

3. Bitext SA

www.bitext.com

Madrid, Spain, is a hot spot for search and advanced text processing. Most North Americans think of Madrid and recall an Ernest Hemmingway novel. Bitext wants the association to hook into natural language processing from “the bits and text company.”

Bitext’s founder and CEO is the lean, handsome Antonio S. Valderrábanos, a graduate of Universidad Autonoma de Madrid. With a PhD in Linguistics, Mr. Valderrábanos founded Bitext in 2001 launched the company to carry out consulting services on language technologies after his long experience in this field in IBM, with Wordperfect and Novell. By 2004, the company’s principal focus was NLP or natural language processing technology.

Item	Quick Facts
Product	NaturalFinder
Price	Begins at 23,000 Euros
Technology	Natural language search system
Key Feature	Performs automatic synonym and query expansion; supports NLP queries
Purpose	Allow a user to interact using sentences in English or Spanish and other languages such as Basque and Catalan
Clients	RENEF, Ministry of Defense
Company	Bitext SL, Madrid, Spain
Contact	info@bitext.com

Table 8: Quick Look at Bitext SA

Bitext’s linguistic technology can also be applied to computer-assisted translation environments. One of the leading companies in this area, Atril, developers of Déjà Vu X, uses Bitext’s linguistic technologies to improve results for searching in translation memory databases. In addition, Bitext has participated in computer-assisted translation projects funded by the European Commission.

Bitext is a privately-held firm. The firm’s customers include:

- RENFE (the Spanish Railroad Company)
- Public Administration National Institute, Spanish Government
- Ministry for the Presidency, Spanish Government
- Ministry of Defense, Spanish Government
- TYPSA
- Sitesa, Grupo EP, distributor of Google Search Appliance in Spain.

In addition, Bitext has developed adaptors to link its NLP technology with dtSearch (a Microsoft-centric search system) and the Google Search Appliance.

The screenshot shows the Bitext.com website header with the logo and tagline 'THE BITS AND TEXT COMPANY'. Below the header is a navigation bar with links for 'DEMOS', 'Live', 'Live-EN', 'BOE', and 'GOOGLE', along with a language dropdown menu set to 'Español'. A search bar contains the text 'information on tourist attractions in Liverpool or Leeds'. Below the search bar are buttons for 'NF+Live' and 'Live'. A 'Start' button is visible below the search bar. The search results are displayed in a list format with the following items:

- LIVE with NaturalFinder**
- Hull Tourist information**
Hull Tourist Information. Kingston upon Hull, England is an ... Hull was placed as the fifteenth leading tourist attraction, beating Leeds, Liverpool and Cambridge.
- Travel Australia Tourist Attraction**
McCarthy Country Tourism Tourist Information Centre Tourist Attraction Cnr Hume Highway and Congressional Drive, Liverpool ... Advertised Tourist Attractions can be found on Page 1 ...
- Liverpool Tourist Attractions**
Tourist Attractions in Liverpool, England ... Select a tourist attraction in Liverpool from the list below, or ... City Information; Photo Gallery; Maps; Hobbies ...

Figure 25: Bitext NLP Adds Functionality to Microsoft's Live Search

The Bitext NLP system adds functionality to Microsoft's Live.com search. Note that Bitext integrates with SharePoint as well. The user can enter a complex query without worrying about Boolean operators and query syntax. The system delivers results that are more specific to the query. Bitext, when executing a Boolean OR, sees term occurrence as significant. The first result in this list points to a site with information about both Leeds and Liverpool.

The Bitext Data Suite

The core functionality of NaturalFinder resides in what the company calls its DataSuite. These are subsystems that perform the heavy lifting required to process content and queries in the NLP system.

DataGrammar

DataGrammar is the subsystem that interprets natural language. It is built into the Bitext system. One feature of the DataGrammar subsystem is that it can “learn” as it processes content; for example, when an unknown phrase appears in a document, DataGrammar recognizes this phrase and adds it to the knowledgebase in the system. Learning is automatic and for most general content does not require the intervention of a subject matter expert.

DataLexica

This is a built-in lexical database. It uses linguistic stemming; that is, the subsystem removes inflections from words. The Bitext stemmer makes context-based decisions. Bitext asserts that its approach is “more than an intelligent stemmer.” For lemmatization and conjugation: DataLexica returns the lemma or root of words along with morphological information. For example, given the Spanish word *casa* DataLexica returns the following morphological information to the system:

- The root *casa* as a feminine singular noun
- The form is the third person of the present tense in indicative mood of the verb *casar* and *casarse*)
- Other forms of the verb *casar* such as *casando*, *casado*, *casada*, *casados*, *casadas*, *caso*, *casas*, etc.

Feature	Beyond Search Comment
Knowledgebase Support	Includes a lexicon, thesauri and knowledge base
Query Types	Natural language
Visualization	None. Third-party tools may be integrated with the API
Entity Extraction	Built in via proprietary algorithms and a knowledgebase
Platforms Supported	Linux and Windows
Export	The API allows expert functions to be defined
Third-Party Support	Can be integrated with third-party systems
Vertical Support	Builds for English, Spanish, and other languages available
Analytic Functions	None

Table 9: Technical Highlights for Bitext

DataSpell

DataSpell is the built-in spelling correction mechanism. It includes more than three million correctly codified words, not including proper names in Spanish in this count. DataSpell determines whether or not a word is correct in a specific language and, if it is not correct, it suggests alternatives. For example, for the Spanish word *immobiliario*, DataSpell offers *inmobiliario*. DataSpell is configurable, and it can be integrated into a wide range of third-party applications, ranging from search engines to enterprise resource planning systems.

DataNet

The DataNet subsystem houses the rules regarding semantic relationships. The subsystem discovers relationships automatically and performs synonym expansion. DataNet makes use of existing thesauri, taxonomies, and ontologies. For example, a user can specify that the words *auto* and *coche* are related because of their similar

meaning; or we can specify that Italy is part of *Europe*. An administrative interface is provided to allow the system administrator to customize DataNet's relationship tables.

System Developer's Kit (SDK)

The SDK makes it possible for a licensee to integrate NaturalFinder components into another application or search engine. The SDK includes libraries for Windows and Linux, sample code, and documentation.

Technology

Bitext's system is available both for Linux and Windows. The company also offers a version of the system suitable for use as a hosted service.

NaturalFinder supports a wide range of documents and file types. These include documents, Web pages, and structured data. In addition, the system can make use of existing metadata, thesauri, and ontologies.

The system can process hundreds of thousands of words per second at the lexical level. Grammatical processing handles thousands of sentences per second.

The minimum recommended memory for the server running NaturalFinder is 256MB of RAM. The SDK makes it possible for licensees to add support for other languages.

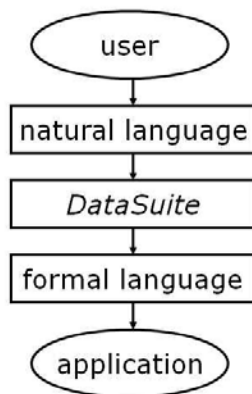


Figure 26: The Bitext data flow is straightforward.

The System in Use

Bitext has customers in Spain, Canada, and Germany. A representative installation is RENFE's use of dtSearch and Bitext. dtSearch is a Windows-centric key word searching system developed by a firm in the Washington, D.C. area. Bitext engineers integrated its NLP system with the dtSearch system.

"Our users at RENFE have been very pleased with the resulting application. Users particularly liked the speed, reliability and precision of searches, and the overall ease-of-use of the application," said Mr. Valderrábanos. "For Bitext, this agreement proves

that its linguistic technology makes a difference in the content management systems of large corporations.”

Bitext's DataSuite for RENFE includes DataLexica, which consists of a large and complete lexical database containing more than three million words classified according to their linguistic features.

At INAP, a digital library in Spain, Bitext installed NaturalFinder. The system was able to index content from different servers in different file formats. The user was able to search one or more of the collections from a single natural language interface. Bitext technology supported federating the content and providing users with the NLP interface. Other features of the INAP installation were filtering by document type, and support for a stringent security system.

Upside

The upside for Bitext's NaturalFinder includes:

- Built in knowledgebase, lexicon, and semantic mappings. These are supplemented with a knowledgebase administrative interface.
- Runs on Linux or Windows
- Supports NLP
- Supports voice or spoken queries when integrated with speech-to-text applications

Downside

The downside for the Bitext system includes:

- The system requires careful set up and configuration. Adequate bandwidth, computational resources, storage, and random access memory are essential for system performance.
- The API for Windows Live lacks some features; for example, the API only shows the first 250 results that Microsoft returns and will not process queries longer than 20 words. Check with Bitext for the functions available in the API.
- Hits with multiple occurrences of terms can be ranked above hits that are directly about a query, for example, in a search for city information.
- The system's relevancy improves with longer queries. Some users enter two to three word queries or prefer clicking on suggested links or categories to discover information without having to formulate a query.

Net-Net

Bitext illustrates the interest in NLP and linguistic search in Europe. Along with Exalead, PolySpot, and Sine Qua Non, entrepreneurial activity in rich text processing is increasing. The Bitext system can add NLP to almost any search system. If you have a search system and want to add automatic entity extract, NLP, and other functions to your existing system, Bitext is definitely worth a look. For companies in Spain, Bitext may well be the NLP system of choice.