

# InTiMe Plataforma de Integración de Recursos de PLN

## *InTiMe* *Integration Platform of NLP Resources*

**José Manuel Gómez**

Departamento de Lenguajes y Sistemas Informáticos  
Universidad de Alicante  
Carretera Sant Vicent del Raspeig s/n  
03690 Sant Vicent del Raspeig (Alicante)  
jmgomez@dlsi.ua.es

**Resumen:** La plataforma InTiMe (INtegration of Tools and corpora In the text-MEss project) es un ambicioso proyecto del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante. Nace con la idea de integrar, en una misma plataforma, gran parte de los recursos utilizados actualmente en Procesamiento del Lenguaje Natural (PLN). De esta forma, cualquier investigador incluido en la plataforma tendrá acceso inmediato, independientemente del sistema operativo que use o de su ubicación, a todas las herramientas y corpus integrados en el sistema. También será capaz de dar a conocer, si así lo desea, a toda la comunidad científica los nuevos recursos desarrollados en sus investigaciones. Evitando así que los investigadores tengan que desarrollar herramientas ya existentes, ahorrando tiempo y recursos y centrando los esfuerzos en actividades más novedosas. Como veremos en el presente artículo, InTiMe agilizará la compartición del conocimiento y el uso de los recursos generados en PLN aumentando la productividad sin tener que cambiar la metodología de trabajo.

**Palabras clave:** Herramientas PLN, Integración recursos, InTiMe, PLN

**Abstract:** The InTiMe platform (INtegration of Tools and corpora In the text-MEss project) is an ambitious project of the Department of Languages and Computer Systems at the University of Alicante. Born with the idea of integrating, in a single platform, almost of the resources currently used in Natural Language Processing (NLP). Thus, any researcher included in the platform will have immediately access, regardless of the operating system he use or his location, all the tools and corpora integrated in the system. It will also be able to disclose, if he so wish, to the entire scientific community developed new resources in his investigations. Avoiding so that researchers need to develop tools that already exist, saving time and resources and focusing efforts on newer activities. As we will see in this article, InTiMe expedite the sharing of knowledge and the use of resources generated in PLN increasing productivity without changing the methodology of work.

**Keywords:** NLP tools, resource integration, InTiMe, NLP

## 1. *Introducción*

La investigación se basa, principalmente, en la idea de compartir conocimientos, herramientas y corpus que permitan a los investigadores aunar sus esfuerzos para lograr metas mayores. En áreas de investigación como el Procesamiento del Lenguaje Natural (PLN) esto adquiere una mayor importancia pues las soluciones a los problemas que se plantean hoy en día se basan en la combinación

de diversos recursos. Por lo tanto, un investigador debe ser capaz de conocer los recursos disponibles, saber utilizarlos correctamente y, a su vez, dar a conocer su propio trabajo.

Es muy común que cada recurso lo desarrolle diferentes personas que tienen intereses muy concretos y es inusual que piensen en una futura integración de su trabajo con el resto de recursos (Graça, Mamede, y Pereira, 2006). Es más, cuando se intenta integrar

todos los recursos y herramientas es cuando surgen los problemas de cómo se van a comunicar las aplicaciones entre sí y cómo van a procesar los distintos corpus.

El problema crece cuando en un grupo de investigación existe personal investigador temporal y no se ha planteado ninguna política de integración de estos recursos. Esta situación se resume en que los grupos de investigación disponen de gran cantidad de herramientas, aplicaciones y corpus (tanto propios como ajenos), en diferentes lenguajes de programación y sistemas operativos, con formatos de salida y entrada particulares, de localización muchas veces difícil puesto que dependen de la persona o personas que los han generado, y de reutilización compleja ya que requiere un esfuerzo adicional para integrarlos en otros desarrollos (Monteagudo y Cuetto, 2005).

Para resolver estos problemas, muchos grupos de investigación han decidido, a lo largo de su vida, aplicar alguna metodología de integración de recursos de PLN que diera a conocer a sus propios miembros los recursos disponibles. Algunos de estos proyectos de integración son parciales pues sólo tienen en cuenta algún aspecto concreto: o bien se centran en un tipo de recurso o en un dominio específico. Entre estos proyectos se pueden destacar el BancTrad (Badia et al., 2002), que proponen un formato estándar para la integración de corpus etiquetados paralelos junto con herramientas para acceder a él; el Emdros text database system (Petersen, 2004), el cual es un motor de base de datos para el análisis y la recuperación del texto analizado o anotado; el Natural Language Toolkit (Bird y Loper, 2004) que es un conjunto de bibliotecas y programas para el procesamiento simbólico y estadístico del lenguaje natural; y el Festival speech synthesis system (Taylor, Black, y Caley, 1998), el cual es un framework para construir sistemas de síntesis del habla.

También están los proyectos que únicamente definen protocolos o formatos para la comunicación entre distintos procesos de PLN, como el Annotation Graphs Toolkit (Maeda et al., 2001) que es una implementación del formalismo de Grafos Anotados de (Bird y Liberman, 2001), y el más influyente trabajo en éste área: la arquitectura Atlas (Bird et al., 2000), que generaliza el trabajo de (Bird y Liberman, 2001) para permitir el uso de señales multidimensionales. El pro-

blema de estos formatos es que no permiten la separación de la información en capas teniendo que cargar, en cada proceso, todas las anotaciones previas. También podemos encontrar el sistema EMU (Cassidy y Harrington, 2001) que está enfocado, específicamente, en tratamiento del habla.

Existen otros proyectos de integración que intentan abarcar tanto la especificación de los corpus y datos como de las herramientas en una única plataforma. El ejemplo más destacado lo podemos encontrar en el proyecto GATE (Cunningham, Wilks, y Gaizauskas, 1996; Bontcheva et al., 2004), que permite añadir módulos en Java de forma muy sencilla y rápida aunque requiere de más trabajo en caso de otros lenguajes de programación. GATE, además, define un formato de datos basado en la arquitectura TIPSER (Grishman, 1996) y en el Annotation Graphs Toolkit. Un sistema muy similar al GATE es el UIMA (Ferrucci y Lally, 2004), que está basado en el proyecto TEXTTRACT (Neff, Byrd, y Boguraev, 2004) de IBM. Pero, al igual que GATE, exige un cambio en la metodología de los grupos de investigación que pretendan usar la plataforma UIMA. Otros trabajos menos conocidos son los realizados por (Graça, Mamede, y Pereira, 2006) con una propuesta cliente/servidor que unifica las herramientas de PLN utilizando repositorios etiquetados con información multicapa que elimina la necesidad de cargar toda la información en cada proceso; y el trabajo de (Monteagudo y Cuetto, 2005), otra herramienta cliente/servidor que, a través de un middleware, unifica las herramientas y establece su propio formato de datos para comunicar los distintos procesos.

Aunque hay bastantes herramientas, protocolos y formatos que te permiten integrar herramientas y recursos, todos fallan en algún aspecto. Algunos de ellos son muy específicos y únicamente abarcan un conjunto de recursos de un área concreta del PLN (Badia et al., 2002; Petersen, 2004; Bird y Loper, 2004; Taylor, Black, y Caley, 1998). Otros no permiten muchos tipos de datos, por ejemplo, se centran en datos de texto o de habla únicamente (Maeda et al., 2001; Bird y Liberman, 2001; Bird et al., 2000; Cassidy y Harrington, 2001). También están los que obligan a trabajar en algún lenguaje informático concreto, un sistema operativo, plataforma, o que obligan a cambiar la metodología de trabajo











