The Emerald Research Register for this journal is available at
www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0022-0418.htm

The Laurin thesaurus

A large, multilingual, electronic thesaurus for newspaper clipping archives

Gregor Retti and Birgit Stehno

The Laurin Project, Department of German Language, Literature and Literary Criticism, University of Innsbruck, Innsbruck, Austria

Received 3 October 2003
Revised 6 January 2004
Accepted 8 January 2004

Keywords *Information retrieval, Thesaurus construction, Indexing, ISO 9000 series*

Abstract *This paper describes the Laurin thesaurus, which is used for indexing and searching in the Laurin system, a software package for digital clipping archives. As a multilingual thesaurus it complies with the corresponding standards, though presenting some approaches going beyond some of the standards' recommendations. The Laurin thesaurus integrates all kind of indexing terms, not only keywords, but proper names as well. The system of categories and relationships is described in detail.*

1. Introduction

The Laurin thesaurus was developed during the R&D-project Laurin (1998-2000) (Mühlberger, 2000) and enhanced in the follow-up project Laurin+ (2000-2002). As part of the Laurin system (Calvanese *et al.*, 2001) the Laurin thesaurus is used by the electronic clipping archive of the *Innsbrucker Zeitungsarchiv*(see http://iza.uibk.ac.at/). The *Innsbrucker Zeitungsarchiv* has collected clippings from the field of literary criticism from German-speaking newspapers and journals since 1960. It comprises a paper archive of approximately one million clippings, but has completely switched to an online clipping archive in 1999 using the Laurin system. The aim of the Laurin project was to create a software package for clipping archives which would allow them to digitise entirely the clipping, indexing, storing, and retrieval of the archived material. A multilingual thesaurus was planned to be a major part of the system, which should be compliant with standards, support manual and even automatic indexing and act as the key device for retrieving clippings from the database.

2. Standards and language

The definitions in ISO 2788 (ISO, 1986) clearly document the dependency of thesauri on natural languages by defining an indexing language as "a controlled set of terms selected from natural language" and a thesaurus as "the vocabulary of a controlled *indexing language* [...], formally organized so that the a priori relationships between concepts [...] are made explicit". An indexing term consistently is defined as "the representation of a concept [...]" and so are preferred term and non-preferred term. The

This study is a result of the Laurin project, a research and development project co-funded by the European Commission (4th Framework Programme, "Telematic Applications for Libraries"; LB-5629/A) and the Austrian Federal Ministry for Science and Traffic as well as the follow-up project Laurin+ funded by the Austrian Federal Ministry of Traffic, Innovation and Technology. Furthermore, the authors want to express their gratitude to The Getty Information Institute for the permission to use the Getty Thesaurus of Geographic Names™ (see http://laurin.uibk.ac.at/).

Journal of Documentation
Vol. 60 No. 3, 2004
pp. 289-301
© Emerald Group Publishing Limited
0022-0418
DOI 10.1108/00220410410534194

relationship between preferred and non-preferred terms is understood to be the equivalence relationship:

> This is the relationship between preferred and non-preferred terms where two or more terms are regarded, for indexing purposes, as referring to the same concept.

This relationship is commonly designated as the "USE" or "USE FOR" relationship. It covers three different types of equivalence: synonyms, quasi-synonyms and "upward posting". While synonymy is a well-known phenomenon in linguistics and other disciplines, quasi-synonyms as well as upward posting are rather confined to the domain of thesaurus construction. Quasi-synoyms are "terms whose meanings are generally regarded as different in ordinary usage, but they are treated as though they are synonyms for indexing purposes" (ISO, 1986, p. 14). The technique of upward posting is the aggregation of one or more specific terms under a broader term that in turn becomes the preferred term for such a set. It should be noted, that the standard recommends using quasi-synonyms and upward posting, which "should generally be avoided", only in the fringe area of the field covered by the thesaurus. As far as synonyms are concerned the equivalence relationship notably differs from the other relationships defined in ISO 2788 (ISO, 1986) as it relates terms with the same meanings while the other relationships relate terms with different meanings, i.e. concepts.

In the Laurin thesaurus, the notion of thesaurus entry is based on the linguistic sign theory that considers every sign to be made up of two components: the meaning (*signifié*) and some sort of expression (*signifiant*) (Saussure, 1967). Each thesaurus entry consists of a semantic part – the concept – and one or more labels which represent the concept. Thus, the equivalence relation is incorporated into the thesaurus entry. This approach allows extending a thesaurus entry to cover more than one language as labels from different languages are simply added to the entry. For each language covered one and only one label is marked as preferred while the others are marked as non-preferred. This information is needed for thesaurus display and thesaurus browsing – and, as a matter of fact, to ensure compliance with the standards (ISO, 1986, 1985). It should be noted, that the model of the thesaurus entry in the Laurin thesaurus has been adapted from the Getty Thesaurus of Geographic Names™ (Harpring, 1998). The Getty Thesaurus of Geographic Names™ flags place names as either "vernacular" or "other" and "current" or "historical":

> TGN contains vernacular and English names of places, as well as variant names in other languages and historical names for some places (Harpring, 1998, p. 19).

Thus, the place names are not assigned to a specific language. A place name is vernacular if that name belongs to a language spoken by the inhabitants of that place, currently or sometimes in history. Therefore, the Getty Thesaurus of Geographic Names™ does not meet the recommendations of ISO 5964 (ISO, 1985) as it does not mark the place names as belonging to specific languages. During the construction of the Laurin thesaurus a large part of the Getty Thesaurus of Geographic Names™ was imported. Languages were assigned semi-automatically to the imported data, e.g. by processing all thesaurus entries in the hierarchy of one country – given that this country has one dominant language only like France, Germany, or Bulgaria – and

setting the language attribute accordingly for all labels marked as vernacular and current (see Figure 1).

There are some well-known problems regarding the degree of equivalence between terms from different languages when constructing a multilingual thesaurus (ISO, 1985, p. 7f; Aitchison *et al.*, 2000, p. 140f). Non-equivalence, i.e. a term from one language does not have any equivalent, not even partially or not exactly, in another language, is by far the most difficult case to resolve. While ISO 5964 (ISO, 1985, p. 18f.) proposes either to adopt the term as a loan term defined by a scope note, or to coin a new term in the target language. In the Laurin thesaurus the slot for the label in a specific language is simply left empty if such a label does not exist. Thus, a thesaurus entry for the classical example "*Berufsverbot*" has only one label marked as German: *Berufsverbot*. The thesaurus entry, on the other hand, is still available for indexing and retrieval in any language, as it will simply be displayed using the available label from another language. This approach has the advantage of not introducing unknown new words and expressions, i.e. "loan terms", "putative 'translations'", or "artificial inventions" (ISO, 1985, p. 18f), into the thesaurus. It is rather unlikely that a user will be looking for such words or expressions in a thesaurus. If someone is looking for a thesaurus entry, which does not have a label in a certain language, but has one in another language, the user may probably use the known label in that language anyway. Therefore, the question is not whether a certain concept may be translated into another language, but whether the target language can supply a known subject term for the concept in question. While this may not be a common case for subject terms (ISO, 1985, p. 18f) it seems to be a regular phenomenon when proper names of an organisation or persons

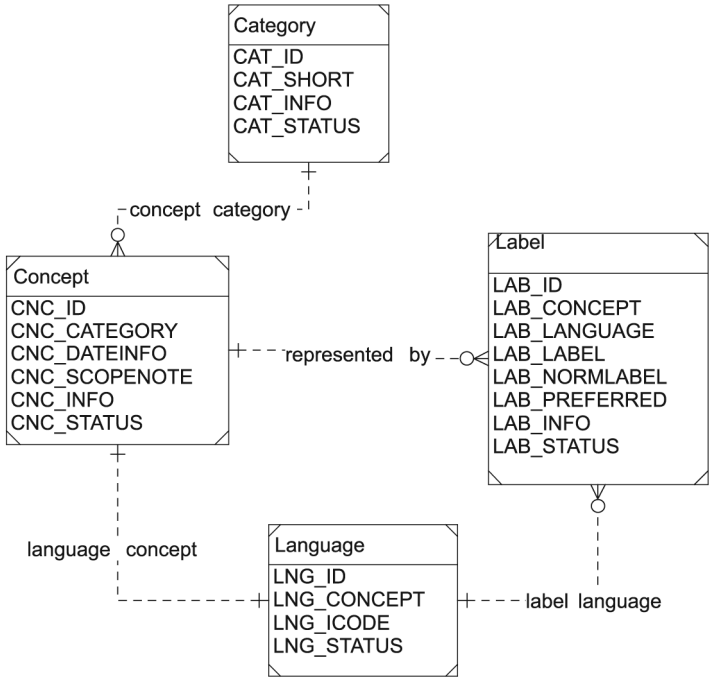

Figure 1.
Implementation schema for concepts and labels in the Laurin thesaurus

are concerned as only very few proper names do have translated variants in multiple languages. ISO 5964 (ISO, 1985, p. 24f) recommends for this case the use of the untranslated term in the target language as well. There may be good reason to do so in practice, but the idea of having an untranslated term marked with the language it has not been translated to seems somehow awkward. The reason for this not being recorded as an annoyance, on the contrary, finding it as a recommendation of an ISO standard, might be sought in the fact that the attribute "language" does not apply as a basic and defining category to proper names.

Single-to-multiple equivalence is another problem in multilingual thesauri:

> A concept represented by a term in the source language is not recognized as a single idea by the users of the target language. Instead it is regarded as consisting of two or more different concepts, each of which is represented by its own specific term (ISO, 1985, p. 12f).

The standard offers four different solutions for this problem, favouring the one which "achieves equivalence for all terms without the need for loan or coined terms". It is best presented through an example: the English word "skidding" corresponds to the German words "*Rutschen*" and "*Schleudern*", which in turn mean "skidding (forwards)" and "skidding (sideways)". The equivalence between source and target language is reached by establishing "skidding"/"*Rutschen*+ *Schleudern*" as broader terms, and "skidding (forwards)"/"*Rutschen*" and "skidding (sideways)"/"*Schleudern*" as narrower terms as shown in Figure 2.

The Laurin thesaurus provides a similar yet slightly more user-friendly solution to this problem. The concept <skidding> is represented by the English term "skidding" and the German term "*Rutschen*+ *Schleudern*" as preferred terms. Additionally the German terms "*Rutschen*" and "*Schleudern*" are added as non-preferred terms. For the concept of <skidding sideways> the English term "skidding sideways" and the German term "*Schleudern*" are used as preferred terms, while "skidding" is a non-preferred English term for the concept. The concept <skidding forward> is treated analogously. Figure 3 gives a picture of this solution.

What looks at a first glance like an overload of labels has a clear advantage when it comes to searching the thesaurus and selecting the correct entry for indexing or retrieval. As the preferred terms are always displayed and these terms should be the most accurate, the user is guided to pick the term closest to the concept he/she is looking for.

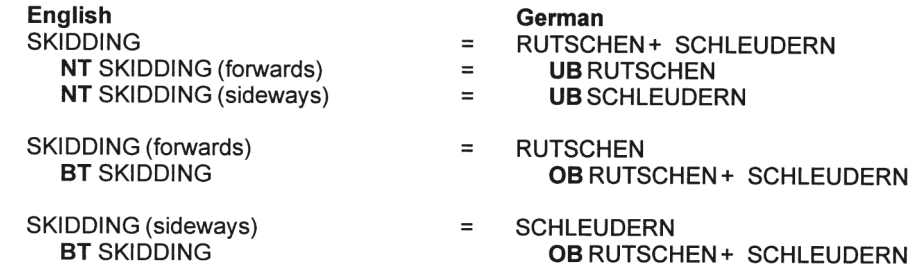

Figure 2.
Single-to-multiple equivalence solution according to ISO 5964

<skidding>
skidding (Engl.; pref. term)
rutschen + schleudern (Germ.; pref. term)
rutschen (Germ.; non-pref. term)
schleudern (Germ.; non-pref. term)

<skidding (sideways)>
skidding sideways (Engl.; pref. term)
skidding (Engl.; non-pref. term)
schleudern (Germ.; pref. term)
BTG <skidding>

<skidding (forwards)>
skidding forwards (Engl.; pref. term)
skidding (Engl.; non-pref. term)
rutschen (Germ.; pref. term)
BTG <skidding>

Figure 3.
Single-to-multiple equivalence solution in the Laurin thesaurus

3. Categories and relations

The Laurin thesaurus includes all terms that are used for indexing. This is also true for proper names of persons and institutions that are often excluded from thesauri (ISO, 1985). Each thesaurus entry is assigned to one and only one out of six basic categories: "keywords" (KEY), "time keywords" (TIM), "persons" (PER), "institutions" (INS), "geographical names" (GEO), and "literary and artistic works" (LAW). These categories should not be misunderstood as facets (Aitchison *et al.*, 2000, p. 68f). While faceted classifications use only one characteristic or principle at a time to establish groups of concepts, the categories of the Laurin thesaurus form a rather simple classification with one level only. They serve to determine which relations may be used with a specific thesaurus entry – and which relations must be present to qualify the thesaurus entry not to be an orphan. Five of the six categories where initially defined during the set-up and construction of the thesaurus. The sixth category, "literary and artistic works", was added after deploying the Laurin system at the *Innsbrucker Zeitungsarchiv* to meet the special needs of this clipping archive, which focuses strongly on fiction, book and theatre reviews.

Currently the Laurin thesaurus comprises 18 different relations to link the thesaurus entries together (see Table I).

The set of relations includes the relationships described in ISO 2788 (ISO, 1986) as well as additional relation types (e.g. "broader term work", "location", "creator"). An augmented set of thesaurus relationships has proved to be helpful for indexing and retrieval purposes. First of all, it allows for a more precise definition of a thesaurus entry, thus assisting the user to understand better the context of a concept. This may lead to a greater correspondence in the choice of indexing terms by indexers and searchers which operate "at different levels of specificity, and at different times" (Tudhope *et al.*, 2001). Second, a richer set of thesaurus relations increases the retrieval potential by suggesting terms for refining a query or query expansion to the searcher (user). In approaches with automatic query expansion a richer set of relation types allows for a "finer grained automated reasoning" since undesired search results ("noise") can be reduced (Tudhope *et al.*, 2001).

Abbr.	Relation	Description
BTG	Broader term generic	Generic relationship (ISO, 1986, p. 15f)
BTI	Broader term instance	Instance relationship (ISO, 1986, p. 17)
BTL	Broader term location	Special case of BTP to be used with "geographical names" only
BTO	Broader term object	"A discipline or field of study and the objects or phenomena studied" (ISO, 1986, p. 18)
BTP	Broader term partitive	Hierarchical whole-part relationship (ISO, 1986, p. 16)
BTW	Broader term work	"An occupation and the person in that occupation" (Aitchison *et al.*, 2000, p. 64)
CRE	Creator	Associative relationship between a "literary and artistic work" and a "person" (or "institution"), who has created it
DRV	Derived from	Associative relationship between a "literary and artistic work" based somehow on another "literary and artistic work"
GEO	Geographical type	Special case of BTI to be used with "geographical names" and "keywords"
LOC	Located in/at	Associative relationship to link any concept to a particular place, i.e. a "geographical name"
MEM	Member	Associative relationship between a "person" or an "institution", who is a member of another "institution"
RT	Related term	Unspecific associative relationship (ISO, 1986, 17f.)
SCR	Secondary creator	Associative relationship between a "literary and artistic work" and a "person" (or "institution"), who was involved in the process of creating it but is not the creator
SOC	Social	Associative relationship between two "persons" with a close social relation
SUC	Successor of	Associative relationship between "institutions" (and some "geographical names"), e.g. when one "institution" merges into another or splits up into several new "institutions"
TIM	Time	Associative relationship to link any concept to a particular period of time or point of time, i.e. a "time keyword"
TT	Top term	Special relationship to form the top level hierarchy of the thesaurus
USE	Use	Useful in cases where the usage of a concept for indexing is discouraged to point to the preferred thesaurus entry

Table I.
Relations of the Laurin Thesaurus

Relations and categories are restricted regarding their combinability. For example, it simply does not make sense to allow a thesaurus entry of the category "person" to be combined with a relation "broader term generic". Therefore, syntactical rules have been defined for the Laurin thesaurus, which control the combinability of relations and categories. Most relations are directional, but the Laurin thesaurus does not explicitly define relations for reciprocal entries. Thus, a rule for combinability is a plain

statement like "KEY → BTG → KEY", which means that a thesaurus entry of the category "keyword" may be related to another thesaurus entry of the same category by the relation "broader term generic". This syntactical model was implemented in the relational schema of the Laurin thesaurus as shown in Figure 4.

With an SQL-statement like the one below all possible relations for a given set of two concepts can be retrieved from the database:

```
SELECT DISTINCT rfc_relationtype
  FROM rt_for_category
  WHERE rfc_categoryx = (SELECT cnc_category FROM concept WHERE
  cnc_id = 'ID_OF_CONCEPT_1')
  AND rfc_categoryy = (SELECT cnc_category FROM concept WHERE
  cnc_id = 'ID_OF_CONCEPT_2');
```

For example, for the keywords <English> and <languages> the result would be BTG, BTI, BTO, BTP, BTW and RT, while for <English> and the geographical name <United Kingdom> it would be LOC and RT (see Table II).

3.1 Keywords (KEY)

The category "keyword" is by far the most complex and heterogeneous category. Thesaurus entries of this category represent concrete and abstract things, actions, processes, events, properties, etc. Furthermore, they relate to each other in a number of different ways, thus making up the richest and most complicated part of the Laurin thesaurus, coming close to what may be called an ontology (Gilchrist, 2003). The thesaurus entries of the category KEY have been arranged under 15 top terms or

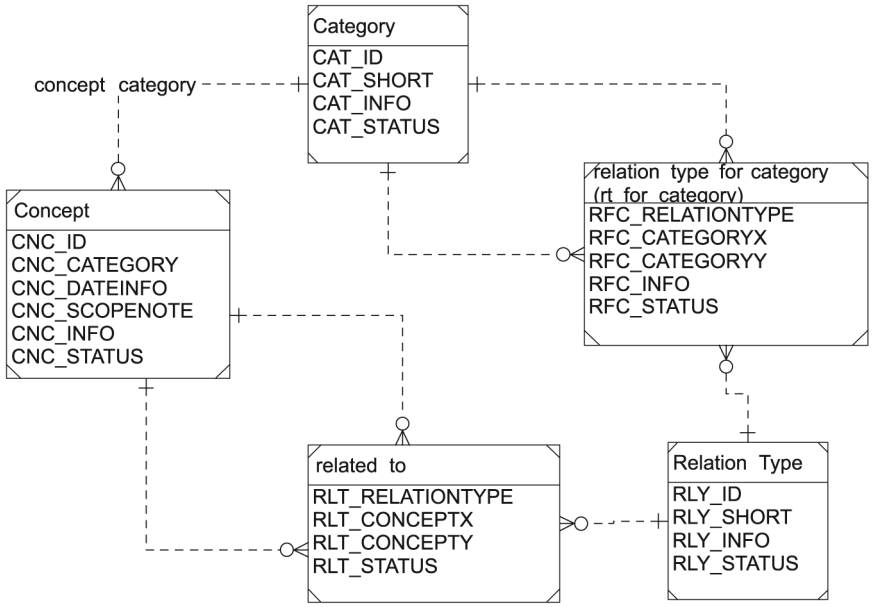

Figure 4.
Implementation schema for the syntactical rules for categories and relationships in the Laurin thesaurus

chapter headings, which have been derived from the "IPTC subject reference system" (IPTC, n.d.):

(1) < culture> ;
(2) < education> ;
(3) < health> ;
(4) < legislation, justice and implementation> ;
(5) < politics> ;
(6) < science> ;
(7) < economy> ;
(8) < environmental issues> ;
(9) < disasters and accidents> ;
(10) < social issues> ;
(11) < religion> ;
(12) < lifestyle and leisure> ;
(13) < human interest> ;
(14) < sport> ; and
(15) < weather> .

The IPTC system – available in several languages – has been especially designed to be used with news items and, therefore, fitted very well to the needs of the Laurin thesaurus regarding the top terms. The remaining hierarchy of the IPTC system had to be re-organised due to its rather arbitrary structure before the items could be integrated into the Laurin thesaurus.

The main hierarchically structuring relations for thesaurus entries of the category "keyword" are the generic relationship, "broader term generic" (BTG), and the hierarchical whole-part relationship, "broader term partitive" (BTP) (ISO, 1986, p. 15f). Furthermore, individual entities other than persons, institutions, or geographical places are attached using the instance relationship, "broader term instance" (BTI) (ISO, 1986, p. 17). It is impossible to arrange the resulting large and poli-hierarchical trees of thesaurus entries under the chapter headings just mentioned using BTG or BTP without violating the semantics of these relations. Therefore, a special relationship, "top term" (TT), has been introduced to link the highest levels of the hierarchy to a

Table II. Categories and relationships of the Laurin thesaurus

	KEY	TIM	PER	INS	GEO	LAW
KEY	BTG, BTP, BTI, BTO, BTW, (TT)	TIM	CRE	CRE	LOC	
TIM		BTP				
PER	BTW, BTI	TIM	SOC	MEM	LOC	
INS	BTW, BTI	TIM		BTP, MEM, SUC	LOC	
GEO	GEO	TIM		MEM	BTL, SUC	
LAW	BTI	TIM	CRE, SCR	CRE, SCR		BTP, DRV

chapter heading in case one of the "broader term" relations could not be applied. An example can be seen in Figure 5.

When defining additional relations for the Laurin thesaurus it was one goal to avoid the uncontrolled use of the associative relationship due to their vagueness and "somewhat inconsistent application" (Aitchison *et al.*, 2000, p. 66; ISO, 1986, p. 17f; Maniez, 1988). Therefore, the associative relationship "a discipline or field of study and the objects or phenomena studied" (ISO, 1986, 18) has been redefined as a hierarchical relation "broader term object" (BTO). BTO may be paraphrased as "x is the object of (study) field y", e.g. <language> BTO <linguistics>. Another associative relationship "an occupation and the person in that occupation" (Aitchison *et al.*, 2000, p. 64) was redefined as "broader term work" (BTW). When applied to thesaurus entries of the category "keyword" the relation BTW is useful to link *nomen agentis* to the base domain of the derivation, e.g. <philosophers> BTW <philosophy>. It should be noted, that these two relationships do not form multilevel hierarchies – as "broader term ..." may somehow imply. The advantage of relationships like BTW or BTO lies in the systematic approach, which assist the indexing as well as the navigation. On the other hand, it should be thoroughly considered whether to include them when exploiting the relations of the thesaurus in an automatic query expansion, because the precision of the result may suffer substantially.

A thesaurus entry of the category "keyword" can be associated to any entry of another category using the global associative relationship "related term" (RT), although indexers are discouraged from doing so. Nevertheless, a few systematic relationships with focussed semantics have been defined. For some entries, especially those designating individual entities, the location may be an important piece of information. Those keywords are linked to geographical names using the relationship "located at/in" (LOC), e.g. <Chianti> LOC <Tuscany>.

3.2 Time keywords (TIM)

Thesaurus entries of the category "time keyword" make up a rather simple hierarchy using the (BTP relationship The separation of thesaurus entries for periods of time or points in time into a category of their own allows a more precise definition of other thesaurus entries, for which time, mostly historical time of course, is significant. Thus,

culture [KEY] *Kultur, cultura, culture, kultur*

			culture	BTP	arts	KEY
				BTP	cultural heritage	KEY
				BTP	cultural values	KEY
				BTP	languages	KEY
				BTP	nonfiction	KEY
				BTP	script	KEY
				BTP	subculture	KEY
				BTP	traditional culture	KEY
				TT	cultural events	KEY
				TT	cultural institutions	KEY
				TT	cultural periods	KEY
				TT	people	KEY

Figure 5.
Screen-shot of the thesaurus browser of the Laurin system showing the chapter heading <culture>

an associative relationship – not yet to be found in the standards or handbooks obviously due to its vagueness – "an event and the time when it happened" is available: keywords for (historical) events may be related to "time keyword" by means of this relationship (TIM), e.g. <Thirty Years' War> TIM <17th century> . The relation can as well be applied to other categories of the Laurin thesaurus, thus meaning something like "a person, an organisation, a geographical place and the time when it existed".

3.3 Persons (PER)

Proper names of persons are often excluded from thesauri in other indexing systems The main reason for keeping persons' names apart may be, that these entries do not form a hierarchical structure but rather a flat list of controlled vocabulary. The decision to integrated entries of the category "person" into the Laurin thesaurus is motivated by the fact, that those entries can be defined and distinguished with the help of some relations to thesaurus entries of other categories. The BTW-relation ("broader term work") or the BTI-relation ("broader term instance") can be used to express, what the profession of a person is. These relations link the person's entry to a "keyword", e.g. <Benigni, Roberto> BTI <stage directors> and <Benigni, Roberto> BTI <actor> or (<Auster, Paul> BTW <English literature> and <Auster, Paul> BTW <literature of the United States> . With the LOC-relation ("located at/in") linking a PER-entry to one of the category "geographical name" it is possible to describe, where a person lives, works, comes from etc. With these two relations, BTW/BTI and LOC, defined, it is very easy for the user to distinguish between persons with identical or similar names. To serve this purpose the additional information supplied through the relations BTI, BTW and LOC should be limited to more or less well-known or relevant facts. The domain of the collection or archive should as well be considered, when setting up these limits. Therefore, it may or may not be reasonable to include <Borodin, Alexander Porfiryevich> BTW <organic chemistry> or <Conrad, Joseph> BTI <mate> or <Canetti, Elias> LOC <Bulgaria> .

Furthermore, "person"-entries may be associated to each other using the "social"-relationship (SOC), which indicates some close as well as well-known social relation of one person to another, e.g. "spouse of". The SOC-relation is helpful, when a person is only known because of his/her relation to another person and not because of some important deeds or achievements. Finally, thesaurus entries of the category "person" may be related to "institution"-entries by the "member"-relationship (MEM). This relation allows to identify groups of persons belonging to the same "institution", e.g. a political party, a group of artists or writers.

3.4 Institutions (INS)

The thesaurus entries of the category "institution", i e. any kind of legal bodies, groups of persons or organisations, are very similar to those of the category "person". They do not build up hierarchies and they are defined through the relations BTW, BTI and LOC. Entries of the category "institution" may furthermore be associated to each other using the MEM in cases where some independent organisations form a larger one or by the BTP-relation, if one institution is effectively a dependent part of the other. This may even lead to small hierarchical structures.

Another relationship linking "institution" entries to one another is the "successor of" (SUC) relation. It is especially useful when a company or organisation changes its name, merges with other companies or organisations or splits up into a set of new "institutions" which is a rather common phenomenon in business world.

3.5 Geographical names (GEO)

Geographical names are another type of proper name which have been grouped together in a category of their own in the Laurin thesaurus. As mentioned above the Laurin thesaurus has been filled with large parts of the Getty Thesaurus of Geographic Names™ (Harpring, 1998). Therefore, the relations applied to thesaurus entries of the category "geographical name" are derived from the structure of the Getty Thesaurus. Entries link to each other using the relationship "broader term location" (BTL), which can be considered as a special type of the BTP relation. It should be noted, that the Laurin thesaurus has adopted the Getty Thesaurus' solution to deal with the multi-hierarchies made up by topological and political structures, e.g. <French Guiana> links to <South America> as well as to <France> through a BTL-relation. The second relation taken from the Getty Thesaurus is called "geographical type" (GEO). It is used to associate a "geographical name" to one or more "keywords", which describe that "geographical name", e.g. <Chicago> is associated to <city> , <county seat> , <commercial centre> , <inland port> etc. Obviously the "geographical type" relationship is a sub-type of the BTI relation. The keywords used with that relation have also been adopted from the Getty Thesaurus. They have been transformed into proper hierarchies using BTG and BTP relations, as those hierarchies where only implicit available in the schema of the identifiers of the Getty Thesaurus.

Less significant relationships for "geographical names" are the relation SUC, e.g. for a state splitting up into two or more new states, and the relation MEM, which requires the "geographical name" to be some sort of human community like a nation or a city and connects "geographical names" to "institutions", e.g. <Austria> MEM <European community> .

3. 6Literary and artistic works (LAW)

The category "literary and artistic works" (LAW) was added to the Laurin thesaurus after the Laurin system had been adopted by the *Innsbrucker Zeitungsarchiv* to meet the particular requirements of its clipping collection. The original design of the Laurin thesaurus as a general purpose thesaurus is open to such extensions, which serve the individual needs of a special archive or collection. Entries of the category "literary and artistic works" cover novels, movies, plays, songs, paintings, i.e. any kind of unique and named human-created artifacts.

Two relationships are required for entries of the category "literary and artistic works". Each must be connected through a BTI relation to a keyword to define its type and nature, e.g. <*Through the Looking-glass*> BTI <novel> . And it must be associated to a "person" – or in some cases to an "institution" – using the relationship "creator" (CRE), e.g. <*Consider Phlebas*> CRE <Iain M. Banks> . Alternatively the CRE-relation may be replaced by a relationship called "secondary creator" (SCR), if the "person" involved acts like an editor or publisher. It should be mentioned that the new

relationship "creator" (CRE) was as well added to connect keywords with persons or institutions, e.g. for keywords designating a commercial product.

Optional relationships to define a "literary and artistic work" are BTP and "derived from" (DRV). BTP may be used if the "literary and artistic work" in question is part of another "literary and artistic work", e.g. <New testament> BTP <Bible> . The DRV-relationship is especially useful when one "literary and artistic work" has been used to make a new one, e.g. a theatre play based on a book or a movie based on a play like <*Blade Runner*> BTI <movies> , <*Blade Runner*> CRE <Scott, Ridley> , <*Blade Runner*> DRV <*Do Androids Dream of Electric Sheep*> CRE <Dick, Philip K.> .

3. 7 Overview

Table II shows an overview of the different categories and how they may be associated to each other using the relationships in the Laurin thesaurus.

4. Conclusion

As by the time of this writing in September 2003 the Laurin thesaurus at the *Innsbrucker Zeitungsarchiv* contains 256,732 concepts with 324,444 labels. A total of 208,068 thesaurus entries are of the category "geographical names"; 27,350 are "persons"; 10,062 are "literary and artistic works"; 8,098 are "keywords"; 3,077 are "institutions"; and 77 are "time keywords". These entries are associated to each other through 580,568 relations, out of which the most frequently used are 251,844 "geographical type" relations, followed by 210,681 "broader term location", 46,921 "broader term work", 35.035 "located at/in", 18.672 "broader term instance", 10.234 "creator", 4.325 "broader term generic", and 1.447 "broader term partitive". The Laurin thesaurus is constantly growing, especially "persons" and "literary and artistic works" entries are added, reviewed and approved on a daily bases.

Approximately 20,000 clippings are added every year to the database of the *Innsbrucker Zeitungsarchiv*. Thanks to the software tools comprising the Laurin interface suite (Retti, 2003) the whole workflow is widely optimised, but still the maintenance of the thesaurus causes additional workload. Anyway, any kind of controlled vocabulary would do so. Therefore, the only inexpensive alternative would have been a full text only retrieval system, which had never been a serious option during the project. On the other hand the indexing process is accelerated for cases where subject terms are already available, as the indexer can rely on the additional information provided by the thesaurus, e.g. a book review will be indexed with the thesaurus entry for the book only, as this entry in turn is related to the author within the thesaurus. The Web-based information retrieval application of the Laurin system tries to exploit the Laurin thesaurus by including a thesaurus browser into the user interface. But it must be admitted that the function of this device is not *ad hoc* obvious to the casual user, who seems to be rather accustomed to the simple search interfaces provided by today's popular Internet search engines. Of course the adept user takes advantage of the thesaurus browser as a navigation tool. Nevertheless, the cost for the thesaurus maintenance seems to be justified by the high quality indexing enforced by the Laurin system. Possible applications based on the thesaurus can be found beyond searching and browsing the clipping database of the *Innsbrucker Zeitungsarchiv*, e.g. a comprehensive bibliography of newspaper and journal articles about a certain author,

a comparative study about the reception of French authors in Austrian, German, and Swiss print media etc.

References

Aitchison, J., Gilchrist, A. and Bawden, D. (2000), *Thesaurus Construction and Use: A Practical Manual*, 4th ed., Aslib, London.

Calvanese, D., Catarci, T. and Santucci, G. (2001), "Laurin: a distributed digital library of newspaper clippings", *World Wide Web*, Vol. 4 No. 1-2, pp. 5-20.

Gilchrist, A. (2003), "Thesauri, taxonomies and ontologies – an etymological note", *Journal of Documentation.*, Vol. 59 No. 1, pp. 7-18.

Harpring, P. (1998), *User's Guide to TGN: Relational Files Format, Version 1.0*, Getty Information Institute, Los Angeles, CA.

International Organization for Standardization (ISO) (1985), *Guidelines for the Establishment and Development of Multilingual Thesauri, ISO 5964*, 1st ed., ISO, Geneva.

International Organization for Standardization (ISO) (1986), *Guidelines for the Establishment and Development of Monolingual Thesauri, ISO 2788*, 2nd ed., ISO, Geneva.

International Press Telecommunications Council (IPTC) (n.d.), "IPTC subject reference system", available at: www.iptc.org/site/subject-codes/index.html (accessed 31 May 2003).

Maniez, J. (1988), "Relationships in thesauri: some critical remarks", *International Classification*, Vol. 15 No. 3, pp. 133-8.

Mühlberger, G. (2000), "Electronic clipping of articles from printed newspapers – a report on the outcomes of the EU-funded project Laurin", *Library Computing*, Vol. 19, pp. 77-85.

Retti, G. (2003), "The Laurin Interface Suite: a software package for newspaper clipping archives", *Journal of Digital Information Management*, Vol. 1 No. 4, pp. 182-7.

Saussure, F. (1967) Engler, R. (Ed.), *Cours de Linguistique Générale*, Critical edition in three volumes, Otto Harrassowitz, Wiesbaden.

Tudhope, D., Alani, H. and Jones, C. (2001), "Augmenting thesaurus relationships: possibilities for retrieval", *Journal of Digital Information*, Vol. 1 No. 8, available at: http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Tudhope/ (accessed 6 May 2003).