

# English Terminology in CLAT

Michael Carl<sup>1</sup>, Maryline Hernandez<sup>1</sup>, Susanne Preuß<sup>1</sup> and Chantal Enguehard<sup>2</sup>

<sup>1</sup>Institut für Angewandte Informationsforschung  
Martin-Luther-Str. 14, Saarbrücken, Germany  
{carl,maryline,susannep}@iai.uni-sb.de

<sup>2</sup>Institut de Recherche en Informatique de Nantes  
2, rue de la Houssinière, 44322 Nantes, Cedex 3 France  
Chantal Enguehard@irin.univ-nantes.fr

## Abstract

CLAT (Controlled Language Authoring Technology) is a tool that supports authors in the production of technical documents. Much work is currently invested to enhance the linguistic components for the English version of CLAT. In this paper we give a short overview of CLAT. We introduce the terminology tool, its underlying technology and describe the graphical interface in CLAT. The paper summarizes an experiment showing figures of precision and recall and discusses future developments.

## 1. Introduction

CLAT (Controlled Language Authoring Technology) is a tool designed to support technical authors in producing and revising technical documents in various domains. The German version of CLAT (i.e. MULTILINT) is most advanced and already in use in a number of companies (Haller et al., 2002)<sup>1</sup>. Much work is currently invested in its English version to enhance the checking possibilities for grammar and terminology.

In this paper we report on ongoing work to extend the terminology component of CLAT for English. In order to effectively check compliance to corporate language requirements, the CLAT term base encodes information concerning the status of the terms: preferred form, admitted form or deprecated form. Moreover, on the basis of morpho-syntactic analysis and lemmatisation, the tool is able to detect typographical and derivational variants in texts.

Using an abductive mechanism (Carl et al., 2002; Carl et al., 2004) the existing terminology is extended with term variation templates. These templates are produced by combining general knowledge of variation patterns with the terminological entries. Term templates can be further enriched with synonymy relations. When checking terminological consistency, CLAT matches a document against the extended terminology and marks the detected terms or term variants. In case the status of the matched entry is other than preferred or admitted, the term variant or the deprecated term is highlighted and the author is provided

with a message and the preferred base form.

While this technique is described in detail in (Carl et al., 2002; Carl et al., 2004), this paper presents the graphical interface of CLAT. In section 2. we give an overview over CLAT and its GUI. Section 3. discusses the terminology component in more detail. Section 4. outlines an experiment and section 5. discusses the results.

## 2. Controlled-Language Authoring Technology

Controlled-Language Authoring Technology (CLAT) has been developed to suit the need of some companies to automatically check their technical texts for general language and corporate language conventions. Within CLAT, texts are checked with respect to:

- orthographic correctness
- company specific terminology and abbreviations
- general and company specific grammatical correctness
- stylistic correctness according to general and company specific requirements

The orthographic control examines texts for spelling mistakes and proposes alternative writings. The terminology component matches the text against a terminology and abbreviation database where also term variants are detected. The grammar control

---

<sup>1</sup>see also <http://iai.uni-sb.de/iaide/de/clat.htm>

checks the text for grammatical correctness and disambiguates multiple readings while stylistic control detects stylistic weaknesses.

The components build up on each other's output. Their modularity is suited to adapt to different texts and requirements. Besides the described control mechanisms, CLAT also has a graphical front-end as shown in figure 1. The lower part in figure 1 plots an input paragraph while the upper part shows the automatically annotated paragraph with mistakes highlighted. The document structure is shown in the left part of the window. A user can revise a document by clicking through the paragraph symbols. The linguistic engine works in the background performing morphological analysis, lemmatisation, terminology checking, shallow parsing and style control while the GUI marks erroneous segments in the texts with different colors.

The user can click on one of the automatically annotated errors in the upper part of the window to display an explanatory message in the middle part of the screen. A separation is made between mistakes on the word level, on a grammatical level and on a stylistic level. The user can switch between the different levels of analysis by clicking on the appropriate buttons. The text in the lower part can also be edited, re-checked and stored.

### 3. The Terminology Tool in CLAT

As a general rule and to enhance readability, terms in technical documents should be used consistently and in accordance with an authorized terminology. However, people often use different linguistic forms to name the same thing. To cope with this problem, CLAT tries to anticipate and detect variants of terms. In section 3.1. we give a background of the implementation while section 3.2. shows a segment of the graphical interface.

#### 3.1. Architecture

There are basically two ways of matching a candidate variant in a document (or in a list of terms) onto a list of authorized terms: a database approach and a run-time approach:

1. in the run-time approach, a candidate sequence of words in the document undergoes a number of transformations which map it onto an authorized term. The original sequence in the document is then marked as a variant of the authorized retrieved term.
2. in the database approach a limited number of possible variants are generated for each authorized

term. The variants are stored in a database with a link to their authorized terms. A matched sequence of words in the document is marked as a variant of the term from which the database entry was generated.

CLAT's terminology tool implements a database approach. The database approach has the drawback that all possible variants which the tool is supposed to recognize are need be generated and stored in a database. However, since CLAT can store underspecified variants the size of the base only marginally increases compared to the gain of coverage. The outstanding advantage of the database approach is, however, *log-time* retrieval.

The run-time solution transforms and maps a candidate sequence of words in a document onto its authorized forms in the database. Time required for this mapping increases linearly in time with every possible transformation that the sequence in the document undergoes. Jacquemin's *FASTR* (Jacquemin, 2001) implements the run-time approach. Using a set of metarules, Jacquemin remains below this linear limit.

CLAT's terminology tool integrates a rule-based approach and an example-based approach. Rules are used to generate variation templates from authorized terms which are stored in a database. Rules are also used to consolidate the findings of the matching process. The technique underlying this process is described in-depth in (Carl et al., 2004; Carl et al., 2002).

#### 3.2. Graphical Interface

Figure 1 highlights mistakes on the word level. Two types of mistakes are differentiated on the word level: either the word (or word form) cannot be analyzed or is unknown to the system, or it is detected a deprecated form or a term variant.

The word "martensite" is unknown to the system. It is not in the system's dictionary nor in the terminology database. By clicking on one of the instances, the user is prompted the following message in the middle window:

This word is unknown. Does it contain a spelling mistake?

The user can click on an ignore button for the word to be admitted in the paragraph or for the entire document.

There is a further occurrence of "martensite" within the compound noun "rate of martensite reaction". This compound is detected a permutation variant of the term "Reaction rate". The user is prompted the message:

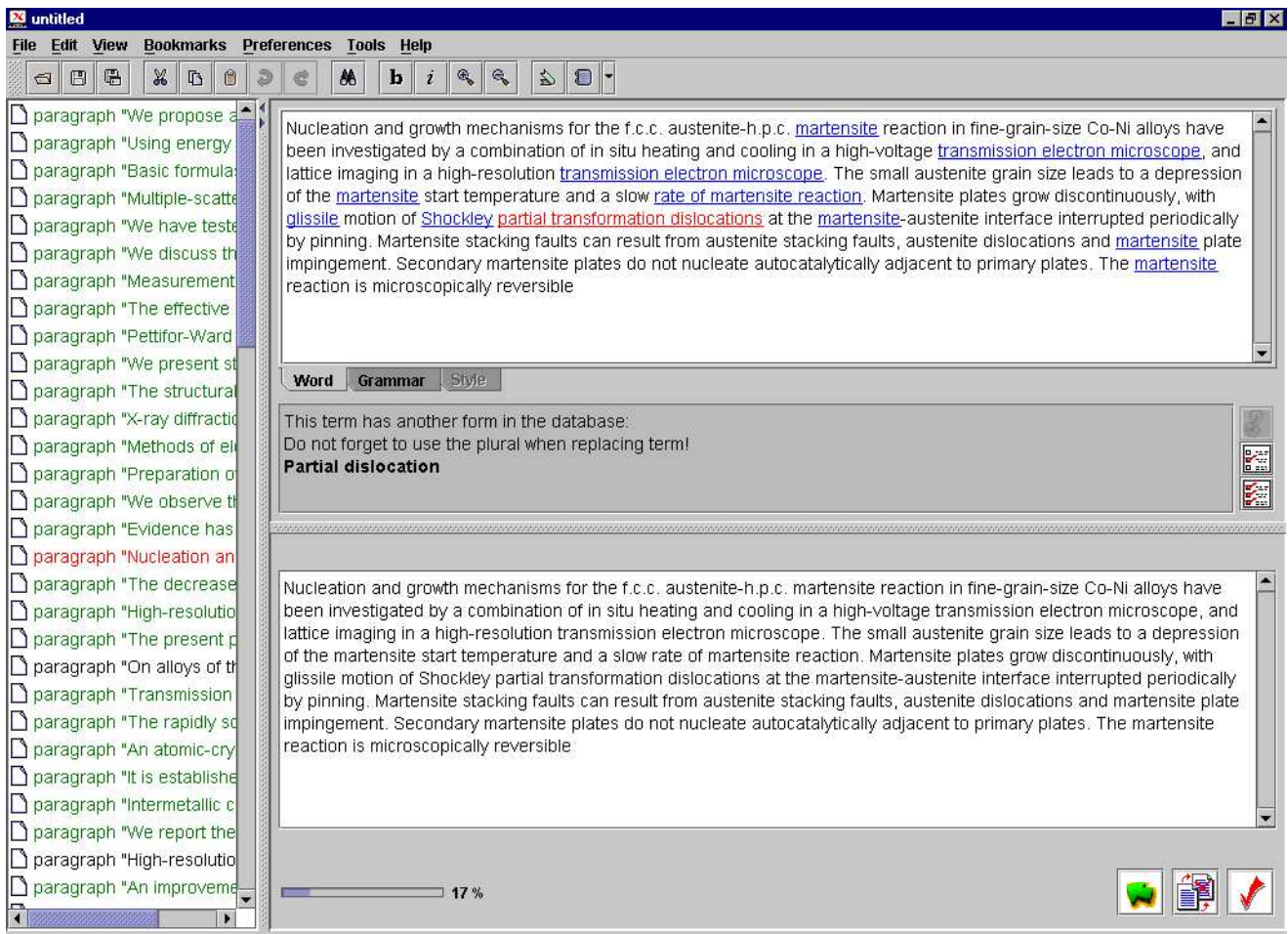


Figure 1: Terminology Checking in CLAT

This term has another form in the database:  
**Reaction rate**

Two occurrences of “transmission electron microscope” are detected as derivation variants of “Transmission electron microscopy”. By clicking on one of the instances, the following message appears in the middle window:

This term has another form in the database:  
**Transmission electron microscopy**

Due to linguistic analysis (Maas, 1996; Maas, 1995), CLAT also detects different inflections in the authorized term and the variant occurrences in the text. For instance “partial transformation dislocations” is detected an insertion variant of the authorized term “Partial dislocation”. By clicking on the marked variant in the upper window, the following message appears in the middle of the screen:

This term has another form in the database:  
 Do not forget to use plural when replacing term  
**Partial dislocation**

Since the variant in the text shows a different agreement in number than the authorized form in the database, the author is hinted to adjust to plural when replacing the term.

Some variants can be traced back to a number of authorized terms. The term “state structure”, for instance, is the head of two authorized term, “Liquid state structure” and “Amorphous state structure”. The author is prompted the following message:

This term has another form in the database:  
**Liquid state structure**  
**Amorphous state structure**

#### 4. Experiment

In an experiment we evaluated and compared the terminology tool based on a collection of 1280 abstracts on the chemistry of metals. Terms and variants of terms in the abstracts were manually annotated in a previous project (Enguehard, 2003). A terminology database was provided containing 6602 authorized base terms.

	<i>Fastr</i>	Syrete	<i>T</i>	<i>TV</i>
Recall	64.63	70.44	66.43	68.42
Precision	89.05	98.81	96.70	94.33

Table 1: Recall and Precision for Syrete, FASTR and CLAT

The abstracts were automatically annotated using three different term recognition systems: Syrete<sup>2</sup>, *FASTR* (Jacquemin, 2001) and CLAT’s terminology tool. Two versions of the CLAT’s terminology tool were used. In the version *T* only the lexemes of the 6602 authorized base terms were indexed. The version *TV* contained the 6602 terms plus 22367 variation templates. These variation templates were generated from 12 variation patterns to detect reduction, insertion and permutation variants (see (Carl et al., 2004) for a similar experiment) so that for each base term on average four variation templates were indexed. Results of the four experiment in terms of precision and recall are shown in table 1.

## 5. Discussion

Despite the excellent results in table 1 for all four settings, a number of open questions remains.

For example, the compound noun “rate of martensite reaction” was annotated a variant of “Reaction rate” in the test text by a specialist in the domain (Enguehard, 2003). However, variants built by several variation mechanisms such as permutation and insertion are not appropriate in all cases. The expression “selectivity of surface processes” is explicitly marked a non-variant of “process selection”. It is, however, unclear what the underlying processes are.

Also “transmission electron microscope” is annotated a variant of “Transmission electron microscopy” in the test text and recovered as such from all four systems. While both compounds built on the same succession of lemmas, i.e. “transmit”, “electron” and “microscope”, they differ in their head words “microscope” vs. “microscopy”. While the former is a thing, the latter a science. Whether such constructions are variants in all instances is doubtful.

It is certainly less doubt if morphological and/or derivational variation occurs in the non-head of the compound. For example, all four systems found “atom displacements” to be a variant of “atomic displacement”. It is likely that this variation process will be much more reliable than derivational variation of the term’s head word.

More in depth investigation is required to uncover similarities in the variation patterns and to examine

the underlying mechanisms of applicability. It is also interesting to see whether the same variation mechanisms apply in an industrial application or whether some templates can be excluded.

One of the major drawbacks to more sophisticated term variation recognition in the English version of CLAT is due to the order of linguistic processing. In the current English CLAT architecture, terms are checked and matched without previous grammatical analysis of the text.

Experience in the German terminology checking has shown that noise can be considerably reduced when word analyses are disambiguated and syntactic tagging on a phrase level has taken place previous to term matching. Term recognition over phrase or, worse, sentence boundaries could thus be excluded.

## 6. Conclusion

In this paper we have presented CLAT, the Controlled Language Authoring Technology and in particular its terminology component. We compare the performance of CLAT’s terminology tool in terms of precision and recall with two similar systems and discuss future development and research directions.

## 7. References

- Carl, Michael, Johann Haller, Christoph Horschmann, Dieter Maas, and Jörg Schütz, 2002. The TETRIS Terminology Tool. *TAL, Structuration de terminologie*, 43(1).
- Carl, Michael, Ecaterina Rascu, and Johann Haller, 2004. Using weighted abduction to align term variant translations in bilingual texts. In *Proceedings of LREC*.
- Enguehard, Chantal, 2003. CoRRecT : Démarche coopérative pour l’évaluation de systèmes de reconnaissance de termes. In *in Proceedings of TALN*.
- Haller, Johann, Horschmann Christoph, Rita Nübel, Ursula Reuther, and Axel Theofilidis, 2002. Sprachtechnologie im Einsatz. Terminologie - Workflow - Sprachkontrolle in der Technischen Dokumentation. Technical report, IAI.
- Jacquemin, Christian, 2001. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.
- Maas, Heinz-Dieter, 1995. Documentation of the features used in MPRO. Software documentation, IAI, Saarbrücken.
- Maas, Heinz-Dieter, 1996. MPRO - Ein System zur Analyse und Synthese deutscher Wörter. In Roland Hausser (ed.), *Linguistische Verifikation, Sprache und Information*. Tübingen: Max Niemeyer Verlag.

<sup>2</sup><http://www.sciences.univ-nantes.fr/info/perso/permanents/enguehard/>