

Ein automatisiertes Verfahren zur Sicherstellung der konventionsgerechten Bezeichnung von Modellelementen im Rahmen der konzeptionellen Modellierung

Jörg Becker¹, Patrick Delfmann², Sebastian Herwig³, Łukasz Lis⁴, Andrea Malsbender⁵ und Armin Stein⁶

Abstract: Eine kritische Voraussetzung für den erfolgreichen Einsatz von fach-konzeptionellen Modellen ist ihre Verständlichkeit und Vergleichbarkeit. Modelladressaten müssen in die Lage versetzt werden, den mit den Modellen vermittelten Inhalt eindeutig zu erkennen. Dies verlangt das Vorhandensein eines gemeinsamen Begriffsverständnisses unter den Modellierern. Insbesondere für Modellklassen, für deren Modellelemente sich satz- bzw. phrasenorientierte Bezeichnungspraktiken etabliert haben, bildet die Herstellung eines solchen gemeinsamen Begriffsverständnisses eine besondere Herausforderung. Der vorliegende Beitrag adressiert diese Problematik und präsentiert einen linguistischen Ansatz zur Formalisierung von Bezeichnungskonventionen für konzeptionelle Modelle. Die Konventionen werden auf Basis eines Domänenbegriffsmodells und natürlichsprachlicher Syntax formalisiert und im Rahmen der Modellierung automatisiert durchgesetzt. Der Schwerpunkt dieses Beitrags liegt auf der Analyse der durch Modellierer eingegebenen Modellelementbezeichnungen und dem im Falle von Konventionsverletzungen sich anschließenden automatischen Vorschlägen von konventionskonformen Alternativbezeichnern.

1 Motivation

Empirische Studien zeigen, dass sich arbeitsteilig erstellte konzeptionelle Modelle erheblich in ihren Bezeichnern unterscheiden können, selbst wenn der gleiche Sachverhalt adressiert wird [HS06]. Darüber hinaus ergeben sich solche Variationen auch dann, wenn die Modelle durch die gleichen Personen zu unterschiedlichen Zeitpunkten erstellt werden. Die für die Nutzung von konzeptionellen Modellen notwendige Vergleichbarkeit ist also nicht per se gegeben. Als Folge gestaltet sich die Analyse solcher Modelle – bspw. für Integrations- oder Benchmarkingzwecke – gemeinhin äußerst aufwändig [PS00; VTM08]. Die in den Bezeichnern auf unterschiedliche Art und Weise explizierte

¹ European Research Center for Information Systems (ERCIS), Westfälische Wilhelms-Universität Münster, Leonardo-Campus 3, 48149 Münster, Deutschland becker@ercis.uni-muenster.de

² European Research Center for Information Systems (ERCIS), Westfälische Wilhelms-Universität Münster, Leonardo-Campus 3, 48149 Münster, Deutschland delfmann@ercis.uni-muenster.de

³ European Research Center for Information Systems (ERCIS), Westfälische Wilhelms-Universität Münster, Leonardo-Campus 3, 48149 Münster, Deutschland herwig@ercis.uni-muenster.de

⁴ European Research Center for Information Systems (ERCIS), Westfälische Wilhelms-Universität Münster, Leonardo-Campus 3, 48149 Münster, Deutschland lis@ercis.uni-muenster.de

⁵ European Research Center for Information Systems (ERCIS), Westfälische Wilhelms-Universität Münster, Leonardo-Campus 3, 48149 Münster, Deutschland malsbender@ercis.uni-muenster.de

⁶ European Research Center for Information Systems (ERCIS), Westfälische Wilhelms-Universität Münster, Leonardo-Campus 3, 48149 Münster, Deutschland stein@ercis.uni-muenster.de

Information ist in geeigneter Weise zu vereinheitlichen, um die Vergleichbarkeit der Modelle herzustellen.

Im Kontrast zu bestehenden Ansätzen (Abschnitt 2) präsentiert der vorliegende Beitrag einen linguistischen Ansatz zur Formalisierung von Bezeichnungskonventionen für konzeptionelle Modelle, der Ambiguitäten in den Modellelementbezeichnern bereits während der Modellkonstruktion verhindert. Die Konventionen werden auf Basis eines Domänenbegriffsmodells und natürlichsprachlicher Syntax formalisiert und im Rahmen der Modellierung automatisiert durchgesetzt. Der Schwerpunkt dieses Beitrags liegt auf der Analyse der durch Modellierer eingegebenen Modellelementbezeichnungen und dem im Falle von Konventionsverletzungen sich anschließenden automatischen Vorschlägen von konventionskonformen Alternativbezeichnern. Im Rahmen des vorliegenden Beitrags wird die Funktionsweise des Ansatzes am Beispiel der englischen Sprache verdeutlicht. Hierzu werden zunächst verwandte Ansätze analysiert und vom vorliegenden Ansatz abgegrenzt (Abschnitt 2). Da eine konzeptionelle Vorstellung des Ansatzes bereits Gegenstand vorangegangener Beiträge war [DHL09a; DHL09b], wird auf die Konzeption nur kurz eingegangen (Abschnitt 3), bevor die formale Realisierung des Ansatzes detailliert dargestellt wird (Abschnitt 4). Die Anwendbarkeit des Ansatzes wird kurz anhand seiner Implementierung als Modellierungssoftware gezeigt (Abschnitt 5), bevor der Beitrag mit einem Überblick über bisherige Erfahrungen mit der Anwendung des Ansatzes und einem Ausblick auf weiteren Forschungsbedarf schließt (Abschnitt 6).

2 Verwandte Ansätze

In der Vergangenheit ist eine ganze Reihe von Ansätzen entwickelt worden, die das Problem der mangelnden Vergleichbarkeit von Bezeichnungen in konzeptionellen Modellen adressieren. Diese lassen sich durch zwei Dimensionen klassifizieren. Die erste Dimension unterteilt die Ansätze in *Ex-post*- und *Ex-ante*-Ansätze: Einerseits werden existierende Modelle untersucht, konfliktäre Bezeichner identifiziert und diese durch entsprechende Verfahren aufgelöst. Andererseits werden Verfahren vorgeschlagen mit dem Ziel, das Auftreten konfliktärer Bezeichner von vornherein zu verhindern. Die zweite Dimension klassifiziert die Ansätze nach der Struktur der betrachteten Bezeichner. Zum Einen werden ausschließlich *einzelne Wörter* als Bezeichner zugelassen, zum Anderen werden *Satzstrukturen, d. h. Phrasen*, analysiert. Darüber hinaus existiert eine Reihe von verwandten Ansätzen, die eine Normierung der natürlichen Sprache zum Ziel haben, deren Anwendungsgebiete sich jedoch *außerhalb der konzeptionellen Modellierung* befinden.

Frühe Ansätze der 1980er und 1990er Jahre adressieren bspw. das Problem der Datenbankintegration und setzen in einem ersten Schritt bei der Integration der Datenbankschemata an [BL84; BKK91; LB01; RB01]. Diese *einzelwortbezogenen Ex-post-Ansätze* fokussieren Datenmodellierungssprachen, häufig insbesondere Dialekte des Entity-Relationship-Modells (ERM [Ch76]). Durch den Vergleich von Bezeichnungen der Schemaelemente werden Ähnlichkeiten identifiziert, wobei betont wird, dass ein solcher

Vergleich ausschließlich manuell unter Einbeziehung der Schemakonstrukteure erfolgen kann.

Phrasenbezogene Ex-post-Ansätze untersuchen nicht einzelne Wörter als Bezeichner von Modellelementen, sondern sogenannte Konzepte. Konzepte umfassen mehrere Domänenbegriffe, die üblicherweise in Ontologien [Gr93; Gu98] abgelegt und durch semantische Beziehungen verbunden sind. So kann bspw. der Begriff „Rechnung“ mit dem Begriff „prüfen“ in Beziehung gesetzt werden, so dass bereits in der Ontologie ausgedrückt ist, dass Rechnungen geprüft werden können. Über diese Konzepte werden die Ähnlichkeitsbeziehungen zwischen Modellen hergestellt [Hö07; EKO07; Sa07].

Einzelwortbezogene Ex-ante-Ansätze, deren Ziel es ist, konfliktäre Bezeichner bereits im Vorfeld zu vermeiden, schlagen sogenannte Namenskonventionen vor. Namenskonventionen werden i. A. in Form von Glossaren oder auch Ontologien spezifiziert, die die für ein Modellierungsprojekt oder eine Modellierungsdomäne gültigen Begriffe enthalten. Während der Modellierung werden diese Vorgaben genutzt, um konventionsgerechte Modelle zu erstellen und so konfliktäre Bezeichner zu vermeiden. Entsprechende Ansätze werden bspw. von [Gr04; BDW07] für Prozessmodelle vorgeschlagen.

Phrasenbezogene Ex-ante-Ansätze wurden insbesondere in den 1990er Jahren entwickelt und fordern neben einer Standardisierung der gültigen Domänenbegriffe auch eine Standardisierung der zur Benennung von einzelnen Modellelementen erlaubten Satzstrukturen [Ro96; Ku00]. Die gültigen Begriffe werden dabei in sogenannten Fachbegriffsmodellen [KR98] abgelegt. Die Standardisierung der Satzstrukturen erfolgt durch textuelle Empfehlung. Ein alternativer Ansatz definiert Geschäftsobjekte sowie Einrichtungen auf diesen und führt sie als Anweisungen in Prozessmodellen zusammen [NZ98].

Das Problem konventionsgerechter Sprachverwendung prägt auch einige *Ansätze ohne Fokus auf konzeptionelle Modellierung*, die nicht direkt das Anwendungsgebiet der konzeptionellen Modellierung adressieren. [Or97] schlägt eine Normierung der natürlichen Sprache mithilfe eines Normlexikons der Fachbegriffe sowie einer auf Satzbauplänen basierenden Normgrammatik vor. Den Einsatzzweck einer solchen Normsprache sieht er dabei beim sogenannten *methodenneutralen Fachentwurf* von Informationssystemen im Übergang von fachlichen, natürlichsprachlichen Aussagen der Benutzer zu diagrammsprachlicher Darstellung des Entwurfs (z. B. in Form konzeptioneller Modelle). Dem Ziel der Anforderungserhebung widmen sich auch einige Ansätze aus dem Bereich des Requirements Engineering. [MK98] schlagen hierfür bspw. die Erstellung eines vorkonstruierbaren Glossars vor. Dies kann nicht nur manuell erfolgen, sondern auch in Form teilautomatisierter Extraktion aus natürlichsprachlichen Dokumenten. Ähnlich wie bei [Or97] werden darüber hinaus Möglichkeiten (teil-)automatisierter Überführung in konzeptionelle Modelle diskutiert.

Grundsätzlich lassen sich alle vorgestellten Ansätze der ersten vier Gruppen zur Herstellung der Vergleichbarkeit konzeptioneller Modelle anwenden. Speziell für Modellierungssprachen, die nicht einzelne Wörter sondern Sätze oder Phrasen als Bezeichner verwenden (z. B. Prozessmodellierungssprachen), sind jedoch insbesondere die einzelwort-

bezogenen Ansätze nicht verwendbar, da sich konfliktäre Bezeichner in Form von Satzstrukturen auf diese Weise nicht vermeiden lassen. Die hier vorgestellten einzelwortbezogenen Ex-ante-Ansätze wurden zwar für Prozessmodelle entwickelt, betrachten jedoch innerhalb der Satzstrukturen der Bezeichner ausschließlich einzelne Begriffe.

Hinsichtlich der Analyse von Satzstrukturen in Modellbezeichnern sind phrasenbezogene Ansätze flexibler. Ex-post-Ansätze sind in der Lage, mangelnde Vergleichbarkeit durch Gegenüberstellung der Modelle und Analyse ihrer Bezeichnungsstrukturen aufzulösen. Die vorherige Vermeidung konfliktärer Bezeichner erscheint hier allerdings zielführender, da dies eine aufwändige Analyse der fertiggestellten Modelle obsolet macht. Freilich ist eine vorherige Vermeidung konfliktärer Bezeichner nur dann möglich, wenn nicht bereits Modelle existieren, die ohne ausreichende Berücksichtigung von Bezeichnungskonventionen erstellt wurden. Die aufgeführten phrasenbezogenen Ex-ante-Ansätze bieten daher aus Sicht der Autoren eine vielversprechende Grundlage für die Formulierung eines Ansatzes zur Herstellung der Vergleichbarkeit konzeptioneller Modelle. Die Ansätze von [Ro96; Ku00] erlauben die Formulierung von Phrasenstrukturen einerseits und Begriffskonventionen andererseits. Es fehlt allerdings eine Formalisierung, die für eine automatisierte Durchsetzung von Namenskonventionen notwendig ist. Letztere ist für den Erfolg von Namenskonventionen kritisch. Der Ansatz von [NZ98] ist formalisiert, betrachtet jedoch ausschließlich die Benennung von Funktionen und Ereignissen Ereignisgesteuerter Prozessketten (EPK [KNS92]) mit einer definierten Menge an Phrasentypen. Obwohl die verwandten Ansätze zur Sprachnormierung nicht direkt auf das hier vorliegende Problemfeld übertragen werden können, liefern sie durchaus interessante Lösungsansätze. Die im Folgenden verwendeten Phrasenstrukturkonventionen entsprechen bspw. in etwa den Satzbauplänen von [Or97].

Der vorliegende Ansatz ist durch (1) seine Fokussierung auf konzeptionelle Modellierung, (2) seine Anwendbarkeit auf beliebige Modellierungssprachen, (3) die explizite Berücksichtigung von Satzfragmenten als Modellelementbezeichner, (4) seine Formalisierung und (5) die hierdurch automatisierbare Sicherstellung der konventionstreuen Modellierung von existenten Ansätzen abzugrenzen.

3 Konzeptionelle Lösung

Als konzeptionelle Basis dient ein Ordnungsrahmen für formalisierte Bezeichnungskonventionen (vgl. Abbildung 1, vgl. im Folgenden auch ausführlich [DHL09b]). Bezeichnungskonventionen basieren auf zwei wesentlichen linguistischen Bestandteilen, die abhängig von einem gegebenen Modellierungskontext (z. B. Domäne, Projekt, Unternehmen) zu spezifizieren sind. Einerseits unterliegt der natürlichsprachliche Teil jeder Modellierungssprache linguistischen Regeln. Hierbei handelt es sich neben gültigem Flexionsverhalten von Wörtern auch um die Einhaltung von syntaktischen Regeln innerhalb von Satzfragmenten. Andererseits wird durch die Modellierungsdomäne ein spezieller Domänenwortschatz vorgegeben, dessen ausschließliche Nutzung bei der Modellierung zu gewährleisten ist. Er beinhaltet eine Menge von gültigen Begriffen, die zur Nutzung

innerhalb der Modellbezeichnungen zur Verfügung stehen und repräsentiert damit eine Untermenge des durch die natürliche Sprache gegebenen Gesamtwortschatzes. Dabei sind die gültigen Substantive, Verben, Adjektive und Adverbien festzulegen, da diese die Domänensemantik enthalten. Andere Wortklassen wie z. B. Präpositionen sind generell domänenneutral und daher stets gültig.

Analog zur Einschränkung des Domänenwortschatzes ist auch die Syntax der zu verwendenden Bezeichnerphrasen dem Modellierungskontext entsprechend einzuschränken. Diese sogenannten Phrasenstrukturkonventionen werden auf Basis vordefinierter Phrasenstruktureregeln mittels sogenannter Phrasentypen verwirklicht (z. B. *<Verb, Imperativ>* *<Substantiv>*). Ähnlich wie Satzbaupläne bei [Or97] spezifizieren sie die erlaubte Anordnung von Wortarten innerhalb einer Phrase sowie deren jeweilige Flexion. Solche Phrasentypen repräsentieren eine Teilmenge der Syntax einer natürlichen Sprache (bspw. Englisch). Zur Beherrschbarkeit der hochkomplexen Syntax natürlicher Sprachen und zur Schaffung einer gemeinsamen Sprachbasis schränkt der Ansatz die Syntax ein.

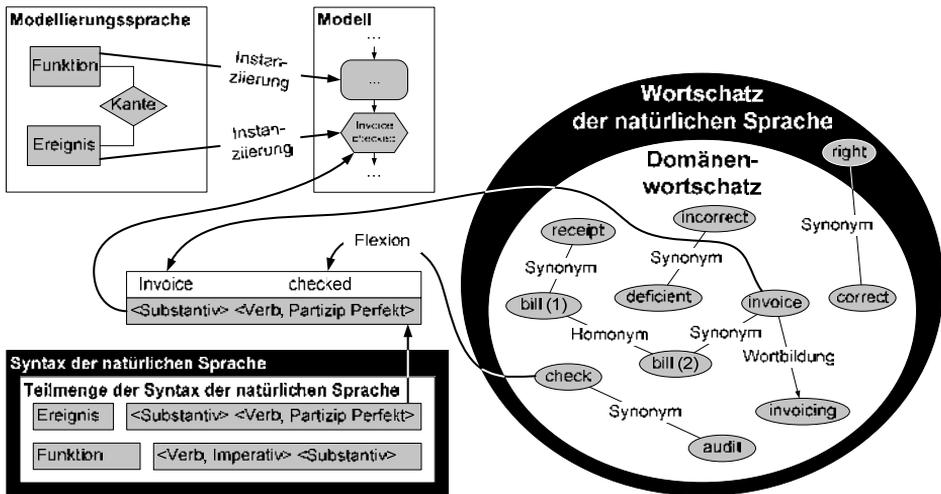


Abb. 1: Ordnungsrahmen für formalisierte Bezeichnungskonventionen

Die jeweiligen Konventionen sind abhängig von der Modellierungssprache. So werden etwa Funktionen einer EPK mit Verrichtungen bezeichnet (z. B. *<Verb, Imperativ>* *<Substantiv>*, konkret bspw. „Check invoice“), während Bezeichnungen von Ereignissen Zustände angeben (z. B. *<Substantiv>* *<Verb, Partizip Perfekt>*, konkret bspw. „Invoice checked“) [Ro96, Ku00]. Für jeden Modellelementtyp der jeweils verwendeten Modellierungssprache ist mindestens eine Phrasenstrukturkonvention zu definieren. Die konkreten Ausprägungen von Phrasenstrukturkonventionen werden mithilfe der gültigen Wörter aus dem Domänenwortschatz realisiert, die entsprechend ihrer Stellung im Satz bzw. der Bezeichnerphrase zu flektieren sind (vgl. Abbildung 1). Grundsätzlich lassen sich so beliebige syntaktische Muster formulieren. Diese Muster unterliegen ausschließlich der Restriktion, dass sie eine Teilmenge einer natürlichen Sprachsyntax repräsentieren müssen (d. h. Ketten von flektierten Wortformen, deren Instanzen entweder einen

gültigen natürlichsprachlichen Satz oder eine gültige natürlichsprachliche Phrase ergeben). Welche Muster gültig sein sollen, ist im Rahmen der Konventionendefinition festzulegen.

Zusammenfassend wird eine gemeinsame Sprachbasis für Modellierungsprojekte geschaffen, die auf einer formalen, natürlichsprachlichen Grammatik basiert. Deren Nicht-terminalsymbole bilden die Phrasenstrukturen, und deren Terminalsymbole werden durch die in die Phrasen eingesetzten und korrekt flektierten Wörter repräsentiert.

Die Bezeichnungskonventionen, d. h. der Domänenwortschatz und die Phrasenstrukturkonventionen, sind einmalig für jeden Modellierungskontext zu spezifizieren, wobei auf ggf. bestehenden Konventionen aufgebaut werden kann. Idealerweise sollte diese Aufgabe von einem Gremium von Modellierungs- und Domänenexperten vor Beginn des Modellierungsprojekts wahrgenommen werden, die in sowohl der Domäne üblichen Begriffe kennen als auch beurteilen können, welche Phrasenstrukturen für welche Modellierungssprachen geeignet sind.

Die Anwendbarkeit des vorgestellten Rahmens kann nur dann sichergestellt werden, wenn er als Modellierungswerkzeug implementiert bzw. in ein solches integriert wird. Trotz einer solchen Werkzeugunterstützung erscheint den Autoren die notwendige Auswahl von Phrasenstrukturen durch den Modellierer und die anschließende Auswahl von gültigen Domänenbegriffen (bspw. über Drop-down-Listen) sowie deren Flexion und Einsetzen in die Phrase jedoch wenig komfortabel und auch nicht effizient. Der Modellierer sollte in die Lage versetzt werden, in gewohnter Weise Modellelementbezeichner einzugeben. Die Analyse dieser Bezeichner hinsichtlich ihrer Konformität mit den Bezeichnungskonventionen und das Vorschlagen alternativer Bezeichnungen im Falle der Konventionsverletzung werden deswegen durch ein automatisiertes Verfahren realisiert.

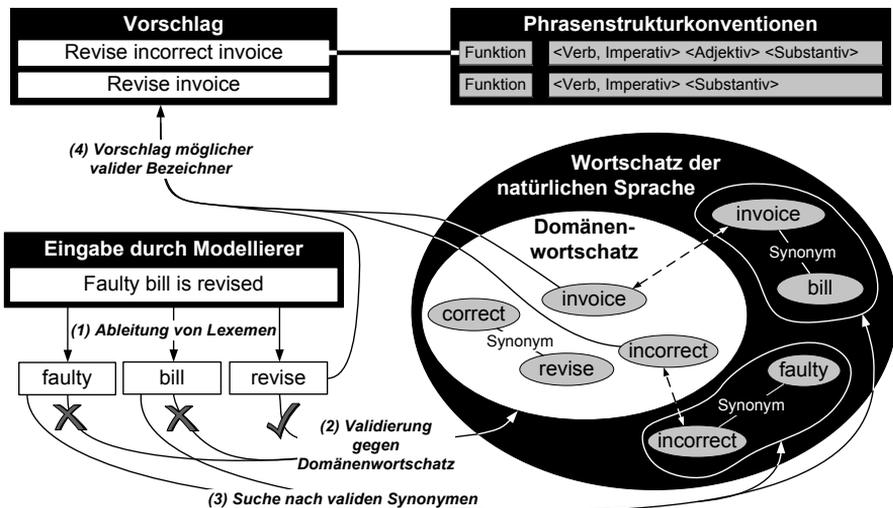


Abb. 2: Abgleich von Modellelementbezeichnern mit Modellierungskonventionen

Während der Modellierung werden die eingegebenen Bezeichnungen simultan mit den Bezeichnungskonventionen abgeglichen (vgl. im Folgenden Abbildung 2). Einerseits findet dieser Abgleich bezüglich der verwendeten Syntax statt. Andererseits werden die verwendeten Begriffe bezüglich ihrer Gültigkeit überprüft (2). Im Falle einer positiven Validierung wird die eingegebene Bezeichnung als konventionsgerecht angesehen und für das Modellelement übernommen. Genügt die eingegebene Bezeichnung den Konventionen nicht, werden dem Modellierer alternative Wörter bzw. Phrasen zur Verfügung gestellt. Im Falle ungültiger Wörter werden im Gesamtwortschatz der natürlichen Sprache Synonyme gesucht und mit dem Domänenwortschatz abgeglichen (3). Gibt der Modellierer bspw. das Wort „bill“ ein und entspricht dies nicht den Konventionen, so wird die Menge der Synonyme zu „bill“ mit dem Domänenwortschatz abgeglichen. Ein entsprechend gültiges Synonym ist bspw. „invoice“, das dem Modellierer als Alternative angeboten wird. Genügen alle Wörter den Konventionen, wird geprüft, ob die eingegebene Phrase den Phrasenstrukturkonventionen entspricht. Ist die Phrase nicht korrekt, so werden die zur Verfügung stehenden, bereits auf Gültigkeit überprüften Wörter in die möglichen Phrasen eingesetzt (4). Der Modellierer kann auf dieser Basis eine Entscheidung treffen, welche der aufgelisteten Phrasen den gewünschten Inhalt am besten wiedergibt. Durch die Möglichkeit, innerhalb des Domänenwortschatzes die einzelnen Beschreibungen der verwendeten Wörter nachzuschlagen, werden Missverständnisse auf Seiten des Modellierers ausgeschlossen.

Zur Realisierung eines solchen automatisierten Verfahrens sind neben der formalen Spezifikation gültiger Phrasen und Domänenbegriffe folgende technische Komponenten notwendig:

- **Linguistischer Parser (LP):** Linguistische Parser analysieren die Struktur eines Satzes oder einer Satzphrase und liefern die syntaktische Beschreibung der Phrasenstruktur sowie die Grundformen der verwendeten Wörter (die sogenannten *Lexeme*) zurück. Die konkrete Phrasenstruktur einer eingegebenen Satzphrase kann so gegen die erlaubten Phrasentypen und die verwendeten Lexeme gegen das Domänenlexikon validiert werden. Das hierfür entwickelte automatisierte Verfahren wird im Abschnitt 4 ausführlich präsentiert.
- **Lexikalische Dienste:** Entsprechen die eingegebenen Wörter nicht den Konventionen, so wird ein *allgemeines Lexikon der verwendeten natürlichen Sprache (AL)* nach entsprechenden Synonymen durchsucht, die wiederum mit dem Domänenwortschatz abgeglichen werden. Auf diese Weise können gültige Alternativen zu nicht validen Wörtern gefunden und vorgeschlagen werden. Voraussetzung hierfür ist, dass in der Datenbasis des Lexikons ein umfassender Wortschatz sowie entsprechende Synonymbeziehungen zwischen den Wörtern hinterlegt und abrufbar sind. Entsprechen die verwendeten Wörter oder Phrasen nicht den Konventionen, so ist eine alternative Phrase mit gültigen Wörtern zu konstruieren. Hierfür ist es notwendig, diese Wörter entsprechend ihrer Stellung im Satz automatisiert mithilfe eines *Flexionsdienstes (FD)* zu beugen. Wird ein Wort verwendet, welches nicht den Konventionen entspricht, wohl aber ein Wort,

das von diesem abstammt oder umgekehrt, so kann das gültige Wort mithilfe eines *Wortbildungsdienstes (WBD)* ermittelt werden.

4 Automatisierung

Grundlage für die Durchsetzung von Modellierungskonventionen bildet ein teilweise durch Nutzerinteraktion gestütztes automatisiertes Verfahren, das auf den Domänenwortschatz und die Phrasenstrukturkonventionen sowie die in Abschnitt 3 angesprochenen externen technischen Komponenten zugreift. Die Realisierung wird im Folgenden schrittweise mithilfe einer differenzierten Betrachtung erläutert.

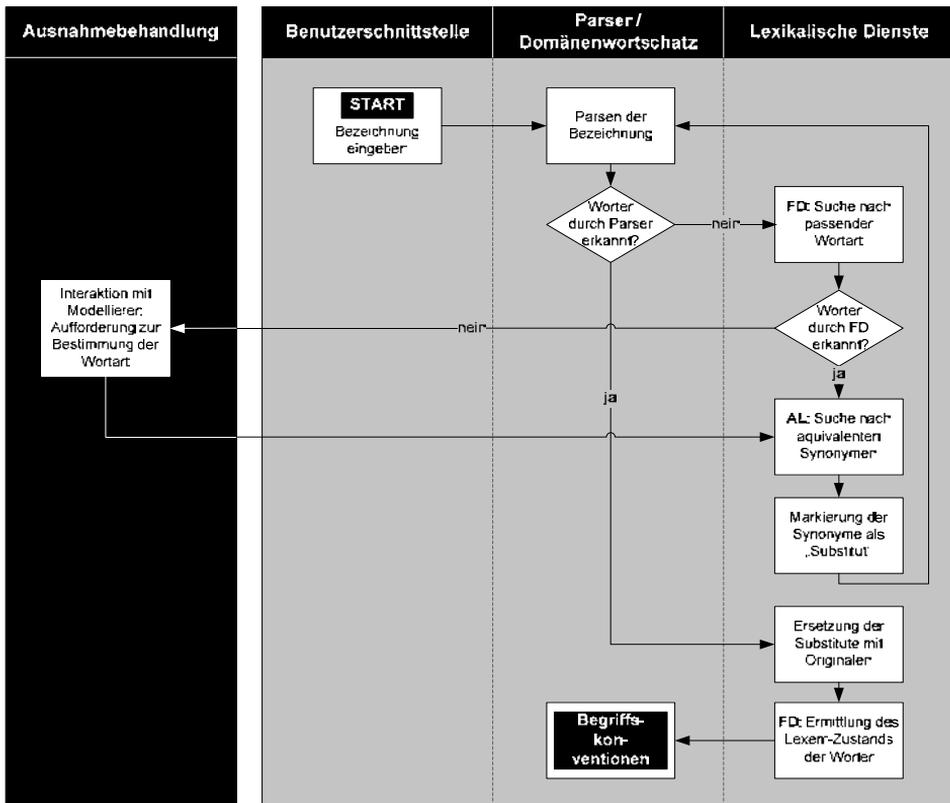


Abb. 3: Analyse der Benutzereingabe

Nach Eingabe der Bezeichnung eines Modellelements wird die Phrase mithilfe des Parsers in ihre Wortbestandteile zerlegt (vgl. Abbildung 3). Die einzelnen Wörter werden hinsichtlich ihrer Wortart und -form durch den Parser klassifiziert. Sollte der Parser ein Wort nicht erkennen, so wird das betroffene Wort an den Flexionsdienst übergeben. Mithilfe dieses Dienstes wird seine Wortart bestimmt. Ist auch durch den Flexionsdienst

keine eindeutige Bestimmung möglich, so findet eine Ausnahmebehandlung bezüglich des betroffenen Wortes statt und der Einsatz des Modellierers wird erforderlich. Kann durch den Flexionsdienst die Wortart bestimmt werden, so wird im allgemeinen Lexikon der natürlichen Sprache ein passendes Synonym ausgewählt und in der identischen Flexionsform in die Bezeichnung als Substitut eingesetzt. Diese Ersetzung erweist sich als unumgänglich, da der Parser ansonsten ggf. fehlerhafte Informationen bezüglich der Wortart und -form der anderen abhängigen Wörter der Phrase zurückliefert. Nach Ersetzung des nicht erkannten Wortes wird die modifizierte Bezeichnung erneut geparkt.

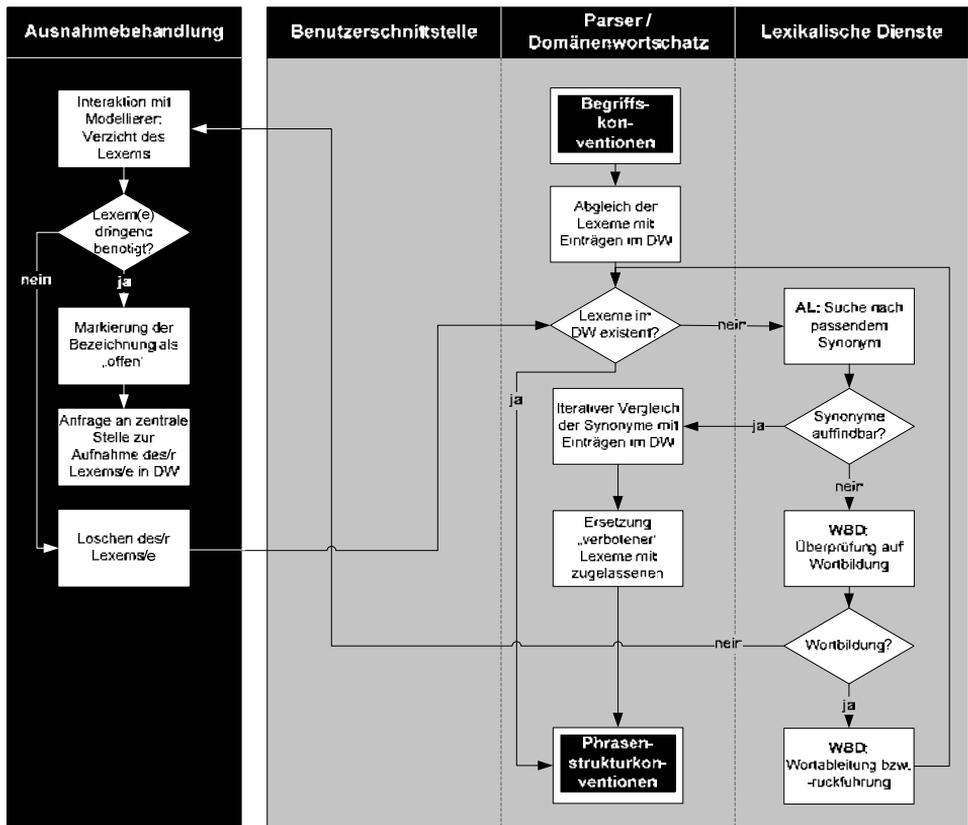


Abb. 4: Abgleich der Benutzereingabe mit dem Domänenwortschatz

Werden alle Wörter durch den Parser erkannt oder bereits nicht erkannte Wörter durch Substitute ersetzt, so werden die identifizierten Wortarten und -formen für die spätere Verwendung zwischengespeichert. In einem weiteren Schritt werden ggf. bestehende Substitute durch ihre Originale ersetzt. Die Wörter der Phrase werden durch den Flexionsdienst in ihre Lexemform transformiert und wiederum zwischengespeichert. Auf Basis dieser Lexeme wird der Abgleich der Begriffskonventionen im zweiten Schritt des Verfahrens durchgeführt. Dieser richtet sich an die domänenspezifischen Begriffe, die

durch die in der Phrase verwendeten Substantive, Verben und Adjektive repräsentiert sind. Der Abgleich mit dem Domänenwortschatz wird somit ausschließlich für diese Wortarten durchgeführt (vgl. Abbildung 4).

Die zuvor zwischengespeicherten Wörter werden mit den Einträgen im Domänenwortschatz abgeglichen. Ist ein Wort nicht Bestandteil des Domänenwortschatzes, wird mithilfe des allgemeinen Lexikons der verwendeten natürlichen Sprache nach zugehörigen Synonymen gesucht. Die aufgefundenen Synonyme werden im Anschluss ebenfalls mit den Einträgen im Domänenlexikon abgeglichen. Bei Auffinden eines passenden Synonyms wird das gemäß den Konventionen nicht erlaubte Wort durch das Synonym ersetzt.

Werden im allgemeinen Lexikon keine passenden Synonyme gefunden, so wird von einer Wortbildung ausgegangen. Eine Wortbildung bezeichnet die Abstammung eines Wortes von einem anderen. Bspw. stammt das Substantiv „approval“ vom Verb „approve“ ab. Ob eine Wortbildung vorliegt, wird mithilfe des Wortbildungsdienstes überprüft. Handelt es sich um eine solche, so werden entsprechend abstammende Wörter bzw. das zu Grunde liegende Wort ermittelt. Die resultierenden Wörter erfahren im Anschluss ebenfalls einen Abgleich mit den Einträgen im Domänenwortschatz. Handelt es sich laut Wortbildungsdienst nicht um eine Wortbildung, findet eine Ausnahmebehandlung statt. Hierbei wird mit dem Modellierer interagiert und ein möglicher Verzicht des Wortes erfragt. Stimmt der Modellierer einem solchen Verzicht zu, wird das betroffene Wort gelöscht und das Verfahren mit den restlichen Wörtern weiter durchlaufen. Anderenfalls wird die gesamte Bezeichnung als „offen“ markiert und eine Anfrage an eine zentrale Stelle – das bereits in Abschnitt 3 erwähnte Expertenteam – durchgeführt. Diese zentrale Stelle entscheidet, ob von Modellierern angefragte Erweiterungen mit in die Konventionen aufgenommen werden oder nicht. Im Falle der Löschung eines nicht validen Wortes wird die Bearbeitung der Bezeichnung mit den restlichen Wörtern fortgesetzt.

Sind alle Wörter mit dem Domänenwortschatz abgeglichen und dementsprechend als konventionstreu bestätigt, wird im dritten Teil des Verfahrens die Anpassung der Bezeichnung an die gegebenen Phrasenstrukturen vollzogen. Diese erfordert in einem ersten Schritt einen Abgleich der verfügbaren Wortarten mit den gegebenen Phrasenstrukturkonventionen (vgl. Abbildung 5). Sind zur Realisierung der Phrasenstrukturen nicht alle benötigten Wortarten vorhanden, findet erneut eine Ausnahmebehandlung statt. Bei dieser wird der Modellierer aufgefordert, das fehlende Wort der vorgegebenen Wortart zu ergänzen. Das dabei eingegebene Wort wird ebenfalls mit den Begriffskonventionen abgeglichen. Es wird zu der bereits bestehenden Liste der verfügbaren Wörter ergänzt und durchläuft den zweiten Schritt des Verfahrens erneut.

Sind alle Wortarten zur Realisierung der möglichen Phrasenstrukturen vorhanden, wird umgekehrt geprüft, ob bei der Realisierung möglicher Phrasen einige Wörter nicht verwendet werden können, d. h. ob die vom Modellierer ursprünglich eingegebene Phrase zu viele Wörter enthält. Der Modellierer wird auf einen möglichen Verlust hingewiesen und bezüglich der Notwendigkeit dieser Wörter befragt. Verzichtet der Modellierer auf überzählige Wörter, werden sie aus der Liste der verfügbaren gelöscht. Anderenfalls

wird die Bezeichnung als „offen“ markiert und eine Anfrage an die bereits erwähnte zentrale Stelle zur etwaigen Neuaufnahme eines Phrasentyps mit einer ausreichenden Zahl an Wörtern in die Phrasenstrukturkonventionen gestellt.

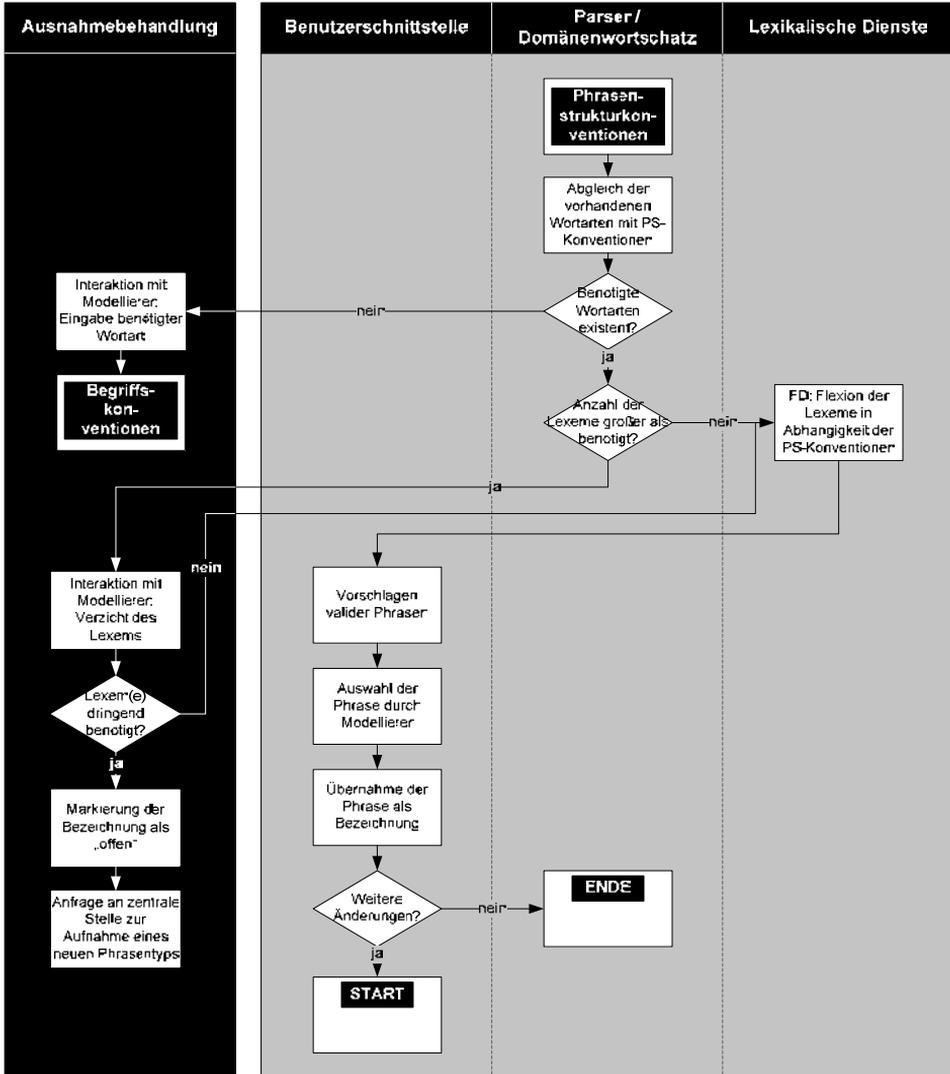


Abb. 5: Abgleich der Benutzereingabe mit den Phrasenstrukturkonventionen

Nachdem die notwendigen Wortarten hinsichtlich der Phrasenstrukturen abgeglichen sind, werden sie mithilfe des Flexionsdienstes in die jeweils benötigten Wortformen flektiert und an die Phrasenstruktur angepasst. Dabei werden dem Modellierer alle syntaktisch möglichen Phrasen angeboten. Existieren exemplarisch zwei Substantive (z. B.

„invoice“ und „receipt“) und ein Verb (z. B. „check“), und die Phrasenstrukturkonvention hat folgende Erscheinung: $\langle \text{Verb}, \text{Imperativ} \rangle \langle \text{Substantiv} \rangle$, so ergeben sich die Möglichkeiten „Check invoice“ und „Check receipt“.

Diese Möglichkeiten werden dem Modellierer in Form einer Liste präsentiert, woraufhin dieser eine Auswahl trifft. Die ausgewählte Phrase wird in die Bezeichnung übernommen und der Modellierer nach weiteren Änderungen gefragt. Sind weitere Änderungen gewünscht, wird die gesamte Phrase an den Startpunkt des Verfahrens übergeben und erneut untersucht. Wünscht der Modellierer keine weiteren Änderungen, wird die Bezeichnung als valide gekennzeichnet und die Ausführung abgeschlossen.

Die Komplexität dieses Verfahrens vermittelt einen Eindruck des Aufwands, der mit einer manuellen Sicherstellung der Konventionstreue von Modellelementbezeichnungen verbunden wäre. Durch die Automatisierung bleibt diese Komplexität für den Modellierer jedoch verborgen, da sie im Hintergrund abläuft. Der Modellierer erhält nur dann Hinweise und Alternativvorschläge für die Benennung eines Modellelements, wenn er eine Konvention verletzt.

5 Technische Realisierung

Eine technische Umsetzung des hier vorgestellten Ansatzes liegt als Modellierungsprototyp vor. Der Modellierer wird durch Popup-Fenster auf etwaige Konventionsverletzungen hingewiesen und erhält konventionstreu Alternativvorschläge (vgl. Abbildung 6).

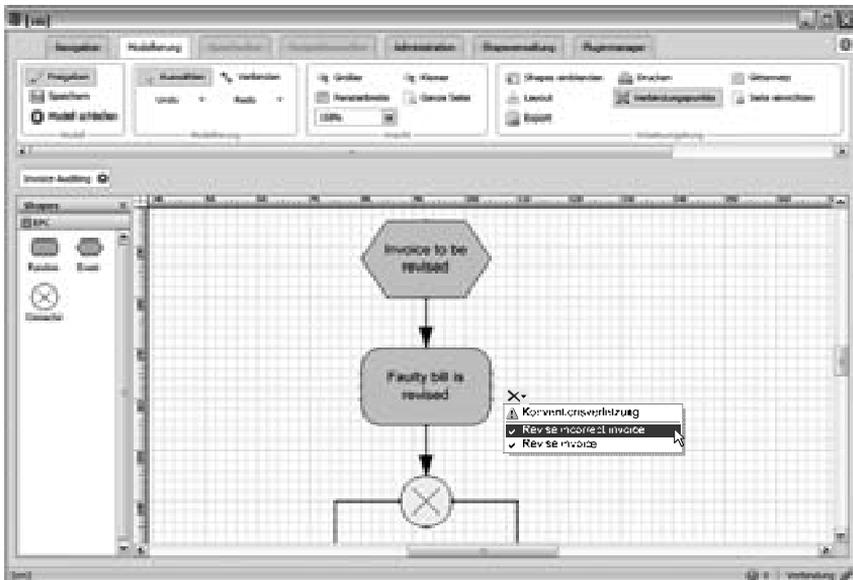


Abb. 6: Umsetzung des Ansatzes als Modellierungswerkzeug

Das im Modellierungswerkzeug realisierte Verfahren greift auf die bereits in Abschnitt 3 angesprochenen technischen Komponenten über Schnittstellen zu. Es existiert eine Vielzahl solcher Dienste, deren Funktionsumfang, Schnittstellenkonfiguration und Leistung erheblich variieren. Gemäß den Anforderungen des vorgestellten Ansatzes wurden diese Dienste analysiert und entsprechend geeignete ausgewählt:

- **Linguistischer Parser:** Als Parser mit umfangreicher Funktionalität stehen der *IPS-Parser* [We92], der *PET-Parser* [Ca00], der *XIP-Parser* [Xe09] und der *Stanford-Parser* [KM03] zur Verfügung. Zur Unterstützung des vorgestellten Ansatzes ist es notwendig, dass die grammatikalische Struktur der Phrase als Anordnung von Wortarten mit Angabe derer Flexionsformen durch einen Parser identifiziert und in strukturierter Form zur Validierung gegen erlaubte Phrasentypen bereitgestellt werden. Solange diese Ausgabe mit dem Spezifikationsformat der Phrasentypen kompatibel ist, sind die interne Arbeitsweise des Parsers und der dort eventuell verfolgte grammatikalische Formalismus für das hier vorgestellte Verfahren irrelevant. Die Parser *XIP* und *PET* sowie der *Stanford-Parser* bieten eine umfassende und strukturierte Ausgabe von grammatikalischen Eigenschaften. Beim *XIP*- und *PET-Parser* erfolgt dies in Form einer Datei. Der *Stanford-Parser* stellt hingegen eine native Schnittstelle zur Verfügung, über die sowohl Analysen angestoßen als auch die Ergebnisse abgefragt werden können. Entgegen dem *XIP*- und *IPS-Parser*, die ausschließlich als webbasierte Demonstrationen verfügbar sind, ist beim *Stanford-Parser* hierdurch auch die Möglichkeit gegeben, diesen eng in Softwareumgebungen zu integrieren. Um mit einem simultanen Abgleich der eingegebenen Bezeichnungen mit den Bezeichnungskonventionen einen durchgängigen Modellierungsfluss gewährleisten zu können, ist zudem eine geringe Antwortzeit notwendig. Von den untersuchten Parsern weisen sowohl der *PET-Parser* als auch der *Stanford-Parser* eine Verarbeitungsgeschwindigkeit auf, die einen durchgängigen Modellierungsfluss zulässt. Anhand der dargestellten Anforderungen und aufgrund der freien Verfügbarkeit hat sich der *Stanford-Parser* für den Modellierungsprototyp als am besten geeignet erwiesen.
- **Lexikalische Dienste:** Mit *GermaNet* [KL02], wortschatz.uni-leipzig.de und *canoo.net* sind Lexika für das Deutsche und mit *WordNet* [Fe98] das Englische verfügbar, die einen umfassenden Wortschatz sowie Synonymbeziehungen bereitstellen. Im Rahmen der prototypischen Realisierung wurde der vorgestellte Ansatz exemplarisch auf die englische Sprache angewendet. Als allgemeines Lexikon wird der weit verbreitete und sprachwissenschaftlich fundierte Dienst *WordNet* eingesetzt, der neben seinem Synonymdienst Wortbildungsfunktionalitäten bereitstellt. Darüber hinaus liefert *WordNet* auch Flexionsinformationen, eine automatisierte Durchführung der Flexion wird jedoch nicht unterstützt. Da im Englischen eine zumeist regelbasierte Flexion möglich ist, konnte diese Funktionalität in Eigenentwicklung realisiert werden.

6 Fazit und Ausblick

Die Einhaltung von Konventionen bei der Benennung von Elementen in konzeptionellen Modellen ermöglicht die Realisierung eines gemeinsamen Begriffsverständnisses. Modelladressaten können auf diese Weise den mit den Modellen vermittelten Inhalt eindeutig und einheitlich wahrnehmen. Nach Auffassung der Autoren kann die faktische Durchsetzung solcher Konventionen nur dann erfolgen, wenn sie bereits während der Modellierung automatisiert stattfindet.

Ferner darf eine effektive Unterstützung zur Einhaltung von Modellierungskonventionen den Modellierungsprozess nicht maßgeblich beeinträchtigen. Vielmehr muss die Modellierung in gewohnter Form stattfinden können, ohne dass der Modellierer sich selbst um die Einhaltung der Konventionen zu kümmern hat. Das in Abschnitt 4 vorgestellte Verfahren übernimmt diese Aufgabe und verlagert den mit der Einhaltung von Modellierungskonventionen verbundenen Aufwand in den Hintergrund.

Die Ergebnisse erster Tests des in Abschnitt 5 vorgestellten Modellierungsprototyps zeigen, dass eine performante und effektive Modelliererunterstützung realisiert werden konnte. Konkret wurden den Tests ein Domänenwortschatz mit ca. 200 Lexemen und ca. 15 linguistische Patterns zu Grunde gelegt. Zukünftige Tests werden auch wesentlich größere Wortschätze berücksichtigen. Es hat sich allerdings gezeigt, dass die Hauptrechenzeit nicht auf den Abgleich des Wortschatzes entfällt (der in Datenbankform vorliegt, so dass bekannte Optimierungen, bspw. Indizes, auch bei relativ großen Wortschätzen greifen), sondern auf die Recherche im allgemeinen Lexikon.

Das entwickelte Verfahren reduziert das Eingreifen des Modellierers auf ein Minimum. Wenn die eingegebenen Bezeichner mit den festgelegten Konventionen konform sind, werden diese direkt akzeptiert. Anderenfalls werden dem Modellierer den Konventionen entsprechende Alternativvorschläge unterbreitet, aus denen lediglich ein passender ausgewählt werden muss. Befürchtungen, der Modellierungsprozess könne durch lange Antwortzeiten der linguistischen Dienste und häufige Störungen in Form von alternativen Bezeichnungsvorschlägen zu sehr beeinträchtigt werden, konnten nicht bestätigt werden. Vielmehr bewegen sich die Antwortzeiten der Dienste im Untersekundenbereich. Die aus der Ermittlung der Alternativvorschläge resultierende Zeitverzögerung ist minimal und auf Seiten des Modellierers kaum wahrnehmbar. Das automatische Vorschlagen alternativer Bezeichner wird ersten Beobachtungen zufolge nicht als störend empfunden, sondern sogar begrüßt.

Obwohl die ersten Ergebnisse zufriedenstellend sind, steht eine umfangreiche Evaluation noch aus. Dafür sind in erster Linie Laborexperimente geplant, in denen die gleiche Modellierungsaufgabe von drei Gruppen (ohne Konventionen, mit papierbasierten Konventionen, mit Werkzeugunterstützung) bearbeitet wird. Dabei werden die Bearbeitungszeiten der einzelnen Gruppen verglichen. Zudem werden die resultierenden Modelle mit den Konventionen abgeglichen. In diesem Rahmen ist auch der zusätzliche Aufwand zu ermitteln, den eine nachträgliche Anpassung der Modelle nach sich zieht, welche ohne methodische Unterstützung zur Einhaltung der Modellierungskonventionen erstellt wur-

den. Darüber hinaus ist die Frage der Nutzerakzeptanz zu stellen. Obwohl erste Tests gezeigt haben, dass die Modellierer entgegen evtl. zu befürchtenden Akzeptanzbarrieren das „Erzwingen“ konventionstreuer Modelle sogar begrüßen, muss dieser Aspekt mittelfristig zum Gegenstand konkreter Fallstudien gemacht werden.

Der Umgang der zur Verfügung stehenden Parser mit Nominalphrasen (z. B. „bill of material“) stellt bisher noch ein Problem dar. Die Erkennung solcher Phrasen in gegebenen Satzfragmenten wird von den getesteten Parsern nicht unterstützt. Konkret müssten solche Nominalphrasen zusammengefasst und komplett an einen Synonymdienst übergeben werden, um valide Alternativphrasen zu finden. Eine entsprechende Erweiterung des Ansatzes ist geplant. Konkret könnte die Erkennung solcher Nominalphrasen analog zu den Phrasenstrukturkonventionen durch syntaktische Muster realisiert werden.

In der momentanen Implementierung des Ansatzes ist pro Bedeutung exakt ein Lexem vorgesehen. Zwar lässt das Domänenlexikon die Definition von Synonymen zu, als gültig gilt jedoch immer das als „dominant“ markierte Lexem. Im Weiteren ist zu prüfen, ob diese eher strenge Verfahrensweise beizubehalten oder zu lockern ist. Alternativ könnten bspw. auch die nicht dominanten, im Domänenlexikon als synonym definierten Lexeme als gültig angesehen werden.

Tests bzgl. der Implementierung des Ansatzes auf Basis der *deutschen* Sprache haben sowohl Erschwernisse als auch Erleichterungen gezeigt. Erleichterungen bestehen bspw. darin, dass anstatt von Substantivketten (wie im Englischen üblich) zusammengesetzte Substantive verwendet werden. Dies macht ihre Identifikation als Einheit einfacher. Erschwernisse bestehen bspw. in der automatisierten Flexion bzw. Rückführung in die Lexemform. Hier kommt ausschließlich tabellenbasierte Flexion in Frage, da sich der Großteil der deutschen Wörter nicht nach festen Regeln flektieren lässt. Entsprechende (kostenpflichtige) Onlinedienste stehen jedoch auch für die deutsche Sprache zur Verfügung (z. B. *canoo.net*). Eine vollständige Realisierung des Ansatzes auf Basis der deutschen Sprache ist mittelfristig geplant.

Literaturverzeichnis

- [BDW07] Born, M.; Dörr, F.; Weber, I.: User-friendly semantic annotation in business process modelling: In: Weske, M., Hacid, M.-S., Godart, C. (Hrsg.): Proceedings of the International Workshop on Human-Friendly Service Description, Discovery and Matchmaking (Hf-SDDM 2007) at the 8th International Conference on Web Information Systems Engineering (WISE 2007). Nancy 2007, S. 260-271.
- [BKK91] Bhargava, H. K.; Kimbrough, S. O.; Krishnan, R.: Unique Name Violations, a Problem for Model Integration or You Say Tomato, I Say Tomahto. ORSA Journal on Computing 3 (1991) 2, S. 107-120.
- [BL84] Batini, C.; Lenzerini, M.: A Methodology for Data Schema Integration in the Entity Relationship Model. IEEE Transactions on Software Engineering 10 (1984) 6, S. 650-663.

- [Ca00] Callmeier, U.: PET – A Platform for Experimentation with Efficient HPSG Processing Techniques. *Natural Language Engineering* 6 (2000) 1, S. 99-108.
- [Ch76] Chen, P. P.-S.: The Entity-Relationship Model – Toward a Unified View of Data. *ACM Transactions on Database Systems* 1 (1976) 1, S. 9-36.
- [DHL09a] Delfmann, P.; Herwig, S.; Lis, L.: Konfliktäre Bezeichnungen in Ereignisgesteuerten Prozessketten – Linguistische Analyse und Vorschlag eines Lösungsansatzes. In: *Proceedings des 8. GI-Workshops EPK 2009: Geschäftsprozessmanagement mit Ereignisgesteuerten Prozessketten*. Berlin 2009.
- [DHL09b] Delfmann, P.; Herwig, S.; Lis, L.: Unified Enterprise Knowledge Representation with Conceptual Models – Capturing Corporate Language in Naming Conventions. In: *Proceedings of the 30th International Conference on Information Systems (ICIS 2009)*. Phoenix, Arizona, USA, 2009.
- [EKO07] Ehrig, M.; Koschmider, A.; Oberweis, A.: Measuring Similarity between Semantic Business Process Models. In: *Roddick, J. F.; Hinze, A. (Hrsg.): Proceedings of the Fourth Asia-Pacific Conference on Conceptual Modelling (APCCM) 2007*. Ballarat 2007, S. 71-80.
- [Fe98] Fellbaum, C. (Hrsg.): *WordNet: An Electronic Lexical Database*. Cambridge 1998.
- [Gr04] Greco, G.; Guzzo, A.; Pontieri, L.; Saccà, D.: An ontology-driven process modeling framework. In: *Galindo, F.; Takizawa, M.; Traunmüller, R. (Hrsg.): Proceedings of the 15th International Conference on Database and Expert Systems Applications (DEXA 2004)*. Zaragoza 2004, S. 13-23.
- [Gr93] Gruber, T. R.: A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 5 (1993) 2, S. 199-220.
- [Gu98] Guarino, N.: Formal Ontology and Information Systems. In: *N. Guarino (Hrsg.): Proceedings of the 1st International Conference on Formal Ontologies in Information Systems*. Trento 1998, S. 3-15.
- [Hö07] Höfferer, P.: Achieving business process model interoperability using metamodels and ontologies. In: *Österle, H.; Schelp, J.; Winter, R. (Hrsg.): Proceedings of the 15th European Conference on Information Systems*. St. Gallen 2007, S. 1620-1631.
- [HS06] Hadar, I.; Soffer, P.: Variations in conceptual modeling: classification and ontological analysis. *Journal of the AIS* 7 (2006) 8, S. 568-592.
- [KL02] Kunze, C.; Lemnitzer, L.: GermaNet – representation, visualization, application. In: *Proceedings of the LREC 2002 Conference, Volume V.*, S. 1485-1491.
- [KM03] Klein, D.; Manning, C. D.: Accurate unlexicalized parsing. In: *Proceedings of the 41st Meeting of the Association for Computational Linguistics – Volume 1*. Sapporo 2003, S. 423-430.
- [KNS92] Keller, G.; Nüttgens, M.; Scheer, A.-W.: Semantische Prozeßmodellierung auf der Grundlage „Ereignisgesteuerter Prozeßketten (EPK)“. In: *A.-W. Scheer (Hrsg.): Veröffentlichungen des Instituts für Wirtschaftsinformatik 89*. Saarbrücken 1992.
- [KR98] Kugeler M.; Rosemann, M.: Fachbegriffsmodellierung für betriebliche Informationssysteme und zur Unterstützung der Unternehmenskommunikation. In: *Informationssystem Architekturen*. Fachausschuss 5.2 der Gesellschaft für Informatik e. V. (GI), 5

- (1998) 2, S. 8-15.
- [Ku00] Kugeler, M.: Informationsmodellbasierte Organisationsgestaltung. Modellierungskonventionen und Referenzvorgehensmodell zur prozessorientierten Reorganisation. Berlin 2000.
- [LB01] Lawrence, R.; Barker, K.: Integrating Relational Database Schemas using a Standardized Dictionary. In: Proceedings of the 2001 ACM symposium on Applied computing (SAC). Las Vegas 2001.
- [MK98] Mayr, H. C.; Kop, C.: Conceptual Predesign – Bridging the Gap between Requirements and Conceptual Design. In: Proceedings of the 3rd International Conference on Requirements Engineering (ICRE'98) 1998, S. 90-100.
- [NZ98] Nüttgens, M.; Zimmermann, V.: Geschäftsprozeßmodellierung mit der objektorientierten Ereignisgesteuerten Prozeßkette (oEPK). In: Maicher, M.; Scheruhn, H.-J. (Hrsg.): Informationsmodellierung – Branchen, Software- und Vorgehensreferenzmodelle und Werkzeuge. Wiesbaden 1998, S. 23-36.
- [Or97] Ortner, E.: Methodenneutraler Fachentwurf. Stuttgart 1997.
- [PS00] Phalp, K.; Shepperd, M.: Quantitative analysis of static models of processes. Journal of Systems and Software 52 (2000) 2-3, S. 105-112.
- [RB01] Rahm, E.; Bernstein, P. A.: A Survey of Approaches to Automatic Schema Matching. The International Journal on Very Large Data Bases 10 (2001) 4, S. 334-350.
- [Ro96] Rosemann, M.: Komplexitätsmanagement in Prozeßmodellen. Methodenspezifische Gestaltungsempfehlungen für die Informationsmodellierung. Wiesbaden 1996.
- [Sa07] Sabetzadeh, M.; Nejadi, S.; Easterbrook, S.; Chechik, M.: A Relationship-Driven Framework for Model Merging, Workshop on Modeling in Software Engineering (MiSE'07). 29th International Conference on Software Engineering, Minneapolis 2007.
- [VTM08] Vergidis, K.; Tiwari, A.; Majeed, B.: Business process analysis and optimization: beyond reengineering. IEEE Transactions on Systems, Man, and Cybernetics 38 (2008) 1, S. 69-82.
- [We92] Wehrli, E.; Clar, R.; Merlo, P.; Ramluckun, M.: The IPS system. In: Boitet, C. (Hrsg.): Actes du quinzième colloque international en linguistique informatique. Coling 1992, S. 870-874.
- [Xe09] Xerox Incremental Parser (XIP). <http://orchid.xrce.xerox.com>. 12.09.2009.

