# Capítulo 1

# Generación de lenguaje natural.

La generación de lenguaje natural (NLG) es un una rama la lingüística computacional y la inteligencia artificial encargada de estudiar la construcción de sistemas computacionales capaces de producir texto en español o cualquier otra lengua humana a partir de algún tipo de representación no-lingüística de la información a comunicar. Estos sistema combinan conocimientos tanto del lenguaje en cuestión cómo del dominio de aplicación para producir automáticamente documentos, reportes, mensajes o cualquier otro tipo de textos.

Dentro de la comunidad desarrolladora e investigadora de la NLG hay un cierto consenso sobre la funcionalidad lingüística general de un sistema de NLG. En este trabajo se optó por seguir la metodología más comúnmente aceptada, propuesta por Reiter y Dale[**?**]. A continuación describiremos brevemente los aspectos mas importantes de esta metodología y en capítulos posteriores desarrollaremos más en profundidad en los puntos mas relevantes para nuestro trabajo.

## 1.1. Análisis de requerimientos

El primer paso en la construcción de cualquier sistema de software, incluyendo los sistemas de generación de lenguaje natural, será el de realizar un análisis de requerimientos y a partir de ahí generar una especificación inicial del sistema.

Para el análisis de requerimientos, Reiter y Dale [RD00] proponen realizar un *corpus* de textos de ejemplo y a partir de ellos obtener una especificación para el sistema a desarrollar. Estos ejemplos estarán compuestos por una colección de datos de entrada del sistema con sus respectivas salidas (texto en lenguaje natural). Estos textos deberán estar redactados por un humano experto y deberían caracterizar todas las salidas posibles que se espera que el sistema genere.

En el capítulo 1 profundizaremos más sobre este tema, describiremos y analizaremos el *corpus de descripciones* utilizado para este trabajo.

## 1.2. Tareas de la generación de lenguaje natural

Dentro de la comunidad desarrolladora e investigadora de la generación de lenguaje natural, hay cierto consenso sobre las tareas que deben llevarse a cabo para, a partir de los datos de entrada generar texto final en lenguaje natural.

Reiter y Dale distingue siete tareas que deben llevarse a cabo a lo largo de todo el proceso: *"determinación del contenido"*, *"estructuración del documento"*, *"agregación"*, *"lexicalización"*, *"generación de expresiones de referencia"*, *"realización lingüística"* y *"realización de la estructura"*.

### 1.2.1. Determinación del contenido.

Esta tarea será la encarga de determinar que información debe ser comunicada en el texto final. Generalmente este proceso consiste en el filtrado y resumen de los datos de entrada. El filtrado o selección consiste en en la elección de un subconjunto de los datos disponibles para ser comunicados en el texto final. El resumen, por otro lado, será necesario cuando los datos de la fuente de información son demasiado detallados para comunicarlos directamente o si la información importante es una generalización de los datos dados en lugar de los datos en sí.

En algunos casos se podrán utilizar los datos de entrada tal y como son proporcionados por la fuente de información, pero generalmente, se tendrá que procesar los mismos para poder utilizarlos posteriormente generando alguna estructura interna que permita representar las entidades, conceptos y relaciones del dominio de aplicación.

Hay algunos investigadores como Evans[**?**] que no consideran a ésta etapa como una tarea de la NLG, pero indudablemente se trata de un problema que en algún momento hay que resolver.

### 1.2.2. Estructuración del documento

La tarea de estructuración del documento consiste en agrupar y ordenar la información a comunicar de forma que el texto resulte coherente y no como una generación de textos ordenados al azar.

Los datos pueden ser agrupados conceptualmente para ser presentados de acuerdo a la información que comunican. Por ejemplo, podríamos agrupar los elementos de modo que en un párrafo se encuentre toda la información referida a un mismo tema y en el siguiente párrafo la información relativa a otro; o podríamos tener un primer párrafo con información general seguido de otro que detalle algún elemento del primero.

También podríamos establecer relaciones retóricas o de discurso entre los elementos o grupos de elementos del texto. Relaciones como *ejemplificación*, *contraste* o *elaboración*, entre otras, podrían relacionar elementos del texto.

### 1.2.3. Lexicalización

La lexicalización es el proceso de elegir las frases y palabras adecuadas con el fin de comunicar la información requerida. En esta etapa se deberá establecer como se expresa un significado conceptual concreto, descrito en términos del modelo del dominio, usando elementos léxicos en lenguaje natural. Por otro lado, estos elementos podrían estar asociados a más de una frase o palabra. Por ejemplo, un elemento podría estar relacionado con varias palabras sinónimas. En estos casos la lexicalización deberá incluir una subtarea para elegir el término correcto. En sistemas multilingüe, los elementos harán referencia a una palabra en cada lengua.

### 1.2.4. Generación de expresiones de referencia

En esta tarea se determina que expresiones se deben usar para referirse o identificar a las distintas entidades del dominio de aplicación. Una misma entidad del dominio de aplicación podría ser referida de distintas formas. Será la etapa de generación de expresiones de referencia la encargada de elegir que expresiones usar para describir una las mismas de modo que el lector pueda identificarla en un contexto dado. La descripción que se elija para hacer referencia a una entidad por primera vez (referencia inicial) dependerá de la razón por la cual se introduce a esta entidad y que se pretende comunicar en posteriormente en el texto. Si se vuelve a hacer referencia a la entidad después de haber aparecido una vez (referencia posterior) la preocupación será la de poder diferenciarla de las otras entidades con las que se podría confundir, pero sin que resulte un texto poco fluido.

### 1.2.5. Agregación

El proceso de agregación se encarga de combinar varios elementos informativos con el fin de conseguir un texto más fluido y legible. La agregación decide que partes de las estructuras se pueden combinar para realizarlas como oraciones complejas para que se pueda generar un texto conciso, cohesionado y que a la vez el significado del texto se mantenga casi igual que sin agregación.

### 1.2.6. Realización lingüística

Es el proceso de convertir las representaciones abstractas del texto en texto real. Al igual que los textos no son secuencias de oraciones ordenadas al azar, las oraciones no son secuencias de palabras ordenadas al azar. Cada lengua está definida por un conjunto de reglas gramaticales que especifican lo que

es una oración bien formada en esa lengua. Estas reglas determinan tanto la morfología, que se ocupa de como se forman las palabras (genero, numero, etc.), la sintaxis, que trata de cómo se forman las oraciones y la ortografía (TODO completar). La realización lingüística consiste en aplicar alguna caracterización de estas reglas de la gramática a una representación abstracta para producir un texto que sea sintáctica y morfológicamente correcto y ortográficamente correcto.

### 1.2.7. Realización de la estructura

Esta etapa se encarga de convertir estructuras abstractas como párrafos y secciones en texto comprensible por el componente de presentación del documento. Por ejemplo, la salida del sistema de NLG podría ser código LaTeX para luego ser post-procesado, en este caso sería esta etapa la encargada de agregar delimitadores y comandos de LaTeX para generar el documento.

## 1.3. Arquitectura para la generación de lenguaje natural.

# Bibliografía


[CAF+11] Maximiliano Cristí, Pablo Albertengo, Claudia Frydman, Brian Pluss, and Pablo Rodriguez Monetti. Applying the test template framework to aerospace software. In *Proceedings of the 2011 IEEE 34th Software Engineering Workshop*, SEW '11, pages 128–137, Washington, DC, USA, 2011. IEEE Computer Society.

[CM09] Maximiliano Cristiá and Pablo Rodríguez Monetti. Implementing and applying the stocks-carrington framework for model-based testing. In *Proceedings of the 11th International Conference on Formal Engineering Methods: Formal Methods and Software Engineering*, ICFEM '09, pages 167–185, Berlin, Heidelberg, 2009. Springer-Verlag.

[CP10] Maximiliano Cristiaánd Brian Plüss. Generating natural language descriptions of z test cases. In *Proceedings of the 6th International Natural Language Generation Conference*, INLG '10, pages 173–177, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[RD00] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA, 2000.

[SC96] Phil Stocks and David Carrington. A framework for specification-based testing. *IEEE Trans. Softw. Eng.*, 22(11):777–793, November 1996.

[Spi92] J. M. Spivey. *The Z Notation: A Reference Manual*. Prentice Hall International (UK) Ltd., Hertfordshire, UK, UK, 1992.