Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

# Wörterbuchstrukturen II

Claudia Kunze

kunze8@googlemail.com, kunze@cl.uni-heidelberg.de

Computational Lexicography

# Themen

- Was sind Wörterbuchstrukturen?
- Analyse von Wörterbuchstrukturen
- Parsing von Wörterbuchartikeln
- Kodierung von Wörterbuchartikelstrukturen
- Standardisierung von Wörterbuchartikelstrukturen

In dieser Einheit wird es vor allem um die Kodierung von WB-Artikelstrukturen gehen.

# Kodierung von Wörterbuchartikelstrukture

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

## Auszeichnung von Dokumenten und Texten mittels so genannter MARKUP-SPRACHEN

- verschiedene Dokumente vom selben Typ können einheitlich verarbeitet werden, z.B. auch Wörterbuchartikel
- Programme können auf die Inhaltssegemente der mit Markup versehenen Dokumente zugreifen
- die in lexikalischen Ressourcen enthaltenen Inhaltssegmente sind dafür einheitlich zu kennzeichen
- Datenfelder der Einträge einer lexikalischen Ressource sollten inhaltsbezogen und somit nachvollziehbar bezeichnet werden

# Strukturbeschreibende Auszeichnung

Methode zur Kennzeichnung von Textteilen nach ihrer Funktion für das Textganze

- Trennung von logischer Struktur und Layout von Texten und ihren Bestandteilen
- Bezeichnung der Auszeichnungselemente (engl. TAGS) soll helfen, aus dem Namen eines Textteils auf dessen Inhalt zu schließen
- Semantik der Namen, die man Auszeichnungselementen gibt, sollten für zukünftige Benutzer der Dokumente in einer Dokumentation niedergelegt werden
- z.B. Orientierung an verbreiteten Namenskonventionen oder ggf. an Standards für die Benamung von Auszeichnungselementen eines WB-Artikel-Segments

# Dokumenttypdefinition (DTD)

Struktur einer Klasse gleichartiger Dokumente wird in einer Dokumenttypdefinition (kurz: DTD) oder in einem Dokumentschema beschrieben.

- DTD hat die Form einer kontextfreien Grammatik
- Parsing der mit (einer spezifischen) DTD konformen Dokumente möglich
- Verwendung existierender DTDs, wenn diese zu erstellende Dokumente in geeigneter Weise spezifizieren, durch Taggen der Dokumentteile mit geeigneten Element-Namen
- sonst: Dokumentstrukturen müssen (neu) beschrieben und in DTD oder Schema formalisiert werden

# Auszeichnungselemente



Auszeichnungselement (TAG) als Kennzeichner wird direkt in den Text eingefügt, umschließt einen Textteil (mit gleichnamigem Anfangs-Tag und End-Tag) und beschreibt die Funktion dieses Textteils. Tags werden konventionell in spitze Klammern eingeschlossen, das End-Tag erhält obendrein einen Slash (= /) vor den Namen.

(1) <Grußformel Language="de" Style="unpersönlich"> Sehr geehrte Damen und Herren </Grußformel>

- Tags können Attribute zur weiteren Spezifikation der umschlossenen Textelemente enthalten
- Beispiel: Attribute des Textelements *Grußformel* –> *Sprache* und *Stil*
- Beispiel-Grußformel erhält die Werte *deutsch* und *unpersönlich*

# Verwendung einer DTD

## Strukturbeschreibende Auszeichnungen für die Erstellung neuer Texte

- Bestimmen des Wurzelelements, das zumeist den Typ des Dokuments bezeichnet (z.B. *Wörterbuch*)
- Bestimmen der Art und Namen der Textelemente, die die Struktur des Textes tragen (die Trägermenge) sowie deren Abfolge (insg. das INHALTSMODELL des Textes)
- Zurückgreifen auf bereits existierende DTD zum gewünschten Textmodell, ggf. Anpassen dieser DTD
- Beispiel: Festlegung von Inhaltsmodellen bzw. Mikrostrukturen unterschiedlicher WB-Artikeltypen im *e*lexiko als verbindliche Schemata für die Redakteure
- Redaktionsleitung kann in DTD festgeschriebene Regeln durch den Einsatz geeigneter Tools durchsetzen

# Verwendung einer DTD

Analyse eines bereits existierendes Dokument und nachträgliche Auszeichnung mit inhaltsbeschreibenden Tags:

- Dokumentation der Textstruktur nötig, z.B. zur nachträglichen Digitalisierung eines existierenden Print-WB
- lexikographisches Manual des Lexikographenteams (mit Festschreibung der Mikrostruktur) nachfragen
- sonst: Rekonstruktion der Wörterbuchartikel-Struktur aus typographisch markierten Artikeln in der expliziten Form einer DTD
- Inkonsistenzen in den Daten erschweren diese Aufgabe, wenn z.B. eine als obligatorisch klassifizierte Angabe in einem Artikel fehlt
- Abweichungen: Kodierungsfehler der ausführenden LexikographInnen oder bei der Rekonstruktion der Dokumentstrukturen noch nicht erfasste Strukturvariante?

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

# Rekonstruktion eines WB-Artikels

- Textlayout: **Gummi** 1. *n*
- HTML-Struktur: Gummi 1. n
- Analyse als XML-Struktur:

```
<eintrag id="1850">
<hom id="1850_1">
<lemma>
<gestalt>Gummi</gestalt>
</lemma>
<formkommentar>
<wortart> Substantiv </wortart>
<genus> Neutrum </genus>
</formkommentar>
</hom>
</eintrag>
```

# Rekonstruktion eines WB-Artikels

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Was wir bisher erreicht haben:

- Wir haben die Beschreibung der logischen Struktur des Artikels vom typographischen Layout (das gleichwohl wichtige strukturelle Hinweise lieferte) abgekoppelt;
- wir haben diese logische Struktur, die Angabetypen, anhand einer konkreten Instanz rekonstruiert;
- wir haben von der textuellen Erscheinung der Angaben abstrahiert: Das *n* als Kürzel für das Genus ist eine für Printwörterbücher typische Textkompression. Wir verwenden im rekonstruierten Fragment die ausgeschriebene Version des Namens;
- wir haben implizite Information – hier die Angabe der Wortart (WA) – explizit gemacht;
- wir haben den sequenziellen Ursprungstext in eine hierarchische Form gebracht. So sind die Angaben zu Wortart und Genus als „Formkommentar“ zusammengefasst.

# DTD des WB-Artikels

Die genannten Informationen sind ausreichend, um eine Dokumenttypdefinition (DTD) für dieses Artikelsegment und ggf. recht viele weitere Artikelsegmente zu erstellen:

```
<!ELEMENT eintrag (hom+)>
<!ATTLIST eintrag id ID #IMPLIED>
<!ELEMENT hom (lemma, formkommentar)>
<!ATTLIST hom id ID #IMPLIED>
<!ELEMENT formkommentar (wortart, genus?)>
<!ELEMENT lemma (#PCDATA)>
<!ELEMENT wortart (#PCDATA)>
<!ELEMENT genus (#PCDATA)>
```

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

# Strukturelemente dieser DTD

- Die ELEMENTDEKLARATION, bestehend aus dem Elementnamen und dem Inhaltsmodell des Elements. Im obigen Beispiel werden z.B. Inhaltsmodelle für die Elemente *hom* und *formkommentar* festgelegt.
- Das INHALTSMODELL. Das einfachste Inhaltsmodell ist beliebiger Text (#PCDATA = 'Parsed Character Data'). Ein Inhaltsmodell kann aber auch aus den Namen weiterer Elemente bestehen. Im Inhaltsmodell werden ferner die Anordnung der Inhaltselemente sowie deren Vorkommensbedingungen festgelegt. Die Anordnung der Elemente wird durch die Anordnung der Elementnamen im Inhaltsmodell wiedergegeben. In unserem Beispiel folgt auf die obligatorische Wortartangabe (WA, Element *wortart*) eine fakultative Genusangabe (GA, Element *genus*).
- Die VORKOMMENSBEDINGUNGEN. Ein Element muss entweder genau einmal vorkommen (keine Markierung), oder es kann keinmal oder genau einmal vorkommen (markiert durch ein Fragezeichen), oder es kann beliebig oft vorkommen (markiert durch einen Stern, den sog. KLEENE STAR), oder es kann beliebig oft, mindestens aber einmal, vorkommen (markiert durch ein Pluszeichen). Diese sog. Iterationsoperatoren werden auch im Kontext regulärer Sprachen verwendet.
- Die ATTRIBUTDEKLARATION. Attribute werden als Liste zu einem Element deklariert. Für jedes Attribut werden der Name, sein Datentyp oder Wertebereich und die Optionalität bzw. Obligatheit der Angabe spezifiziert. Ein wichtiger, auch mehrmals in unserem Fragment verwendeter Datentyp ist der Identifier (ID). Für diesen Datentyp gilt: jeder ID-Wert darf pro Dokument nur einmal vergeben werden.

# Konversion des GermaNet in ein XML-Form

Komplexeres Fallbeispiel der Konversion der Datenstruktur des deutschen Wortnetzes GermaNet in ein XML-Format und Entwicklung einer DTD:

- eigene Entwicklung am Seminar für Sprachwissenschaft - Kenntnis des Aufbaus und der Struktur der Ressource;
- GermaNet als wichtige Online-Ressource für das NLP;
- Konsistenz der Daten ist weitgehend gesichert, durch Konversion in andere Formate - daher können strukturelle Aspekte im Mittelpunkt der Betrachtung stehen;
- Visualisierung der GermaNet-Datenstruktur in einem Entity-Relationship-Graphen
- Betrachtung der relevanten GN-Arbeitsdateien (Lexicographers' Files)

# GN: Strukturelle Relationen und Trägermer

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

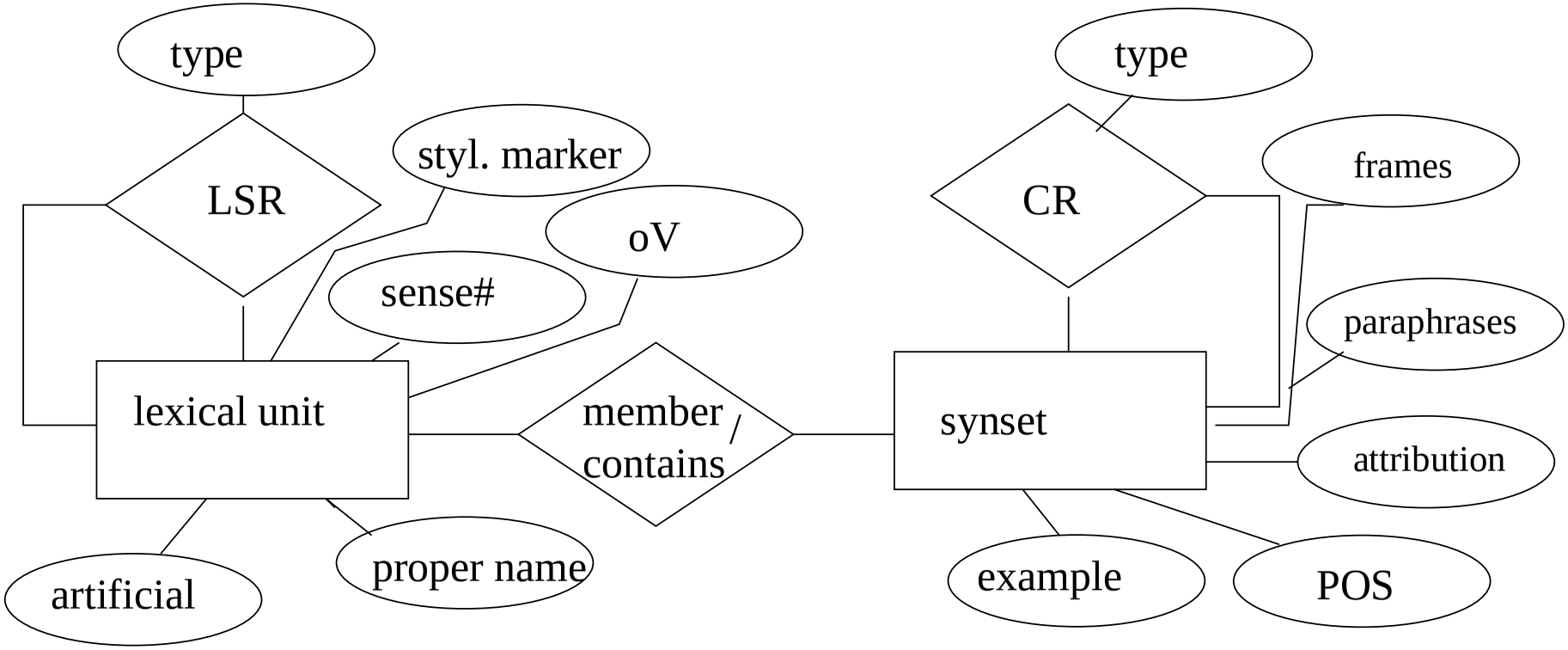
- Zentrale Repräsentationseinheiten für KONZEPTE stellen in GermaNet wie in anderen Wortnetzen die SYNSETS dar. GermaNet repräsentiert sowohl Konzepte (Knoten) als auch Relationen (Kanten) zwischen diesen Konzepten.
- Ein Synset besteht aus einer Menge von LEXIKALISCHEN EINHEITEN ('lexical units'), mindestens aber einer. Wir haben bereits im Kapitel "Lexikalische Semantik" eine lexikalische Einheit definiert als aus einer Form und einer Bedeutung bestehend.
- Synsets sind wortartenhomogen, d.h. dass sie ausschließlich lexikalische Einheiten einer Wortart enthalten. Ein Gliederungsaspekt ist deshalb der nach Wortarten: Nomen-Synsets, Verb-Synsets, Adjektiv-Synsets.
- Synsets können durch verschiedene KONZEPTUELLE RELATIONEN miteinander verknüpft werden (Hyperonymie-Hyponymie; Teil-Ganzes; kausale Beziehung etc.). Die Relation der Hyperonymie / Hyponymie bildet das hierarchische Gerüst des Wortnetzes. Innerhalb dieser Beziehung ist „multiple Vererbung" möglich, d.h. dass ein Konzept mehrere übergeordnete Konzepte haben kann.
- Synsets können genauer charakterisiert werden durch: eine BEDEUTUNGSPARAPHRASE ('gloss'), ein oder mehrere BEISPIELE ('examples') sowie eine ATTRIBUTION, in der zumeist auf Abweichungen zwischen grammatischem und natürlichem Geschlecht hingewiesen wird.

# GN: Strukturelle Relationen und Trägermer

- Verb-Synsets werden grammatisch charakterisiert durch die SUBKATEGORISIERUNGSRAHMEN (‚subcat frames'), in denen die beteiligten lexikalischen Einheiten auftreten können.
- Lexikalische Einheiten können verknüpft werden durch die LEXIKALISCH-SEMANTISCHE RELATION der Antonymie und durch die morphologisch motivierte Relation der Pertonymie.
- Lexikalische Einheiten können markiert sein als ORTHOGRAPHISCHE VARIANTE (einer anderen lexikalischen Einheit), als STILISTISCH MARKIERTE FORM oder als EIGENNAME.

# Entity-Relationship-Graph

Die Struktur des Wortnetzes wird durch die folgende graphische Abbildung veranschaulicht: Die zentralen Elemente sind die OBJEKTE (*Synsets* und *Lexical Units*), die RELATIONEN (konzeptuelle Relationen verbinden die *Synsets* miteinander, lexikalisch-semantische Relationen verbinden die *Lexical Units* miteinander) und die ATTRIBUTE, die sowohl die Objekte als auch die Relationen identifizieren bzw. charakterisieren.




CR=conceptual relation; LSR=lexical−semantic relation; oV=orthographic variant

# Umsetzung in DTD

Umsetzung dieser Struktur (oder auch ähnlicher Strukturen) in eine DTD nach folgenden Prinzipien:

- Objekte werden als Elemente modelliert;
- Relationen werden als *Links* modelliert; Links werden in XML besonders behandelt und erhalten eine eigene Spezifikation (XLink);
- identifizierende Attribute werden als Attribute der Elemente, die sie identifizieren, modelliert;
- charakterisierende Attribute werden als Elemente innerhalb des Inhaltsmodells der Elemente, die sie charakterisieren, modelliert.

Da es in einem Entity-Relationship-Diagramm allerdings keine Möglichkeit gibt, identifizierende von charakterisierenden Attributen zu unterscheiden, und da es ebenso keine verbindlichen Richtlinien für die Verwendung von Elementen und Attributen in DTDs gibt, liegt die Umsetzung von Attributen des ER-Modells in eine DTD im Ermessen der DTD-Designer.

# GN DTD für Synsets und Lexical Units

```
<!ELEMENT synsets       (synset)+>
<!ELEMENT synset        ((lexUnit)+, attribution?,
                frames?, paraphrases?, examples?)>

<!ATTLIST synset        id              ID      #REQUIRED
                        wordClass       CDATA   #IMPLIED
                        lexGroup        CDATA   #IMPLIED>
<!ELEMENT lexUnit       (orthForm)+>


<!ATTLIST lexUnit       id              ID      #REQUIRED
                        StilMarkierung  (ja|nein)    "nein"
                        sense  CDATA            #REQUIRED
                        orthVar         (ja|nein) "nein"
                        artificial      (ja|nein)  #REQUIRED
                        Eigenname      (ja|nein) #REQUIRED>
<!ELEMENT orthForm      (#PCDATA)>
```

# Fortsetzung DTD

```
<!ELEMENT paraphrases   (paraphrase)+>
<!ELEMENT paraphrase    (#PCDATA)>
<!ELEMENT examples      (example)+>
<!ELEMENT example       (text, frame*)>
<!ELEMENT frames        (frame)+>
<!ELEMENT attribution   (#PCDATA)>
<!ELEMENT text          (#PCDATA)>
<!ELEMENT frame         (#PCDATA)>
```

Für die zweite DTD, die Dokumente beschreibt, in denen Relationen zwischen Synsets und Lexical Units abgelegt sind, sei auf die Darstellung im Lehrbuch verwiesen. Diese DTD orientiert sich an der XLink-Spezifikation.

# Ausschnitt aus GN-Lexicographers' File

```
{ ?geistspezifisch,                                    }
{ ?intelligenzspezifisch,              ?geistspezifisc
{ [klug, dumm,!] intelligent, ?intelligenzspezifis
{ clever,                                klug,@ }
{ [scharfsinnig, nomen.Kognition:Scharfsinn,\]
                 klug,@ ('mit Scharfsinn') }
{ [einfallsreich, nomen.Kognition:Einfall,\]
                 klug,@ }
{ kreativ,                 einfallsreich,@ }
{ weise,                   klug,@ ('mit Weisheit') }
{ schlau,                  klug,@ }
{ hell,                    klug,@ ("ein heller Kopf")
.....
```

# Charakteristika der Datenkodierung

- Die Daten sind in einer einfachen Textdatei kodiert; die LexikographInnen bearbeiten die Daten unmittelbar in diesem File, was die Gefahr von Fehlkodierungen in sich birgt. Die Daten müssen deshalb vor ihrer Konvertierung auf ihre KONSISTENZ geprüft werden;
- Relationen werden mithilfe von Symbolen dargestellt; der Skopus der Relationen ist implizit gegeben. Vor dem Relations-Symbol steht der Name des Verweisziels;
- Verweise (zu anderen Synsets oder Lexical Units) sind direkt und ausschließlich an der Verweisquelle kodiert, sie sind also Bestandteil der Synsets bzw. lexikalischen Einheiten, von denen sie ausgehen;
- Substrukturen werden durch Klammerung dargestellt;
- einige Attribute werden durch Symbole dargestellt, die unmittelbar an die Repräsentation einer lexikalischen Einheit angehängt werden. Dies erschwert die Suche nach lexikalischen Einheiten, da diese Symbole von der Wort-Zeichenkette wieder abgetrennt werden müssten;
- es gibt weder für Synsets noch für lexikalische Einheiten eindeutige Schlüssel („unique identifier“).

# Konversion der Daten in XML

- Die Konsistenz der Daten wird geprüft.
- Relationen werden zunächst explizit kodiert. Erst in einem zweiten Konversionsschritt werden die Relationen aus dem Informationsgefüge der Synsets und Lexical Units gelöst und in einer eigenen Datenstruktur repräsentiert.
- Substrukturen werden durch das (hierarchische) Inhaltsmodell der Elemente repräsentiert.
- Die Attributsymbole werden von der Form des graphischen Repräsentanten einer Lexical Unit oder eines Synsets getrennt und explizit als Attribute der entsprechenden Elemente kodiert.
- Im Zuge der Konversion wird für jedes Synset und für jede Lexical Unit ein eindeutiger Kennzeichner („identifier“) vergeben, die auch der Referenzierung der Elemente in Verweisen dienen.

# Vorteile der Konversion

XML-Datenrepräsentation besser geeignet als bisherige:

- für den Zugriff von Anwendungsprogrammen auf die Daten als lexikalische Ressource (via standardisierte *Application Programme Interfaces*, APIs),
- für die Verknüpfung der lexikalischen Ressource mit anderen lexikalischen Ressourcen, die etwa detailliertere Angaben zur Form und Funktion der Lexical Unit beitragen könnten,
- für die Verknüpfung des deutschen Wortnetzes mit Wortnetzen anderer Sprachen,
- für die Konversion in andere web-fähige Formate, was die Verwendung der Ressource als ontologische Ressource für die Entwicklung des „Semantic Web“ geeignet macht.

# Beispielsynset

```
<synset id="vKommunikation.524" wordClass="verben">
<lexUnit Eigenname="nein" artificial="nein"
         id="vKommunikation.524.lesen2"
orthVar="nein" sense="2" stilMarkierung="nein">
<orthForm>lesen</orthForm>
</lexUnit>
<frames>
<frame>NN.PP</frame>
</frames>
<paraphrases>
<paraphrase>Vorlesungen halten</paraphrase>
</paraphrases>
<examples>
<example>
<text>Er liest [über] englische Literatur.</text>
</example>
<example>
<text>Der Autor liest aus seinen Werken.</text>
</example>
</examples>
</synset>
```

# Standardisierung von Wörterbuchartikelstr

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Internationale Bemühungen um die Standardisierung von Wörterbuchartikelstrukturen, der verwendeten Angabetypen und Wertebereiche dieser Angabetypen zur

- Wiederverwendbarkeit lexikalischer Beschreibungen
- besseren Kombinierbarkeit unterschiedlicher Arten lexikalischer Ressourcen
- Erstellung kompatibler multilingualer Lexika aus verschiedenen Quellen
- zur Nutzung der lexikgraphischen Daten im NLP

Es gab zahlreiche europäische und weltweite Projekte und Initiativen zu diesem Unterfangen (MULTILEX, GENELEX, PAROLE, ISLE)

# Lexical Markup-Framework

Neuste Arbeitsgruppe zur Etablierung eines nachhaltigen ISO-Standards:

- Berücksichtigung der Ergebnisse von Vorläuferprojekten
- Unterstützung der Anwendung des Standards beim Aufbau neuer lexikalischer Ressourcen
- Unterstützung von Konversionsprozessen existierender lexikalischer Ressourcen in standardkonforme Formate
- dadurch Verknüpfung unterschiedlicher standardkonformer Ressourcen möglich
- Verwenden von ISO-Datenkategorien bei Festlegung des lexikal. Informationsprogramms

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

# LMF-Kern und Erweiterungen

Der Standard definiert den KERN einer lexikalischen Ressource – Lexikon bzw. lexikalische Datenbank – und die Mikrostruktur eines lexikalischen Eintrags. Desweiteren werden fünf für die Sprachtechnologie zentrale Resso urcen als Erweiterungen des Kerns festgeschrieben:

- Maschinenlesbare Wörterbücher;
- Morphologische Ressourcen für sprachtechnologische Anwendungen;
- Syntaktische Ressourcen für sprachtechnologische Anwendungen;
- Semantische Ressourcen für sprachtechnologische Anwendungen;
- Mehrsprachige Ressourcen;
- Muster für mehrgliedrige lexikalische Einheiten.

# LMF-Kern

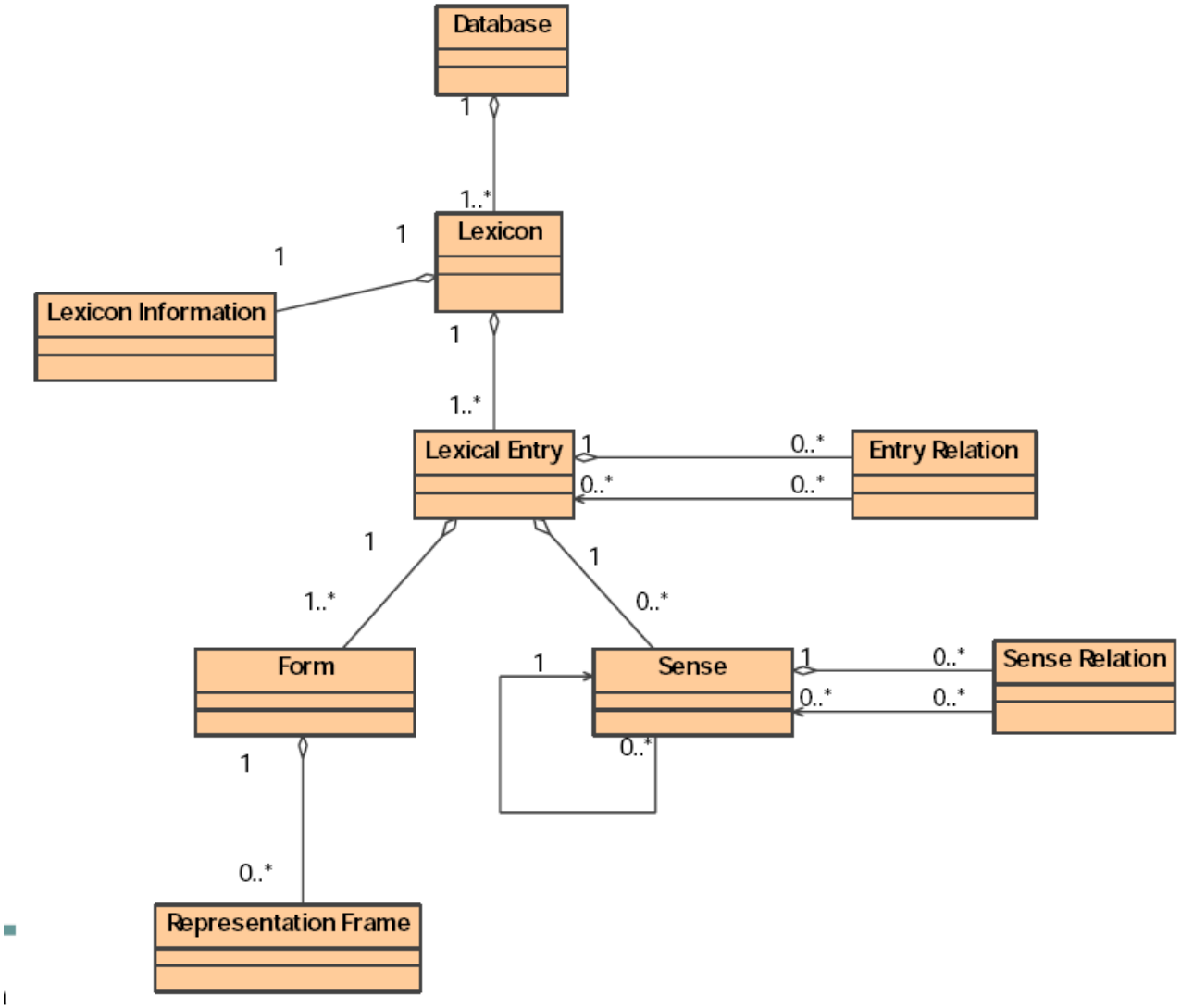



Abbildung 1: Kernmodul des Lexikonmodells

# Beispielintrag *clergyman*



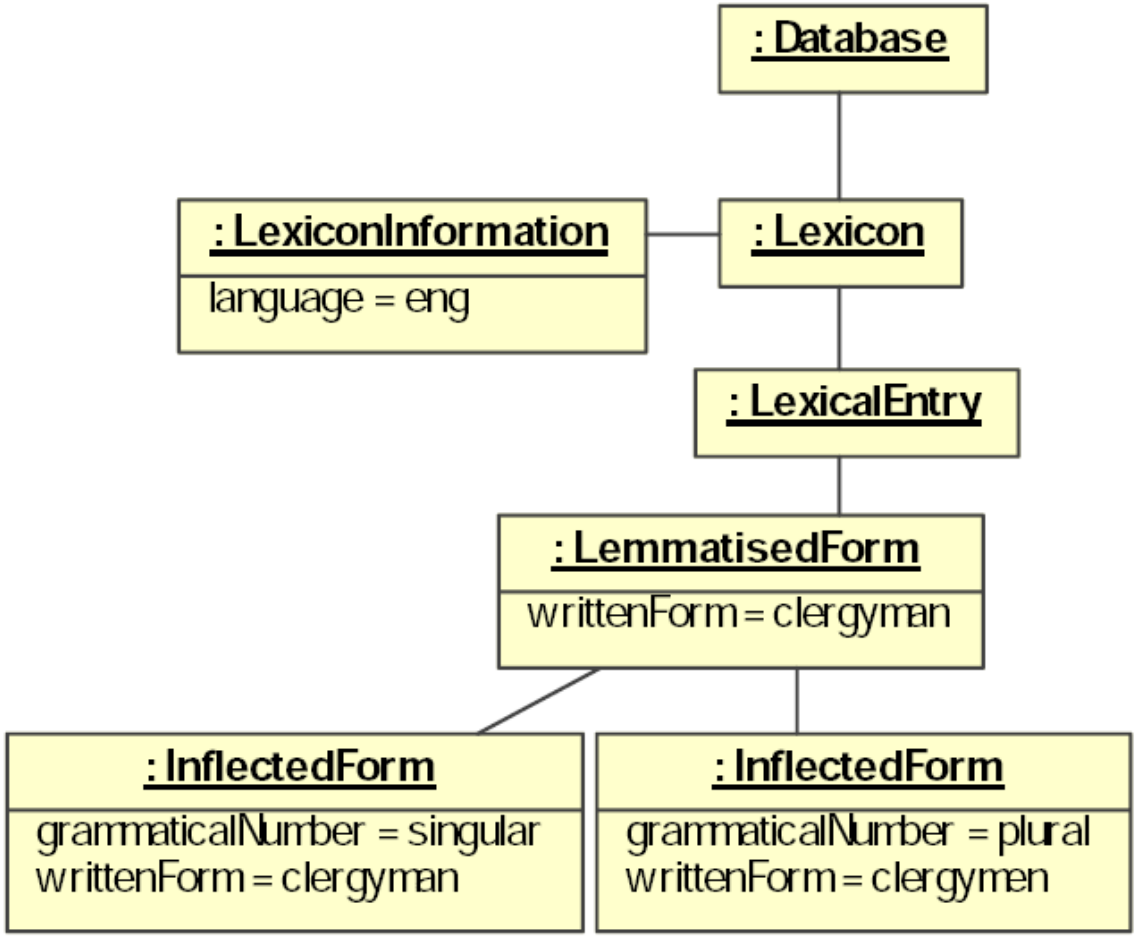



Abbildung 2: Ein einfaches Beispiel

# Beispiel Representation Frame

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

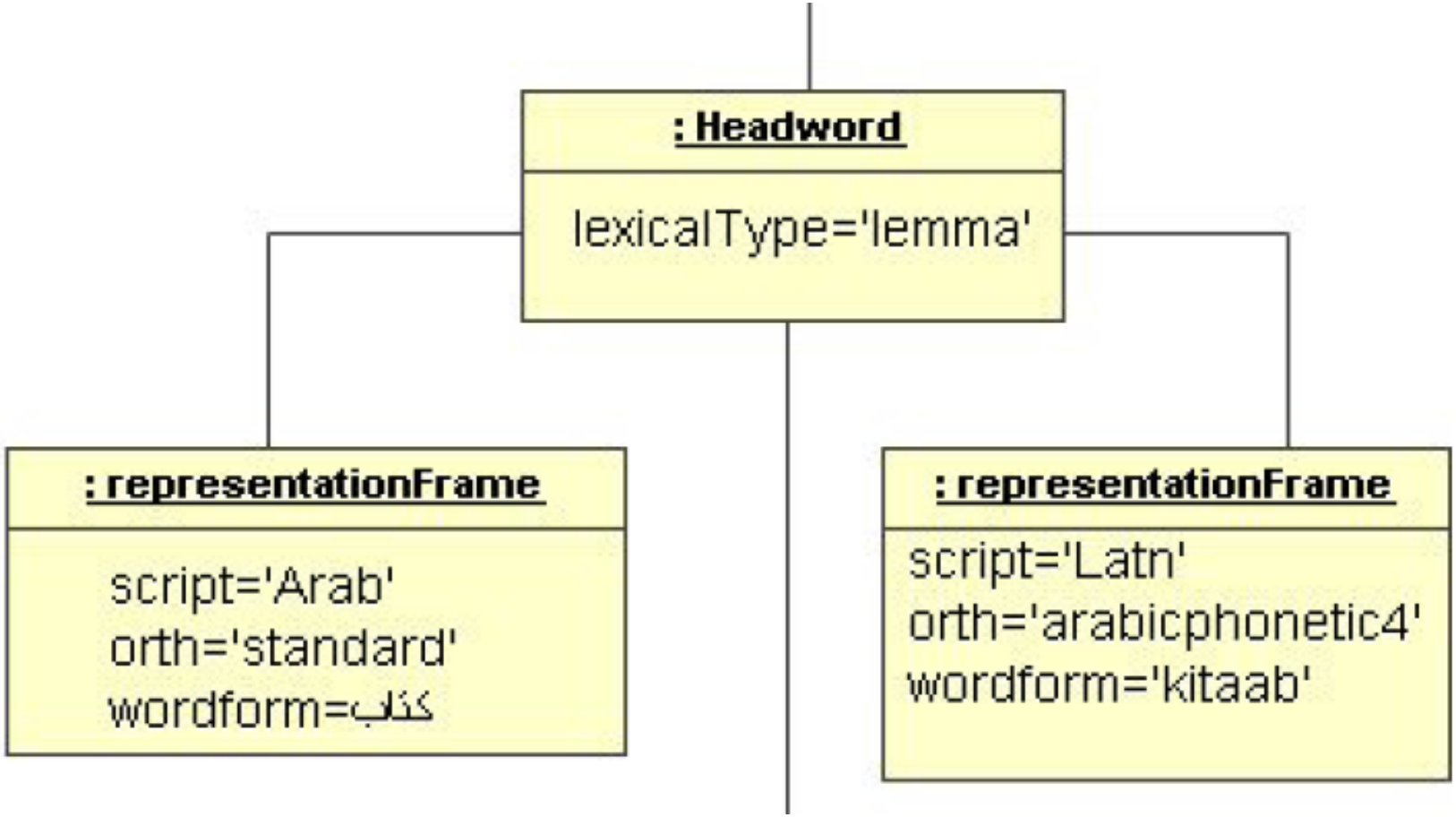



Abbildung 3: Beispiel für die Verwendung des Representation Frame

# Morphologische Erweiterung

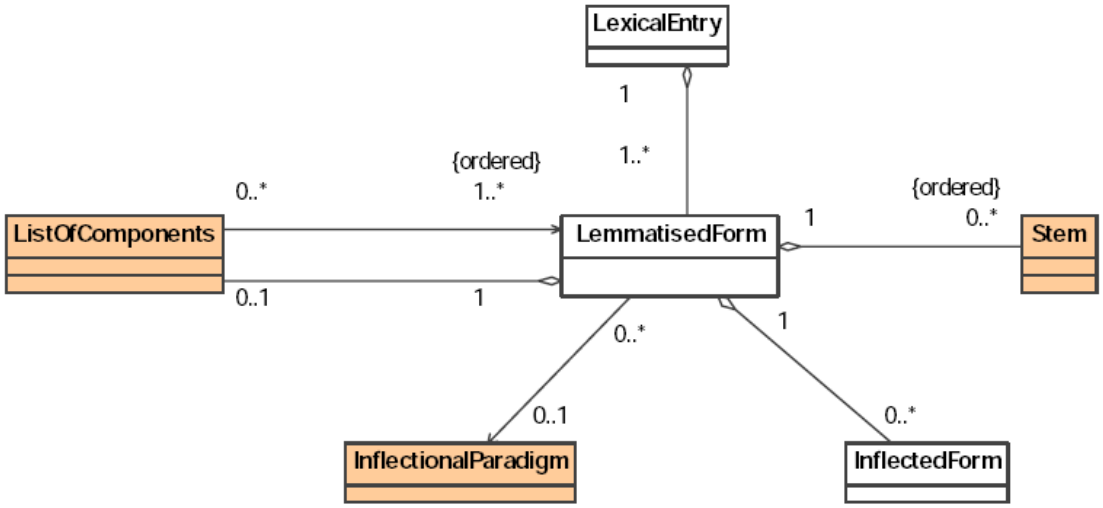



Abbildung 4: Erweiterungen für die morphologische Beschreibung

# Multilinguale Erweiterung

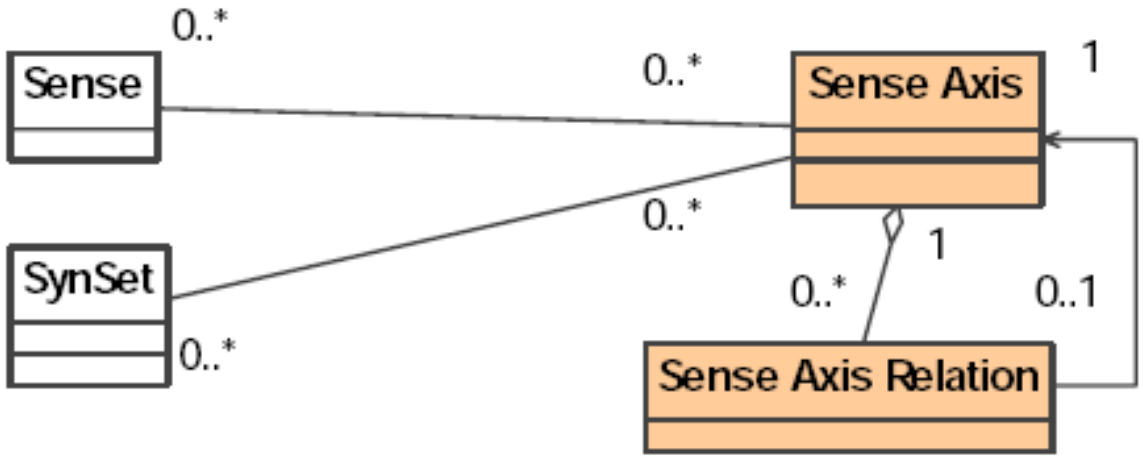



Abbildung 5: Erweiterung für multilinguale Ressourcen

# Ein Beispiel

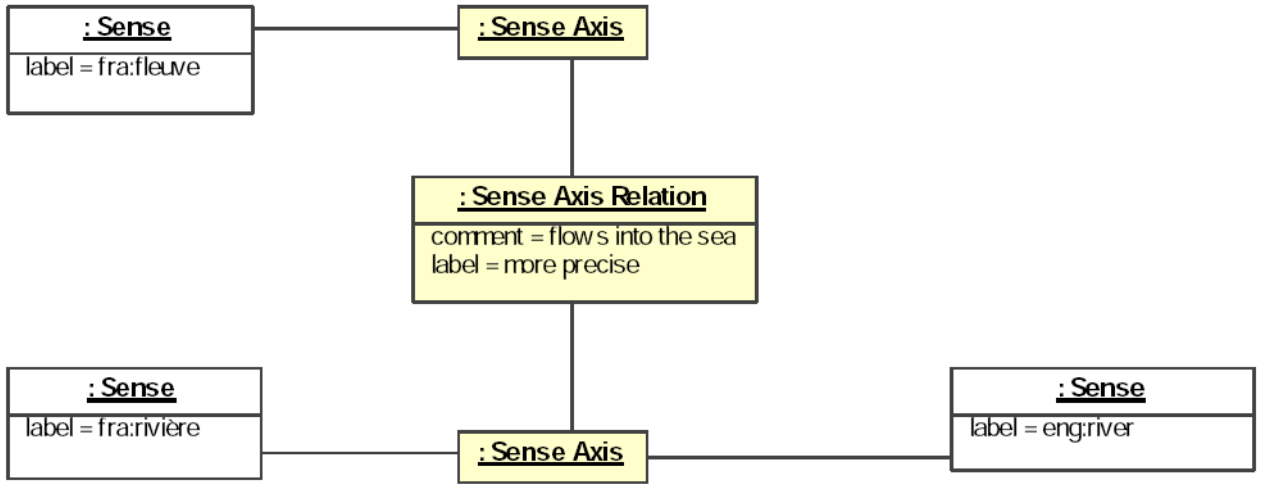



Abbildung 6: Erweiterung für multilinguale Ressourcen, ein Beispiel

# Erstellung einer neuen Ressource

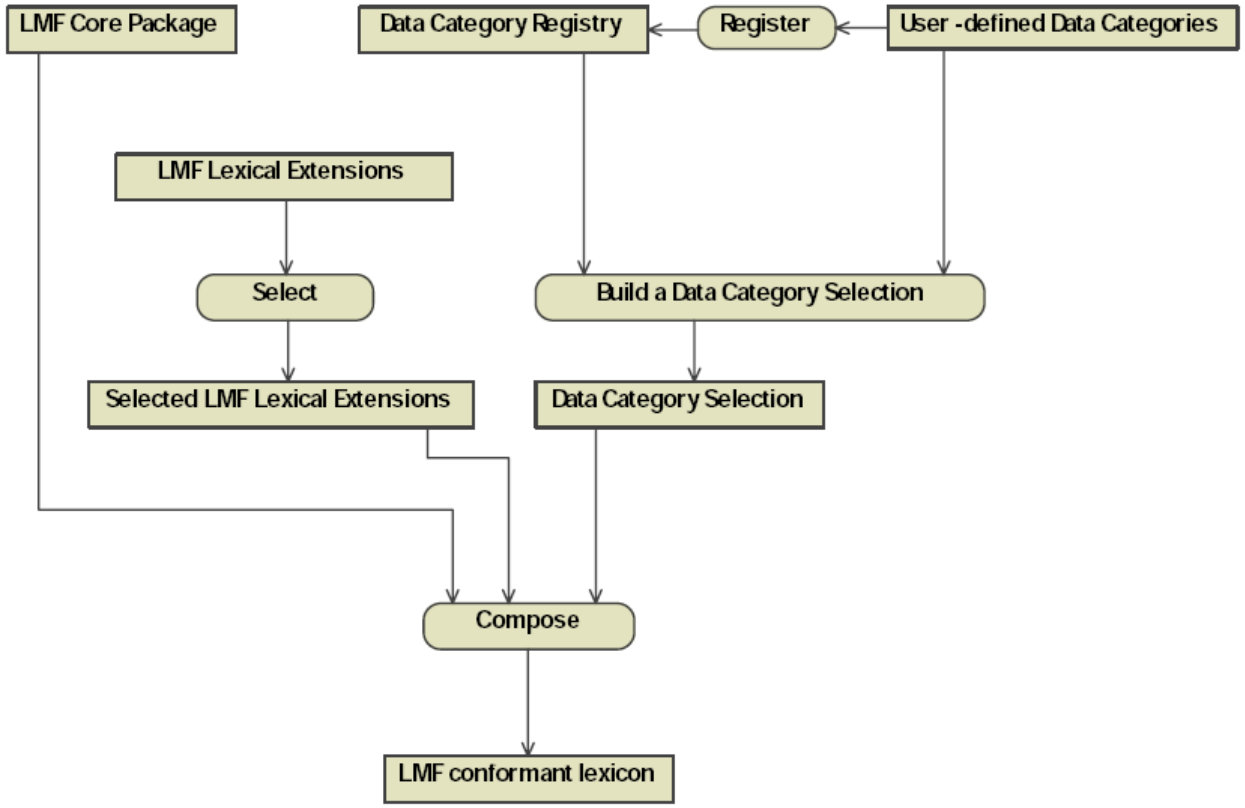



Abbildung 7: Verlaufsdiagramm des Designprozesses für eine neue lexikalische Ressource

# Umsetzung des Modells

Implementierung und Durchsetzung des Modells als Standard umfasst drei Aspekte:

1. Die Verabschiedung des Vorschlags als Standard durch die ISO-Gremien;
2. Anleitungen für die Anwendung des Standards bei der Erstellung neuer Ressourcen;
3. Verfahren für die Konvertierung bestehender Ressourcen in das Format, das der Standard vorgibt.

Der erste Punkt bleibt abzuwarten. Zum zweiten Punkt macht der Standardentwurf konkrete Vorschläge, insbesondere zur Definition einer standardkonformen Lexikonstruktur. Der Prozess umfasst:

- Auswahl des Kernmoduls und der notwendigen Erweiterungsmodule,
- Auswahl der benötigten Datenkategorien aus dem Datenkategorie-Register,
- Beschreibung und Registrierung von Datenkategorien, die (noch) nicht im Datenkategorie-Register zur Verfügung stehen,
- Zusammenfügung dieser Elemente zu einer standardkonformen Artikelstruktur.

Am Ende dieses Designprozesses wird dann vermutlich ein XML Schema stehen. Zum drittem Punkt präsentiert Francopoulo sechs Fallstudien zu unterschiedlichen Ressourcetypen.