Abschlussbericht des Studienprojekts

WAT
WDG - Analyse und Transformation

von Ineta Sejane und Wiebke Wagner
am Institut für Deutsche Sprache, Mannheim

WDG
Wörterbuch zur deutschen Grammatik

WDG enthält zahlreiche Informationen über Morphosyntax der Lemmata (ca. 150.000), ihre Zugehörigkeit zum Fachwortschatz und dialektale Verbreitung. Es werden Homographen und mehrdeutige Wörter unterschieden.

Ausgangspunkt fürs Projekt:

Textdateien (ca. 32 MB txt) sollen in ein modernes Format überführt werden, das lesbar ist und weiterverarbeitet werden kann.

Beispiel der Originaldaten

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

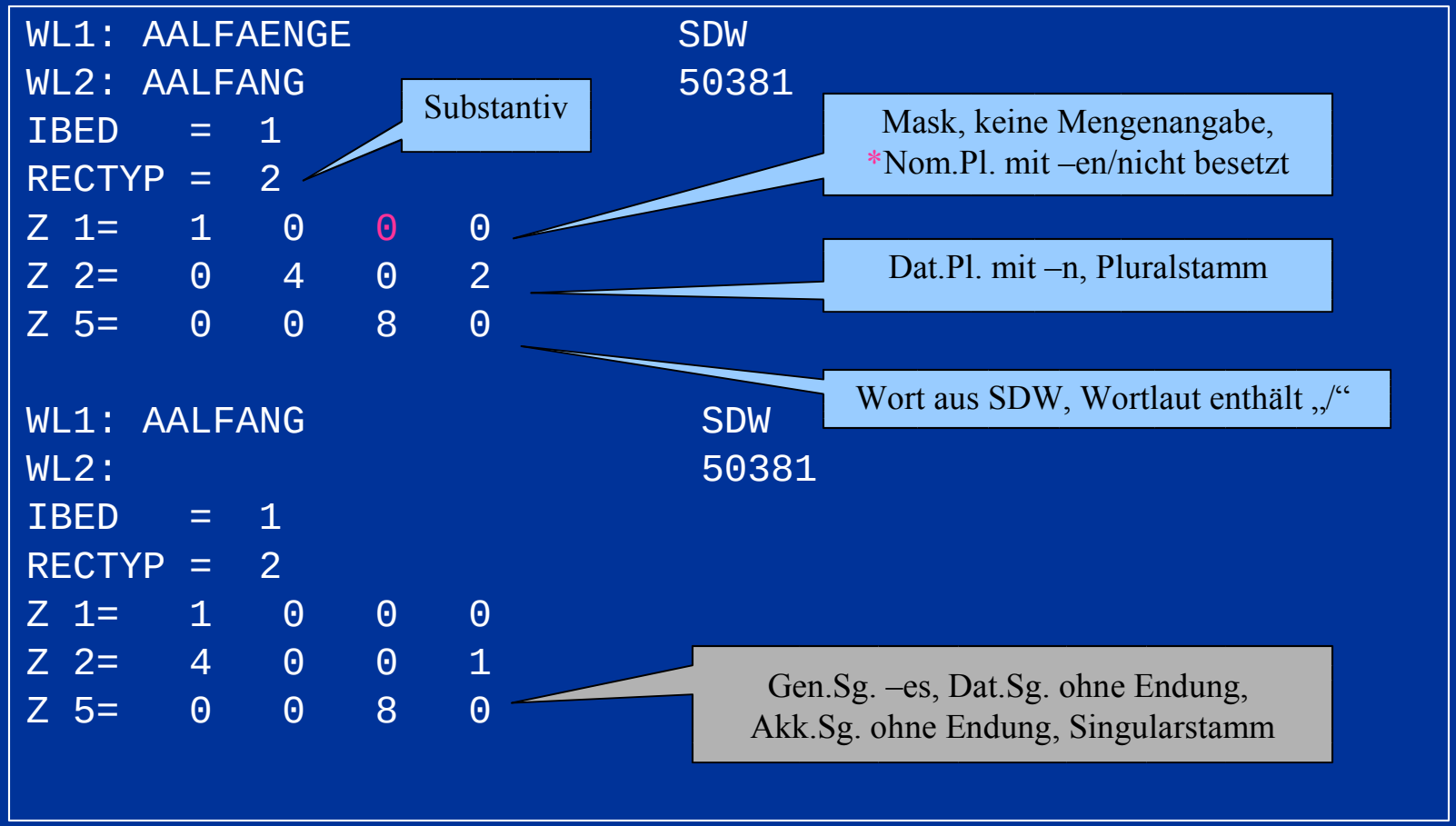
WL1: AALFANG SDW
WL2: 50381
IBED = 1
RECTYP = 2
Z 1= 1 0 0 0
Z 2= 4 0 0 1
Z 5= 0 0 8 0

Substantivmaske

Z1	Genus	Mengenangabe	Nom.Pl.	-
Z2	Genitivendung	Dativendung	Akkusativendung	Numerusangabe
Z3	Namen, Appos.	Präp. Valenz	Inf. Anschluss	dass-, Fragesatz
Z4	Präp. Valenz	Präp. Valenz	-	Teil einer festen Sequenz
Z5	-	-	Wortschatzauswahl	-
Z6	Zeitsubstantiv Kompositabildung	Gewichtung	Fuge	
Z15	Fachgebiet	intern definiert	intern definiert	intern definiert

kommt mehrfach vor

Beispiel Substantiv



Beispiele der Einträge

```
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

Zielsetzung des Projekts

- Originaldaten erhalten, erschließen und ergänzen
  - Darstellung der Originaldaten in XML und HTML
  - Daten validieren und Fehler beheben. Wenn nicht möglich, vermerken für weitere Bearbeitung
  - neue Dokumentation erstellen

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.

Aufbereitung von WDG

Die Bearbeitung bis zur Endversion in XML-Format

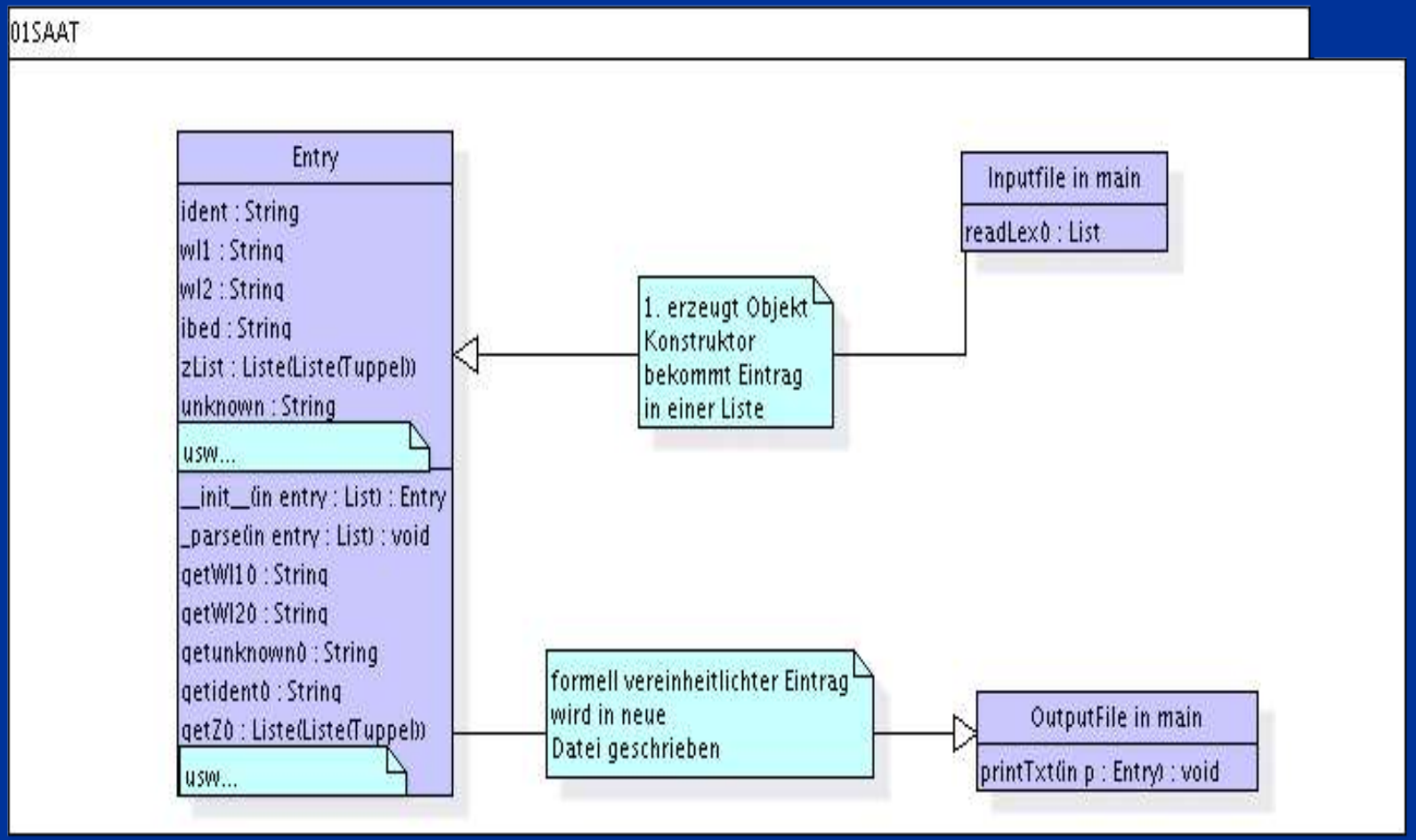
- erfolgt durch fünf unterschiedliche Programme
- generiert vier Zwischenversionen in txt-Format

01WAT - 1:1-Übertragung

Vereinheitlichung im Format ohne Veränderung der Daten:

- Pro Zeile eine Informationseinheit
- Topik und Informationsteil durch „:“ getrennt
- In Q-Zeilen Klammern, Kommata, Leerzeichen zwischen den Elementen entfernt. Zuweisung innerhalb der Elemente durch =-Zeichen
- Zuweisung einer ID
- Topik *unknown* für alle übrigen Elemente

Klassendiagramm 1



02WAT - Lemmakorrektur

Neues Lemmaformat: WL2: AB **(PRP)** **AB**

Ziel: Die Lemmavarianten sollen mit einem Korpus abgeglichen werden, um die wahrscheinlichste Variante zu ermitteln.

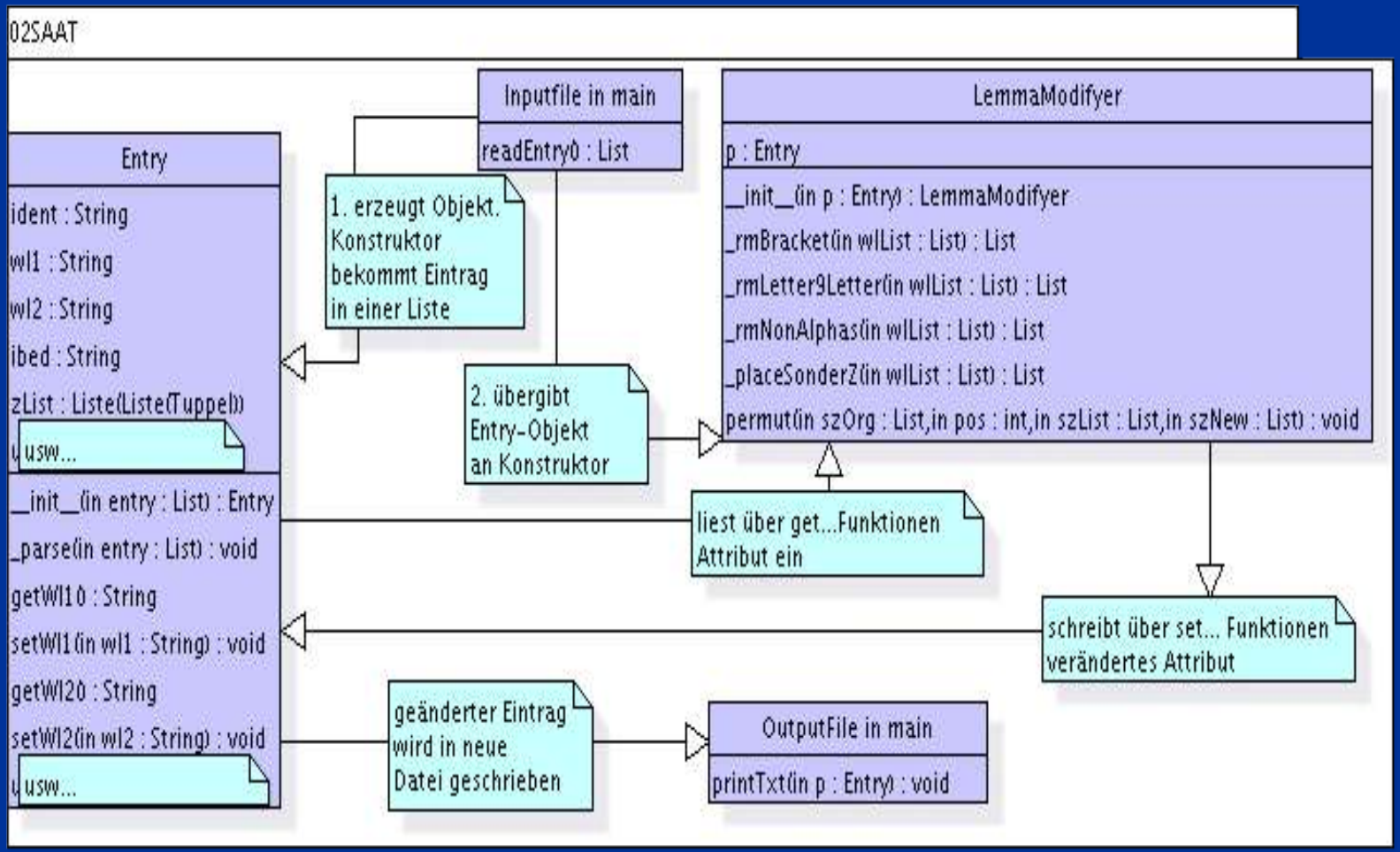
Umlaute und ß: in WDG nur als AE*, OE, UE, SS*

- für jedes mögliche Sonderzeichen wird eine Lemmavariante ergänzt, ggf. wird durchpermutiert, z.B. WL1: FEUERTUER FEUERTÜR FEÜRTUER FEÜRTÜR

Lemmakorrektur 2

- Nicht-alphanumerische Zeichen und Zahlen innerhalb der Lemmata werden in einer zusätzlichen Lemmavariante getilgt:
  - z.B. WL2: AB **(PRP)**, WL2: AB **(POP)**
  - z.B. WL1: KONTROL**9**LAMPE, WL1: WOL**9**LAPPEN
  - z.B. WL1: AUGENMASS**<**, WL1: ENGPAESS**<<**E
  - z.B. WL1: DOMINO**]**SPIEL, WL1: FOLIO**]**BAND
  - ferner getilgt: &, „“, '

Klassendiagramm 2



03WAT - Handarbeit

mit Zahlen kodierte Kurzform des eigentlichen Lemmas (gut 80 Einträge)

- WL1: ABGAEB, WL2: 1 BEN
- WL1: ABGESEGELT, WL2: ^@ 4 GELN
- Neue Lemmaversion mit ausgeschriebenem Infinitiv angehängt.

04WAT – Abgleich der Lemmata

Eine Inputliste aller Wortlemmata wird mit den Sprachkorpora des IDS abgeglichen.

Eine Outputliste enthält die Lemmata mit ihrer Trefferanzahl in den Korpora.

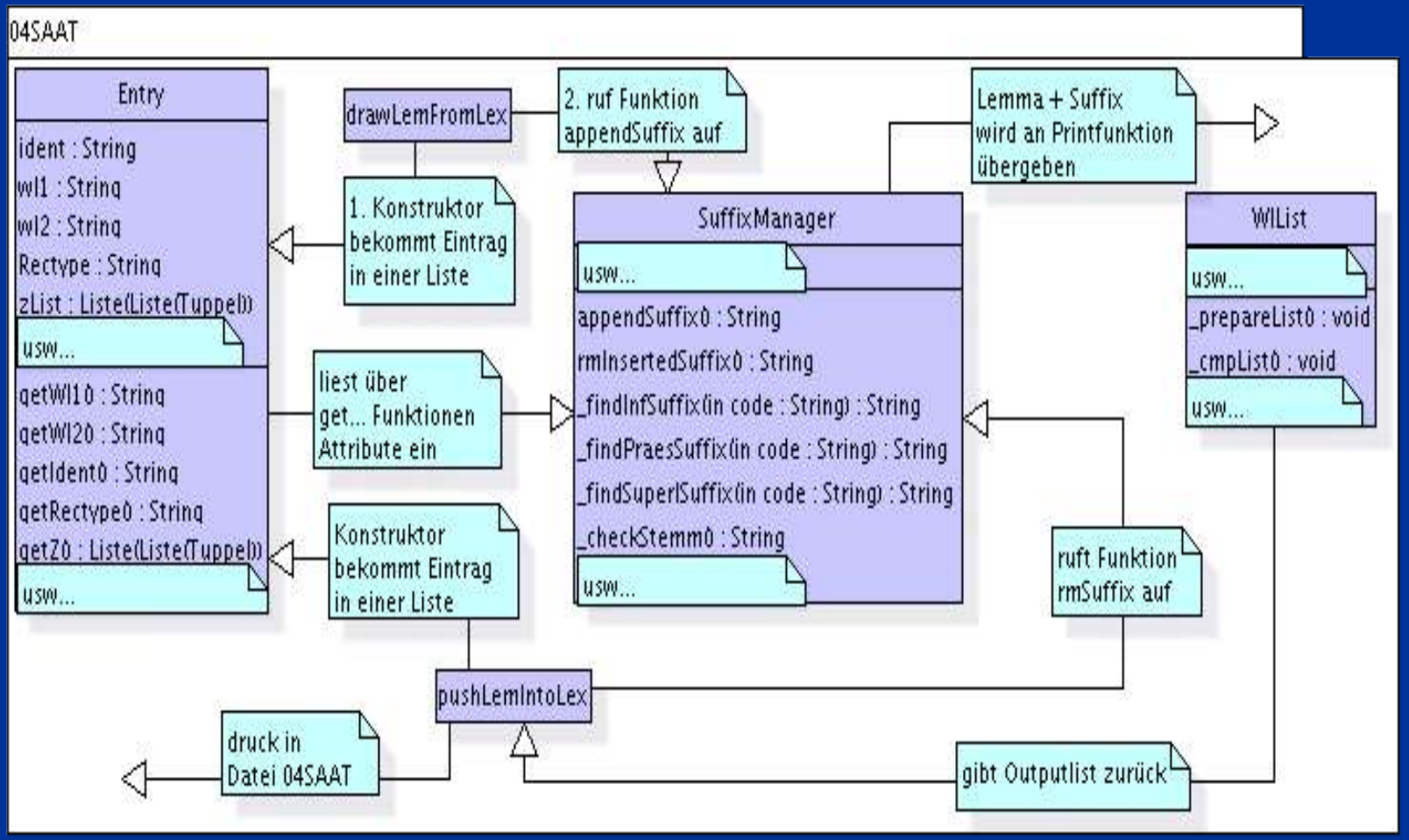
Vorbereitungen:

- Endungen bei nackten Stämmen (gaeb, geb, ...)
- Kodierung zum Rücktransfer der Liste in das Wörterbuch

Ziele:

- Überprüfung der Existenz des Lemmas
- Auswertung der Lemmavarianten

Klassendiagramm 3



05WAT – Datenkorrektur und XML

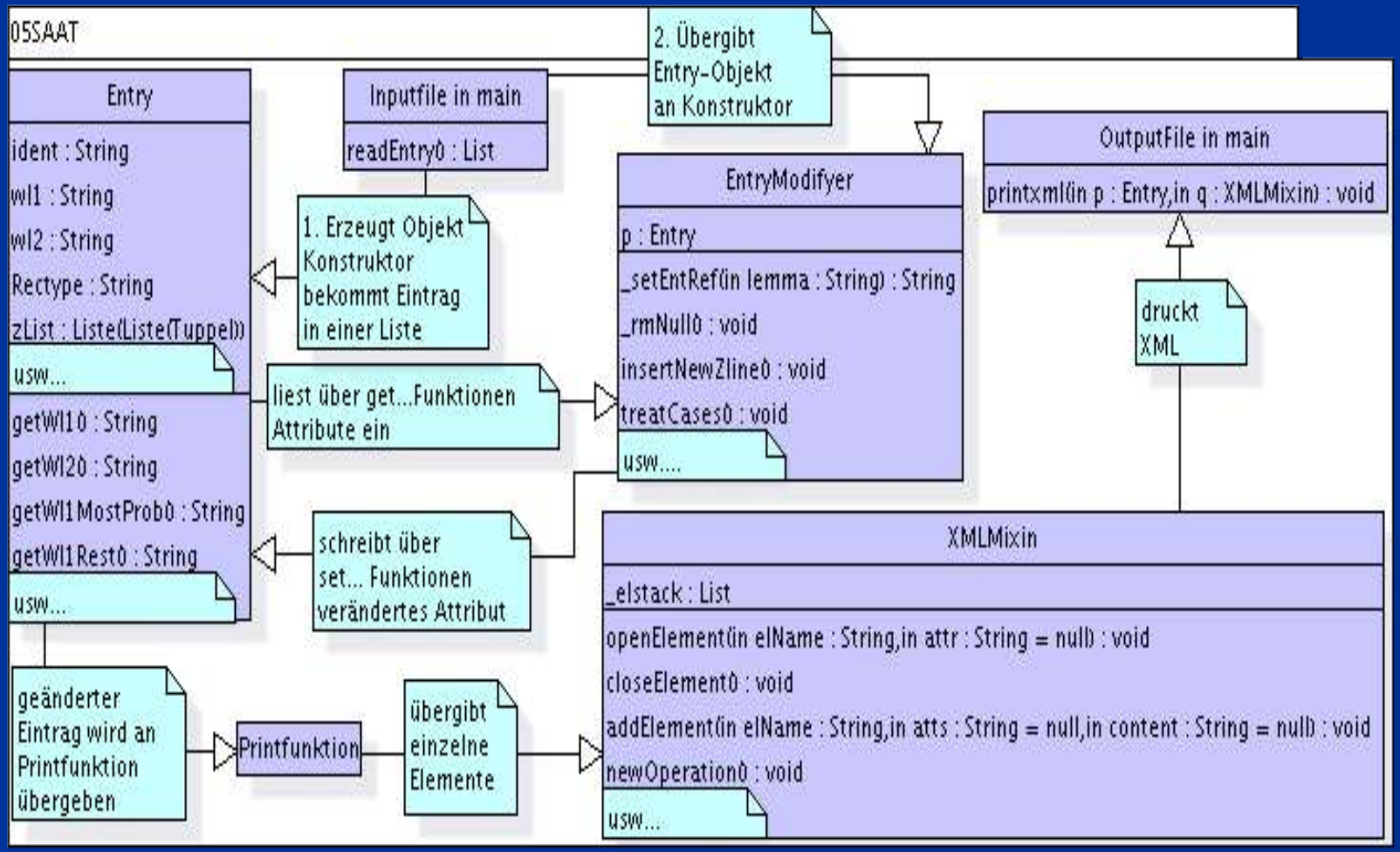
Vorbereitungen zur XML-Transformation:

- Entity Referenzen für XML-Metazeichen setzen
- Tilgen des Steuerzeichens ^@

Weitere Datenkorrekturen:

- Groß-und Kleinschreibung der Hauptwortklassen
- Einfügen mutmaßlich gelöschter Z-Zeilen

Klassendiagramm 4



XML-Struktur

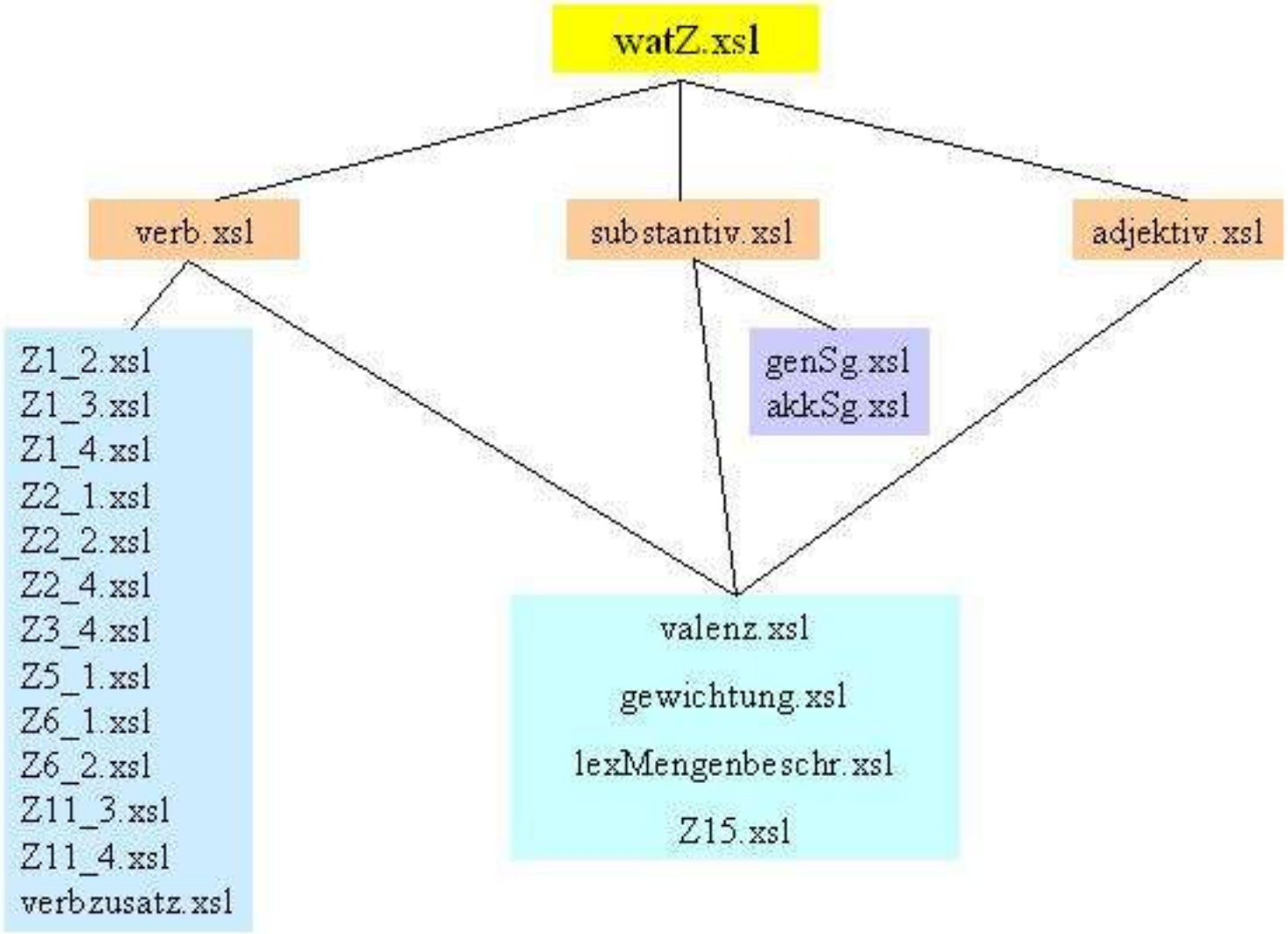
```
<entry checked="0" id="46402">
<wl1>
<original lemma="fuenfstimmig" freq="0"></original>
<mostProb lemma="fünfstimmig" freq="27"></mostProb>
<wl1Rest lemma="fünfstimmig" freq="27" id="1"></wl1Rest>
</wl1>
<ibed nr="4"></ibed><rectyp typ="3"></rectyp>
<Z new="no" id="6"><s1 info="0"></s1><s2 info="0"></s2>
<s3 info="8"></s3><s4 info="0"></s4>
</Z>
<author name="SDW"></author>
<date date="30681"></date>
</entry>
```

Interpretation der Daten durch Stylesheets

- dtd
- Stylesheets
- Zentral: watZ.xsl

  - Verweise auf weitere benötigte Stylesheets
  - Einführen einer Variable „XPath“
  - Auslesen der allgemeinen Daten (WL1, WL2, IBED, RECTYP)
  - Weiterleitung zu entsprechenden Stylesheets nach RECTYP

Schema der Stylesheets für Hauptwortklassen mit Z



Evaluation

WAT wurde nach drei unterschiedlichen Aspekten evaluiert:

1) Vergleich der XML-Version mit der Originalversion
2) Vergleich des Stylesheets mit der Originaldoku
3) Vergleich der Daten mit der linguistischen Realität

Evaluationskorpus: 100 Einträge von einem Zufallsgenerator ausgewählt.

XML-Daten vs. Originaldaten

- Originaldaten vollständig vorhanden: ja
- Groß- und Kleinschreibung korrekt: ja
- Element *mostProb* enthält höchste Trefferanzahl: ja
- Element *wlRest* enthält alle Lemmaversionen außer dem Original: ja
- Lemma mit der höchsten Trefferfrequenz ist das richtige: nein bei AUSFLUSS AUSFLUß (neue Rechtschreibung!)
- Die selbst generierte Form ist die wahrscheinlichste: in 29% der Fällen
- Systematischer Fehler in den Q-Zeilen ->musste wegen DTD bereinigt werden.
- Lemmaauswahl: nur 18 Einträge hatten mehr als 20 Treffer.

Stylesheets vs. Doku

Die Information wird entsprechend der Doku und den Stylesheets vollständig ausgelesen.

Notwendige Korrekturen:

- Bei Singularstämmen darf Z1/3 = 0 (Nom. Pl. Endung) nicht interpretiert werden.

Daten vs. Linguistische Realität 1

Richtige Einträge: 68

Fehlerhafte Einträge: 32 davon:

- 8 abtrennbare Präfixe
- 7 Belegungen, die es nicht gib
- 6 Dative

Daten vs. linguistische Realität 2

- Präfixlänge bei Verben
  --> kann im Programm korrigiert werden
- Z1 Zeile bei Nomen fehlt häufig (eliminiert?)
  --> kann evt. eingefügt werden, wenn Genitiv-, Dativ- und Akkusativendung = endungslos ist.
- Dativendung Sg./Pl. oft falsch
- Unentschlüsselte Information (Wort aus SDW)
- Komplemente oft falsch *herreist* – *Akkusativkomplement möglich

Daten vs. linguistische Realität 3

- Alle Einträge mit definierter 0 potenziell fehlerhaft:
  1) *Interlunium* (Sg-Stamm): *Nom.Pl. –*en*
  Glücksfälle (Pl-Stamm): *Nom.Pl. –*en*
  --> kann in den Stylesheets korrigiert werden
  2) *Froschblut* Sg u Pl-Stamm (da mit 0 belegt)
  3) Kompositabildung oft fraglich (mit 0 kodiert), z.B. *feinsinnig*