2. ¿Por qué usas comandos externos del sistema operativo para realizar algunas funciones del programa?

Preguntas del oponente

3. ¿ Qué ventaja presenta el uso del árbol de documentos en la interfaz?
4. ¿ Se programó el sistema tomando en cuenta la posibilidad de su extensión posterior? ¿Cómo se logra?

TEXTLYNX PLUS

Ampliación de funciones para un analizador avanzado de textos

AUTOR: Víctor Annier Barrios Cañizares
TUTOR: Lic. Manuel Llanes Abeijón