A ReSTful Web Service for Accessing a Multilingual and Multi-format Syntactic Parser

NEBHI, Kamel, NERIMA, Luka, WEHRLI, Eric

Abstract

To increase the accessibility and usability of language resources and tools, this paper proposes a ReSTful Web Service for accessing the multilingual Fips parser. Our services offer some widely known and used linguistic data formats such as the TEI and the PASSAGE formats. In the current version, the services are available for English, French, German, Italian, Spanish and Greek. Users can access all of these technologies and data formats simply by using a browser while developers can create Web applications based on these resources.

Reference


Available at:
http://archive-ouverte.unige.ch/unige:24069

Disclaimer: layout of this document may differ from the published version.
A ReSTful Web Service for Accessing a Multilingual and Multi-format Syntactic Parser
Kamel Nebhi, Luka Nerima, Eric Wehrli
LATL – Department of linguistics
University of Geneva
{kamel.nebhi, luka.nerima, eric.wehrli}@unige.ch

Abstract: To increase the accessibility and usability of language resources and tools, this paper proposes a ReSTful Web Service for accessing the multilingual Fips parser. Our services offer some widely known and used linguistic data formats such as the TEI and the PASSAGE formats. In the current version, the services are available for English, French, German, Italian, Spanish and Greek. Users can access all of these technologies and data formats simply by using a browser while developers can create Web applications based on these resources.

Keywords: Language Resources and Tools, Web Service, ReSTful Architecture

I. Introduction and Motivation

Researchers always tried to develop language resources and tools (LRT). While in most cases, these technologies are available by download only and/or restricted to a particular platform or environment, several types of LRT can be successfully wrapped as Web services (WS) as a way of “democratization” of language technologies. Three of the best known language infrastructures are HoG [1], Language Grid [2] and WebLicht [3]. Such WS provide access to various language services like machine translation or analyzers. Other services allow access to resources for example semantic resources like WordNet-type semantic Lexicons [4].

For all of these platforms, Web servicization of Natural Language Processing (NLP) tools and resources allow:

- easy access of LRT for non-experts through a Web 2.0 interface;
- interoperability between different tools and resources [5];
- development of Web applications, for example eLearning.

This paper reports our experience integrating Fips, a multilingual parser [6, 7] into a ReSTful Web service in order to provide access to multilingual and multi-format resources.

Section 2 describes the Fips Parser integrated in the WS. Section 3 describes the characteristics of the WS and the available data format. Finally, section 4 is devoted to the creation of a Web application based on the WS.
II. The Fips Parser

Fips is a robust multilingual parser which is based on generative grammar concepts for its linguistic component and object-oriented design for its implementation. It uses a bottom up parsing algorithm with parallel treatment of alternatives, as well as heuristics to rank alternatives.

The syntactic structures built by Fips are all of the same pattern, that is: $[XP \ L \ X \ R]$, where L stands for the “possibly empty” list of left constituents, X for the (possibly empty) head of the phrase and R for the (possibly empty) list of right constituents. The possible values for X are the usual parts of speech Adverb, Adjective, Noun, Determiner, Verb, Tense, Preposition, Complementizer, Interjection.

The parser makes use of 3 fundamental mechanisms: projection, merge and move.

Fips builds enriched phrase-structure representations for a sentence, in which extraposed elements (relative pronouns, clitics, interrogative phrases, etc.) are coindexed with empty constituents in canonical positions (i.e., argument or adjunct positions). For instance, the sentence in (1) below is assigned the syntactic structure in (2), in which the canonical position of object for the verb address is taken by the empty constituent $e$. The latter stands for the trace of the noun issue, which has been extraposed through relativization. The trace $e$, the relative pronoun $\phi$ (a zero-pronoun), and the noun issue are all linked via the index $i$.

1. This too is an issue the Convention must address
2. $[TP [DP This] [VP [AdvP too] is [DP an[NP issue][CP[DP\phi]] [TP [DP the [NP Convention]] must [VP address [DEPE]]]]]]$

III. Characteristics of the Web Service

III.1 A ReSTful Web Service

We implemented the system as a "Representational State Transfer" (ReST) Web service because of its simplicity and an environment clearly focused on the World Wide Web [8]. Whereas the SOAP approach uses full-blown remote objects with remote method invocation and encapsulated functionality, ReST deals with existing well-known standards (HTTP, XML, URI and MIME) and needs only infrastructure that has already become ordinary [9].

The ReSTful approach is basically composed of four concepts [10] :

- the use of the "Uniform Interface" : all resources can be manipulated using the HTTP protocol and the method : PUT, GET, HEAD, POST, DELETE;
- the identification of resources via URI (Uniform Resource Identifier) : each resource can be uniquely identified and addressed;
- the operation are stateless : the client is responsible for sending all the information required for a successful method execution.
the use of standards like HTML, XML or JSON.

III.2 Access to resources – Examples

The Data representation is done using standards (UTF-8, XML, etc.). Applications can be operated directly by URI. Tables 1, 2 and 3 describe resources and request parameters\(^1\). Figure 1 gives an example of request and response in French of the Analyze resource.

<table>
<thead>
<tr>
<th>Table 1. Available Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Resources</strong></td>
</tr>
<tr>
<td>Analyze</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Request parameters for Analyze</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameters</strong></td>
</tr>
<tr>
<td>ln</td>
</tr>
<tr>
<td>text</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3. Request header for Analyze</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Header</strong></td>
</tr>
<tr>
<td>Accept</td>
</tr>
</tbody>
</table>


\(^2\)Currently, baseURI is : http://129.194.19.89
III.3 Data formats

The Fips Web service provides five output formats: Parser – a condensed X-bar representation of the syntactic structure of the sentence, XML – a full X-bar structure in XML format, Tagger – a standard part of speech tag output, XML TEI and PASSAGE. The two latter formats are described below.

III.3.1 Tei

The Text Encoding Initiative (TEI) was established in 1987 as a constantly maintained XML standard for encoding textual data [11]. TEI P5 is a very rich toolbox proposing several modules to conceptualize textual data. The TEI P5 analysis module3 provides the way to represent syntactic constituency. In Figure 4, the WS build an XML tree of the input sentence with the element <s>(entence), <phr>(ase) and <w>(ord)4. Attributes can contain other information like:

- `@lemma` provides a lemma for the word, such as an uninflected dictionary entry form
- `@type` provides additional interpretative information about the category (for example Part-of-Speech)

---


The PASSAGE syntactic representation was designed for eponymous evaluation campaigns of French parsers [12]. It has evolved over the years until its current version PASSAGE 2 [13]. Since it is primarily intended for the evaluation of parsers which rely on a wide range of linguistic theories, it was defined as the minimal common ground to reasonably describe linguistic phenomena without presupposing any specific theory. PASSAGE addresses the syntactic aspect of the language, more precisely the identification of constituents and the outline of the syntactic relations between constituents. With respect to these properties we can say that this representation comes relatively close to dependency structures. Even if it was created on the basis of French grammar, it is suitable for syntactic schemes of other languages, possibly with some extensions or adaptations. In short, even if it is not the ultimate syntactic representation, it is a reasonable one and suitable for a large range of applications. Last but not least, as all the major parsers of French participated to the PASSAGE campaign, this format became a standard in the French research community.

For all these reasons we included this format in the Fips Web service. A parsed sentence in PASSAGE representation is in XML. The output is formed of:

- the individual words of the sentence and their automatically assigned IDs; these IDs are used to refer to the words in the constituents and syntactic relations;
- the syntactic constituents and their IDs; the lemmas;
- the syntactic (binary) relations

For instance, the sentence

3. Paul eats an apple

is represented in PASSAGE as

```xml
<Sentence id="E1">
  <T id="E1W0T0" start="1" end="5">Paul</T>
  <T id="E1W1T0" start="6" end="10">eats</T>
  <T id="E1W2T0" start="11" end="13">an</T>
  <T id="E1W3T0" start="14" end="19">apple</T>
  <G id="E1G0" type="GN">
    <W id="E1W0" tokens="E1W0T0" pos="properNoun" lemma="Paul" />
  </G>
  <G id="E1G1" type="NV">
    <W id="E1W1" tokens="E1W1T0" pos="verb" lemma="eat" />
  </G>
</Sentence>
```

5 The authors of PASSAGE choose XML format in order to easily process the outputs of the parsers for evaluation. Of course this format is also adequate for a web service. A DTD can be found on the Web site of PASSAGE [http://atoll.inria.fr/passage/documents.fr.html](http://atoll.inria.fr/passage/documents.fr.html)
III.4 Evaluation of Fips

The Fips multilingual parser has been developed for 6 languages (English, French, German, Italian, Spanish and Greek). A comparative evaluation has been conducted to show how the various implementations of Fips compare with respect to a near identical corpus, the European Parliament corpus [14]. We parsed approximately 1 million words in each language (300'000 for Greek). The Table 4 shows the results:

<table>
<thead>
<tr>
<th>Language</th>
<th>German</th>
<th>English</th>
<th>Spanish</th>
<th>French</th>
<th>Greek</th>
<th>Italian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of symbols</td>
<td>1106559</td>
<td>1075246</td>
<td>1228240</td>
<td>1350522</td>
<td>343461</td>
<td>1181785</td>
</tr>
<tr>
<td>Unknown words</td>
<td>10685</td>
<td>5852</td>
<td>9165</td>
<td>5643</td>
<td>6788</td>
<td>9006</td>
</tr>
<tr>
<td>Number of sentences</td>
<td>47058</td>
<td>41488</td>
<td>46216</td>
<td>45694</td>
<td>13328</td>
<td>44124</td>
</tr>
<tr>
<td>% of complete analyses</td>
<td>66,54%</td>
<td>75,68%</td>
<td>71,4%</td>
<td>75,97%</td>
<td>51%</td>
<td>70,76%</td>
</tr>
<tr>
<td>Speed (word/second)</td>
<td>17</td>
<td>38</td>
<td>132</td>
<td>83</td>
<td>196</td>
<td>112</td>
</tr>
</tbody>
</table>

IV. Web application based on the Web Service

IV.1 FipsColor : an eLearning application based on the Web Service

FipsColor is an example of a Web application based on the Fips Web service⁶ [15]. The Fips parser maps a sentence into a syntactic structure reflecting lexical, grammatical and thematic information. The FipsColor displays such structures with color representations.

⁶ We request the Fips Web service using the Client URL (cURL) library of PHP. Then, we use an XSLT stylesheet to transform the result in HTML.
This tool is used to highlight particular aspects of French, such as the lexical ambiguity of certain words. This is illustrated in Figure 2. In this example, “est” is a verb (“is”) and a noun (“east”). The different colors allow to clearly distinguish the correct grammatical category. FipsColor also highlights the grammatical functions.

This online interactive tool has been developed for primary education and can be used freely.

![Figure 2. Screenshot of the Web application FipsColor](image)

V. Conclusion and Further Work

This paper presents a ReSTful Web service, which aims at increasing the accessibility and usability of multilingual and multi-format language resources developed at LATL. We believe that these services can be useful not only for researchers in linguistics, but also for other disciplines which rely on syntactically analyzed text. This Service also enables the creation of Web Applications like FipsColor.

In a future version of the Web service, we will add other languages: Romanian, Japanese and Portuguese are under development. We plan to include Semantic Web technology in order to provide search and reasoning as well.
References


