

Verarbeitungsschritte des terminologischen Netzes

Dokumentation

Projekt
**Hypertextualisierung auf
textgrammatischer Grundlage**
(www.hytext.info)

Eva Anna Lenz

2004

Für die Eingabe und Verwaltung des terminologischen Netzes (TermNet) nutzen wir das Werkzeug K-Infinity, das uns freundlicherweise von der Firma Intelligent Views (<http://www.i-views.de/web/>) zur Verfügung gestellt wurde. Dieses terminologische Netz wird anschließend automatisch in eine XML Topic Map (XTM) konvertiert. Ein solches standardisiertes, XML-basiertes Austauschformat für Wissensnetze hat den Vorteil, dass es mit verschiedenen Werkzeugen (z.B. der Programmiersprache XSLT) weiterverarbeitet werden kann und – perspektivisch – als Wissensressource auch von anderen Projekten genutzt werden kann. Wir nutzen das XTM-Format, um darauf Inferenzen durchzuführen und eine HTML-Präsentation eines erweiterten Glossars inklusive einer SVG-Visualisierung von Teilen des Netzes zu erzeugen. Das Glossar ist mit den Korpustexten in beide Richtungen verlinkt.

Die Modellierung und Eingabe des terminologischen Netzes mit K-Infinity erfolgt manuell, alle nachfolgenden Schritte laufen vollautomatisch ab. Das vorliegende Dokument beschreibt diese Schritte.

Aus K-Infinity heraus wird das terminologische Netz zunächst nach XML exportiert und über eine Kette von Verarbeitungsschritten weiterverarbeitet, über die Konvertierung in das Topic-Map-Format und die Durchführung von Inferenzen bis hin zur Erzeugung von HTML- und SVG-Dateien. Da alle diese Schritte vollautomatisch ablaufen, kann der Prozess auch nach einer Änderung des terminologischen Netzes problemlos erneut angestoßen werden.

Die nachfolgend aufgeführten und in der Grafik „Workflow aus Entwicklersicht“ visualisierten Verarbeitungsschritte sind aus Gründen der Arbeitsorganisation und der Wartbarkeit des Codes in einzelne Programme zerlegt worden. Die einzelnen Perl- und XSLT-Skripte müssen aber nicht einzeln aufgerufen werden. Nach dem Export aus K-Infinity und der Speicherung unter dem Namen `TermNet.xml` genügt der Aufruf des folgenden Perl-Skripts, das die Einzelprogramme nacheinander aufruft:

```
perl ../export2svg.pl TermNet.xml
```

Genaueres zum Aufruf des Skripts ist diesem selbst zu entnehmen.

Dieser vereinfachte Workflow ist in einer zweiten Grafik („Workflow aus Anwendersicht“) ebenfalls dargestellt.

Es folgt die Dokumentation der einzelnen Schritte dieses Workflows aus Entwicklersicht.

1. Export aus K-Infinity

Menü Werkzeuge → Datenexport → XML-Export
Speicherung unter dem Namen `TermNet.xml`

2. Transformation in das Topic-Map-Format (XTM)

Die Transformation des Exportformats von K-Infinity nach XTM erfolgt mit dem ebenfalls von der Firma von intelligent views zur Verfügung gestellten Stylesheet `ki2tm.main.xsl`. Dieses wurde noch leicht verändert (Entfernung der Doctype-Deklaration, Entfernung kleinerer Fehler, siehe Anhang).

Aufruf:

```
saxon TermNet.xml ../kinfinity2tm/ki2tm-veraendert/
ki2tm.main.xsl
> TermNet-XTM-raw.xtm
```

3. TermNet-Anpassungen

Das Stylesheet von K-Infinity übersetzt zwar in eine syntaktisch korrekte Topic Map, dennoch sind – in unserem Fall – noch einige Anpassungen notwendig, die u.a. aus den unterschiedlichen Modellierungsparadigmen von K-Infinity vs. Topic Maps resultieren (gerichtete Relationen vs. ungerichtete Relationen mit Rollen, Vorhandensein eines Wurzelements in K-Infinity, teilweise Überführung von Attributen in Skopen). Zudem enthält die Topic Map noch weitere K-Infinity-eigene Topics, die in unserem TermNet nicht auftauchen sollen. Schließlich setzt der K-Infinity-Export einen Merging-Prozess mit einer K-Infinity-eigenen weiteren Topic Map voraus, den wir aus pragmatischen Gründen nicht durchführen (u.a., um nicht noch weitere Topics zu erhalten, die nicht Bestandteil des TermNet sind); stattdessen werden die Verweise auf die externen Topics entfernt bzw. ersetzt. (Zu den Details siehe Anhang dieses Dokuments.)

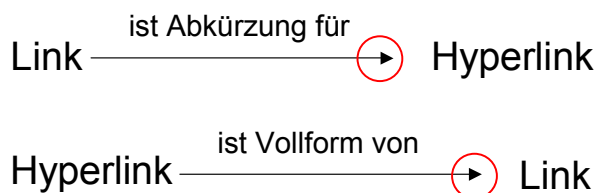
Aufruf:

```
saxon TermNet-XTM-raw.xtm ../XTM-raw2XTM-ok.xslt
> TermNet-XTM-ok.xtm
```

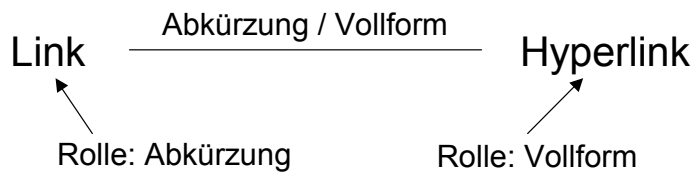
4. Überführung gerichteter Relationen in Assoziationen mit Rollen

Bei der Modellierung von Wissensnetzen lassen sich zwei Sichtweisen auf Relationen unterscheiden: Die Betrachtung aller Relationen als *zweistellig und gerichtet* versus deren Betrachtung als *ungerichtet mit Vergabe von Rollen*.

Werden Relationen als zweistellig und gerichtet betrachtet, können Sie durch Pfeile visualisiert werden. Zu vielen Relationen kann dann auch eine Umkehrrelation (Konverse) angegeben werden, bei symmetrischen Relationen sind Relation und Konverse identisch. Diese Sichtweise ist leicht verständlich. Ein weiterer wichtiger Vorteil liegt darin, dass es für beide Leserichtungen – Relation und Konverse – eine natürlichsprachliche Benennung gibt, aus der i.d.R. auch syntaktisch korrekte Sätze generiert werden können. Ein Beispiel aus dem HyTex-TermNet:



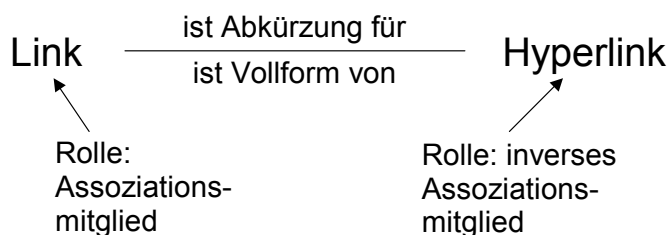
Die Relation „ist Abkürzung für“ hat hier die Konverse „ist Vollform von“. Diese Sichtweise findet sich in vielen Repräsentationsformalismen für Wissensnetze wieder, zum Beispiel in RDF(S) und auch in dem von uns benutzten Wissensnetz-Verwaltungswerkzeug K-Infinity. Andere Repräsentationsformalismen – darunter die Topic Maps – beruhen auf einer anderen Sichtweise für Relationen, bei der Relationen ungerichtet sind. Um Eindeutigkeit herzustellen, welche der durch die Relation verbundenen Entitäten welche Rolle in der Relation spielt, werden Rollen für die an der Relation beteiligten Entitäten eingeführt. Im Beispiel werden die Entitäten „Link“ und „Hyperlink“ durch die Rollen „Abkürzung“ und „Vollform“ gekennzeichnet:



Bei symmetrischen Relationen wie z.B. „Bruder“ können die Rollenindikatoren wegfallen. Im Falle der Topic Maps gilt es als guter Stil, die Relationen richtungsneutral – möglichst mit einem Substantiv – zu benennen, da eine Benennung wie „ist Abkürzung von“ im Falle einer ungerichteten Relation irreführend ist. Der Hauptvorteil dieser Sichtweise auf Relationen besteht darin, dass auch mehrstellige Relationen repräsentiert werden können. Es gibt jedoch keine allgemein bewährte Visualisierung.

Generell können gerichtete Relationen immer auf ungerichtete Relationen mit Rollen abgebildet werden, was für den umgekehrten Fall bei mehrstelligen Relationen nicht möglich ist. Will man gerichtete Relationen jedoch *automatisch* in ungerichtete überführen, so lassen sich bei der Neu-Benennung der Relationen (durch Substantive) und der Erzeugung von Rollenamen aus der ursprünglichen Benennung nur Heuristiken anwenden. Bei der Verarbeitung des TermNet benötigen wir aber eine solche automatische Überführung, da K-Infinity mit gerichteten Relationen, Topic Maps aber mit ungerichteten arbeiten. Dieser Schritt wird durch das Stylesheet `XTM2XTM-roles.xslt` geleistet.

Dabei liegen die Relationen in der Topic Map nach Schritt 2 (Transformation in das Topic-Map-Format (XTM)) bereits in der folgenden Form vor:



D.h. bei der Konvertierung nach Topic Maps durch das Stylesheet von K-Infinity wurden die beiden ursprünglichen Relationsnamen als gleichwertige Namen beibehalten und die Rollen „Assoziationsmitglied“ und „inverses Assoziationsmitglied“ eingefügt. Daraus lässt sich natürlich die ursprüngliche Leserichtung wieder eindeutig ableiten, diese Repräsentation wird der Idee der Rollenvergabe jedoch nicht gerecht und ist für Verwender der Topic Map u.U. nur schwer nachvollziehbar.

Das Stylesheet `XTM2XTM-roles.xslt` verwendet nun Heuristiken, um die Rollenbezeichnungen und Relationsnamen aus den (deutschsprachigen) Benennungen herzuleiten. Es liefert einen sinnvollen Relationsnamen und zwei Rollenbezeichnungen, wenn die ursprünglichen, d.h. gerichteten, Relationensnamen und – bei nicht-symmetrischen Relationen – ihre Konversen einem der beiden folgenden regulären Ausdrücke genügen:

- (1) `ist (.*)? [A-ZÄÖÜ].*` (für | von | als | zu)
- (2) `hat [A-ZÄÖÜ].*`

(Dabei steht `□` für ein Leerzeichen.)

Mögliche Relationsnamen, die verarbeitet werden können, sind also z.B. das Relationsnamenpaar

ist Abkürzung für
ist Vollform von

oder das Relationsnamenpaar

ist fremdsprachliches Äquivalent zu
ist Lehnwort zu

oder das Relationsnamenpaar

hat Teil
hat Ganzes

Daraus werden dann im ersten Fall der Relationsname Abkürzung - Vollform und die Rollenbezeichnungen Abkürzung und Vollform generiert, im zweiten Fall der Relationsname fremdsprachliches Äquivalent - Lehnwort und die Rollenbezeichnungen fremdsprachliches Äquivalent und Lehnwort.

Das Stylesheet nutzt die Tatsache, dass Substantive im Deutschen groß geschrieben werden, um Beispiele wie oben von Relationsnamen wie ist ähnlich zu zu unterscheiden, bei denen die Extraktion einer Rollenbezeichnung fehlschlagen würde.

In den folgenden Fällen nimmt das Stylesheet keine Veränderung vor:

- wenn eine symmetrische Relation vorliegt (z.B. ist orthographische Variante von)
- wenn der Relationsname dem regulären Ausdruck (1), die Konverse aber dem regulären Ausdruck (2) genügt (oder umgekehrt), wie z.B. bei ist Mitglied von und hat Mitglied (in diesem Fall könnte nur eine Rolle extrahiert werden, was keinen Vorteil hat)
- wenn nur der Relationsname, nur die Konverse oder keines von beidem einem der beiden regulären Ausdrücke genügt

Aufruf des Scripts:

```
saxon TermNet-XTM-ok.xtm ../XTM2XTM-roles.xslt  
> TermNet-XTM-ok-roles.xtm
```

5. Durchführung von Inferenzen und Überprüfungen

- Inferenz der Relation der Bedeutungsähnlichkeit (bei WordNet: Synonymie) zwischen Lexemen, die demselben Konzept zugeordnet sind
- Inferenz der Relation der Disjunktivität aus gleichen Attributwerten: Wenn zwei Konzepte kohyponym sind und dasselbe Attribut mit demselben Attributwert besitzen, dann sind sie disjunkt, d.h. zwischen ihnen soll eine Relation vom Typ Diskunktivität eingeführt werden.
Beispiel: Die Konzepte mit den Namen *1:1-Link und *1:n-Link sind beides direkte Hyponyme von *Link. Beide besitzen das Attribut Differenzierungskriterium mit dem Wert Stelligkeit. Sie werden durch die Disjunktivitäts-Relation miteinander verbunden (da jeder Link entweder ein 1:1-Link oder ein 1:n-Link ist, aber nie beides gleichzeitig sein kann).
- Inferenz der sprachkontaktbedingten Lexemkonkurrenz:
 - a) Wenn zwei verschiedene Lexeme durch die Relation ist_Lehnwort_zu mit demselben englischen Lexem (angezeigt durch Skopus) verbunden sind, wenn es also mehrere Lehnübersetzungen desselben englischen Lexems gibt, dann wird zwischen

den deutschen Lexemen die symmetrische Relation `ist_Lokalisierungsvariante_von` eingeführt.

- b) Wenn ein deutsches Lexem über die Relation `ist_Lehnwort_zu` mit einem englischen Lexem (angezeigt durch Skopus) verbunden ist und ein anderes deutsches Lexem über die Relation `ist_Lehnübersetzung_von` mit demselben englischen Lexem verbunden ist, dann wird zwischen den beiden deutschen Lexemen die Relation `konkurriert_sprachkontaktbedingt_mit` eingeführt.

Beispiel:

Link `ist_Lehnwort_zu` Link (engl.)

Verweis `ist_Lehnübersetzung_von` Link (engl.)

→ Link `variiert_sprachkontaktbedingt_mit` Verweis

- Überprüfung, ob es zu jedem Lexem mindestens ein Konzept gibt und zu jedem Konzept mindestens ein Lexem, ggf. Ausgabe von Warnungen.

Details: siehe Anhang.

Aufruf:

```
saxon TermNet-XTM-ok-roles.xtm ../XTM-ok2XTM-inferiert.xslt  
> TermNet-XTM-inferiert.xtm 2> errors.txt
```

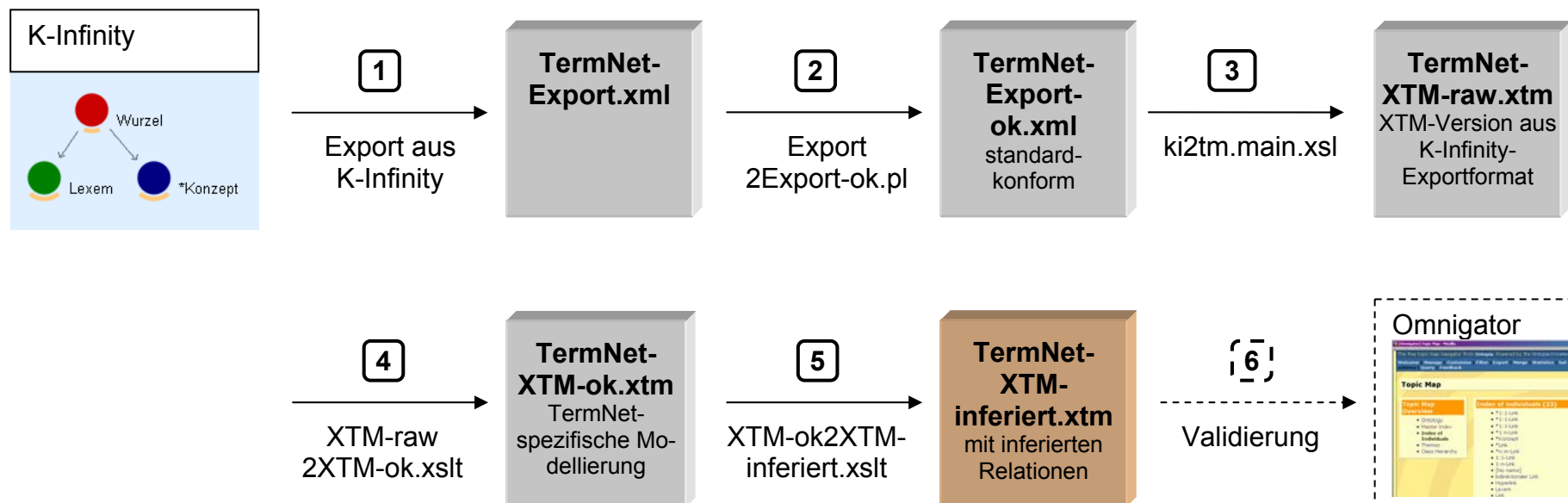
Die zusätzlich erzeugte Datei `errors.txt` enthält die bei Inkonsistenzen ausgegebenen Warnmeldungen.

6. Validierung mit dem Omnigator (optional)

Dieser Schritt ist nur in der Entwicklungsphase zur Qualitätskontrolle der Stylesheets und für einen Überblick über das terminologische Netz sinnvoll. Der Omnigator der Firma Ontopia (<http://www.ontopia.net/>) nimmt syntaktische Überprüfungen der Topic Map vor, auch solche, die nicht durch die Validierung mit der XTM-DTD abgedeckt werden, zum Beispiel werden ins Leere laufende Referenzen gefunden. Zudem kann durch die Navigation im Omnigator noch mal eine manuelle Überprüfung der Modellierung vorgenommen werden, z.B. kann die Klassenhierarchie angezeigt werden oder eine Liste aller Topics, die nicht als Typen oder Rollen dienen (d.h. für die TermNet-Topic-Map: eine Liste aller Lexeme und Konzepte).

7. Visualisierung der Topic Map in SVG. (Hier nicht näher beschrieben.)

Workflow aus Entwicklersicht



Workflow aus Anwendersicht



Anhang

1. Veränderungen des Stylesheets `ki2tm.main.xsl` von intelligent views

(Schritt 3)

- Keine Ausgabe der Dokumenttyp-Deklaration, da saxon damit sonst Probleme bekommt.
- Einfügen der XTM-Namensraum-Deklaration.
- Änderung der XSLT-Version auf 1.1 (da tree-valued-variables benutzt werden).

2. Aufgaben des Stylesheets `XTM-raw2XTM-ok`

(Schritt 4)

- Entfernen der `mergeMap`-Direktiven (**fertig**)
- Löschen von überflüssigen Topics: (**fertig**)
 - a) Entfernen des K-Infinity Topics mit dem Namen `export to NetNavigator`
 - b) Entfernen aller Topics mit dem Namen `Klassifikator`
 - c) Entfernen des Topics mit dem Namen `Wurzel` und Entfernen aller Assoziationen, bei denen ein Assoziationsmitglied das Topic mit dem Namen `Wurzel` ist
- Entfernen aller `<instanceOf>`-Elemente von Topics, die auf das Topic mit der ID `on-topsi.xtm#ot.metadata` verweisen. Die Topics selbst werden nicht gelöscht. (Topics von diesem Typ sind anscheinend Attribute, z.B. das Topic mit dem Namen `Differenzierungskriterium`). (**fertig**)
- K-Infinity superclass-subclass-Relation durch eigene ersetzen: (**fertig**)
 - a) Einfügen eigener Topics für die superclass-subclass-Relation und deren Rollen (mit PSIs)

```
<topic id="superclass-subclass">
  <subjectIdentity>
    <subjectIndicatorRef

xlink:href="http://www.topicmaps.org/xtm/1.0/core.xtm/#superclass-
subclass"/>
  </subjectIdentity>
  <baseName>
    <baseNameString>Oberbegriff - Unterbegriff</baseNameString>
  </baseName>
</topic>

<topic id="superclass">
  <subjectIdentity>
    <subjectIndicatorRef

xlink:href="http://www.topicmaps.org/xtm/1.0/core.xtm/#superclass"/>
  </subjectIdentity>
  <baseName>
    <baseNameString>Oberbegriff</baseNameString>
  </baseName>
</topic>
```

```

<topic id="superclass">
  <subjectIdentity>
    <subjectIndicatorRef

xlink:href="http://www.topicmaps.org/xtm/1.0/core.xtm/#subclass"/>
    </subjectIdentity>
    <baseName>
      <baseNameString>Unterbegriff</baseNameString>
    </baseName>
  </topic>

```

b) Ersetzen der Verweise in topicRef-Elementen:

```

ontopsi.xtm#superclass-subclass → #superclass-subclass
ontopsi.xtm#superclass          → #superclass
ontopsi.xtm#subclass            → #subclass

```

- Ersetzen der superclass-subclass-Relation zwischen dem Topic mit dem Namen Konzept und allen Konzepten sowie zwischen dem Topic mit dem Namen Lexem und allen Lexemen durch die Relation „WordNet-Klasse-Domänen-Instanz“. (fertig)

a) Einführung der Topics für den Assoziationstyp und die zugehörigen Rollen:

```

<topic id="wordnet-klasse-domaenen-instanz">
  <baseName>
    <baseNameString>WordNet-Klasse-Domänen-Instanz</baseNameString>
  </baseName>
</topic>

<topic id="wordnet-klasse">
  <baseName>
    <baseNameString>WordNet-Klasse</baseNameString>
  </baseName>
</topic>

<topic id="domaenen-instanz">
  <baseName>
    <baseNameString>Domänen-Instanz</baseNameString>
  </baseName>
</topic>

```

b) Ersetze in allen Assoziationen vom Typ superclass-subclass, in denen ein Mitglied das Topic mit dem Namen Konzept oder das Topic mit dem Namen Lexem ist, den Assoziationstyp und die Rollen entsprechend.

- Ersetzen des Attributs „Sprache“ mit dem Wert „EN“ durch einen Skopus. (nicht fertig)
Dazu werden alle Topics gesucht, die Topic-Anker mit resourceData-Elementen mit Verweis auf den Topic-Anker-Typ mit dem baseName „Sprache“ haben. Die Topic-Anker werden gelöscht, dafür wird folgender Skopus (mit Verweis auf einen PSI) für den baseName des Topics eingeführt:

```

<scope>
  <subjectIndicatorRef xlink:type="simple"
    xlink:href="http://www.topicmaps.org/xtm/1.0/
    language.xtm#en"/>
</scope>

```

Anschließend wird das Topic mit dem baseName „Sprache“ gelöscht.

- Ersetzen der von K-Infinity generierten IDs für die Topics „Lexem“ und „Konzept“ durch die IDs „lexem“ und „konzept“ zur effizienteren Weiterverarbeitung. Die Ersetzung geschieht auch in den Verweisen auf diese Topics. (erledigt)

```
<topic id="lexem">
  <baseName>
    <baseNameString>Lexem</baseNameString>
  </baseName>
</topic>

<topic id="konzept">
  <baseName>
    <baseNameString>Konzept</baseNameString>
  </baseName>
</topic>
```

- Wenn die Assoziation vom Typ "lexikalisiert" ist, wird die Rolle "iviews.xtm#assoc-member" durch Verweis auf das Topic "Konzept" ersetzt, die Rolle "iviews.xtm#inverse-assoc-member" durch Verweis auf das Topic "Lexem".

3. Aufgaben des Stylesheets XTM2XTM-roles.xslt

(fertig, alles oben beschrieben)

4. Aufgaben des Stylesheets xtm-ok2xtm-inferiert.xslt

(Schritt 5)

Dieses Stylesheet führt einerseits Inferenzen aus, andererseits nimmt es Konsistenzprüfungen vor und gibt ggf. Warnmeldungen aus.

- Inferenz der Bedeutungsähnlichkeits-Relation: (fertig)
Alle Lexeme, die durch die lexikalisiert-Relation mit demselben Konzept verbunden sind, werden untereinander durch die Relation „Bedeutungsähnlichkeit“ verbunden. Für diese wird ein neues Topic eingeführt:

```
<topic id="bedeutungsaeahnlichkeit">
  <baseName>
    <baseNameString>Bedeutungsähnlichkeit</baseNameString>
  </baseName>
</topic>
```

- Inferenz der Relation der Disjunkтивität (fertig)
Wenn zwei Konzepte kohyponym sind und dasselbe Attribut mit demselben Attributwert besitzen, dann sind sie disjunkt, d.h. zwischen ihnen soll eine symmetrische Relation vom Typ Alternative (Disjunktivitäts-Relation) eingeführt werden.
Beispiel: Die Konzepte mit den Namen *1:1-Link und *1:n-Link sind beides direkte Hyponyme von *Link. Beide besitzen das Attribut Differenzierungskriterium mit dem Wert Stelligkeit. Sie werden durch die Disjunktivitäts-Relation miteinander verbunden (da jeder Link entweder ein 1:1-Link oder ein 1:n-Link ist, aber nie beides gleichzeitig sein kann).

Übertragen auf die Topic Map heißt das:

Wenn zwei Topics über die subclass-superclass-Relation mit demselben Topic verbunden sind und beide ein <occurrence>-Element haben, das über <instanceOf> auf denselben Typ verweist und denselben <resourceData>-Wert hat, dann füge eine Assoziation zwischen den beiden Topics vom Typ **Alternative** ein.

Dazu wird folgender Assoziationstyp eingeführt:

```
<topic id="alternative">
  <baseName>
    <baseNameString>Alternative</baseNameString>
  </baseName>
</topic>
```

- Inferenz der sprachkontaktbedingten Lexemkonkurrenz: (angezeigt durch „Lokalisierungsvariante“)
- c) Wenn zwei verschiedene Lexeme durch die Relation `ist_Lehnwort_zu` mit demselben englischen Lexem verbunden sind, wenn es also mehrere Lehnübersetzungen zu demselben englischsprachigen Lexem gibt, dann wird zwischen den deutschen Lexemen die Relation `konkurriert_sprachkontaktbedingt_mit` eingeführt.
- d) Wenn ein deutsches Lexem über die Relation `ist_Lehnwort_zu` mit einem englischen Lexem verbunden ist und ein anderes deutsches Lexem über die Relation `ist_Lehnübersetzung_von` mit demselben englischen Lexem verbunden ist, dann wird zwischen den beiden deutschen Lexemen die Relation `konkurriert_sprachkontaktbedingt_mit` eingeführt.

Beispiel:

```
Link ist_Lehnwort_zu Link (engl.)
Verweis ist_Lehnübersetzung_von Link (engl.)
→ Link variiert_sprachkontaktbedingt_mit Verweis
```

Dies bedeutet für die Topic Map:

Suche alle englischen Lexeme, d.h. alle Lexeme, deren `baseName` mit dem folgenden Skopus versehen ist:

```
<scope>
  <subjectIndicatorRef xlink:type="simple"
    xlink:href="http://www.topicmaps.org/xtm/1.0/
    language.xtm#en"/>
</scope>
```

Sammele zu jedem englischen Lexem alle damit verbundenen Lexeme (Verbindung durch die Relationen `ist_Lehnwort_zu` und `ist_Lehnübersetzung_von`, dies ist aber ohnehin gewährleistet.), und verbinde diese paarweise durch die Relation `variiert_sprachkontaktbedingt_mit`. Da diese Relation symmetrisch ist, werden keine Rollen benötigt.

Es wird folgender Relationstyp eingefügt:

```
<topic id="sprachkontaktbedingte_lexemkonkurrenz">
  <baseName>
    <baseNameString>sprachkontaktbedingte Lexemkonkurrenz
  </baseNameString>
  </baseName>
</topic>
```

- Überprüfung, ob es zu jedem Lexem mindestens ein Konzept gibt und zu jedem Konzept mindestens ein Lexem. Falls das nicht der Fall ist, wird eine Warnmeldung ausgegeben.