

## Research Centre Europe

### A Deep Robust Parsing Environment For Contextual Entity Recognition And Knowledge Extraction

Extracting the meaning from very large quantities of unstructured content is an ever growing need. It calls for computational text analysis that goes far beyond simple keyword search to recognize relations between words.

Two major requirements for a parsing environment are:

- Deep parsing: to extract dependencies between words that may be distant from each other, perhaps in different sentences.
- Robust parsing: to provide linguistic analysis for real-world documents (e.g., Web pages, newspapers, scientific literature, encyclopedia).

Xerox Incremental Parsing (XIP) is Xerox licensable technology for deep robust parsing.

### XIP Rule Formalism

The XIP rule formalism offers superior expressive power to handle rich and fine-grained dependency descriptions. The formalism enables recognition of linguistic relations between any number of words or constituents as a consequence of structural, topological and/or lexical conditions, at sentence and text levels.

XIP benefits from an incremental organization of the rules, which enables a broader range of strategies for applying various sets of rules.

Linguistic descriptions are organized in modules, depending on their depth level. Modularity facilitates the maintenance of linguistic data and makes the system easily customizable or reusable.

### XIP Parsing Engine

The XIP Parsing Engine offers high computational performance, allowing one to process more fine-grained linguistic phenomena while keeping fast response times (ca. 2,000 words per second in English).

### XIP Grammars

Xerox has developed deep functional dependency grammars for XIP with especially high coverage in English and French. It offers text output under the form of annotated chunks and sets of dependencies (see Figure 1 for an example in English).

XIP grammars are also being developed for a number of other languages, including German, Japanese and Chinese. XIP grammars can be customized to tackle specific requirements, thanks to the modular and incremental nature of the XIP rule formalism.

*Caldwell's resignation had been expected for some time.*

#### Chunks:

```
SC{NP{Caldwell 's resignation} FV{had}} NFV{been expected}
PP{for NP{some time}}.
```

#### Dependencies:

```
QUANTD(time,some)
NMOD_PRE(resignation,Caldwell)
SUBJ_PRE(had,resignation)
NUCL_PASSIVE_VLINK(been,expected)
NUCL_PERFECT_VLINK(had,been)
VMOD_DURATION_POST(expected,time)
PREPD_DURATION(for,time)
MAIN_PERFECT_PASSIVE(expected)
```

Figure 1: XIP parsing example (in English) with its text output

### Major XIP Applications

Major applications include contextual entity recognition, lexical and structural disambiguation, coreference resolution and more globally knowledge extraction (e.g., fact extraction). XIP is also used in a commercial product for spoken language analysis.

### XIP Interfacing Options

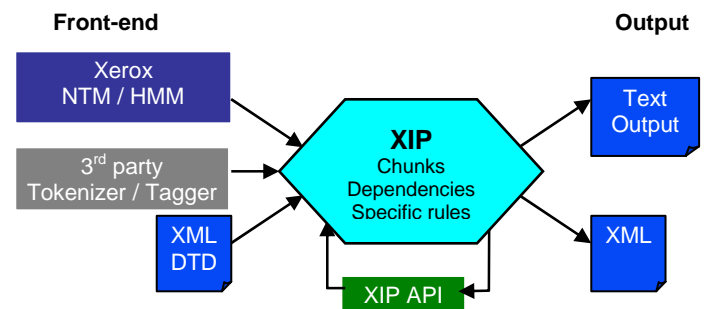


Figure 2: XIP interfacing options

XIP offers the advantage of accepting various types of inputs (see Figure 2). XIP processes raw text through Xerox's state-of-the-art tokenization, normalization, morphological analysis and Part-of-Speech tagging. XIP can be used as a deep parser for front-end shallow parsers or taggers. XIP can process XML marked-up documents to enrich the mark-up with deeper linguistic analysis. Finally, XIP can be iterated through its API.

XIP is available for evaluation, research and commercial (incl. OEM) licensing for Windows, Linux and Solaris.