ANALEC: a New Tool for the Dynamic Annotation of Textual Data

Frédéric Landragin, Thierry Poibeau and Bernard Victorri

LATTICE-CNRS

École Normale Supérieure & Université Paris 3-Sorbonne Nouvelle

1 rue Maurice Arnoux, F-92120 Montrouge, France

E-mail: frederic.landragin@ens.fr, thierry.poibeau@ens.fr, bernard.victorri@ens.fr

Abstract

We introduce ANALEC, a tool which aim is to bring together corpus annotation, visualization and query management. Our main idea is to provide a unified and dynamic way of annotating textual data. ANALEC allows researchers to dynamically build their own annotation scheme and use the possibilities of scheme revision, data querying and graphical visualization during the annotation process. Each query result can be visualized using a graphical representation that puts forward a set of annotations that can be directly corrected or completed. Text annotation is then considered as a cyclic process. We show that statistics like frequencies and correlations make it possible to verify annotated data on the fly during the annotation. In this paper we introduce the annotation functionalities of ANALEC, some of the annotated data visualization functionalities, and three statistical modules: frequency, correlation and geometrical representations. Some examples dealing with reference and coreference annotation illustrate the main contributions of ANALEC.

Keywords: cyclic annotation, visualization, data query
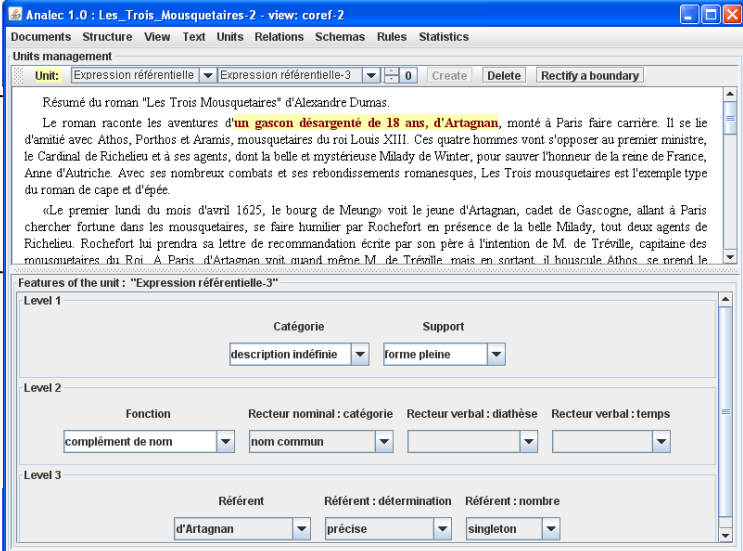
1. Introduction

Corpus annotation is often described as a linear process in which one defines an annotation scheme, annotates data according to this scheme and then exploits these annotations for linguistic studies. Most annotation tools are developed taking this assumption as granted and do not provide any facilities to adapt the annotation scheme, which is often necessary during the annotation process. Contrary to this assumption, we think that designing an annotation scheme is a major issue that requires a 'trial and error' strategy.

the scheme is too complex and time consuming for practical applications). In this situation, researchers and annotators need versatile tools, allowing them to refine the annotation scheme, rename or merge some features, create new features, annotate a new part of the corpus or add new annotations to an already partially annotated corpus, etc.

It is thus clear that data annotation is not a static process but a dynamic one, which depends on corpus visualization and data queries. ANALEC (Victorri, 2011) [1] is a new annotation tool: its goal is to allow researchers to dynamically build their own annotation scheme, using the

Corpus visualization zone

Corpus annotation zone: In this example, the features are classified using three levels (some features can be edited, others not depending on the 'view').

Menu bar:
Document: management of corpus (loading, merging annotations, concatenating several corpus, etc.).
Structure: management of the annotation structure.
View: management of data accessibility.
Units, Relations, Schemas: management of annotated data.
Rules: not implemented yet.
Statistics: computations.

Figure 1: Annotating one unit with several annotation levels.

Our experience shows that corpus annotation is cyclic and requires to gradually refine the annotation scheme, either because an intermediate version of it is not complete or because it is not tractable (this is typically the case when

possibilities of partial annotation, dynamic scheme revision, data querying and visualization during the

---

[1] Based on the architecture of ANALOR (Avanzi et al., 2008).

annotation procedure itself. Automatic annotation (i.e. annotations produced by other tools like part of speech taggers) can also be integrated to help corpus annotation. In this paper, we have chosen to present the main functionalities of ANALEC taking co-reference annotation as an example. In the four sections of this paper we will show how to annotate a corpus using ANALEC, how to visualize and update the annotated data, how to search for correlations, and how to exploit geometrical representations to help finding trends and specificities in the corpus. We discuss the differences between ANALEC and other annotation tools before reaching the conclusion of the paper.

the screen shot).

For co-reference annotation, referring expressions are defined as units, and co-reference chains as schemas. This is the result of several attempts to annotate these objects using different strategies (for example, modeling co-reference chains as a series of relations between text units is possible but not as clear as schemas, especially for subsequent calculations). So, while defining our annotation scheme, we had to delete some features, add new ones, rename or merge some others, etc. All this is possible with ANALEC, which implements various control mechanisms (merging two different fields may erase previous information or may create a conflict

This window allows the user to navigate in the annotation scheme.

Each type, each element and each value of an element can be renamed and deleted.

When deleting, ANALEC checks the presence of annotated data and informs the user whenever necessary. When renaming, ANALEC checks the presence of an already existing item with the same name. In this case, renaming implicates merging annotations. Merging two elements or two values is done in a dynamic and ergonomic manner.
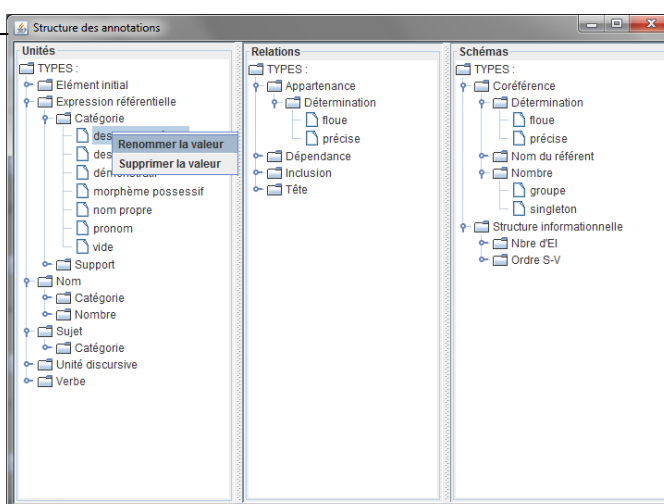
The three parts of this window correspond to the three kinds of annotations: units, relations, and schemas.

In this example, there are 6 types of units, 4 types of relations, and 2 types of schemas.

Figure 2: The dynamic management of the annotation scheme.

2. Annotating text using ANALEC

ANALEC is an open source piece of software dedicated to text annotation (Figure 1). More precisely, it is dedicated to manual annotation and is therefore especially useful for semantic and pragmatic annotation, where automatic tools are still not accurate enough. All the examples we take in this paper deal with the annotation of semantic roles of some discourse units or noun phrases. Some of them deal with referring phenomena, and consist in the annotation of referential expressions and of links between several referential expressions.

There are three main objects that can be manipulated using ANALEC: units, i.e. chunks of texts, relations between text units and schemas to group complex sets of units and relations. Figure 2 highlights these three objects: the first column corresponds to 'units', i.e., text chunks that are marked and enriched with annotations; the second one to 'relations', i.e., links between units (note that annotations can be added to each link); the third one to 'schemas', i.e., sets of units, relations and possibly other schemas, which can also include specific annotations. Some features depend on other features (for ex., tense and mood are relevant only for verbs, not for other linguistic units): ANALEC makes it possible to define different annotation levels, some levels depending on others (on Figure 1, the different levels are visible at the bottom of

between two different possible values, ANALEC detects these kinds of problems and interact with the user whenever necessary).

Lastly, annotated corpora are often difficult to visualize due to the high number of annotations, most of the time piled the ones on top of the others. In order to solve this issue, we have imagined the notion of 'view'. For the same corpus and the same set of annotations, several 'points of view' on the data can be defined, in order to make pieces of information appear or not in the annotation interface (bottom part of Figure 1). Views can be saved in separated files, as well as the annotation schemes. Moreover, some features may be locked (e.g. for non modifiable properties of a view).

3. Annotation visualizations with ANALEC

In traditional annotation tools, units are usually emphasized with colours, links are materialized by arrows between units, and schemes (if they exist) are represented by boxes or chains that include all related units and relations. ANALEC makes it possible to get new visualization modes that fit better with the kind of data one wants to observe.

For example, in the case of co-reference chains, one wants to be able to immediately observe the succession of referring units, along with their referent. Figures 3, 4 and 5 provide three different views of chained referring units,
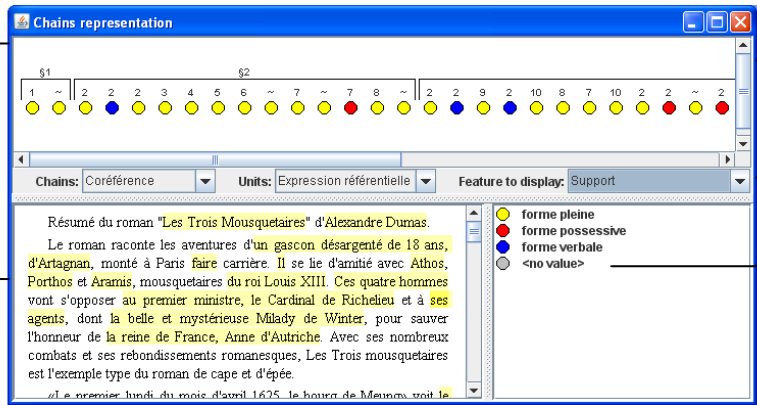
Figure 3: Linear representation of references in a text.

their link with text and their structure in paragraphs. We are currently developing tools to exploit this information and automatically extract specific patterns from these data: to the best of our knowledge, this has never been done on a large scale before and would be very difficult and time consuming, if it had it to be done manually.

The use of graphs like the one in Figure 4 is interesting because points, groups of points and colours are immediately interpretable. This kind of figures can reveal very quickly some trends in the annotated data that several tables of numbers will not reveal so easily.
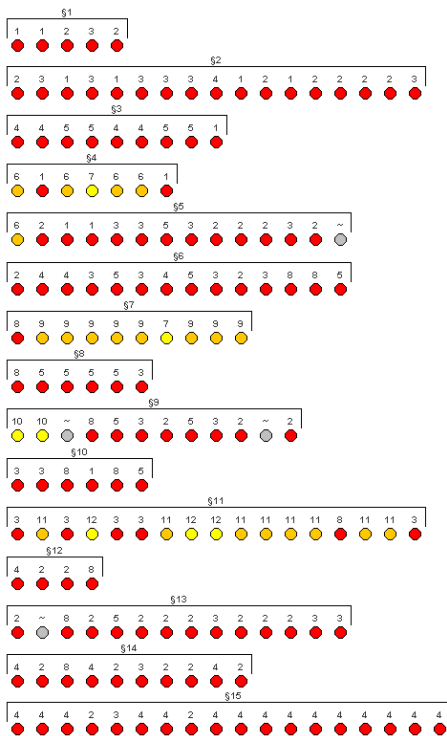

Figure 4: Representation of references per paragraph.

Whereas the colours in Figure 3 were linked to one of the features of isolated referring units, in Figure 4 they correspond to the length of reference chains and provide some information on the way the characters are mentioned in each section of the text (red corresponds to a co-reference chain with more than 10 elements, orange to a chain with 4 to 10 elements, yellow to a chain with 2 or 3 elements, and grey to singletons).

Figure 5 corresponds to a short story from Maupassant where all referring expressions have been annotated and grouped into reference chains. The colours are here linked to the types of referring expressions: pink for a proper name, yellow for a possessive determiner or pronoun, and so on.

4. Searching for correlations between features

ANALEC includes three kinds of computation and visualization modules to analyse annotated data: frequencies, correlations, and geometrical representations. Most annotation tools provide frequency information, so we will rather focus on correlations and geographical relations here. Identifying correlations is automatic in ANALEC using the correlations window introduced in Figure 6. The user can choose a unit and two different features, and ANALEC automatically computes a table displaying their correlations. The cells in the table are coloured following the result of a classical chi-squared test. The red-coloured cells are then identified as significant, and the user may have a look at them. The examples corresponding to one cell are displayed when one clicks on the cell (see the concordance at the bottom of Figure 6) [2].

5. Geometrical representations of annotated data

Lastly, ANALEC provides advanced graphical visualization tools based on various kinds of correspondence analyses. The idea is to offer graphical representations of annotated data, where units sharing similar contexts appear together, while other ones appear isolated. It makes it easy for the linguist to identify canonical examples, more interesting ones, and rare events (here again, inspecting these examples may help to discover either interesting facts or inconsistencies in the annotation).

For example in Figure 7, each point in the graph is linked

---

[2] This is especially useful to correct inconsistencies in the annotation, making the tool being really dynamic.

Personnages :
1. Mr. Lantin
2. Mme. Lantin
3. Mère de Mme. Lantin
4. Groupe incluant Mr. et Mme. Lantin
5. Premier bijoutier
6. Commis du premier bijoutier
7. Deuxième bijoutier
8. Flâneurs
9. Deuxième épouse de Mr. Lantin

Dét Possessif
GN Défini
GN Démonstratif
GN Indéfini
GN Possessif
N ou GN sans dét
Nom Propre
Pron Démonstratif
Pron Indéfini
Pron Interrogatif
Pron Pers Anaphorique
Pron Pers Déictique
Pron Relatif
Forme atténuée

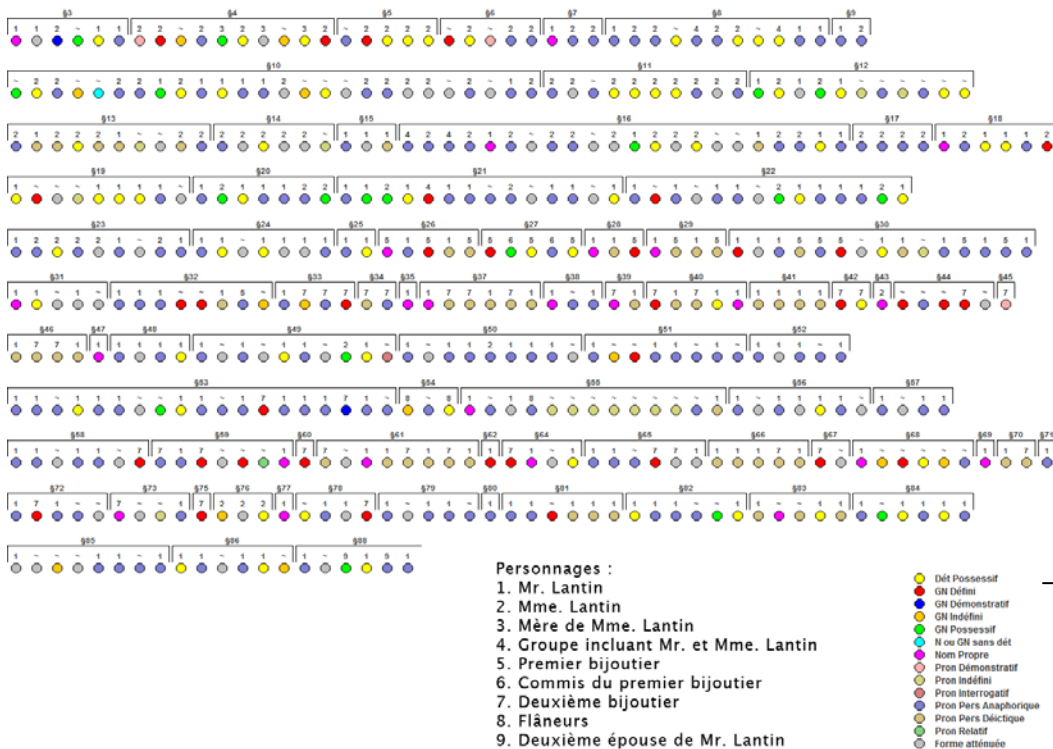Captions : number code (referents) and colour code (types of referring expressions).

Figure 5: The complete references chain of a short story.

to a linguistic unit, and some interesting groupings may appear depending on the features taken into account. This module is especially useful to check that all the elements of a certain type are grouped together or not. Like with the correlations window (and the frequency window), each click on a point or a group of points updates the concordance field at the bottom of the window.

This module can also be used to observe the relative weight of different features during the analysis of linguistic phenomena. For example, one can observe on Figure 8 that grammatical functions and thematic roles of referring expressions are somehow correlated in the annotated corpus. Although this is not new, ANALEC makes it possible to observe and prove linguistic hypotheses in vivo, directly from the observation of the data.
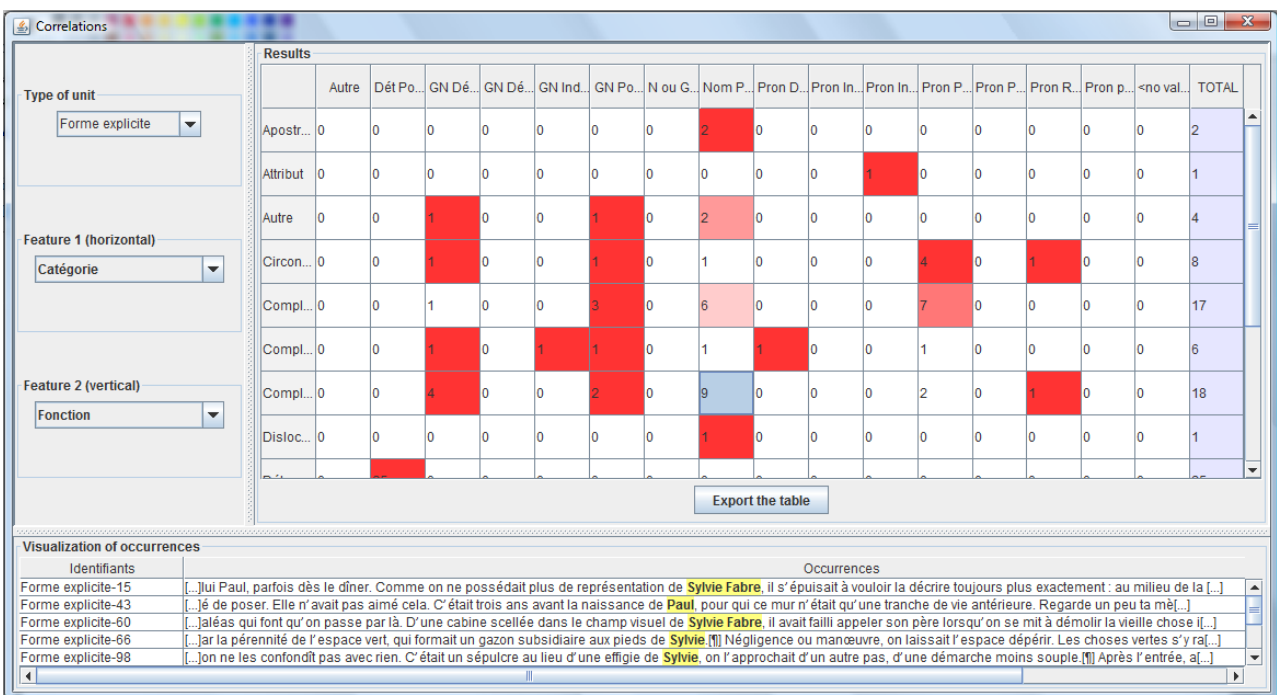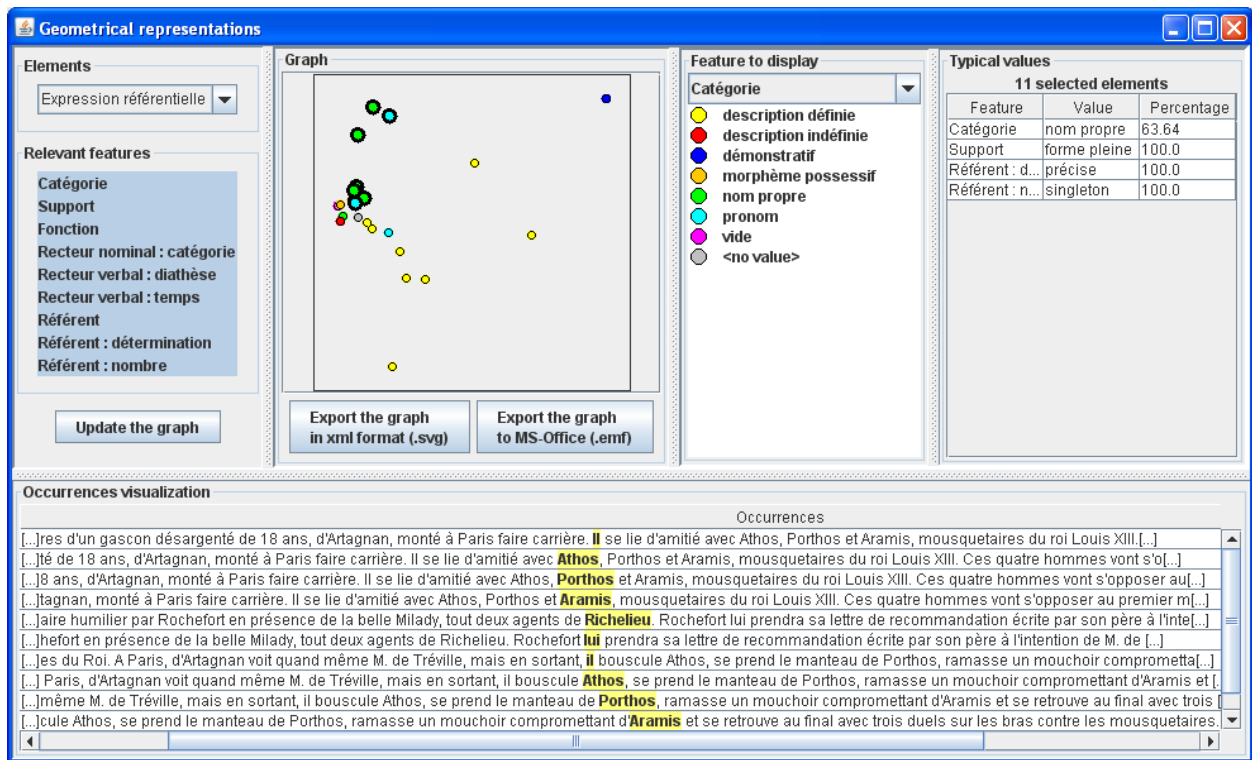

Figure 6: ANALEC correlation window.

360

Figure 7: ANALEC geometrical representation window.

6. Comparisons with previous works

It is possible to find various annotation tools that can be compared with ANALEC, but none implements the same range of functionalities in a compact piece of software. We especially examined MMAX (Müller & Strube, 2001), WordFreak (Morton & Lacivita, 2003), Glozz (Mathet & Widlöcher, 2009; 2011) and Gate (Cunningham *et al.*, 2002), among several others. None of these platforms offers the easy-to-use, integrated and visual environment implemented in ANALEC. They all require to a priori define the annotation scheme, using an XML file most of the time. With ANALEC, even if all the information is encoded using XML, graphical interfaces make it possible to define a model without for the end user the need to know XML or even notions such as document type definition, XML elements or labelled tree. Our experiments with various groups of linguists have proven that people who are sometimes not at ease with computers can easily use this tool. This is not the case with a platform like Gate, which includes a wide range of facilities for text annotation as well, but is much more complex to use for non experts.

7. Conclusion

We have introduced a new annotation tool called ANALEC that includes various facilities for the manual and semi-automatic annotation of corpora. This tool is already used by dozens of end-users for various corpus-based linguistics studies (among other, co-reference resolution, discourse structure analysis,

lexical semantics of prepositions, etc.). It is an open source piece of software, the external format used for annotation is XML to ensure the compatibility with various other tools and platform, including Glozz that provides complementary modules for visualization (esp. for relations). Graphical interfaces make it possible to use the tool even without having any knowledge of XML or related technologies. ANALEC is developed using Java and J2EE so that it is multi-platform. It is available for free at: http://www.lattice.cnrs.fr/Analec.

Future work concerns the development of new facilities to help the integration of ANALEC into other tools. We are currently studying the possibility to deliver ANALEC as a plug-in so that this integration would be relatively straightforward.

8. Acknowledgements

9. References

Avanzi, M., Lacheret-Dujour, A., Victorri, B. (2008). ANALOR, A Tool for Semi-Automatic Annotation of French Prosodic Structure, In *Proceedings of the 4th Conference on Speech Prosody*, Campinas, Brazil,
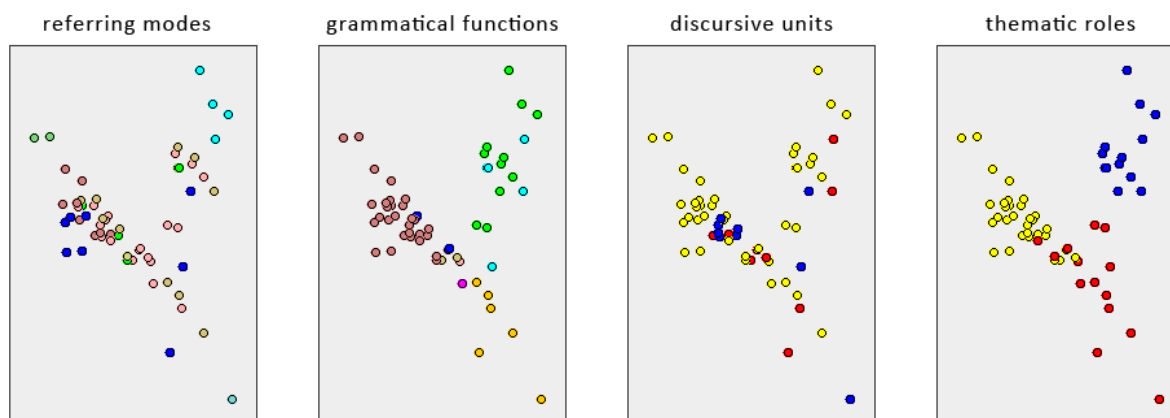
Figure 8: Compared geometrical representations for four properties.

pp. 119−122.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, pp. 168−175.

Dipper, S., Götze, M., Stede, M. (2004). Simple Annotation Tools for Complex Annotation Tasks: an Evaluation. In *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora*, Lisbon, Portugal, pp. 54−62.

Mathet, Y., Widlöcher, A. (2011). Glozz Annotation Platform, http://www.glozz.org/

Morton, T., LaCivita, J. (2003). WordFreak: An Open Tool for Linguistic Annotation. In *Proceedings of Human Language Technology (HLT) and North American Chapter of the Association for Computational Linguistics (NAACL)*, Edmonton, pp. 17−18.

Müller, C., Strube, M. (2001). MMAX: A tool for the Annotation of Multi-modal Corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Seattle, pp. 45−50.

Victorri, B. (2011). ANALEC download Web page, UUhttp://www.lattice.cnrs.fr/Analec.

Widlöcher, A., Mathet, Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN)*, Senlis, France.