WDG Analyse und Transformation (WAT)

Studienprojekt von Ineta Sejane und Wiebke Wagner

am Institut für Deutsche Sprache Mannheim

WDG

Das Wörterbuch zur deutschen Grammatik WDG enthält ca. 150 000 Lemmata. Es ist in txt-Format abgespeichert und ergibt 32 MB.

In den Einträgen stecken zahlreiche Informationen zur Morphologie und Syntax der Wörter sowie zur Zugehörigkeit zum Fachwortschatz, dialektaler Verbreitung. Es werden Homographen und mehrdeutige Wörter unterschieden.

Substantiv

Z1	Genus (0,1,3)	Mengenangabe(1)	Nom.Pl.(0-9, 19, 31)
	-		
Z2	Genitivendung(0-9)	Dativendung(0-3, (+4))	Akkusativendung(0-2)
	Numerusangabe(0-2)		
Z3	Namen, Apposition(1-13)	Präp. Valenz(max.3)***	
	Inf. Anschluss (0,1)	dass-, Fragesatz(0,1)	
Z4	Präp. Valenz(max.3)	Präp. Valenz(max.3)	-
	Teil einer festen Sequenz(1)		
Z5	-	-	Wortschatzauswahl(0,1,2,4,8)	-
Z6	Zeitsubstantiv(1)	Gewichtung(1-4)	Fuge
	(0,1,2,4,8,16,?32)	Kompositabildung(0,1,?2)	
Z15	Fachgebietsmarkierung(1,2,4)	intern definiert**	intern definiert
	intern definiert		

*Bei den orange gekennzeichneten Einträgen handelt es sich um Werte, die durch Addition gewonnen werden.

** Von BSA, DTJ und TRANSIT intern definiert

*** kommt mehrfach vor

Beispiel Substantiv

```
WL1: AALFAENGE                    SDW
WL2: AALFANG                      50381
IBED    =  1
RECTYP  =  2
Z 1=    1    0    0    0
Z 2=    0    4    0    2
Z 5=    0    0    8    0

WL1: AALFANG                      SDW
WL2:                              50381
IBED    =  1
RECTYP  =  2
Z 1=    1    0    0    0
Z 2=    4    0    0    1
Z 5=    0    0    8    0
```

Substantiv

Mask, keine Mengenangabe, *Nom.Pl. mit -en

*Gen.Sg. ohne Endung, Dat.Sg. ohne Endung, Dat.Pl. mit –n, Akk.Sg. ohne Endung, Pluralstamm

Wort aus SDW, Wortlaut enthält „/“

Gen.Sg. –es, Dat.Sg. ohne Endung, Akk.Sg. ohne Endung, Singularstamm

Beispiel Substantiv vs. Adjektiv

```
WL1: AACHENER                    LUCKHARDT
WL2:                             41088
IBED    =  1
RECTYP =  2
Z 1=    1    0    4    0
Z 2=    2    0    0    0

WL1: AACHENER                    LUCKHARDT
WL2: AACHEN                      171088
IBED    =  1
RECTYP =  3
Z 1=    0    4    0    0
Z 3=    0    0    0    1
Z 6=    0    0    0    1
Z 7=    1    0    0    0
```

Mask, keine Mengenangabe, Nom.Pl. ohne Endung

Gen.Sg. mit -s, Dat.Sg. ohne Endung, Akk.Sg. ohne Endung, Singular- und Pluralstamm

Adjektiv

keine Valenz, kann weder als Adverb noch als Prädikativ gebraucht werden, kein Nebensatzanschluss möglich, Adjektiv der Art und Weise (default)

Infinitivanschluss nicht möglich, nicht flektierbar

darf nicht an Kompositabildung teilnehmen

Teil einer festen Sequenz

Beispiel Funktionswortklasse

```
WL1: AAO.                    KREBS
WL2:                         10880
IBED   =  9
RECTYP = 32
Q( 1): (  0)  jwk   = 21,  jstw  =  0,  jbed  =  9
Q( 2): jwkbin=  0,   16,   0,   0
Q(18): ksemkl=  2,   0,    0,   0
Q(19): duerfte nicht besetzt sein!   9   0   0   0
```

Funktions-wortklasse

Adverb, Funktionswort, Bedeutungsnummer des Funktionswortes

semantische Klasse Ort

im Speicher ist Q19 leer definiert

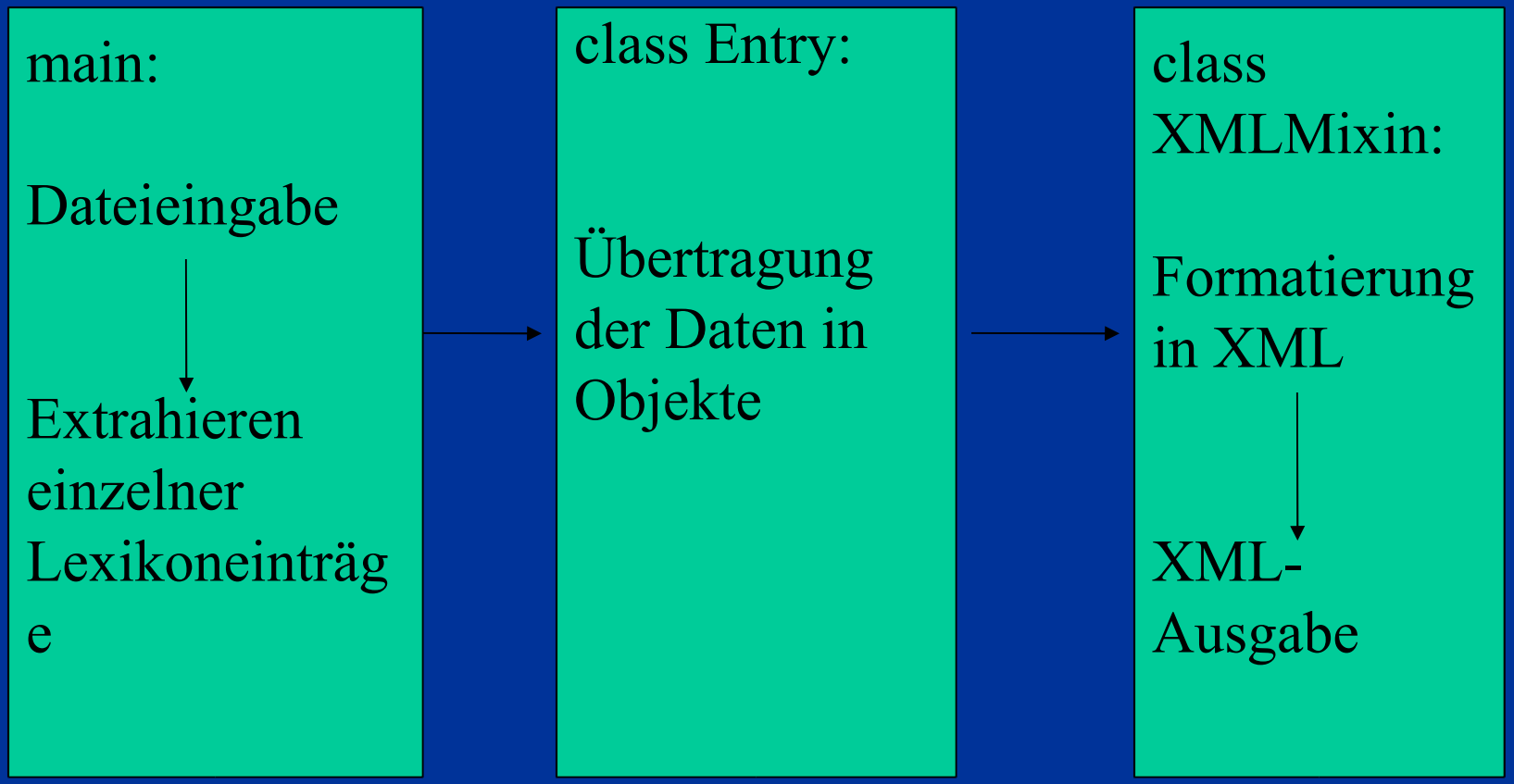
Überführung in ein XML-Format

Um jeglichem Informationsverlust vorzubeugen, sollen die Originaldaten 1:1 in ein XML-Format überführt werden. Damit soll eine verlässliche Einheitlichkeit der Struktur erreicht werden, um eine maschinelle Weiterverarbeitung zu ermöglichen. Hierfür ist notwendig:

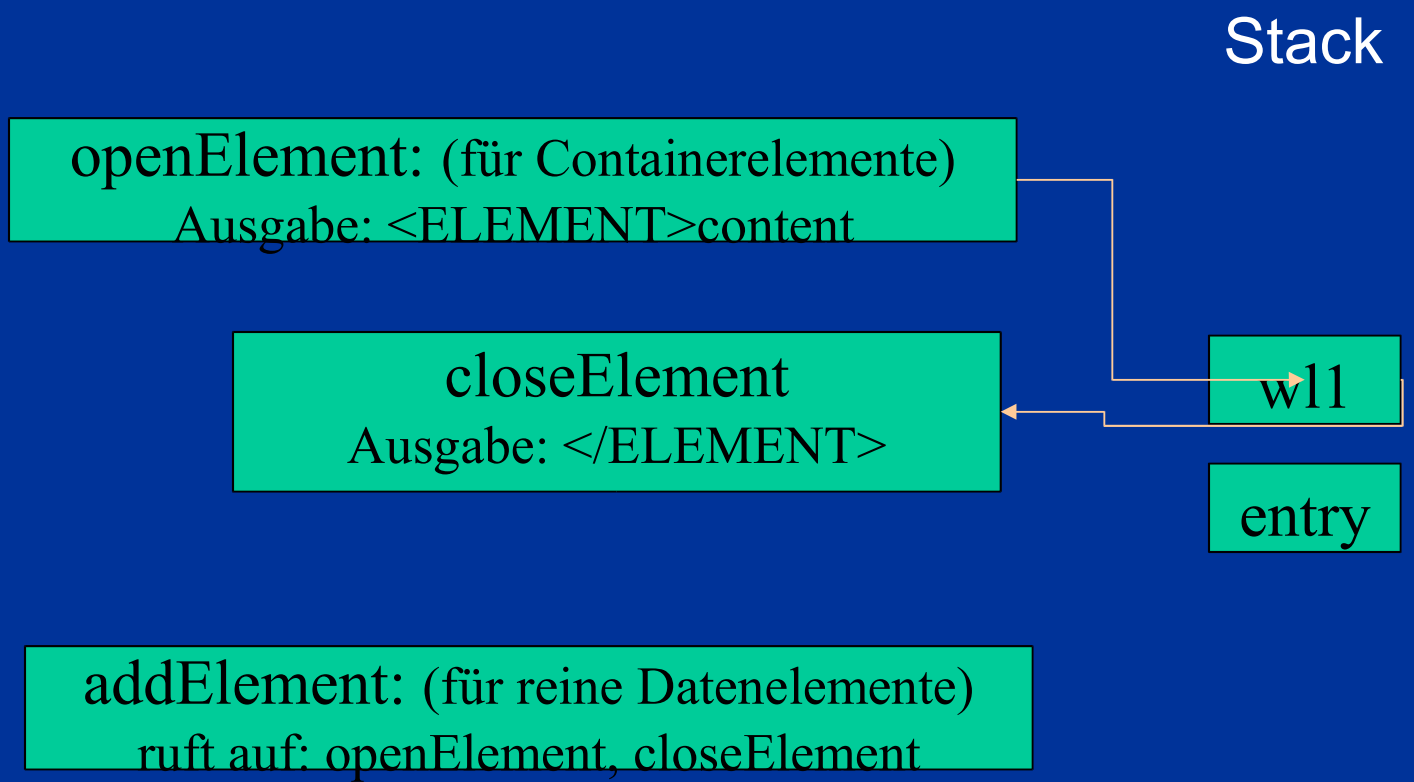
- das Erstellen einer DTD
- die Implementierung eines robusten Parsers, der die Einträge einheitlich strukturiert in XML-Code überführt. Nicht erkannte Zeilen werden als unbekannte Elemente mitaufgenommen.
- Erstellen eines Style Sheets (XSLT)

Die Daten benötigen eine übersichtliche Dokumentation.

Parser


main:
Dateieingabe
Extrahieren einzelner Lexikoneinträge
class Entry:
Übertragung der Daten in Objekte
class XMLMixin:
Formatierung in XML
XML-Ausgabe

class XMLMixin



XML-Struktur

```
<entry id="0000001">
    <wl1>
        AALT
    </wl1>
    <wl2>
        AALEN
    </wl2>
    <ibed>
        4
    </ibed>
    <rectyp>
        1
    </rectyp>
    <z2>
        <s1>
            1
        </s1>
        <s2>
            0
        </s2>
    </z2>
</entry>
```

Bearbeitung der Daten

- Bereinigung der Lemmata z.B.
  - *DOMINO]SPIEL*
  - *DEUTSCHE& DEMOKRATISCHE& REPUBLIK*
  - *BLAESHUEHNER r*
  - Q(19): duerfte nicht besetzt sein! 9 0 0 0
- Groß- / Kleinschreibung
- Worte mit unterschiedlichen Stammformen werden gesondert aufgeführt. Eine Verschmelzung mit der Grundform als Lemma ist angestrebt.

Probleme

(1) Weiterverarbeitung von Abkürzungen

(2) Umlaute

(3) ß vs. ss (entsprechend der neuen Rechtschreibregelung)

(4) f vs. ph (entsprechend der neuen Rechtschreibregelung)

(5) Sonderzeichen (z.B. Accents)

(6) Überschreitung der Eingabebegrenzung (z.B. *A LA BONNE HE*)

Korpora

- Lösung: Abgleich gegen Korpusdaten
  - Alle Problemeinträge werden gesucht (Lemmata mit *oe, ae, ue, ss, ph,* etc. z.B. *Autoeinfahrt*)
  - Lemma wird mit Korpusdaten verglichen. (Dazu steht COSMAS zur Verfügung)
  - Eine statistische Analyse entscheidet, ob der Umlaut gesetzt wird oder die Buchstabenfolge belassen wird (80%).

Mögliche Weiterverwendung der Daten

- Vollformgenerator von Verben, Nomen, NPs
- Wörterbücher einzelner Wortarten
- Valenzwörterbücher
- Statistiken über grammatische Eigenschaften (z.B. wieviele Prozent der Nomen bilden den Plural mit -en/-e, etc.)
- Wortlisten von Wörtern mit bestimmten grammatischen Eigenschaften, z.B von starken Verben, von Verben mit *dass*-Satzanschluss, von transitiven Verben , etc.

Validierung der Daten?

- Inwiefern stimmen die Wörterbuchangaben mit der Realität überein, z.B.
  - Satzanschluss bei Nomen und Verben
  - regierte Präpositionen (***bis zu*** *einem Jahr*)
  - morphologische Formen
  - mögliche Kompositabildung