

Introduzindo o Corpógrafo

um conjunto de ferramentas para criar corpora especializados e comparáveis e bases de dados terminológicas

Cet article décrit le Corpógrafo, un environnement disponible en ligne qui a pour vocation la recherche de corpus. Il permet la compilation de textes en plusieurs formats, la constitution et l'analyse de corpora, l'extraction de terminologie et la création de bases de données terminologiques. Il offre également la possibilité de créer des relations sémantiques et de produire des ontologies.

El Corpógrafo es un entorno disponible en línea para la gestión de corpus. Permite la compilación de textos en varios formatos, la elaboración y el análisis de corpus, la extracción de terminologías y la creación de bases de datos terminológicas. También brinda la posibilidad de crear relaciones semánticas y producir ontologías.

O Corpógrafo

Há vários anos que a Linguateca, um centro de recursos distribuído para o processamento computacional da lín-

gua portuguesa, com pólos em Oslo, Lisboa, Braga e Porto, tem vindo a criar uma vasta selecção de corpora em português com um leque de ferramentas de pesquisa linguística associadas. Desde Janeiro de 2003, o Polo

CLUP da Linguateca tem vindo a desenvolver pesquisa no uso de corpora especializados comparáveis para o estudo e a extracção de terminologia. Criámos, para este efeito, o Corpógrafo, um conjunto de ferramentas disponível 'online' para quem estiver interessado em pesquisar autonomamente. O Corpógrafo permite colecionar textos em vários formatos, formar e analisar corpora, extrair terminologia e criar bases de dados terminológicas com a possibilidade de codificar relações semânticas e ontologias.

O Corpógrafo (Sarmento & Maia, 2003; Sarmento, Maia & Santos, 2004) é, assim, uma plataforma de pesquisa sobre corpora especializados que surge da necessidade de integrar no mesmo ambiente todo um conjunto de operações e de processos, anteriormente realizados utilizando várias ferramentas ou sistemas cujo acesso era muitas vezes restrito ou difícil. O Corpógrafo oferece ao utilizador, através de uma simples interface na rede (Web), a possibilidade de compilar e pesquisar os seus próprios corpora (a partir de documentos em formato PDF, Ms-Word, PostScript, RTF ou HTML) sem que para isso seja necessário ter conhecimentos especiais de informática. O Corpógrafo complementa a oferta de corpora publicamente oferecidos pela Linguateca (veja-se os projectos AC/DC e Compara, ou os corpora jornalísticos CETEM-Público e CETENFolha), possibilitando a construção e pesquisa em corpora pessoais e específicos em áreas especializadas, para utilizadores com interesses nas áreas da linguística, tradução, terminologia ou engenharia do conhecimento.

Para a área da linguística, o Corpógrafo possibilita pesquisas de concordâncias e colocações, assim como estudos de frequências de n-gramas. Para tarefas associadas à tradução e à engenharia do conhecimento, o Corpógrafo possui funcionalidades avançadas de pesquisa terminológica, directamente integrada num sistema de base de dados para uma fácil organização dos termos extraídos. As capacidades de pesquisa terminológica (fundamentalmente em português e inglês, mas também em espanhol, francês, italiano e alemão) são complementadas com módulos de identificação de definições dos termos extraídos e de reconhecimento de possíveis relações semânticas entre os conceitos.

Actualmente o Corpógrafo foi experimentado por cerca de duzentas pessoas e é utilizado regularmente por quarenta, localizadas maioritariamente em Portugal e no Brasil, embora também seja utilizado por vários investigadores noutros países da Europa. Na FLUP, decorrem várias teses de doutoramento e mestrado e projectos de terminologia utilizando o Corpógrafo e pesquisando nas áreas de engenharia mecânica, engenharia electrónica, geografia, genética, neuroanatomia, engenharia da linguagem, etc.

Prevedemos, com a entrada numa nova fase de funcionamento do Corpógrafo em Novembro de 2004 e a continuação do seu desenvolvimento com a pesquisa em curso, dar um salto qualitativo e quantitativo ao nível de trabalho possível com este ambiente. Em particular, planeamos desenvolver sistemas de procura mais inteligentes vocacionados para uma área especializada, e iniciar trabalho na detecção semi-automática de relações semânticas especializadas.

Belinda Maia

Faculdade de Letras da Universidade do Porto

Luís Sarmento

Faculdade de Engenharia da Universidade do Porto,
Polo CLUP da Linguateca

Diana Santos

Responsável pelo projecto Linguateca
Polo de Oslo no SINTEF

Referências

- MAIA, BELINDA & LUÍS SARMENTO 2003 *Constructing comparable and parallel corpora for terminology extraction - work in progress*, Poster presentation at Corpus Linguistics 2003, Lancaster U.K. (Winners of 1st prize).
- SARMENTO, LUÍS BELINDA MAIA & DIANA SANTOS. "The Corpógrafo - a Web-based environment for corpora research". In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of LREC 2004* (Lisboa, Portugal, 26-28 May 2004), pp. 449-452.