Getting Started with ROSANA

Roland Stuckardt
D-60433 Frankfurt am Main, Germany
e-mail: roland@stuckardt.de

10th April 2003

1 Synopsis

This document briefly describes how to install and use the ROSANA System (Version 5.1, October 2002) and the sample data (texts, parses, keys) under a Linux environment. ROSANA should work under other operating systems as well, provided that Common Lisp is available.

2 Licensing conditions

ROSANA is made available on condition that the user unconditionally accepts the "**License Agreement for ROSANA**" as stated in appendix A of the document at hand. In using ROSANA or any component that comes along with the ROSANA distribution, the user acknowledges that he/she has read through and understood the License Agreement for ROSANA and unconditionally accepts it.

3 Prerequisites

- Common Lisp (e.g., Xanalys Lispworks 4.2.0 for Linux, or Allegro Common Lisp of Franz, Inc.)
- a reasonably equipped PC (500 MHz, 128 MB RAM should do)
- basic knowledge of Lisp
- in-depth knowledge of anaphor resolution and the respective formal evaluation issues
- the willingness as well as the ability to cope with an experimental software system for which virtually no documentation is available

4 Installation

1. Unpack the gzipped distribution file in a suitably chosen installation directory (subsequently referred to as `$ROSANAPATH`):

   ```
   > tar xvzf ROSANA51.tgz
   ```

   ROSANA 5.1 unpacks to two subdirectories: `$ROSANAPATH/ROSANA51/System`, which contains the Common Lisp sources, and `$ROSANAPATH/ROSANA51/Data`, which contains the sample data.

2. Change to the source code directory:

   ```
   > cd ROSANA51/System
   ```

3. Adaptation of the installation path: in file *schnittstellen-definitionen.lisp*, change the constant ROSANAHOME according to the context of the local installation.
   E.g. if `$ROSANAPATH = ~/CommonLisp` then modify the constant definition to

   ```
   (defconstant ROSANAHOME "~/CommonLisp/ROSANA51")
   ```

4. Starting your Common Lisp environment: in case you are using Xanalys Lispworks 4.2.0 for Linux, type

   ```
   > lispworks-4200 &
   ```

5. Compiling the Common Lisp sources of ROSANA: load the file *compile.lisp* in a Lisp listener; this should compile the sources and generate the respective binaries (ufsl, fasl, depending on the Common Lisp environment you use):

   ```
   CL-USER 1 > (load "compile")
   ```

5 Using ROSANA under Xanalys Lispworks

If you are using Xanalys Lispworks 4.2.0 for Linux, a simple graphical interface is provided.

1. *Starting the graphical version of ROSANA:* load the file *capi-rosana.lisp*:

   ```
   CL-USER 2 > (load "capi-rosana")
   ```

   During first-time use, a copyright and licensing message will be displayed. Please read carefully through the "**License Agreement for ROSANA**" that comes along with this distribution (cf. appendix A). Upon acceptance of the licensing conditions, ROSANA will be loaded and initialized, and a frame titled ROSANA will be opened which displays the ROSANA listener. The ROSANA frame offers some basic means to process the sample data and to modify the processing and trace settings of ROSANA.

2. *Processing sample data:* in the `Resolve` menu, select one of the sample data suites, e.g. `PressReleases [32-66]`. This will paste a function call into the ROSANA-listener buffer. Execute this call in the ROSANA listener:

```
SEM 1 > (PressReleases-32-66)
```

This invokes the anaphor resolution process. The results are written to the respective directory of the selected sample data. In the above case, the name of the directory is `$ROSANAPATH/ROSANA51/Data/PressReleases` .
Alternatively, one may use the anaphor resolution interface function *dpe-anares*, which makes available a more comprehensive set of options (cf. section 8).

3. *Browsing the generated output files:*

   (a) Switch to the directory of the processed sample input suite:

   ```
   > cd $ROSANAPATH/ROSANA51/Data/PressReleases/
   ```

   (b) Examine the various output files which, as far as the respective output options are selected (cf. section 10), have been generated by ROSANA:

   - *pr.ref*: depending on the selected options, this file displays the anaphor resolution output and/or the traces;
   - *pr.ana*: this file contains the coreference classes as identified by ROSANA (used for scoring purposes);
   - *pr.sco*: depending on the selected scoring options, this file displays the evaluation results of ROSANA in various evaluation disciplines (occurrences, coreference classes, immediate antecedents, lexical anchors).

   Cf. section 7 for a comprehensive list of the sample data file types.

4. *Experimenting with the output settings:* you may experiment with the different trace options offered in the `AR Output` menu. These settings determine the comprehensiveness of the generated output (antecedents, candidates, salience computation, syntactic disjoint reference verification etc), as written to the *.ref* file. (Cf. section 9 for a description of the most important options.)

5. *Experimenting with the processing settings:* further options are offered in the `Processing Traces` and `Processing Settings` menus. Perhaps of use are the following `Processing Traces` options:

   - whether the scoring results are written to a *.sco* file (default) or to the listener buffer,
   - whether the anaphor resolution output is written to a *.ref* file (default) or to the listener buffer,
   - whether a more detailed output regarding the scoring of
     - coreference classes
     - occurrences
     - lexical anchors

     will be generated.

   Cf. sections 10 and 11 for further details.

6 Using ROSANA under other Lisp environments

ROSANA should work under any Common Lisp Environment. In fact, it was originally developed under Allegro Common Lisp for Linux of Franz, Inc, and virtually no modifications were necessary to make it run under Xanalys Lispworks. Under environments other than Xanalys Lispworks, no graphical user interface will be available. The above described functionality, however, is still fully available. It may be accessed through function calls issued directly in the Lisp listener.

1. *Starting the listener version of ROSANA:* load the file *rosana.lisp*:

   ```
   CL-USER 2 > (load "rosana")
   ```

   During first-time use, a copyright and licensing message will be displayed. Please read carefully through the "**License Agreement for ROSANA**" that comes along with this distribution (cf. appendix A). Upon acceptance of the licensing conditions, the binaries will be loaded, and ROSANA will be initialized to some default settings regarding the anaphor resolution output, the processing traces, and the processing options.

2. *Change to Package* `SEM`*:*

   ```
   CL-USER 3 > (in-package "SEM")
   ```

3. *Processing sample data:* anaphor resolution may be performed by issuing one of the sample calls (as output by the help function *(help-ar)*). E.g. the predefined call

   ```
   SEM 4 > (PressReleases-32-66)
   ```

   starts ROSANA on documents 32 to 66 of the Press Releases corpus. (During implementation of the ROSANA system, documents 1 to 31 were used for training, and documents 32 to 66 served as the test set.) Further predefined calls are *(PressReleases-1-31)*, *(PressReleases-1-66)* (i.e. anaphor resolution on the complete corpus), and *(Mozart-1-3)*. Alternatively, one may use the interface function *dpe-anares*, which makes available a more comprehensive set of options (cf. section 8).

4. *Browsing the generated output files:* as described in section 5.

5. *Experimenting with the output, trace, and processing settings:* these settings may be changed by directly typing in the function calls `(configure-ar-output)`, `(configure-processing-traces)`, and `(configure-processing-settings)`. E.g.

   ```
   SEM 5 > (configure-ar-output)
   ```

   allows the user to interactively select among the set of available anaphor resolution output options.

6. *Getting help:* the available sample calls and configuration options might be redisplayed by typing

   ```
   SEM 6 > (help-ar)
   ```

7 About the sample data

ROSANA comes with a set of sample corpora (document collections), comprising ASCII versions of the texts, the respective parses (as determined by the robust parser of Timo Järvinen and Pasi Tapanainen ([1])), the respective key data (coreference annotations), and some supplementary information. At current, two (copy-righted!) corpora are provided:

- Press Releases: 66 documents, 24,712 words,
- Mozart Operas: 3 documents, 2,522 words.

The different kinds of knowledge are distinguished through filename suffixes. Files used as input to anaphor resolution and evaluation/scoring are:

- *filename.txt:* the numbered collection of source documents;
- *filename.par:* parsing results as determined by the Dependency Parser of English ([1], commercially distributed by Conexor Oy, Helsinki);
- *filename.pos:* token-numbered parsing results;
- *filename.vof:* inflected forms for which morphological information is needed;
- *filename.mor:* morphological information as determined by EngTwoL (English two-level morphology, commercially distributed by Lingsoft, Helsinki);
- *filename.ngp:* supplementary (general and corpus specific) lexical information;
- *filename.key:* intellectually gathered key data (coreference classes, referred to during evaluation);
- *filename.tag:* intellectually gathered key data (coreference classes, immediately annotated in *.pos* file, synchronized with (i.e. referentially equivalent to) *filename.key*).

Files generated during anaphor resolution and scoring are:

- *filename.ana:* the coreference classes as identified by ROSANA (referred to during scoring);
- *filename.ref:* anaphor resolution output as determined by the settings in the `AR Output` menu (only available if the processing traces option *Write AR Output to .ref File* has been chosen);
- *filename.sco:* depending on the selected scoring options, this file displays the evaluation results of ROSANA in various evaluation disciplines (as far as selected: occurrences, coreference classes, immediate antecedents, lexical anchors) (only available if the processing traces option *Write Scoring Results to .sco File* has been chosen).

8 Calling ROSANA manually

Instead of using one of the predefined shortcuts, anaphor resolution might instead be started manually, having available a more comprehensive set of options. A typical call of the interface function *dpe-anares* looks as follows:

```
SEM 7 > (dpe-anares
         "PressReleases/pr.pos"     ;; (1) the (parsed) document collection
         nil                        ;; (2) non-default .mor file (optional)
         "PressReleases/pr.key"     ;; (3) key data to be used for scoring
         (dnum-nach-dname '(32 66)) ;; (4) specificaton of a document sub-
                                    ;;     set to be processed (optional)
```

The most important parameters are: parameter *(1)*, which specifies the (parsed) document collection to be processed; parameter *(3)*, which specifies the key data file to be used for scoring (if set to *nil*, no scoring will take place); parameter *(4)*, which selects for resolution a subset of documents of the collection (if not provided or set to *nil*, *all* documents of the collection will be processed). Parameter *(4)* refers to the document number tags that are provided in the *.pos* files (e.g. *"% DOKUMENT 1"*); the function *dnum-nach-dname* serves as a tool for generating respective tag reference lists for the documents to be selected, e.g.

- by specifying *(dnum-nach-dname '(32 66))*, documents with numbers 32 to 66 (i.e. an interval) will be processed;
- by specifying *(dnum-nach-dname nil '(1 7 18))*, exactly the documents with numbers 1, 7, and 18 will be processed.

A further parameter (*(5)*, not shown in the above example) allows to explicitly specify the data directory prefix under which it is looked for the *.pos* and *.key* files, and in which the result data is written. If ROSANAHOME has been set to `$ROSANAPATH/ROSANA51` (cf. section 4), this directory prefix defaults to `$ROSANAPATH/ROSANA51/Data`.

9 Anaphor resolution output options

In the graphical user interface of the Xanalys Lispworks version of ROSANA, the lower part of the `AR Output` menu provides a number of options through which the anaphor resolution output might be configured according to the specific needs. The most important options are:

- *Candidates:* displaying the list of ranked candidates among which the antecedent occurrence is selected;
- *Salience:* displaying information about the computation and assignment of salience / preference weights;
- *Antecedents:* displaying the selected antecedent occurrence;
- *Syntactic Disjoint Reference:* displaying details regarding the robust verification of the syntactic disjoint reference conditions (binding principles A, B, C);

- *Parse Tree Results:* displaying the (possibly fragmentary) surface-syntactic information that constitutes, in particular, the base for the verification of the syntactic disjoint reference conditions;

- *Compact Output:* if selected, occurrences (anaphors, candidates, antecedents) will be displayed in a compact format.

The upper part of the menu provides shortcuts to several useful combinations of the above atomic options.
The *configure-ar-output* function of the non-grphical interface makes available an even more comprehensive set of combined, non-exclusive (i.e. combinable) output offers.

10 Processing trace options

Some options that might be of particular use are:

- *Progress of Processing:* if selected, progress information will be displayed in the ROSANA listener buffer;

- *Verbose Coreference Class Scoring:* if selected, the scoring output will provide detailed information regarding the coreference classes scoring (illustrating the model-theoretic coreference scoring scheme of Vilain et al. ([4]) and its extension by Stuckardt ([3]));

- *Verbose Occurrence Scoring:* if selected, the scoring output will provide detailed information regarding the scoring of the identified occurrencs;

- *Verbose Lexical Anchor Scoring:* if selected, the scoring output will provide detailed information regarding the scoring of the lexical anchors that have been determined for pronominal anaphors;

- *Scoring of Immediate Antecedents:* if selected, the performance with respect to the identification of immediate (possibly pronominal) antecedents will be scored as well;

- *Write Scoring Results to .sco File:* if selected, scoring results will be written to a *.sco* file (otherwise they will be output to the ROSANA listener buffer);

- *Write AR Output to .ref File:* if selected, the anaphor resolution output will be written to a *.ref* file (otherwise it will be output to the ROSANA listener buffer).

11 General processing settings

Some options that perhaps may be of use are:

- *Reading Morphosyntactical Information from .mor File:* if selected, morphosyntactical information given in the *.mor* file will be used; otherwise, it is referred to the morphosyntactical information provided by the parser;
  (Definitely, this issue should be documented in more detail. By now, it should suffice

to keep in mind that this option ought to be selected when processing the Mozart data, and unselected when processing the PressReleases data. When using one of the sample cases offered in the `Resolve` *menu of the Xanalys Lispworks ROSANA frame, or when issuing one of the predefined calls of the Common Lisp version of ROSANA, this is automatically taken care of.)*

- *Apposition as Coreference Trigger:* (recommended) if selected, appositive relations will be interpreted as syntactic coreference triggers;
- *Coreference Class Scoring on Correct Occurrences only:* (strongly recommended) if selected, model-thoretic coreference class scoring will be confined to the subset of correctly identified occurrences, i.e. to occurrences in the intersection set of system output and key; through this, the two evaluation disciplines are mutually decoupled in order to avoid certain cases of scoring anomaly (cf. [3, 2]);
- *Determining Lexical Anchors for Pronouns:* if selected, ROSANA tries to determine lexical anchors (i.e. occurrences providing conceptual content, e.g. common nouns or names) for pronominal anaphors; this discipline might be paraphrased as anaphor resolution proper.

References

[1] Timo Järvinen and Pasi Tapanainen. *A Dependency Parser for English.* Technical Report TR-1, Department of General Linguistics, University of Helsinki, 1997.

[2] Roland Stuckardt. *Qualitative Inhaltsanalyse durch Computer - ein uneinlösbarer Anspruch? Untersuchungen zur algorithmischen Textinhaltserschließung am Beispiel der referentiellen Interpretation*, Ph.D. dissertation, Department of Social Sciences, Johann Wolfgang Goethe University Frankfurt am Main. Also: Tenea-Verlag, Berlin, 2000.

[3] Roland Stuckardt. *Design and Enhanced Evaluation of a Robust Anaphor Resolution Algorithm*, In: *Computational Linguistics* 27(4), 2001, 479-506.

[4] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman, *A Model-Theoretic Coreference Scoring Scheme.* In: *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufmann, San Francisco, 1996, 45-52.

A License Agreement for ROSANA

This agreement is made and entered into as of _____ by and between the parties:

a)
Dr. Roland Stuckardt
Im Mellsig 25
D-60433 Frankfurt am Main
Germany

referred to as: "**LICENSER**"

b)

[The user who gained access to ROSANA immediately through LICENSER]

referred to as: "**USER**"

A.1 Product

ROSANA is the Common Lisp implementation of a robust syntactic salience-based anaphor resolution algorithm.

ROSANA comes with two (copy-righted!) sample document collections, comprising ASCII versions of the texts, the respective parses (as determined by the robust parser of Timo Järvinen and Pasi Tapanainen), the respective key data (coreference annotations), and some supplementary information. An evaluation module is included which scores the output of ROSANA according to various evaluation disciplines.

A.2 Copyright

Copyright of ROSANA is to Dr. Roland Stuckardt.

A.3 License

LICENSER grants USER a non-exclusive license to use ROSANA, version as distributed. USER agrees to use ROSANA only for non-commercial, non-profit research purposes; to report changes that USER makes to the programs to LICENSER; and to acknowledge the use of ROSANA in all publications reporting on results produced with the help of ROSANA. Use of ROSANA or products derived from ROSANA for any commercial purposes requires explicit written agreement of LICENSER.

A.4 Non-Disclosure

ROSANA will be held in confidence by USER and will not be disclosed by USER to third parties.
USER shall and will employ all necessary precautions to ensure that no persons or institutions other than persons as are in the employ of USER or in the same research project as USER will get access to ROSANA or parts thereof. Other persons or institutions desiring access to ROSANA should be directed to LICENSER to obtain separate license agreements.

A.5 Fee

This license is granted by LICENSER to USER free of charge.

A.6 Disclaimer

ROSANA and its documentation is provided on an "as is" basis, with no guarantee of its veracity or accuracy. No liability is accepted for any damage caused by its use.