

# Robust language analysis components for practical applications

Timo Järvinen, Mikko Laari, Timo Lahtinen, Sirkku Paajanen, Pirkko Paljakka,  
Mirkka Soininen and Pasi Tapanainen

Connexor Oy  
Koetilantie 3  
00790 Helsinki  
Finland  
info@connexor.com

## Abstract

We present here robust language analysis components used in AthosMail, a multilingual spoken dialogue system for reading of e-mail messages. The components are originally developed for written text and adapted to the speech data and dialogue modelling.

## 1 Introduction

Connexor has specialized in making language analysers for text. During the past years we have invested in multilinguality, and currently we have morphological analysers and syntactic parsers available for ten languages. In addition to multilinguality, our development is focussing on the usability of the analysers by importing new domains, adding customisability and offering more suitable information for different applications.

The Dumas project offered us the opportunity to apply our core technology for the domain of speech processing in an adaptive multilingual environment. In the course of the project, we customized the selected parsers to more informal text types, even to transcribed speech data. Though not directly involved with speech processing or dialogue modelling, we received valuable feedback on a new and challenging application domain from the project.

## 2 Language components for AthosMail

Connexor provided three separate external modules to the text pre-processing component of the AthosMail (Turunen & al., 2004; Jokinen and Gambäck, 2004) application:

- Language identification program,
- Sentence segmentation component, and
- Syntactic parsers for English, Swedish and Finnish.

### 2.1 Language Identification

Connexor's text classification program `cnxlangid` is used for language identification. The program receives a chunk of text and labels it with an

appropriate language code. According to the evaluation in (Gambäck & al., 2003), the language recognition precision rates vary between 96-99.5% and recall 98.5-99.4% for the three AthosMail languages.

### 2.2 Sentence Segmentation and Preprocessing

Connexor's syntactic parser contains a language independent sentence preprocessing and segmentation component which can be run independently. The preprocessing module recognises or guesses various text encoding systems (such as HTML, email, text corpus conventions in various corpora) and transforms them into fixed notation. The segmentation module recognises various conventions for headers in text, sentence boundaries and paragraphs. It cleans the input text, and marks sentences and paragraphs in the cleaned output. The component is extracted originally from the Connexor's FDG parser version 3.7.

In AthosMail, the task of the component is to identify the sentence boundaries within paragraphs. The program adds XML codes at sentence boundaries.

### 2.3 Connexor FDG parsers for English, Swedish and Finnish

Connexor has created text parsers for various languages. The parsers provide morphological, syntactic and semantic information in various levels. Connexor's lexicons and grammars are based on linguistic generalisations and rules. Texts of hundreds of millions of words have been used in testing and further improving the performance of these analysers. The approach is very robust; the parsers are capable of producing analysis of any input, whether well-formed sentence, sentence fragment or just a single word-token. Basically, the reliability of the analysis improves considerably when some context is provided.

In the Dumas project, we worked with the Connexor FDG 3.7 syntactic parsers for English, Swedish and Finnish, and customised them for

spoken language data and for analysing email messages.

The dependency-based model (Functional Dependency Grammar, FDG) was first introduced in (Tapanainen and Järvinen, 1997). It is a direct ancestor of the current commercial product, developed further in Connexor, Machine Syntax<sup>1</sup>, which has proved to be suitable for analysing a variety of languages. Currently, ten languages are supported (English, French, Spanish, German, Swedish, Finnish, Italian, Dutch, Danish and Norwegian) by Connexor.

The latest work in FDG or Machine model has focussed on expanding the limits of this approach, and development of a new model of presentation that is able to represent different linguistic facts in an even more illustrative way.

The programs have been written in the C programming language.

The Connexor FDG used in the Athosmail prototype produces output in Extensible Markup Language (XML) format. This is the default output. The option --dtd prompts the Connexor FDG document type definition (dtd) to the standard output.

The DTD consists of the following definition:

```
<!DOCTYPE fdg:analysis [
  <!ELEMENT analysis (sentence |
  paragraph)* >
  <!ELEMENT paragraph EMPTY >
  <!ELEMENT sentence (token*) >
  <!ATTLIST sentence id ID
  #REQUIRED>
  <!ELEMENT token (text, lemma,
  (depend)?, (tags)*) >
  <!ATTLIST token id ID #REQUIRED>
  <!ELEMENT text (#PCDATA) >
  <!ELEMENT lemma (#PCDATA) >
  <!ELEMENT depend (#PCDATA) >
  <!ATTLIST depend head IDREF
  #REQUIRED>
  <!ELEMENT tags ((syntax)*,
  morpho)* >
  <!ELEMENT syntax (#PCDATA) >
  <!ELEMENT morpho (#PCDATA) >
  ]>
```

Due to the parallel development of the new Connexor Machine Language Model, the analysers for different languages use similar principles of analysis and the tagsets are highly compatible between different languages.

The FDG output consists of sets of five tab-separated fields, illustrated by the analysis of the sentence ‘Do I have any messages from Bob?’ below. The fields in Figure 1 from the left to the right are (i) word position, (ii) text token, (iii) base form, (iv) function with a dependency name and

the number of the head, and (v) tags for morphology and syntax. The morphological tags consist of part of speech labels and subfeatures. Surface syntactic tags indicate phrase-level entities. Functional dependency tags indicate explicit dependency relations between words. It is possible to convert the functional tags as a dependency tree where the nodes correspond to the FDG tokens and the arcs to the functional labels.

1	Do	do	v-ch:>3	@+FAUXV	V PRES
2	I	i	subj:>1	@SUBJ	PRON PERS SG1
3	have	have	main:>0	@-FMAINV	V INF
4	any	any	det:>5	@DN>	N DET
5	messages	message	obj:>3	@OBJ	N NOM PL
6	from	from	mod:>5	@<NOM	PREP
7	Bob	bob	pcomp:>6	@<P	N NOM SG +ind
8	?	?			

Figure 1. A dependency analysis

Tag	Explanation	Examples
+org	Organization, Company	Red Cross, IBM
+loc	Location	USA, Stockholm
+ind	Individual	John, Boris Yeltsin
+name	Name	Beatles
+role	Occupation, Title	managing director, Mister

Table 2. Named entity classification

1	zlenko	subj>2	@NH N UTR SG Indef NOM +ind
2	säga	main>0	@MAIN V ACT IND PAST
3	att	pm>7	@PREMARK CS
4	medlem#skap	subj>7	@NH N UTR SG Indef NOM
5	i	mod>4	@POSTMOD PREP
6	pfp	pcomp>5	@NH N UTR SG Indef NOM +org
7	komma	obj>2	@MAIN V ACT IND PRES
8	att	pm>9	@PREMARK PREP
9	ge	obj>7	@MAIN V ACT INF
10	ukraina	dat>9	@NH N NEU SG NOM +loc
11	vidsträckt	attr>12	@PREMOD A Def
12	utsikt	obj>9	@NH N UTR PL Indef NOM
13	till	advl>9	@PREMARK PREP
14	militär	attr>15	@PREMOD A NEU SG Indef
15	samarbete	pcomp>13	@NH N NEU SG Indef NOM
16	.		

Figure 3. Swedish analysis

<sup>1</sup> <http://www.connexor.com/demos/>

The FDG parsers for English and Swedish utilized in the Dumas project are special versions with the named entity recognition and classification components. We adapted, tested and evaluated the named entity recognition and classification components during the Dumas project, and developed the corresponding components for Swedish and Finnish.

Table 2 shows the named entity classification scheme used in this project. The proper nouns are classified into categories: organisations, locations, individuals and unspecified name as a default category. Roles of individuals such as occupation or title are also recognised.

Consider the analysis of the Swedish sentence: “Zlenko sade att medlemskap i PFP kommer att ge Ukraina vidsträckt utsikter till militärt samarbete” in Figure 3. The names Zlenko, PFP and Ukraina are tagged as +ind, +org and +loc, respectively.

### 3 Corpus Annotation

From our point of view, there is a real challenge in providing informative analyses of e-mails, which often are very informal. However, some informal text types, including transcribed speech, have already been taken into account when developing earlier versions of the analysers (Järvinen, 2003a).

We contributed to the definition of the morphosyntactic coding scheme and finally the manually transcribed WOZ corpus was annotated morphosyntactically by Connexor's tools. The annotation result was stored as the value of the feature FDG in the Annotation Graph (Black & al., 2002).

Our experiences from annotating speech transcriptions show that ideally, various speech phenomena can be tolerated with a small amount of adaptation of the parser originally developed for analysis of standard written language. For example, the analysis of the transcribed utterance below “er, er how many messages do we have in the inbox?” depicts a complete parse tree capturing all the propositional elements, while some elements pertaining to the delivery such as false starts and hesitations are unattached, though recognised as interjections or as sentence adverbials. Therefore, the resulting analysis is compatible with more specific annotation levels presently not fully automatic such as dialogue modelling.

1	er		@DUMMY %EH INTERJ
2	,		
3	er		@DUMMY %EH INTERJ
4	how	How	ad:>5 @AD-A> %E> ADV WH
5	many	many	qn:>6 @QN> %>N DET ABS PL
6	messages	message	obj:>9 @OBJ %NH N NOM PL
7	do	do	v-ch:>9 @+FAUXV %AUX V PRES
8	we	we	subj:>7 @SUBJ %NH PRON PERS NOM PL1
9	have	have	main:>0 @-FMAINV %VA V INF
10	in	in	loc:>9 @ADVL %EH PREP
11	the	the	det:>12 @DN> %>N DET
12	inbox	inbox	pcomp:>10 @>P %NH Heur N NOM SG
13	?	?	

Figure 4. Speech data analysis

### 4 Conclusion

In the DUMAS project we have investigated the adaptability of our technology to a new domain of multilingual speech processing. The evaluation and exploitation of the results is underway. The process of harmonisation of Connexor's linguistic descriptions was completed parallel to the DUMAS project. The project thus benefited not only of the latest, improved versions of the parsers, but of the new Machine language model, that used uniform tagsets and descriptive principles for a variety of languages.

The software modules delivered to the project were tested intensively using large text corpora at Connexor prior to the delivery to the project partners.

One of the evaluation methods listed in the project plan is cross-language multimodal information retrieval, as one of the possible application areas of the techniques developed in DUMAS. Connexor and SICS are jointly participating in CLEF 2004 Cross-Language System Evaluation campaign, in which Connexor's analysers are used for linguistic annotation.

Note that Connexor's tools are used by many researchers and research projects worldwide, including domains such as information extraction, multimodality and machine translation.

The FDG parser has been applied to information extraction by (Yangarber, 2000). The name

recognition components included in the English parser version, utilised in DUMAS, were developed initially for the project BRIEFS (Brief Driven Information Retrieval and Extraction for Strategy)<sup>2</sup>, (Keijola, 2003).

Connexor's parsers have been used in multimodal applications even prior to DUMAS. The Behavior Expression Animation Toolkit (BEAT)<sup>3</sup> developed at MIT, use the FDG parser. Also, (Ma and Kevitt, 2004) have experimented the parser for a multimodal environment.

The English parser has been adapted for Machine Translation purposes in the EU-funded project MLIS-5008 LINGMACHINE<sup>4</sup>, which produced a new branch for the parser, Connexor Machine Semantics (for details, see (Järvinen, 2003b)). The MATADOR<sup>5</sup> Machine Translation project is exploiting the Spanish parser.

Connexor has started commercialising the Dumas project results as the components of the multilingual Machine product family.

## 5 Acknowledgements

Our thanks go to all people involved.

## References

- William J. Black et al. 2002. *Report of the Corpus Analysis*. DUMAS Deliverable D1.2., UMIST, Manchester, England, May.
- Björn Gambäck et al. 2003. *Prototype Text Processor*. DUMAS Deliverable D6.2, SICS, Kista, Sweden, October.
- Kristiina Jokinen and Björn Gambäck, 2004. *DUMAS - Adaptation and Robust Information Processing for Mobile Speech Interfaces*. Proceedings of The 1st Baltic Conference "Human Language Technologies - The Baltic Perspective", Riga, Latvia, pp. 115-120.
- Timo Järvinen. 2003a. *Bank of English and beyond - hand-crafted parsers for functional annotation*. In Anne Abeillé, editor, *Treebanks - Building and Using Parsed Corpora*, pages 43-59. Kluwer, Dordrecht.
- Timo Järvinen. 2003b. *Multi-layered annotation scheme for treebank annotation*. In Joakim Nivre and Erhard Hinrichs, editors, *TLT 2003*. Proceedings of the Second Workshop on Treebanks and Linguistic Theories, pages 93-104, Växjö. Växjö University Press.
- Matti Keijola. 2003. *On Smart and Natural Language Technology Support of Strategy Work*. Academic Dissertation. Helsinki University of Technology, Espoo, Finland.
- Minhua Ma and Paul Mc Kevitt. 2004. *Interval relations in visual semantics of verbs*. Special Issue of Artificial Intelligence and Cognitive Science 2003 (AICS-03), Artificial Intelligence Review.
- Pasi Tapanainen and Timo Järvinen. 1997. *A non-projective dependency parser*. In Proceedings of the 5th Conference on Applied Natural Language Processing, Washington, D.C., pages 64-71, Washington, D.C., April. Association for Computational Linguistics.
- M. Turunen, E-P. Salonen, M. Hartikainen, J. Hakulinen, W.J. Black, A. Ramsay, A. Funk, A. Conroy, P. Thompson, M. Stairmand, K. Jokinen, J. Rissanen, K. Kanto, A. Kerminen, B. Gambäck, M. Cheadle, F. Olsson, M. Sahlgren, 2004. *AthosMail - a multilingual Adaptive Spoken Dialogue System for E-mail Domain*. Proceedings of the COLING Workshop Robust and Adaptive Information Processing for Mobile Speech Interfaces, Geneva, Switzerland.
- Roman Yangarber 2000. *Scenario Customisation for Information Extraction*. PhD Thesis, Department of Computer Science, Courant Institute of Mathematical Sciences, New York University.

---

<sup>2</sup> <http://briefs.cs.hut.fi/>

<sup>3</sup> <http://gn.www.media.mit.edu/groups/gn/projects/beat/>

<sup>4</sup> <http://www.ling.helsinki.fi/kieliteknologia/tutkimus/lingmachine/>

<sup>5</sup> <http://clipdemos.umiacs.umd.edu/matador/main.html>