A Hybrid Example-Based Approach for Detecting Terminological Variants in Documents and Lists of Terms

Michael Carl
RALI/Université de Montréal, Quebec, Canada
carl@iro.umontreal.ca

Johann Haller, Christoph Horschmann and **Axel Theofilidis**
Institut für Angewandte Informationsforschung, Martin-Luther-Straße 14,
66111 Saarbrücken, Germany
{hans;chris;axel}@iai.uni-sb.de

Abstract

The TETRIS project at the Institut für Angewandte Informationsforschung is a MULTILINT follow-up project in collaboration with BMW and a number of smaller German companies. TETRIS provides linguistically intelligent components for specification, production and maintenance of multilingual documents supporting authors for technical documentation and manuals as well as terminologists. The paper presents the TETRIS terminology tool. TETRIS has a prescriptive terminology tool which detects variants of terms in documents and in the lists of terms. We show that a data and example-based approach fruitfully complements a rule-based approach for the detection of terminological variants.

1 Introduction

In the past years, big companies have recognized the importance for creating and maintaining corporate terminology, also with the aim to achieve a distinguished corporate identity. Terminology tools which are integrated into commercial translation memories such as Trados MultiTerm or Star's TermStar are sometimes used to support the creation and maintenance of the terminology. Even though these tools offer easy possibilities to creating and extending the terminology, they provide little help in detecting variants or synonyms of terms in documents.

Since terminology is added to the databases as it occurs in the documents, such an approach could be described as *descriptive*. In fact, as these tools work in the context of translation memories, full-forms of terms and variants are memorized together with their translations without keeping track which form is to be preferred and which forms are legal or deprecated variants.

The main objective of these conventional terminology management tools is the retrieval of as many forms as possible from a source language document in order to have appropriate translations at hand during the translation process. The descriptive approach in terminology management and maintenance has emerged from an uncontrolled authoring and translation environment and is rooted in the way current TM technology works. Many translators (i.e non-specialists) need to deal with texts of which they have no prior knowledge. Such translators need to practice ad hoc terminology management for concepts which they are unfamiliar with. As Wright (Wright and Budin, 1997, p. 148) points out this can lead to disastrous translations.

In a controlled language scenario — such as in TETRIS — this approach is inappropriate and misplaced. TETRIS follows a conceptual approach where each concept is represented by one and only one authorized full form[1]. Any entry contained in the TETRIS terminology is assumed to be an authorized form. Based on these authorized forms, the TETRIS terminology tool (henceforth TTTT) automatically detects a number of variants and notifies authors and/or terminologists of this findings. Therefore, the TETRIS approach can be described as *prescriptive* as its purpose is to standardize and normalize the use of terms according to the authorized terminology (cf. (Wright and Budin, 1997, p. 329)). As every terminology, even a standardized one, undergoes modifications and extensions, TTTT also provides an environment to dynamically elaborate, update and modify a prescriptive terminology.

In this paper, we first outline the aims of TTTT, we give examples of variants detected in TTTT, show how this is helpful in elaborating a prescriptive terminology. In section 3 we discuss previous research on terminological variation. In section 4, we describe the architecture and processing strategies in TTTT. We show how TTTT integrates a rule-based and an example-based component.

2 Aims of TTTT

A tool for checking terms and their variants in documents is at best as good as the terminology it uses. Ideally, a terminology contains one term for each concept in the domain and each relevant concept is represented by one authorized term. In addition, a

[1] A possible relation between full forms and other occurrences like abbreviations, acronyms etc. is only partially implemented in the TETRIS authoring environment.

terminology may contain definitions and/or collocations of terms, a thesaurus, synonymous writings or negative forms of the terms which are to be avoided.

As terminologies grow it becomes increasingly difficult to ensure consistency of terms and a 1-to-1 mapping of linguistic forms and concepts. Conventional terminology tools offer little help in this respect, as they contain essentially collections of forms with little automatic support for detecting inconsistencies, variants and non-obvious redundancies.

The aim of TTTT is to overcome this shortcoming by detecting variants in lists of terms and documents. In this section we give examples of variants that TTTT detected in a list of 16,039 BMW terms; the same variants could also have been detected in documents. In this list, 1896 clusters of terms were found containing two or more forms. 1309 of these clusters contain synonyms, writing and derivational variants and 587 clusters syntactic variants, permutations and omissions. Ideally, each cluster should contain only one authorized form. TTTT suggests the terminologist to revise these entries. In case a cluster contains two (or more) forms which the terminologist actually considers not to be variants of each other, an appropriate flag can be set. Both forms are henceforth considered as distinctive forms denoting different concepts. In this way, TTTT helps a terminologist to conceptually stabilize a terminology, avoid "changes in aspects of his science" (cf. (Polanco et al., 1995)) and establish conventions in the domain.

2.1 Writing and derivational Variants

In the table below, the variants in cluster 1 differs in the use of lower case letters and upper case letters. Cluster 2 is writing variants where both forms on the left and right side differ with respect to a hyphen and an upper case letter. Cluster 3 differs in the use of upper case/lower case letters and the inflected adjective *halbharter* left while the non-inflected form *Halbhart* is used on other variant.

1. *integrierte Universal-Fernbedienung, Integrierte Universal-Fernbedienung*
2. *Getriebeseriennummer, Getriebe-Seriennummer*
3. *halbharter Schaumstoff, Halbhart-Schaumstoff*

2.2 Synonyms

The TETRIS term tool detects synonyms if the term is made up of two or more lexemes. Experience has shown that admitting synonyms for single lexeme terms generates too much noise. Similar to Hamond and Nazarenko (Hamon and Nazarenko, 2001, cf. section 3), TTTT detects synonyms in head and expansion of the compound. Variants in cluster 4 differ in their expansion *lagern halten befestigen aufhängen* while the head *gummi* is identical in all variants. The variants in cluster 5 differ in their head-noun *Einrichtung* vs. *Anlage* expansion *Abgaskontrolle* is identical. Cluster 6 contains variants which differ in both, the head *prüfen, messen, anzeigen* and the expansion *Vorrichtung, Einrichtung, Anlage, Einheit*

4. *Lagergummi, Haltegummi, Befestigungsgummi, Aufhängegummi*
5. *Abgaskontrolleinrichtung, Abgaskontrollanlage*
6. *Prüfvorrichtung, Prüfeinrichtung, Messanlage, Anzeigeneinheit*

2.3 Permutation Variation

Variation by permutation is a kind of syntactic variation. Permutation variation may occur in two directions: a noun phrase may be generated from a compound noun or a compound noun may be generated from a noun phrase. In the examples below, a compound noun is shown on the right side while its paraphrased noun phrase is on the left side. In the cluster 7, the head nouns *Minuspol* is percolated from its end position in the compound noun to the front position in the syntactic construction. In cluster 8 the expansion *Güte* on the left side is on the same time replaced by a synonym *Sicherung.*

7. *Minuspol der Batterie, Batterie-Minuspol*
8. *Sicherung der Güte, Qualitätssicherung*

2.4 Variation by Omission

The term tool also checks variants by omission for entries which have three or more lexemes. Cluster 9 has a three lexeme term and a two lexeme term where the 2 lexeme term is missing the middle component *Licht.* In addition the lexemes in the reduced variant are separated by a hyphen. Cluster 10 consists of a 4 lexeme term and a 3 lexeme term where the latter one is missing the component *Reihe.* Cluster 11 contains three variants. The head noun *Untersuchung* is replaced by the synonym *Kontrolle* and in cluster 12 both omission and synonym substitution occurs in parallel.

9. *Abblendlichtrelais, Abblend-Relais*
10. *4-Zylinder-Reihenmotor, 4-Zylinder-Motor*
11. *Abgassonderuntersuchung, Abgasuntersuchung, Abgaskontrolle*
12. *Luft-Saugventil, Belüftungsdüse*

3 Previous Research on Terminology Variation

A number of researchers have investigated terminological variation for different languages and with different means and goals. Royouté (Royauté et al., 1996) and Jacquemin (Jacquemin, 1996; Jacquemin, 2001) investigate a number of term variations for French:

- (variation by) inflection
  this subsumes derivational variation, such as *acoustic test / acoustic testing* and variation by number, such as *deficiency / deficiencies.*
- (variation by) insertion
  the inserted element modifies the head of the term as e.g. in *thin film* vs. *thin gold film* or the inserted element becomes the head of the term as in *quantum well structure* vs. *quantum structure.*
- (variation by) permutation
  this type of transformation concerns noun-noun structures such as *electron diffraction* which appears as a permuted form in *diffraction of fast electrons*
- (variation by) coordination
  adjectives (and nouns) can be coordinated as in *electron diffraction* vs. *electron and photoelectron diffraction*

Jacquemin finds that "clustering of terms related through coordinations yields classes of conceptually close terms while graphs resulting from insertion denote generic/specific relations." (Jacquemin, 1996, 425)

The research of Royauté is based on the notion of variation and stability of terminological noun phrases. The absence of variation can be interpreted as a sign of the conceptual stabilization of the term and provides a measure for "changes in aspects of sciences" (cf. (Polanco et al., 1995)):

> "While the stability of a term is a sign that the notion which it represents is becoming ordinary, its variation, on the contrary, often reveals a conceptual instability of the notion, underscoring the activity of an emerging or growing sector."

In another piece of work, Hamond and Nazarenko (Hamon and Nazarenko, 2001) seek to detect synonyms in order to help structuring of terminologies. The authors use three rules to detect synonymy relations between terms. An identical mechanism is also implemented in TETRIS. Two compound candidate terms are synonyms if:

- the heads are identical and the expansions are synonymous;
- the heads are synonymous and the expansions are identical;
- the heads are synonymous and the expansions are synonymous;

Through interaction with a terminologist, the authors find that recall is a more useful property than precision: errors are suggestive and may be interesting for terminologists.

4 Functions and Strategies of TTTT

TTTT works — in contrast to commercial terminology tools — on lexemes and lemmas and represents morpho-syntactic information of terms and words. While lexemes are the abstracted skeleton which is shared between the authorized form of a term and its variant, differences in features such as type of derivation, typographical information or the lemma determines the type of the variant.

The heart of TTTT consists of three tools (cf. figure 1). The first tool, TermLint, is designed for a terminologist to evaluate the entries in a list of terms. The components of this tool are listed in the upper part in figure 1. Via a graphical interface, the terminologist can sort, link, modify or annotate these terms.

Another tool is designed for authors. This tool checks documents for the consistent use of terms. The functional components of this tool are listed in the lower part in figure 1. Via a mailing system, authors can communicate with the terminologist in order to dynamically update the terminology.

The third tool abduces from an authorized terminology a set of variants to be recognized in the author's and the terminologist's tools. The components of this tool are listed in the middle part in figure 1.

TTTT presupposes an initial list of terms. These terms might have been proposed by authors or engineers or they might have been automatically extracted from existing corpora or during the authoring process. TETRIS comprises a term mining tool (Hong et al., 2001). Similar to the approach of Jacquemin (Jacquemin, 1996; Jacquemin, 2001), this tool follows the premiss "updating rather than acquiring" and builds upon existing terminology. Word forms which are neither authorized terms nor unauthorized variants and which comply with certain term formation criteria are detected in the background and are a possible source for terminology extension.

4.1 Architecture of TTTT

TermLint (in the upper part in figure 1) checks the consistency and uniqueness of terms. TermLint is designed for and used by terminologists who decide in collaboration with authors which terms are to be included into the company's term pool, which are their authorized forms and which are deprecated variants. TermLint also contains modules for orthographic, grammatical and term formation checking.

TermLint produces term clusters containing two or more entries which it classifies variants of each other as shown in section 2. These term clusters are a valuable help for the terminologist to revise the entries in the terminology and provides a means of deciding the consistency of a terminology. In case

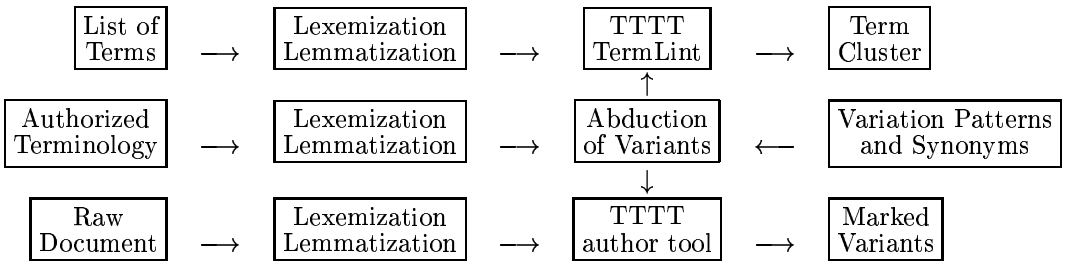

Figure 1: Main components of TTTT. up: Checking Consistency of a List of Terms with TermLint. middle: Compiling a Terminology into a database of variants. down: Checking Consistency of the Terms in Documents.

TermLint clusters two entries which actually represent two different concepts, an appropriate authorization flag can be set such that the entries will henceforth be considered as different terms.

The lower part in figure 1 represents the author's tool which is used 'online' when checking a document for consistent use of terms. The author's tool recognizes variants of terms in texts, highlights them in a document production environment and triggers a message showing the authorized form of the term together with a pre-defined comment. It provides authors with useful hints for using terms in their authorized forms.

Both tools first perform morphological analysis, lexemization and lemmatization of the linguistic expressions in a document (or a list of terms). The stream of linguistically enriched tokens is then matched against a database containing variants of authorized terms. In case a sequence of words in the document matches an entry in the database, an appropriate marker is set pointing to the authorized form.

4.2 An Abductive Approach

TTTT recognizes variants of terms by abduction. According to Streiter (Streiter, 2001), abductive reasoning creates hypothesis which are not logically implied by the premises. Unlike deductive reasoning, abduction is not always correct in all reasoning steps. However, abductive reasoning should be "plausible" in a context and yield correct results in the vast majority. Where deductive inference stops in front of gaps, abduction creates new hypothesis which allow to bridge the gap and continue the inference. As an illustration for abductive reasoning, Streiter gives the following example

Mooney (Mooney, 2000) examines the relation between abduction and induction. Although precise definitions of abduction and induction are still somewhat controversial, he finds:

> "In abduction, the hypothesis is a specific set of assumptions that explain the observations of a particular case; while in induction, the hypothesis is a general theory that explains the observations across a number of cases." (Mooney, 2000, p.183)

Mooney applies abductive learning for theory refinement. Theory refinement is the task to make an existing imperfect domain theory (e.g. an authorized terminology) consistent with a set of data. For him, abduction is primary useful in generalizing a theory to cover more positive examples. For each individual positive example that is not derivable from the current theory, abduction is applied to determine a set of assumptions that would allow it to be proven. In a similar way, TTTT proves the presence of a term by detecting their variants.

4.3 Software Components of TTTT

In TTTT — and in MULTILINT — lists of terms, documents and texts are morphologically analyzed, lemmatized and lexemized by means of MPRO (cf. (Maas, 1996)). The output of MPRO is a sequence of sets of feature bundles which encodes a number of linguistic properties of the words. Variants of terms are automatically generated based on the term's morpho-syntactic properties and the sequences of lemmas and lexemes. The variants are generated by means of a partial rule-based parser KURD (Carl et al., 1997). Terms and their generated variants are stored in an example-based machine-translation system EDGAR which is applied in TTTT for recognizing terms and their variants in texts and documents.

EDGAR maps terms and their variants onto the morphologically analyzed text and generates for each matched variant a partial, shallow derivation tree. While the mother node of this derivation tree contains head information of the authorized term, the leaves describe the variant as it occurred in the document. This structure is passed to KURD to compare the head information of the authorized term in the mother node and the information contained in the daughters. Where there are differences, the occurrences are marked with an appropriate sta-

tus flag indicating the type of variant together with its correct, authorized form. A more detailed description of this process is contained in (Carl et al., 2002).

5 Conclusion

In this paper we have presented a hybrid rule and example-based system for terminology management, the TETRIS Terminology Tool (TTTT). The aim of this tool is to provide an incremental approach for the elaboration of a prescriptive terminology and a means for checking consistent use of this terminology in documents. To accommodate both requirements, TTTT has two interfaces. The terminologist tool is designed to acquire and check new terminology and to update and modify the terminology pool. The author interface checks documents for consistent use of terms. Both interfaces are integrated in the company document flow.

There are a number of unsolved problems in TTTT which will be tackled in the near future. To name just two: Whether or not a given word or sequence of words is a term also depends on the syntactic context in which the word or sequence occurs. For instance, in its nominal reading the expression might be a term but it is not a term when used as a verb. Parsing of the syntactic structure becomes important in order to figure out the syntactic context of a candidate term. Although TETRIS performs a shallow syntactic analysis[2] in its style and grammar checking tools cf. (Haller, 2000), the results of these processing steps are not fully exploited in TTTT.

A more complex processing device is also required when adding discontiguous terms to the terminology, or when special verbs are required in the context of particular terms. Discontiguous terms consist of several parts which are distributed in a sentence, while frame information of verbs select and control their arguments. Also here, in order to relate the different parts of the term or to select appropriate arguments for the verb, a syntactic analysis is required for a proper treatment.

Future research in TETRIS and in TTTT will focus on the exploitation of empirical knowledge, such as technical texts, manuals and/or term lists. These knowledge resources are to be joined with analytical knowledge of the language and integrated into a human supported prescriptive authoring system.

References

Didier Bourigault, Christian Jacquemin, and Mairie-Claude L'Homme. 2001. *Recent Advances in Computational Terminology.* John Benjamins Publishing Company, Amsterdam/Philadelphia.

Michael Carl, Antje Schmidt-Wigger, and Munpyo Hong. 1997. KURD - a Formalism for Shallow Postmorphological Processing. In *Proceedings of the NLPRS*, Phuket, Thailand.

Michael Carl, Johann Haller, Christoph Horschmann, Dieter Maas, and Jörg Schütz. 2002. The TETRIS Terminology Tool. *TAL*, Structuration de terminologie(1).

Johann Haller. 2000. MULTIDOC -authoring aids for multilingual technical documentation. In *First Congress of Specialized Translation*, Barcelona. http://www.iai.uni-sb.de/~munpyo/IAI/bcn.doc.

Thierry Hamon and Adeline Nazarenko. 2001. Detection of synonymy links between terms: Experiment and results. In *in (Bourigault et al., 2001)*, pages 185–208.

Munpyo Hong, Sisay Fissaha, and Johann Haller. 2001. Hybrid filtering for extraction of term candidates from german technical texts. In *TIA-2001.* http://www.iai.uni-sb.de/de/pub.html.

Christian Jacquemin. 1996. A symbolic and surgical acquisition of terms through variation. In *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438.

Christian Jacquemin. 2001. *Spotting and Discovering Terms through Natural Language Processing.* The MIT Press, Cambridge, Massachusetts.

Heinz-Dieter Maas. 1996. MPRO - Ein System zur Analyse und Synthese deutscher Wörter. In Roland Hausser, editor, *Linguistische Verifikation, Sprache und Information.* Max Niemeyer Verlag, Tübingen.

Raymond J. Mooney. 2000. Integrating abduction and induction in machine learning. In P. Flach and A. Kakas, editors, *Abduction and Induction*, pages 181–191, Kluwer Academic Publishers.

Xavier Polanco, Luc Grivel, and Jean Royauté. 1995. How to do things with terms in infometrics: Terminological variation and stabilization as science watch indicators. In *Fifth International Conference of the International Society for Scientometrics and Infometrics*, pages 435–444.

Jean Royauté, Chantal Muller, and Xavier Polanco. 1996. Une approche linguistique infométrique de la validation terminologique pour l'analyse de l'information. In *Informatique & Langue naturelle, ILN'96.*

Oliver Streiter. 2001. Treebank Development with Deductive and Abductive Explanation-based Learning: Exploratory Experiments. In *to appear*, forthcoming.

Sue Ellen Wright and Gerhard Budin, editors. 1997. *Handbook of Terminology Management.* John Benjamins Publishing Company, Amsterdam/Philadelphia.

[2]This syntactic analysis is achieved with a set of KURD rules.