Modellierung eines semantischen Wissensnetzes für lexikographische Anwendungen am Beispiel der Duden-Ontologie

Melina Alexa melina.alexa@bi-media.de, Tel. +49 621 3901-330

Die Duden-Ontologie hat mittlerweile eine mehr als 10-jährige Geschichte, von denen ich hier verschiedene Aspekte vorstellen möchte. Zu Beginn stand die Vision alle Duden-Werke in einer zentralen Quelle zu speichern, aus der heraus alle bisherigen und je nach Bedarf auch neue Werke in verschiedenen Formaten und für verschiedene Medien weitgehend automatisch produziert werden können. Darüber hinaus sollten auch sprachtechnologische Produkte diese Quelle nutzen und so von einer permanenten Pflege und kontinuierlichen Überarbeitung und Ergänzung der zentralen Ressource unmittelbar profitieren können. In diesem Papier werde ich zunächst die Motivation und die Ziele erläutern, die uns zu Beginn des Projektes veranlasst haben, uns in dieses Abenteuer zu stürzen. Aus diesen Motiven und Zielen leiteten sich die Anforderungen an die Datenmodellierung ab. Das daraus resultierende Datenmodell werde ich kurz darstellen und anschließend auf die Implementierung eingehen. Zum Schluss gehe ich auf den Einsatz des Wissensnetzes in der Verlagspraxis ein.

1. Motivation und Ziele

Zu Beginn des Projektes im Jahr 2000 wurden die lexikographischen Daten mit einem SGML-basierten Redaktionssystem titelbezogen erstellt, aus dem heraus die Daten für den Drucksatz erzeugt wurden. Obwohl die Inhalte der verschiedenen Wörterbücher sich zum Teil überlappten, mussten die Daten für jedes Werk extra erfasst werden, eine Wiederverwendung war wegen der Werkbezogenheit des Redaktionsprozesses schwierig. Auch die Pflege war ineffizient, da Änderungen, Ergänzungen und Korrekturen der Daten für jedes Werk einzeln nachgearbeitet werden mussten, wie es z.B. während der Rechtschreibreform sehr häufig vorkam. Und neben der Ineffizienz gesellte sich noch die erhöhte Fehleranfälligkeit durch die Doppelarbeit als gravierender Nachteil hinzu.

Daher kam der Wunsch auf, alle Wörterbuchdaten in einer einzigen lexikographischen Ressource zu speichern und zu pflegen, um damit die unterschiedlichen Produkte und Dienstleistungen möglichst flexibel und medienneutral bewerkstelligen zu können. D.h., es sollte möglich sein, eine Neuauflage eines Dudenwörterbuchs aus der zentralen Ressource automatisch in der aktuellen Version aller in Frage kommenden Wörterbuchartikel zu exportieren und das möglichst ohne weitere manuelle Korrekturen an dem zu druckenden Text. Es sollte auch möglich sein, ein komplett neues Wörterbuch zu definieren, d.h. eine Anzahl von Lemmata nach verschiedenen Kriterien festzulegen und ebenso die Inhalte für jeden Eintrag durch Regeln zu definieren, um so den Wörterbuchtext automatisch erzeugen zu lassen. Ob eine Publikation in elektronischer Form, als gewöhnliches Buch oder als eine Kombination aus Buch plus CD angedacht war: dies sollte keinen zusätzlichen Aufwand erfordern. Damals hatte man bei elektronischen Produkten noch eher an CDs oder Online-Angebote gedacht, heute sind es auch E-Books bzw. Apps. Darüber hinaus sollten auch Wörterbuchdaten für sprachtechnologische Software aus der zentralen Ressource exportiert werden können, um auch solche Anwendungen von der permanenten Pflege der lexikographischen Inhalte profitieren zu lassen.

Uns war damals schon klar, dass man ein solch ehrgeiziges Ziel nur erreichen konnte, wenn die Wörterbuchinformationen formal und explizit in einer Datenbank repräsentiert waren. Dazu musste eine effiziente und konsistente Pflege der lexikographischen Inhalte organisiert und gewährleistet werden. Daher wurde ein ambitioniertes Projekt vom Verlag initiiert, an dem die Firma intelligent views GmbH (http://www.i-views.de/web/) und die Fraunhofer Gesellschaft (http://www.ipsi.fraunhofer.de/ipsi/nav/ipsi_f_profile.html) beteiligt waren. Die speziell auf unsere Bedürfnisse hin angepasste Software wurde von intelligent views bereitgestellt und wird bis heute von dort auch gewartet.

Das Ziel dieses Projektes war der Aufbau einer mächtigen Ressource der deutschen Sprache, die sämtliche Informationen und Informationstypen der Dudenwörterbücher beinhaltet und eine Wiederverwendung der Wörterbuchinhalte ohne Informationsverlust für Print- *und* elektronische Anwendungen ermöglicht. Es sollte auch möglich sein, die Ressource durch neue Informationen über die vorhandenen hinaus zu erweitern. Eine redundanzarme Speicherung sowie effiziente Verwaltung, Aktualisierung und Pflege der Wörterbuchinhalte mussten gewährleistet werden. Denn wichtig war von Anfang an, dass alle Duden-(Print-)Wörterbücher mindestens so effizient produziert werden können wie früher aus den verschiedenen titelbezogenen Datenbanken, d.h., es war eine formale und explizite Repräsentation des gesamten lexikographischen Wissens in den Wörterbüchern nötig.

Das Datenmodell muss auch flexibel und erweiterbar sein, um neue Anforderungen, z.B. Repräsentation von weiteren lexikographischen Informationen, erfüllen zu können. Darüber hinaus werden auch mächtige Datenexportmöglichkeiten benötigt, um den vielfältigen Schnittstellen der bei der Produktion nachfolgenden Prozesse genügen zu können.

2. Datenmodell: Duden-Ontologie

Auf Basis all dieser Überlegungen wurde die sogenannte Duden-Ontologie entwickelt. Grundlage unseres Modells ist eine konzeptbasierte Repräsentation, die die Definition semantischer Relationen zwischen den Konzepten ermöglicht. Die Wörterbuchdaten sind mittels einer generischen Hierarchie-Relation klassifiziert, analog zu einer Ontologie.

Eine Ontologie im informatischen Sinne bietet eine formale Methode, Mengen von Individuen zu strukturieren, wobei die Menge der Individuen die Extension eines Begriffs (*Konzept*) ist. Diese Konzepte sind gemäß einer strikten Hierarchie-Relation verbunden. Das ermöglicht die Faktorisierung von gemeinsamen Informationen auf eine abstraktere Ebene. Die wesentlichen Elemente einer Ontologie sind daher typischerweise:

- eine Klassifikation von Konzepten auf Basis der generischen Hierarchierelation (SUB-CONCEPT_OF-Relation) und
- die Unterscheidung zwischen Individuen und Konzepten, bei der ein Individuum zu einem Konzept durch eine INSTANCE_OF-Relation verbunden ist.

Eine Ontologie modelliert eine Bedeutungswelt, in der die Bedeutungen als Konzepte repräsentiert werden. Konkrete Entitäten, wie konkrete Personen, Organisationen, Institutionen, geographische Orte etc., z.B. *Immanuel Kant*, *EU*, *Mannheim*, *IDS*, *Olympische Spiele Athen 2004*, sind darin die Individuen. Entsprechende Konzepte sind *Europäer/-in*, *Philosoph*, *politische Organisation*, *Großstadt*, *wissenschaftliches Institut*, *moderne Olympische Spiele*.

Unsere Idee für die Duden-Ontologie ist, dasselbe Prinzip für die Wörter selbst zu verwenden und eine Ontologie der ‚Welt der Wörter' zu kreieren. Darin werden die Wörter (*Lemmata*) einer Sprache, in unserem Fall der deutschen Sprache, formal als *Individuen* modelliert. Die morphologischen und Grammatikklassen der Sprache, die die Lemmata klassifizieren, sind in diesem Modell *Konzepte.*

Daraus ergibt sich eine Art morphosyntaktische Ontologie über die Welt der Wörter. Diese könnte man als eine weitere Dimension der ersten Ontologie sehen, die die Bedeutungen und die Weltobjekte repräsentiert.

2.1 Term: die Brücke zwischen Lemma und Konzept

Durch diesen Ansatz entstehen also zwei Ontologien: eine ‚normale' Ontologie und eine grammatische Ontologie, die eine strukturiert die Bedeutungen und die andere die Benennungen. Als Brücke zwischen den zwei Ontologien benutzen wir eine Denotationsrelation, um ein Lemma einer Lesart/Bedeutung oder mehreren Bedeutungen zuzuordnen.

Jede Bedeutung eines Lemmas ist eine Rolle, die das Lemma im ‚Sprachspiel' spielt. Jede Rolle ist als einziges Objekt repräsentiert, dieses nennen wir Term. Ein Lemma hat oft mehr als eine Bedeutung, daher können einem Lemma mehrere Terme zugeteilt werden. Jede Bedeutung eines Lemmas ist durch ein einziges Konzeptobjekt repräsentiert.

Auf der anderen Seite können auch einem Konzept mehrere Terme zugeteilt werden, sodass es häufig mit mehr als einem Lemma verbunden ist. Dadurch entsteht die Synonymie-Relation: Zwei Lemmata sind synonym, wenn sie über verschiedene Terme zu dem gleichen Konzept führen.

Man sieht in Abbildung 1 das Top-Konzept der Bedeutungswelt, *Topic*, und das Top-Konzept der morphosyntaktischen Ontologie, *Benennung*. Die Kluft zwischen den Topics und den Lemmata wird durch die Terme überbrückt. Alle Eigenschaften, die für *Topic*, *Benennung* und *Term* im Modell gemeinsam sind, werden zum *BasisObjekt* faktorisiert. Dadurch, dass man Terme als explizite Objekte speichert, können sie auch die Verlinkung zu Anwendungsbeispielen und Zitaten leisten. Auf diese Weise hat man spezifischere Selektionsmöglichkeiten, da so Beispiele für ein Lemma in einer bestimmten Lesart gefiltert werden können.

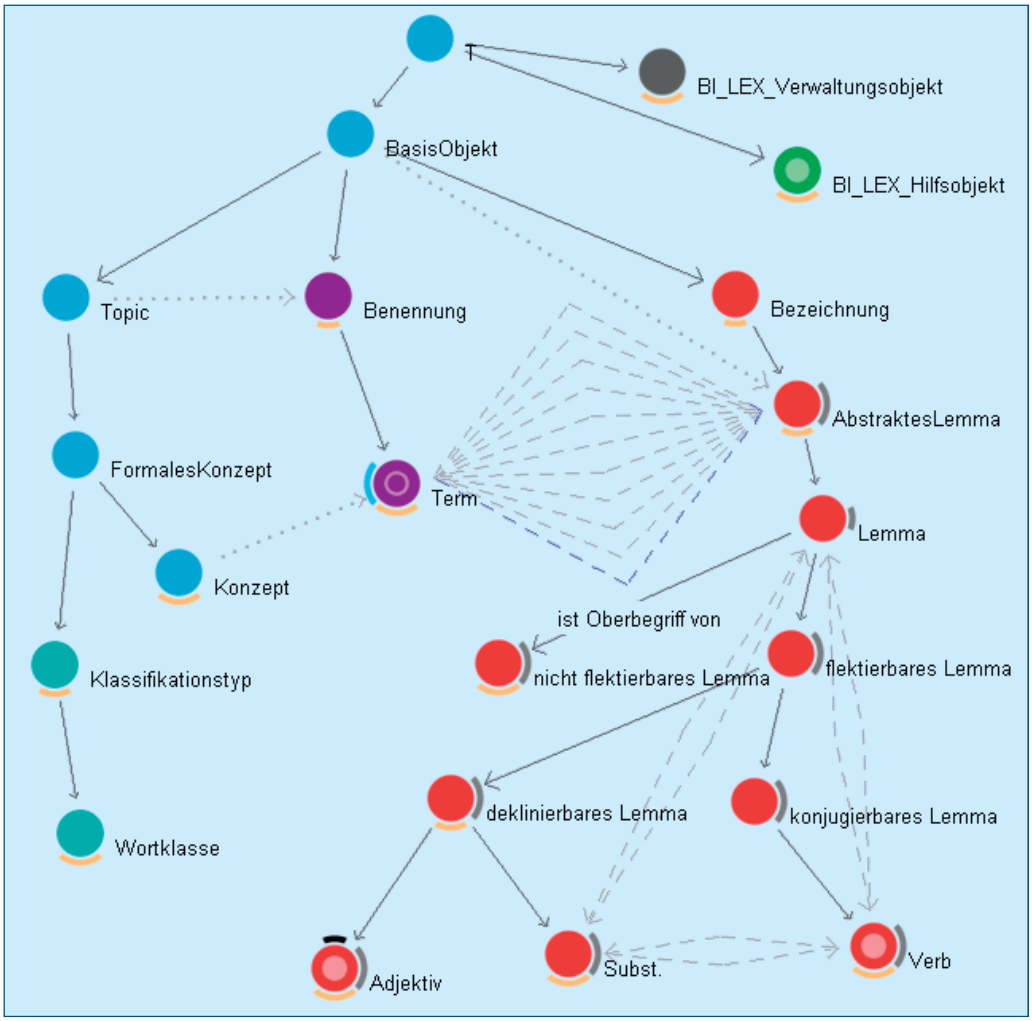

Abb. 1: Top-Level der Duden-Ontologie

Die Wortklassenhierarchie, die die Welt der Wörter gruppiert bzw. klassifiziert, wird in Abbildung 2 gezeigt.

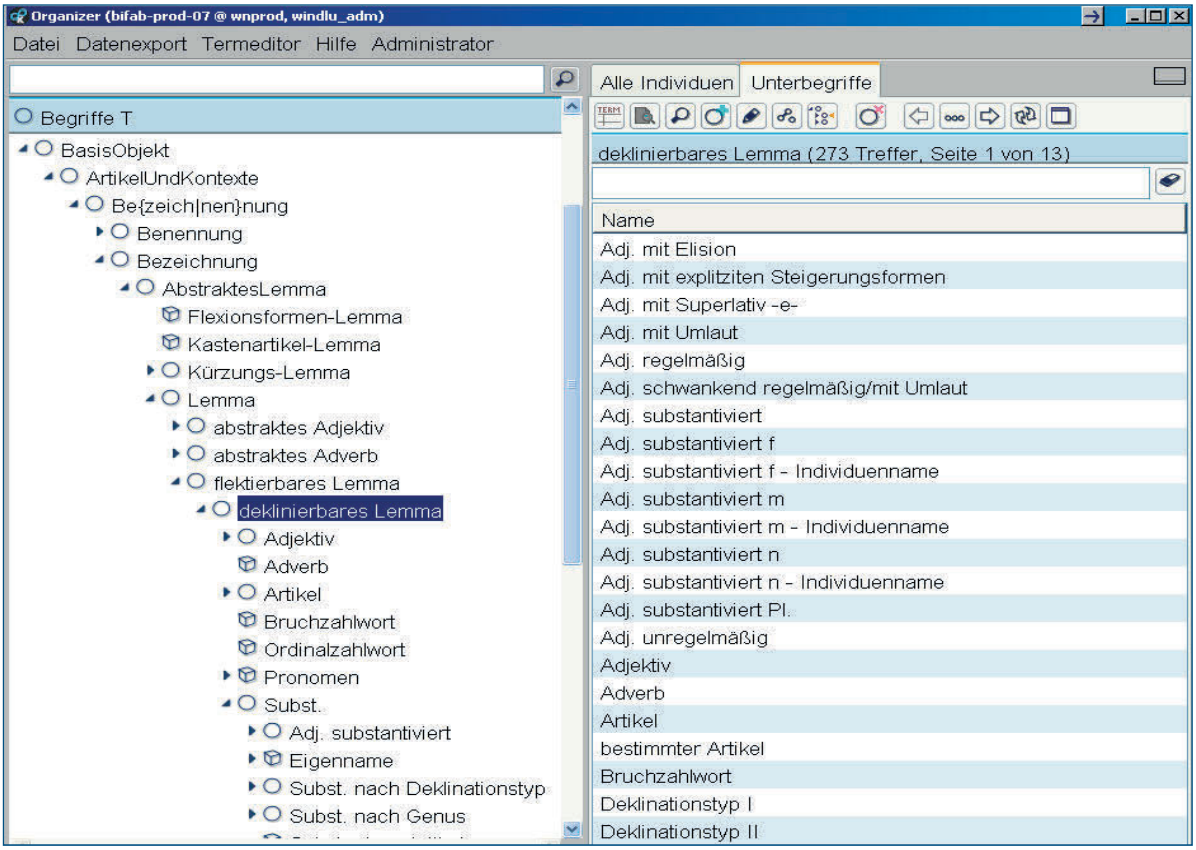

Abb. 2: Wortklassenhierarchie in der Duden-Ontologie

2.2 Implementierung

Die Ontologie ist als Objektnetz repräsentiert. Jedes Konzept steht in Beziehung zu seinen Ober- oder Unterkonzepten. Dadurch können bereits definierte Attribute und Relationen von den allgemeinen zu den spezifischen Konzepten vererbt werden. Des Weiteren gibt es die Möglichkeit, mit multiplen Hierarchien umzugehen; d.h., ein Konzept kann mehrere Oberbegriffe haben, wie am Beispiel der Wortverwendungsklasse in Abbildung 3 zu sehen ist.

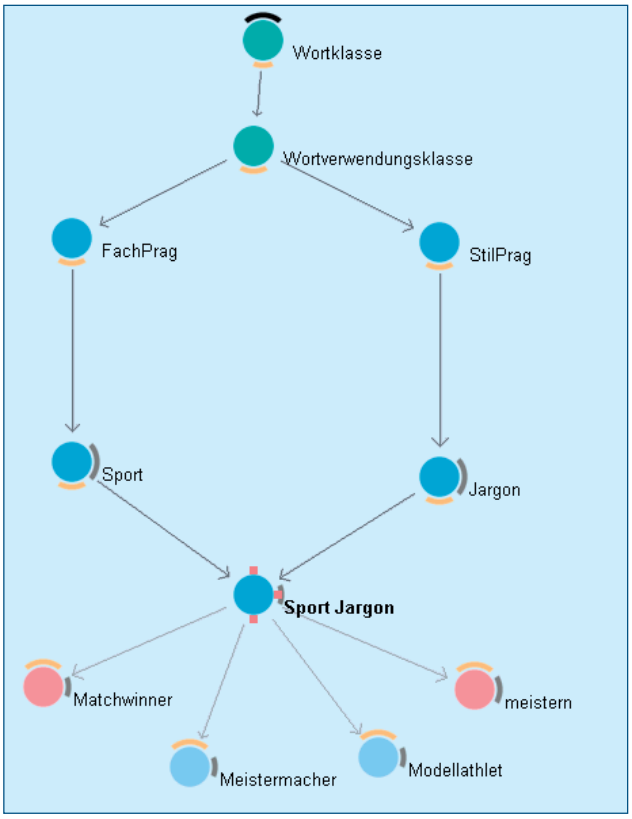

Abb. 3: Beispiel für die Modellierung der Wortverwendungsklasse, Zugehörigkeit der Wörter zu multiplen Hierarchien in der Duden-Ontologie

Das Datenmodell wurde mit der Software K-Infinity von intelligent views implementiert. Die vielfältigen und flexiblen Modellierungsmöglichkeiten sowie die verschiedenen Werkzeuge zur Erzeugung, Bearbeitung, Pflege und Nutzung des Wissensnetzes wurden und werden dabei reichlich genutzt.

3. Import der Wörterbuchdaten ins Wissensnetz

Die ursprünglichen Wörterbuchdaten lagen in SGML vor. Daher bestand der Datenimport aus dem Parsing der SGML-Daten und dem anschließenden Mapping der Elemente und Inhalte in Objekte und Relationen zwischen den Objekten. Typischerweise hat ein Wörterbucheintrag viele verschiedene Informationstypen, u.a. Lemma-Name, Wortklassen-Angaben, Pragmatik- und etymologische Informationen und Anwendungsbeispiele. Für alle diese Informationstypen galt es für den Import entsprechende Mappingregeln zu spezifizieren und dann die resultierenden Objekte und Relationen im Wissensnetz zu generieren.

So wurde jeder Wörterbucheintrag auf ein Lemma-Objekt und die grammatische Information auf eine Grammatikklasse abgebildet. Auch wenn sich das einfach anhört, war dies ein arbeitsintensiver Prozess. Nehmen wir als Beispiel ein Element mit einer spezifischen grammatischen Information: Diese wurde für den Leser des Wörterbuchs geschrieben und nicht für ein Analyse-Programm, um eine systematische Klassifikation in Grammatikklassen zu erreichen. Für solche Fälle mussten wir für den Import in die Duden-Ontologie spezifische Abbildungsprozesse implementieren.

Weitere Herausforderungen waren z.B. die Redewendungen, die wir schließlich als spezifische Lemmatypen behandelt haben und die automatisch während des Imports angelegt wurden. Die Verwendungsbeispiele eines Lemmas mussten einem Termobjekt zugeordnet werden. Da die Terme nur implizit in der Struktur des Wörterbucheintrags vorhanden sind, mussten sie während der Analyse explizit gemacht werden. Auch die Definitionen gehören an die Term-Objekte, d.h., während des Imports musste die jeweilige Bedeutungsvariante ermittelt werden. Wir waren anfangs etwas skeptisch, wie gut das Mapping auf die Terme funktionieren kann. Aber es hat sich herausgestellt, dass die Lesart-Struktur der Wörterbuchartikel gut zu der Termstruktur unseres Modells passt.

3.1 Kompositaanalyse

Leider ist die notwendige Information für ein Mapping nicht immer explizit im Wörterbucheintrag vorhanden, z.B. wird die Grammatikklasse eines Kompositums im Wörterbuch oft nicht angegeben – für den Leser eines Printwörterbuchs ist diese auch nicht zwingend notwendig, denn er oder sie kann sie aus der Angabe der Grammatikklasse für das Grundwort schließen. Für die Zuordnung der Lemmata zu ihren Grammatikklassen aber ist dies vom Modell her zwingend notwendig, z.B. waren im zehnbändigen Duden ca. 50 % der Wörterbucheinträge Komposita, für die eine Lösung gefunden werden musste.

Wir haben eine automatische morphologische Dekomposition durchgeführt, um die Komposita zwischen Grund- und Bestimmungswort zu trennen. Die Ergebnisse dieser Analyse haben wir darüber hinaus genutzt, um diese Relation auch im Netz explizit zu speichern. Wir haben dabei zwei Relationen *hat_Bestimmungswort* und *hat_Grundwort* sowie das Attribut *hat_Fuge* für die Repräsentation der morphologischen Dekomposition definiert.

Die Kompositarelationen für Substantive sind sowohl für Lemmata als auch für Terme, d.h. die einzelnen Lesarten, definiert. Damit Hyperonyme der Komposita verknüpft werden konnten, ist die Termzuordnung unabdingbar gewesen, z.B. ist *Gartenbank* Unterbegriff von *Bank* (als Sitzgelegenheit) und *Investmentbank* ist ein Subkonzept von *Bank* (als Geldinstitut).

3.2 Semiautomatische Extraktion von semantischen Relationen

Ein weiteres Ziel beim Datenimport war es, das Netz mit semantischen Relationen, wie Synonymie, Hyperonymie und Teil-von-Relation zu füllen.

Da jedoch kein explizites Mark-up für solche Informationen in den SGML-Daten vorhanden und eine vollautomatische Akquisition von semantischer Information nicht möglich war, haben wir bestimmte Wörterbucheigenschaften und lexikalische Indikatoren genutzt, um dieses Ziel zu erreichen: Beispielsweise eignen sich Einwortdefinitionen für die Erkennung von Synonymen und auch die Kompositaanalyse selbst eignet sich für die Erkennung von Oberbegriffen, wie eben erläutert. Da die automatischen Verfahren zum großen Teil zwar richtige Ergebnisse liefern, die Extraktion jedoch nicht hundertprozentig korrekte Informationen lieferte, wurden die Ergebnisse nachbearbeitet. Dazu wurden gezielt spezielle Algorithmen und Werkzeuge entwickelt und eingesetzt.

4. Erweiterung des Modells für neue Informationen

Nach den ersten Datenimporten und der ersten Phase unserer Implementierung wurde das Datenmodell um neue Informationen erweitert, z.B. um flektierte Formen (Vollformen), Häufigkeitsklassen und Markierung spezieller Wortschätze, die typischerweise kein Bestandteil von Printwörterbüchern sind, aber z.B. für unser ‚Internetwörterbuch' – Duden online – genutzt werden.

4.1 Vollformen

Im Modell wurden neue Vollformobjekte für Substantive, Verben und Adjektive definiert, die das komplette Flexionsparadigma enthalten. Neu definierte Relationen zwischen den Vollform- und den entsprechenden Lemma-Objekten verknüpfen diese im Wissensnetz.

Somit sind aktuell im Wissensnetz Vollformenobjekte für über 750.000 flektierte Substantive, mehr als 644.000 Verbformen und über 2 Millionen Adjektivformen gespeichert.

4.2 Häufigkeitsklassen

Diese Modellerweiterung betrifft die Information über die Häufigkeitsklasse eines Lemmas. Wünschenswert wäre natürlich, dass die Frequenzanalyse auch für die Terme durchgeführt werden könnte, d.h., sie müsste für jede Lesart die Häufigkeiten ermitteln. Dies könnte nur durch eine semantische Disambiguierung geleistet werden, die jedoch für die deutsche Sprache nicht vorhanden ist. Im Wissensnetz speichern wir daher die Ergebnisse der Frequenzanalyse beim Lemma, somit tragen *Bank*[1] und *Bank*[2] dieselbe Häufigkeitsklasse.

Die Frequenzanalyse wertet dabei das Dudenkorpus (vgl. Münzberg 2011) aus, ein Textkorpus der Gegenwartssprache mit über 2 Milliarden. Wortformen in fünf Textsorten. Die von der Frequenzanalyse ermittelten Angaben zum Wortgebrauch werden in drei Attributen am Lemma gespeichert: *Absolutes Vorkommen* im Korpus, *Korpusrang* nach Häufigkeit und *Frequenzklasse* (selten bis sehr häufig). Daher kann die Häufigkeitsklasse bei den Datenexporten für jedes Lemma problemlos mitgeliefert, als Filterkriterium für die Bearbeitung oder als Angabe für Offline- und Online-Produkte verwendet werden.

4.3 Explizite Markierung von Wortschätzen

Ein anderes Beispiel für eine der aktuellen Modellerweiterungen betrifft die Markierung von Wortschätzen. Konkret umfasst diese die Markierung der Wortschatzzugehörigkeit eines Lemmas mit Hilfe eines Attributes direkt am Lemma. Zurzeit sind der *Grundwortschatz Deutsch als Fremdsprache* und der *Wortschatz des Zertifikats Deutsch* (Goethe-Institut) markiert. Dies ist insbesondere wichtig für Duden online und für neue elektronische und gedruckte Wörterbücher.

5. Steckbrief des Wissensnetzes

Mit dem *Wissensnetz Deutsche Sprache*, wie die Duden-Ontologie auch genannt wird, haben wir also eine umfangreiche linguistische Ressource kreiert. Die dicht geknüpften Daten werden in einer objektorientierten Datenbank gespeichert, die explizit zugreifbares Expertenwissen zu u.a. folgenden Informationstypen enthält:

- Wörter des Deutschen (Stichwörter = Lemmata)
- Rechtschreibung
- Rechtschreibvarianten wie *Dudenempfehlungen* und *Agenturschreibweise*
- Aussprache
- Angaben zur Grammatik
- Bedeutungsangaben
- Synonymie
- Homonymie, Homografie
- Komposita
- Fremdwörter
- Verknüpfung durch semantische Hierarchierelationen
- Flektierte Formen (Vollformen)
- Markierung von Wortschätzen (*Zertifikatswortschatz*, *Duden-Grundwortschatz*)

6. Das Wissensnetz heute in der Praxis

Betrachten wir unseren unmittelbaren Anwendungskontext, die Verlagspraxis, haben wir mit der Duden-Ontologie heute Beeindruckendes erreicht: Importiert wurden bisher die wichtigsten Dudentitel aus unterschiedlichen Wörterbuchreihen, z.B. aus Großwörterbüchern, aus der Reihe „Der kleine Duden“, aus der Schülerdudenreihe oder der Reihe mit den Dudenbänden 1 bis 12. Zu den Wörterbüchern, die im Wissensnetz permanent bearbeitet werden, gehören daher u.a. der Rechtschreibduden und der zehnbändige Duden. Die ins Wissensnetz importierten Titel und Inhalte werden dort integriert gepflegt und aktualisiert. Selbstverständlich werden sie für Neuauflagen und auch für neue und neuartige Produkte verwendet.

Das Wissensnetz ist heute *das* Arbeitswerkzeug der Dudenredaktion und der Duden-Sprachtechnologie, es erlaubt medien- und titelneutrale Bearbeitung von Sprachdaten und titelbezogene Exporte von Sprachdaten für die Buch- und Softwareproduktion. Darüber hinaus ist es die Datenquelle für Duden online und ermöglicht strukturierte Datenexporte zur Überführung in „beliebige“ Formate. Das Wissensnetz bildet dadurch auch die Schnittstelle für unseren Vertrieb für Content- und Datenlizenzgeschäfte und ist nicht zuletzt die Datenquelle für die sprachtechnologischen Produkte aus dem Hause Duden: die „Duden Rechtschreibprüfung“ und den Duden-Thesaurus für Endkunden bzw. die „Duden Proof Factory“ inkl. Thesaurus-Funktionalität für Geschäftskunden.

7. Das Duden-Wissensnetz in einigen Zahlen

Technisch besteht das Wissensnetz heute aus mehr als 1,2 Millionen Individuen-Objekten mit 15,6 Millionen Eigenschaften (Attributen) und 46 Millionen Verknüpfungen untereinander (Relationen).

Inhaltlich umfasst das Wissensnetz insgesamt über 310.000 Lemmata im weiteren Sinne, da auch idiomatische Wendungen in einigen Werken Stichwortstatus haben, darunter über 195.000 Substantive, 22.000 Verben und über 29.000 Adjektive. Des Weiteren gibt es über 1050 verschiedene Wortverwendungsklassen (*umgangssprachlich, veraltend, scherzhaft, derb* etc.) und über 345.000 Definitionen und über 317.000 Ober-/Unterbegriffe.

8. Zusammenfassung und Ausblick

Wir haben mit dem Duden-Wissensnetz eine umfangreiche Ressource über die deutsche Sprache, die eine effiziente Produktion von Print-, Offline- und Online-Wörterbüchern und sprachtechnologischen Anwendungen unterstützt. Die Entscheidungen für die Datenmodellierung und die anschließende Umsetzung wurden in enger Abhängigkeit von unseren Produkten, Anforderungen und Möglichkeiten getroffen: Wir haben ein erweiterbares und integriertes Modell von semantischen und grammatischen Informationen, das die redundanzarme Datenspeicherung unterstützt und unterschiedliche Datensichten mit unterschiedlicher Granularität ermöglicht.

Wir arbeiten kontinuierlich am Ausbau des Wissensnetzes, z.B. durch Ergänzung von weiteren Vollformen und Audio-Daten zur Aussprache. Eine wesentliche Aufgabe für die künftige Arbeit ist die Anreicherung des Wissensnetzes mit zusätzlichen für unsere Produkte und Services wichtigen Informationen, insbesondere für die sprachtechnologischen Anwendungen und für Duden online, dazu zählen z.B. zusätzliches Fachvokabular und die Vertiefung der Vernetzung der semantischen Konzepte mit weiteren Ober- und Unterbegriffen.

9. Literatur

Alexa, Melina/Kreissig, Bernd/Liepert, Martina/Reichenberger, Klaus/Rostek, Lothar/Rautmann, Karin/Scholze-Stubenrecht, Werner/Stoye, Sabine (2002): The Duden Ontology: An Integrated Representation of Lexical and Ontological Information. In: LREC-2002, OntoLex-Workshop 2002: Ontologies and Lexical Knowledge Bases, 27th May 2002, Las Palmas, Canary Islands – Spain.

Münzberg, Franziska (2011): Korpusrecherche in der Dudenredaktion. Ein Werkstattbericht. In: Konopka, Marek, et al. (Hg.): Grammatik und Korpora 2009. Tübingen, S. 181-197. (= CLIP 1).