

## AMTA 2010 Technology Showcase

<b>Company or organization</b>	The Technology Development Group (www.thetdgroup.com)
<b>Products or software</b>	OMPC and LLB Light
<b>Version</b>	OMPC 3.0 / LLB Light 1.0
<b>Available for licensing Y/N</b>	Y
<b>Kind of product (e.g., MT, OCR, Dictionary, etc.)</b>	OMPC = Online Multimedia Parallel Corpora database  LLB Light = Linguist Language Broker, a Dynamic social terminology management application
<b>Description</b>	OMPC is an online web-application that serves as a resource for parallel corpora associated with multimedia files, available for assisting in USG NLP tool creation  LLB Light: is a light desktop application that allows users to seamlessly interface with the main LLB terminology management system
<b>Languages or language pairs supported</b>	OMPC: Arabic, Chinese, English, French, Russian, Spanish  LLB Light: All Unicode languages
<b>Unique about this system</b>	OMPC: This system allows for the import, alignment/segmentation, search and sharing of multimedia-associated parallel corpora in a variety of languages  LLB Light: This application provides users with the ability to quickly store, collaborate on, search and share unique foreign language terms and phrases
<b>New this year</b>	OMPC: This version includes additional parallel corpora across a variety of languages as well as improved search capabilities. Users can now view and download targeted corpus collections.  LLB Light: this brand-new version is a light application that allows users to seamlessly interface with the main LLB terminology management system.
<b>Part of a larger system? Describe</b>	OMPC: encompasses the entire system  LLB Light: interfaces with the main LLB terminology management system
<b>Made up of other systems? Describe</b>	OMPC: Solr 1.3, AppTek-Aligner, HotSpot,  LLB Light: Solr 1.3,
<b>Standards supported</b>	
<b>Used by</b>	OMPC: US-Gov users and associated researchers/academics  LLB Light: US-Gov users and associated researchers/academics
<b>Presentations at AMTA—what/when</b>	Mike O'Malley is presenting a paper on "Challenges of a Distributed Parallel Corpora"
<b>Contact for more information</b>	
<b>Name</b>	Jon Phillips
<b>Telephone</b>	571-262-2699 / 571-525-7232
<b>Email</b>	<a href="mailto:jphillips@thetdgroup.com">jphillips@thetdgroup.com</a>

## OMPC

There is increasing interest and need to have access to a foreign language corpus of original source data and the corresponding translated data that represents what the Intelligence Community (IC) collects, processes and analyzes on a daily basis. Computer based language technologies have made great strides in the past several years, especially the ability to save data in its original form and/or convert it into Unicode based text. This capability makes accessing and archiving the data feasible and of great value to various groups in the IC: the researchers, developers, analysts, instructors and learners of foreign languages.

This type of data archiving is not unique. The Linguistic Data Consortium (LDC) is a successful example in academia. “The LDC was founded in 1992 to provide a new mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies. Based at the University of Pennsylvania, the LDC is a broadly-based consortium that now includes more than 100 companies, universities, and government agencies. Since its foundation, the LDC has delivered data to 197 member institutions and 458 non-member institutions (excluding those who have received data as a non-member and later joined).” However, the data that the LDC collects and archives is not representative of internal IC data.

Like the academic and research community, the IC needs access to an enormous variety of linguistic data — speech, text, lexicons, and grammar — to improve the processes by which it is able to understand and apply automated linguistic solutions to the data. In addition, the IC also has a need for specialized linguistic data based upon its specific and unique mission. Such databases are expensive to create and document. The most valuable data is that which has been validated within the IC. Further value is added by associated human translation (HT), machine translation (MT), and/or metadata. Having a centralized repository for such data is highly desirable because sharing the resources across agencies provides benefits that go beyond just having access to the data. It enriches the systematically documented repository and offers savings for all participants. In addition to the specific benefits to each entity in the Community, the overall benefits in language processing research and development will have a positive impact on the community as a whole.

The Online Multimedia Parallel Corpora (OMPC) System is designed to ingest, segment and align, store and index, and provide newly created and existing parallel corpora (PC) documents at the sentence or paragraph level. In many cases, these parallel corpora are linked to multimedia artifacts; for example they are transcriptions of video broadcasts.

TDG FuzeIn Web Desktop - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://192.168.6.40:8080/Fusion/#

Google

Hi

Fusion Chat

FuzeIn Stream Capture

Federated Research

Personal Terminology Management

Term Vetting

OMPC Asset Upload

OMPC Asset Upload

OMPC Manual Aligner

Asset(s) Listing

Asset Name	File Size
------------	-----------

Assets Desc. & Processing Instructions

Asset Description

Title:

Source:

Summary:

Keyword(s):

Genre(s):

Asset Processing Instructions

Upload Queue Metadata Viewer

Delete	Asset Type
--------	------------

translation

insert Delete Merge

PRC: PLA 2nd Artillery Base Conducts New-Type Informatized Training in Mid-August  
CPP20061120318015 Beijing Jiefangjun Bao (Internet Version-VWW) in Chinese 24 Aug 06 p 2

[Report by Xia Hongqing and special correspondent Wang Yongxiao: "A Second Artillery Base Actively Builds System That Enables Military Training To Be Carried Out Under Conditions of Informatization -- A Batch of New Training Subjects and Methods Are Unveiled on Troop-Training Fields"]

Editor's Note: Since the end of

击了一些正在悄然发生的变

Displaying Segment 1 - 10 of 125

Displaying Segment 1 - 14 of 125

The TDGroup FuzeIn Desktop

- FuzeIn Virtual Keyboard
- Fusion Chat
- Applications
  - FuzeIn Virtual Keyboard
  - Communications
  - Internet
  - Preferences
  - LLB
  - Online Multimedia Parallel Corpora
- Preferences
- Help

Log Out: johnph

Lock Screen

start

C:\Documents and Se... OMPC\_CPR-ORR-DR... Sprint Mobile Broadband TDG FuzeIn Web Des... Untitled (82%) - Pain...

EN Desktop 100% 1:30 PM Thursday 10/2/2008

## LLB Light

The Linguist Language Broker (LLB) is a highly-customized web-based terminology management system. The LLB system includes general terminology management, and also integrates several features that are desired by today's intelligence community (IC) translators.

LLB is a "translator's toolbox," an integrated workspace where translators can collaborate on and research definitions and translations of uncommon foreign-language words or phrases. These can be added to the LLB data-store, allowing IC translators to learn and benefit from their colleagues' knowledge and prior work. The LLB also utilizes links to existing IC dictionaries and terminology systems. LLB further aids translators by filtering their search results (by usage domain, for example) to find the most relevant ones.

LLB Light is a small desktop application that allows user to easily access the robust terminology management functions of the main LLB system.

