WDG - Analyse und Transformation (WAT)

Ineta Sejane, Wiebke Wagner
16.02.05

Institut für Deutsche Sprache
Mannheim
Sommersemester 04 - Wintersemester 04/05

Inhaltsverzeichnis

1 Planung

1.1 Ziel

Das Institut für Deutsche Sprache verfügt über das grammatische Wörterbuch WDG (Wörterbuch zur deutschen Grammatik), das bedingt durch das veraltete Format nicht mehr effizient nutzbar ist. Bei diesem Wörterbuch handelt es sich um ein sehr umfangreiches Werk, das auf Daten beruht, die an der Universität Saarbrücken durch automatische Lemmatisierung von Korpora in den 70er und 80er Jahren erstellt wurden. WDG existiert heute nur noch in Form von eher kryptischen Textdateien. Der Umfang beläuft sich auf knapp 150.000 Einträge. Das Wörterbuch verfügt nur in kleinem Rahmen über semantische Angaben. Dafür liefert es sehr detaillierte morphologische und syntaktische Informationen, sowie lexikalische Angaben über dialektale Begrenzung und fachsprachliche Zugehörigkeit. Die Informationen sind teils in Zahlen und Zeichen verschlüsselt, teils aber auch durch ihre Position innerhalb der Daten festgelegt. Die Kodierung ist dabei nicht einheitlich, die vorliegende Dokumentation leider unvollständig.

In diesem Projekt soll WDG zur besseren Benutzbarkeit neu bearbeitet werden. Dafür werden die Daten zunächst vereinheitlicht und das Wörterbuch in der neu erstellten Version WAT in ein XML-Format transferiert. Ferner wird der Inhalt des Wörterbuchs auf seine linguistische Korrektheit analysiert und ggf. korrigiert.

Auch nach der Bearbeitung sollen die Originaldaten vollständig erhalten bleiben, Korrekturen werden als solche vermerkt. Das Wörterbuch soll in eine Form gebracht werden, die mit modernen Programmen problemlos weiterverarbeitet werden kann.

1.2 Ressourcen

1.2.1 WDG

Statistik:
Einträge insgesamt: 148 507
Verben: 44 402
Nomen: 86 202

Adjektive: 14 757
Präfixe, Fugen, Suffixe: 410

Die kodierte Information

Die Lemmata sind in vier große Klassen eingeteilt: Substantive, Adjektive, Verben und eine Restklasse, genannt Funktionswortklasse. Die ersten drei Klassen sind gut beschrieben und können gut weiterverarbeitet werden. Ganz anders die Restklasse. Sie umfasst alle Wortformen der Wörter, die keine Substantive, Verben oder Adjektive sind, und außerdem solche Wortformen, die nicht von der Flexionsanalyse erkannt werden können, d.h. die nicht von einem Stamm ableitbar sind wie *bin*, *ist*. Diese Klasse ist nicht so gut dokumentiert wie die 3 großen Wortartenklassen. Die Informationen dieser Wortklasse sind zum Teil in einer allgemeinen Maske abgelegt, die nicht wortartenspezifisch differenziert ist, und zum Teil in Teilmasken für die jeweilige Wortklasse, auf die nach Prüfung bestimmter Angaben aus der gemeinsamen Maske geschlossen werden soll.
Jedes Lemma enthält mindestens folgende Daten:

1. WL1 das Textwort
   Das Textwort ist das Lemma. Es handelt sich dabei um jede abweichende oder nicht reguläre Form des Wortes.
   Bei Substantiven ist es der Nominativ Singular. Falls sich der Wortstamm Nominativ Plural davon unterscheidet, wird auch dieser als Textwort aufgenommen, z.B.
   Haus und Häuser, Lexikon und Lexika
   Bei Adjektiven ist es der unflektierte Positivstamm. Falls sich der Komparativstamm und/oder der Superlativstamm davon unterscheiden, werden sie auch als Textworte aufgenommen, z.B.
   schön 1 Eintrag (schön/schöner/schönst)
   arg/ärg 2 Einträge (arg/ärger/ärgst)
   nah/näh/näch 3 Einträge (nah/näher/nächst)
   Bei Verben werden max. folgende Stämme bzw. Formen als Textworte aufgenommen: Infinitivstamm, Präsensstamm, Partizip II, Infinitiv mit „zu“ im Wortlaut, Präteritumstamm, Konjunktiv II-Stamm. Auf jeden Fall werden Infinitivstamm und Präteritumstamm sowohl bei starken als auch bei schwachen Verben angegeben. Falls zwei oder mehr Stämme im Wortlaut übereinstimmen, werden sie nur einmal aufgenommen, z.B.
   geb, gib, gegeben, abzugeben, gab, gäb
   horch, gehorcht, aufzuhorchen, horcht
   absolvier, absolviert

2. Name der Person, die den Eintrag gemacht hat oder Quelle des Eintrags.

3. WL2 die Grundform des Wortes
   Sie wird nur dann eingetragen, wenn der Wortlaut des Textwortes WL1 von der Grundform abweicht, z.B. *Haus* für *Häuser*.

4. Datum des Eintrags

5. RECTYP die Wortklasse
   1 Verb
   2 Substantiv
   3 Adjektiv
   32 Funktionswortklasse

   außerdem bei Präfixen, Suffixen und Fugen:
   4 Präfix
   5 Fuge
   6 Suffix

6. IBED Bedeutungsnummer
   Bedeutungsnummer macht eine Unterscheidung zwischen identisch geschriebenen Textwörtern aus einer oder unterschiedlichen Wortklassen, z.B.
   lieb Adjektivstamm IBED 4
   lieb Infinitivstamm IBED 6
   oder
   überläuf intransitives Verb IBED 6 (Wenn die Milch überläuft, ..)
   überläuf transitives Verb IBED 7 (Wenn es dich überläuft, ..)
   Substantiv: 1-3
   Adjektiv: 4-5
   Verb: 6-8
   außerdem bei Präfixen, Suffixen und Fugen:
   99 - Fugen
   X0 - Präfix
   XY - Suffix
   Die X-Y-Kodierung enthält Information über die Konversion: X ist die Stammwortklasse des Stammes, dem das Präfix bzw. Suffix vor- bzw. nachgestellt wird (1-Verb, 2-Substantiv, 3-Adjektiv, addierbar), Y ist die Ergebniswortklasse (1-Verb, 2-Substantiv, 3-Adjektiv), z.B. beim Suffix -keit (Rectyp=6) wird unter IBED mit 42 kodiert, da es aus Verb und Adjektiv (RECTYP 1+3) ein Substantiv (RECTYP 2) macht.

7. INFO
   Hier wird die ganze morphologische und syntaktische Information der Einträge kodiert. In den 3 großen Hauptwortklassen sind es Z-Zeilen, jede mit 4 Stellen, in der Restklasse Q-Zeilen, auch mit je 4 Stellen. Die Information ist mit Zahlen kodiert, deren Bedeutung der Dokumentation entnommen werden kann. Leider ist die Originaldokumentation nicht vollständig erhalten bzw. nicht up-to-date.

1.2.2 Die Textkorpora des IDS

Die Korpora geschriebener Gegenwartssprache des IDS bilden mit knapp zwei Milliarden Wörtern die weltweit größte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten. Sie enthalten neben einer großen Zahl von Zeitungstexten belletristische, wissenschaftliche und populärwissenschaftliche Texte sowie eine breite Palette weiterer Textarten. Diese Ressource wurde genutzt, um die in WDG vorkommenden Lemmata nach ihrem Vorkommen zu überprüfen, da das Wörterbuch viele Lemmata beinhaltet, die nicht existent sind wie z.B. *&& 2 1EBAN* oder *CITROE5N*. Um solche Elemente zu filtern, wird die Trefferzahl für jedes Wortlemma in den Korpora ermittelt und in das Wörterbuch mit aufgenommen. Die Lemmafrequenz kann zwar nicht als sicheres Indiz für die Existenz oder Nicht-Existenz eines Lemmas gewertet werden. Sie gibt aber doch mit einer großen Wahrscheinlichkeit Aufschluss.

1.3 Methode

Die Aufbereitung des Wörterbuchs erfolgt in mehreren Schritten und bedarf fünf Programmaufrufe. Vor der endgültigen XML-Version 05WAT werden vier Zwischenversionen erstellt.

1.3.1 01WAT – 1:1-Übertragung

Die vorliegenden Quelldateien weisen große Uneinheitlichkeit im Format auf, was die maschinelle Lesbarkeit stark erschwert. Deshalb wurde zunächst eine Version geschaffen, die die Daten selbst unberührt lässt und lediglich die einzelnen Einträge in ihrer Mikrostruktur vereinheitlicht. Zunächst zwei Beispiele eines Eintrags aus dem Original und 01WAT:

WDG Originalversion:
(Aus datenschutzrechtlichen Gründen, können die Beispiele leider nicht angezeigt werden.)

WDG 01WAT:
(Aus datenschutzrechtlichen Gründen, können die Beispiele leider nicht angezeigt werden.)

Veränderungen:

- Jeder Eintrag bekommt eine ID zugewiesen.
- Gibt es mehrere Informationseinheiten innerhalb einer Zeile, so werden die Zeilen getrennt. Das geschieht in der ersten Zeile, in der das Wortlemma1 und ein Name - vermutlich der Autorenname - vermerkt ist und in der zweiten Zeile, in der das Wortlemma2 und ein Datum - vermutlich das Eintragungsdatum - enthalten sind.
- Das Topik einer jeden Zeile wird von dem Informationsteil mit einem Doppelpunkt und einem Leerzeichen getrennt. Keinen Doppelpunkt erhalten die Q-Zeilen und die Z-Zeilen. Sie werden von ihrem Vierer-Tupel durch einen Tab getrennt. Ebenso werden die vier Tupelelemente innerhalb einer Q- und Z-Zeile durch einen Tab unterschieden.
- Die Q-Zeilen sind die weit heterogenste Gruppe des Wörterbuchs. Uneinheitlichkeit besteht in Klammerung von Zahlen und Kommasetzung. Da beide als Trenner nicht nötig sind, werden sie entfernt. Häufig gibt es Zuweisungen innerhalb der Q-Zeilen, die durch ein =-Zeichen ausgedrückt werden. Wenn das =-Zeichen im Original nicht gesetzt ist, wird es nachträglich hinzugefügt.
- Für die Zeilen, die in keines der angelegten Schemen passen, wird das neue Topik *unknown* angelegt. Die betreffenden Einträge finden sich in der Datei „01unknown.txt“ des Ordners „res/teilergebnisse“. Auch sie weisen aber Einheitlichkeit auf. Es handelt sich ausschließlich um Q-Zeilen, in denen die Notiz: *dürfte nicht besetzt sein!* vermerkt ist.
- Da das originale Wörterbuch keine Tabs, dafür sehr viele Leerzeichen verwendet, reduziert sich die Dateigröße in der ersten Version von ca. 30 MB auf ca. 15 MB.

1.3.2 02WAT – Datenkorrektur

Die Lemmata des Originals weisen eine Vielzahl von Zahlen und nicht-alphanumerischen Zeichen auf, die in einem neuen Wörterbuch nur stören. Vermutlich handelt es sich dabei um Metazeichen, die entweder nicht rekonstruierbar sind oder sehr unvollständig gesetzt sind. In der Dokumentation finden sich leider keine Informationen

darüber. Diese Zeichen sollen unter Beibehaltung des Originals in einer neuen Lemmaversion so gut wie möglich getilgt werden. Das originale Lemma bleibt dabei immer an der ersten Stelle der Lemmaauflistung. Die neuen Lemmaversionen werden durch ein Tab-Zeichen getrennt. Der Tab eignet sich deswegen als Trenner, da er in den Wörterbuchdaten nicht vorkommt und deshalb eindeutig ist.
Zur Verdeutlichung noch einmal ein Beispiel:

WDG 01WAT:
(Aus datenschutzrechtlichen Gründen, können die Beispiele leider nicht angezeigt werden.)

WDG 02WAT:
(Aus datenschutzrechtlichen Gründen, können die Beispiele leider nicht angezeigt werden.)

Veränderungen:

- Das Originalwörterbuch enthält eine Gruppe von Einträgen, die hinter dem Lemma in Klammern die Wortart oder eine andere Information enthalten z.B. *AB (PRP)*. Diese Information ist vielfach redundant, da sie sich bereits in den Kodes der Q-Zeilen befindet. Ist sie tatsächlich in den Daten vermerkt, wird sie in einer zusätzlichen Lemmavariante entfernt. Keine neue Variante wird eingefügt, wenn die Information innerhalb der Klammer nicht anderen Orts vermerkt ist, oder nicht interpretiert werden kann.

- Es gibt ferner eine Gruppe, die statt eines dritten gleichen Buchstabens innerhalb des Lemmas eine 9 enthalten, z.B. *RAUMSCHIF9FAHRT*, *WOL9LAPPEN*, etc. Möglicherweise sollte damit markiert werden, dass bei Silbentrennung ein dritter Buchstabe dazukommt. Da nach der heutigen Rechtschreibung ohnehin drei Buchstaben stehen, wird dem Lemma eine neue Lemmavariante angehängt, in der die 9 durch den entsprechenden Buchstaben ersetzt ist.

- Einige Lemmata enthalten nicht-alphanumerische Zeichen. Sofern es sich um einigermaßen verständliche Einträge handelt, wurde eine neue Lemmavariante ergänzt.

  - **Spitze Klammern:** Die sich öffnende spitze Klammer entscheidet vermutlich über Doppel-S und ß, das in WDG ausgeschrieben nicht vorkommt. Zwei Klammern drücken dabei das Doppel-S aus und eine Klammer das ß, z.B.
    WL1: AUGENMASS<
    WL1: FESTAUSSCHUESS<<E

Leider ist diese Markierung sehr unvollständig. Sie kommt nur in den wenigsten Lemmata mit Doppel-S vor, was sie quasi unbrauchbar macht. Deshalb wird jeweils eine neue Lemmavariante angehängt. Die Entscheidung, um welche Schreibung es sich in einem Lemma handelt, wird an anderer Stelle getroffen.

- **Eckige Klammern**: Die sich schließende eckige Klammer scheint ein Trenner zwischen Stamm und Suffix bzw. zwischen zwei Kompositumteilen von Nomina zu sein, z.B.
  WL1: VARIABIL]ITAET
  WL1: KOMMANDO]TURM
  Auch dieser Trenner ist sehr unvollständig gesetzt und deshalb für unsere Zwecke unbrauchbar. Auffällig ist, dass bei den meisten (nicht bei allen!) dieser Lemmata Z5/3 mit einer 8 markiert ist. Laut WDG-Dokumentation bedeutet dies: *Wort aus SDW, Wortlaut enthält "/"*. Die Aussage dieser Information ist allerdings unklar.
- **Kaufmännisches Und:** Wird sehr uneinheitlich gebraucht. Hier einige Beispiele:
  Wl1: DEUTSCHEN-&DEMOKRATISCHEN-&REPUBLIK
  WL2: &&0201LANDGOTT
- **Apostroph:** Der Gebrauch von Apostroph ist vermutlich auf dialektale und umgangssprachliche Wendungen zurückzuführen:
  WL1: 'RUNTERRUTSCH, WL2: 'RUNTERRUTSCHEN
  WL1: 'RAUSKOMM, WL2: 'RAUSKOMMEN
  Da diese Wendungen üblicherweise ohne Apostroph geschrieben werden, bekommt das Lemma eine neue Variante.
- **Dollarzeichen:** Lemmata mit *$*-Zeichen drückten in WDG möglicherweise irgendwelche Regeln aus, die aber weder irgendwo spezifiziert noch nachvollziehbar sind:
  WL1: $CMPRULW2, WL2: Substantiv-Komposita-Regel
  WL1: $CMPWKL.
  Da die Einträge für unsere Version ohnehin überflüssig zu sein scheinen, werden sie nicht behandelt.

- Umlaute und wie bereits erwähnt ß gibt es in WDG nur in umschriebener Form, also *AE, OE, UE, SS*. Um die reguläre deutsche Schreibung zu finden, wird zunächst für jedes mögliche Sonderzeichen eine neue Lemmavariante mit Sonderzeichenschreibung ergänzt. Bei mehreren potenziellen Sonderzeichen wird durchpermutiert, z.B.
  WL1: FEUERTUER FEUERTÜR FEÜRTUER FEÜRTÜR
  Maximal kommen 16 Varianten für ein Lemma zustande.

Lemmata, die mit neuen Varianten ergänzt werden, kopiert das Programm in extra Dateien. Sie sind in dem Ordner „res/teilergebnisse“ unter „02rm....txt“ zu finden. Ferner sind in diesem Ordner einige Textdateien mit den Namen 02wl....tex“ , die von den Programmen in „src/02wdgCorr/check“ generiert worden sind. Diese Textdateien wurden benötigt, um die behandelten Sonderzeichen in den Wortlemmata zu beurteilen.

1.3.3 03WAT – Handarbeit

Das WDG-Original enthält gut 80 Einträge, die in ihrem WL2 - seltener auch in WL1 - eine Kurzform des eigentlichen Lemmas enthalten. Die Kurzform besteht entweder aus Zahl + den letzten paar Buchstaben des Wortes oder, was bei weitem problematischer ist, aus dem Steuerzeichen ^@ + Zahl + den letzten Buchstaben, z.B.
WL1: ABGLITT, WL2: 1 EITEN
WL1: ABGESEGELT, WL2: ^@4 GELN
Die 4 nach dem Steuerzeichen bezeichnet möglicherweise die fehlende Buchstabenanzahl des Infinitivs; allerdings stimmt die zwischengestellte Zahl nicht immer mit der ausgelassenen Buchstabenanzahl überein. Die richtige Form des Lemmas maschinell zu erschließen ist hier unmöglich. Deshalb wird den entsprechenden Einträgen von Hand eine zusätzliche Lemmavariante gegeben, die die vollständige Wortform des Lemmas, meist den Infinitiv, beschreibt, also:
WL1: ABGLITT, WL2: 1 EITEN ABGLEITEN
WL1: ABGESEGELT, WL2: ^@4 GELN ABSEGELN
Da der XML-Parser das Steuerzeichen nicht parsen kann, muss es vor dem XML-Transfer endgültig getilgt werden. Eine Liste von Lemmata, die das Null-Zeichen enthält, findet sich in der Datei „02wlNullCharacter.txt“ in „res/teilergebnisse“. Eine Liste der übrigen von Hand manipluierten Einträge mit Zahlen liegt im gleichen Verzeichnis unter „021rmNum.txt“

1.3.4 04WAT – Abgleich der Wortlemmata

Da viele Lemmata von WDG im Sprachgebrauch nicht vorkommen, wird eine Input-Liste aller Lemmata erstellt, die mit den Sprachkorpora des IDS abgeglichen werden. Es wird für jedes Lemma bzw. für jede Lemmavariante eine Suchanfrage gestartet und die Anzahl der Treffer des betreffenden Lemmas in einer Output-Liste hinter das Lemma geschrieben. So kann festgestellt werden, ob ein Lemma überhaupt vorkommt und welche Lemmavariante am häufigsten vorkommt. Gibt es mehrere Lemmavarianten, so wird die Variante mit der größten Trefferanzahl in der XML-Version 05WAT als das wahrscheinlichste angenommen. Um die Output-Liste anschließend wieder in das Wörterbuch zu transferieren, sind die Lemmata folgendermaßen durch

Leerzeilen getrennt: Lemmavarianten werden durch einen Zeilenumbruch getrennt. Wl1 wird von Wl2 mit einer Leerzeile getrennt, Wl2 wird von Wl1 von zwei Leerzeilen getrennt. Das sieht aus wie folgt:

Input-Liste	**Output-Liste**	
AB OVO	AB OVO 38	Wl1 zwei Leerzeilen
AB UND ZU	AB UND ZU 13018	Wl1 zwei Leerzeilen
ABAENDERLICH ABÄNDERLICH	ABAENDERLICH 0 ABÄNDERLICH 0	Wl1 zusätzliche WL1-Variante zwei Leerzeilen
ABAENDERTE ABÄNDERTE	ABAENDERT 0 ABÄNDERT 39	Wl1 zusätzliche WL1-Variante eine Leerzeile
ABAENDERN ABÄNDERN	ABAENDERN 0 ABÄNDERN 389	Wl2 zusätzliche WL2-Variante zwei Leerzeilen

Bei Verben und Adjektiven vermerkt WDG meist nur den Stamm der Wortform. Deshalb würden Lemmata wie *WL1: GAEB* oder *WL1: AELT* keine Treffer ergeben. Um dies zu vermeiden, wird den Konjunktiv-, Infinitiv-, Präsens-, Komparativ- und Superlativstämmen der Input-Liste künstlich die entsprechende Endung angehängt. Bei *GAEB* wäre das *GAEBE*, bei *AELT* wäre das *AELTER*. In der Output-Liste werden diese Endungen wieder entfernt bei Beibehaltung der Trefferanzahl. Die betroffenen Lemmata finden sich in den Dateien „03wl...Stem.txt“ in „res/teilergebnisse/“
Die Version 04WAT enthält die Ergebnisse aus der Output-Liste. In den Lemmata wird zu jeder Lemmavariante eine Trefferanzahl notiert. Lemma und Treffer sind durch ein μ getrennt, da μ in den Wörterbuchdaten ansonsten nicht vorkommt. Der Trenner μ ist notwendig, um die Daten aus der Datei wieder mühelos einlesen zu können.
Bei der Generierung der Output-Liste gab es Probleme mit dem bereits erwähnten Steuerzeichen Null. Zeilen, die dieses Zeichen enthalten, konnten von dem Steuerzeichen an bis zum Zeilenende nicht dargestellt werden. Die dadurch entstehende Diskrepanz der Listen wird ausgeglichen, indem die Input-Liste und die Output-

Liste miteinander abgeglichen werden. Sind die Lemmata nicht identisch, wird die Zeile der Input-Liste in der Output-Liste übernommen und mit einer Null versehen. Die Null stellt die nicht überprüfte Trefferanzahl dar. Es ist aber kaum davon auszugehen, dass Lemmata mit Null-Charakter Treffer liefern würden. Die manipulierten Zeilen werden in der Datei „03listNotEqual.txt“ in „res/ teilergebnisse/“ aufgezeichnet. In ihren Einträgen lässt sich überprüfen, dass es ausschließlich bei Zeilen mit Null-Charakter bei der Bearbeitung zu Unterschieden gekommen ist. Diese Datei lässt sich am besten auf Konsole oder in einem Konsolen-Editor anschauen, da die herkömmlichen Editoren das Steuerzeichen nicht abdrucken, sondern die Datei ab dem ersten Steuerzeichen leer darstellen.

1.3.5 10WAT – Datenkorrektur und XML-Ausgabe

Bevor die Version 04WAT in XML überführt wird, erfolgen weitere Korrekturen an den Daten, die in der Klasse EntryModifyer bewerkstelligt werden. Diese Klasse bekommt einen kompletten Wörterbucheintrag, der innerhalb der Klasse verändert werden kann. Einige Veränderungen werden automatisch durchgeführt, da sie für die Lesbarkeit des XML-Parsers unabdingbar sind. Andere können im main-Programm selbst aufgerufen werden. Diese Klasse ist beliebig erweiterbar, wenn noch weitere Korrekturen angestrebt sind.

Veränderungen:

- WDG enthält einige XML-Metazeichen wie &, “, ’ und <. Um diese Zeichen für XML-Parser zugänglich zu machen, müssen sie in Entity Referenzen umgewandelt werden.
- Der Null-Charakter, der in 44 Einträgen vorkommt, muss hier eliminiert werden, da auch er für XML-Parser nicht zu lesen ist. Diese Veränderung ist innerhalb von WAT nicht ersichtlich. Die betroffenen Einträge sind aber in der Datei „04rmNullCharakter“ in „res/teilergebnisse“ zu finden.
- Innerhalb der Z-Zeilen besteht das Problem, dass eine Null sowohl eine nicht belegte Stelle darstellen, oder aber auch eine Informationseinheit enthalten kann. Aus dieser Konfusion heraus sind vermutlich einige Z-Zeilen aus WDG getilgt worden, die nur Nullen enthielten. Darauf ist zu schließen, wenn Z-Zeilen auf andere Z-Zeilen Bezug nehmen, die aber nicht vorhanden sind. Ein Beispiel ist bei den Verben die Zelle Z6/3, in der der Verbstamm kodiert ist. Handelt es sich um einen Präteritumstamm, so sollte auch die Zeile Z2 vermerkt sein, da sie das Flexionsparadigma der Präteritumstämme enthält. Ein sehr häufiges Muster (z.B. das von *nahm*) ist in Z2/1 mit null markiert. Wenn diese Zeile fehlt, obwohl es sich um einen Präteritumstamm handelt, wird sie

mit Z2 0 0 0 0 künstlich eingefügt. In dem XML-Dokument lässt sich diese Veränderung an dem Attribut *new* = *"yes"* ablesen. Bei allen originalen Z-Zeilen ist dieses Attribut mit *"no"* kodiert.

- Die Lemmata in WDG enthalten (fast) ausschließlich Großbuchstaben. Die wenigen Einträge mit Kleinbuchstaben sind hier aufgeführt:
  - ID: 4757, WL1: AMTL. WL2: r 1
  - ID: 69999, WL1: INTRIGANT, WL2: r
  - ID: 84147, WL1: MAD, WL2: r
  - ID: 98341, WL1: PHYLETISCH, WL2: r
  - ID: 123010, WL1: saemtlichst, WL2: saemtlich

  Bei den Lemmata der Hauptwortklassen (Verben, Nomen und Adjektiven) wird die Schreibung entsprechend der Rechtschreibregeln korrigiert. Die Funktionswortklasse (Rectyp 32) bleibt in Großbuchstaben belassen.

XML-Ausgabe
Die XML-Struktur der einzelnen Einträge sieht folgendermaßen aus:

- **Element *entry*:** Das Element umschließt einen kompletten Wörterbucheintrag. Es enthält in zwei Attributen zum einen die ID des Eintrags und zum anderen *checked*. *Checked* ist defaultmäßig auf 0 gesetzt und soll eine 1 bekommen, wenn der Eintrag von Hand überprüft und ggf. korrigiert worden ist.

- **Element *wl1*:** Das Element *wl1* enthält weitere Unterelemente:
  1) *original*: Das Element *original* hat ein Attribut *lemma*, in dem das originale WDG-Lemma angegeben ist und ein Attribut *freq*, das die Trefferhäufigkeit in den Korpora darstellt.
  2) *mostProb*: Das Element ist aufgebaut wie *original*. Es enthält aber das Lemma mit der höchsten Trefferanzahl. Haben alle Lemmavarianten 0 Treffer, gibt es dieses Element nicht.
  3)*wl1Rest*: *wl1Rest* enthält alle Wl1-Varianten außer dem Original und kann somit kein Mal, ein Mal oder mehr Mal vorkommen. Es ist ebenfalls so aufgebaut wie Element *original*, hat aber noch das zusätzliche Attribut *id*. Dieses Attribut nummeriert die wl1Rest-Elemente eines Eintrags durch.

- **Element *wl2*:** *Wl2* ist aufgebaut wie *wl1*. Es ist aber nicht bei allen Einträgen verzeichnet.

- **Element *ibed* und *rectyp*:** geben jeweils die in einem Attribut gespeicherte Nummer zurück.

- **Element *Z*:** Das Element *Z* kann mehrfach oder gar nicht vorkommen. Es enthält in einem Attribut die Nummer des Z-Eintrags. Das darauf folgende Vierer-Tupel wird in Form von vier eigenen Elementen *s1* bis *s4* dargestellt. Sie enthalten in ihrem Attribut die informationsgebende Zahl.
- **Element *Q*:** ist aufgebaut wie Element *Z*.
- **Element *author*:** enthält das Attribut *name*, das den mutmaßlichen Autorennamen bzw. die Quelle wiedergibt.
- **Element *date*:** enthält das Attribut *date*, das das mutmaßliche Eintragungsdatum enthält.
- ***unknown:*** Dieses Element beinhaltet alle Zeilen des Wörterbuchs die in keines der oben beschriebenen Schemen passt.

2 Implementierung

2.1 Python-Implementierung

Zur Übersichtlichkeit sollen hier vier Klassendiagramme aufgeführt werden, die grafisch das Zusammenspiel der Klassen und main-Programms darstellen. Einen ausführlichen Überblick erhält man in der Programm-Dokumentation in „doc/programmDoku“.

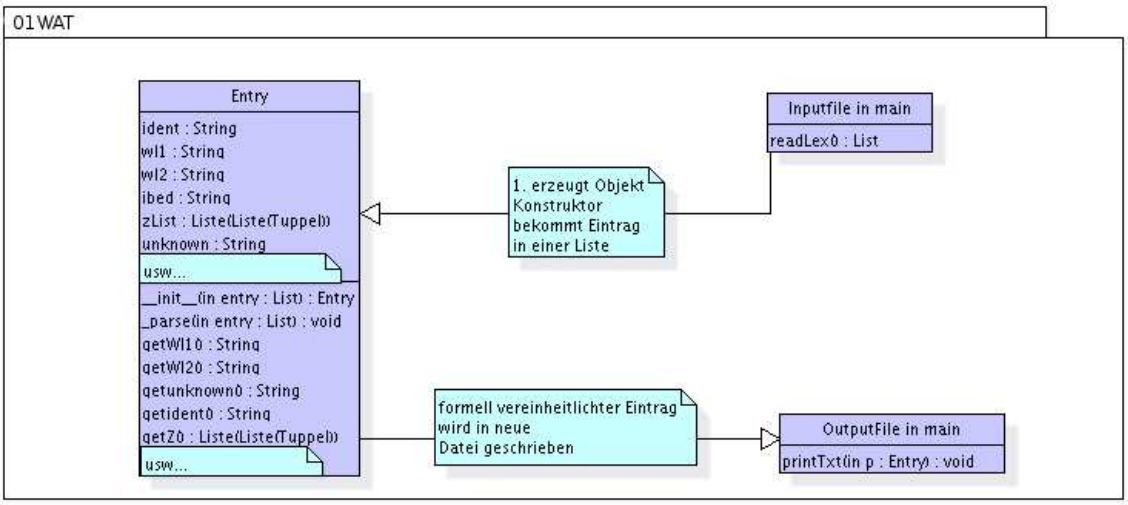

Abbildung 2.1: WAT01 Klassendiagramm

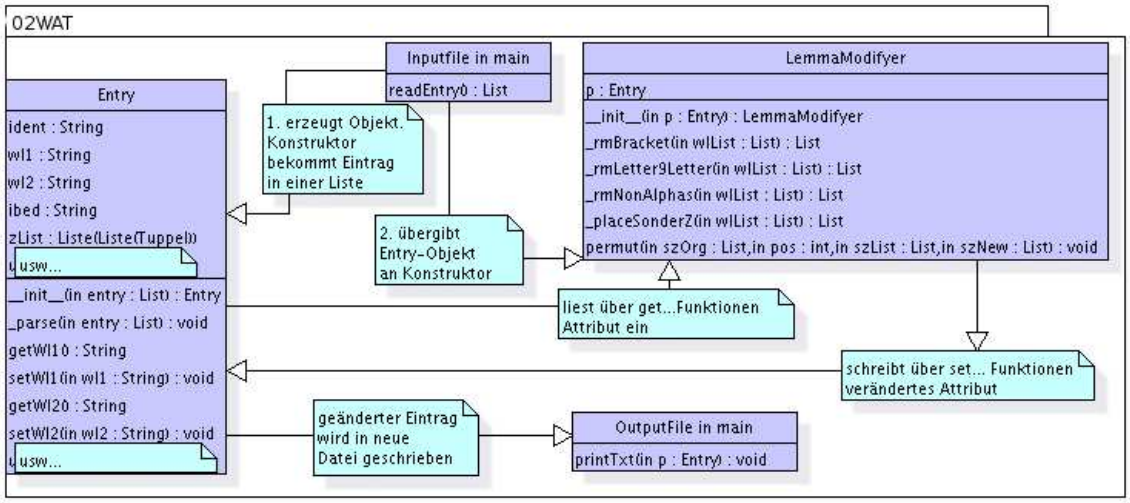

Abbildung 2.2: WAT02 Klassendiagramm

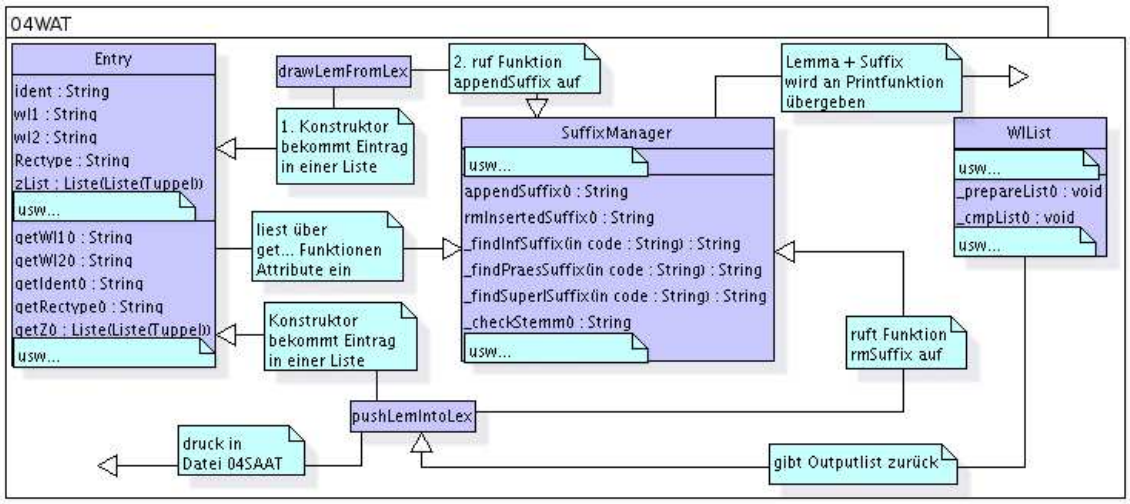

Abbildung 2.3: WAT03 Klassendiagramm

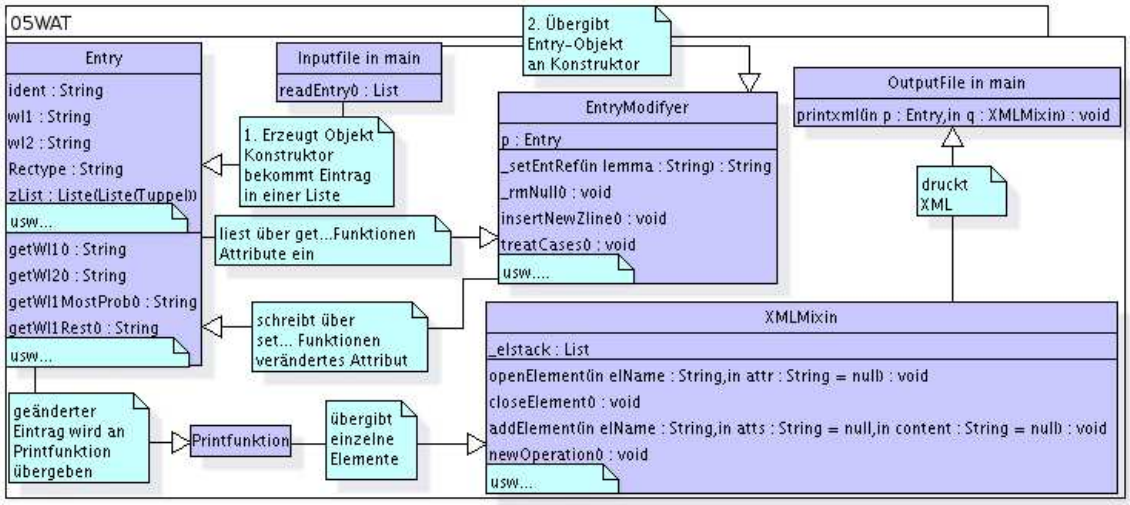

Abbildung 2.4: WAT04 Klassendiagramm

2.2 XSL-Implementierung

Mit Hilfe der Stylesheets werden die Informationen aus den Lemmata der drei Hauptwortklassen ausgelesen. Getrennt davon werden Suffixe, Fugen und Präfixe abgehandelt, die wegen ihrer Struktur und Funktion einen anderen Stellenwert haben; sie können allerdings ohne größeren Aufwand in die Gesamtstylesheets eingegliedert werden.

Die Struktur der Stylesheets für die Hauptwortklassen ist im folgenden Schema dargestellt. Nähere Information dazu befindet sich in der WDG-Dokumentation in „doc/WDG_Doku“.

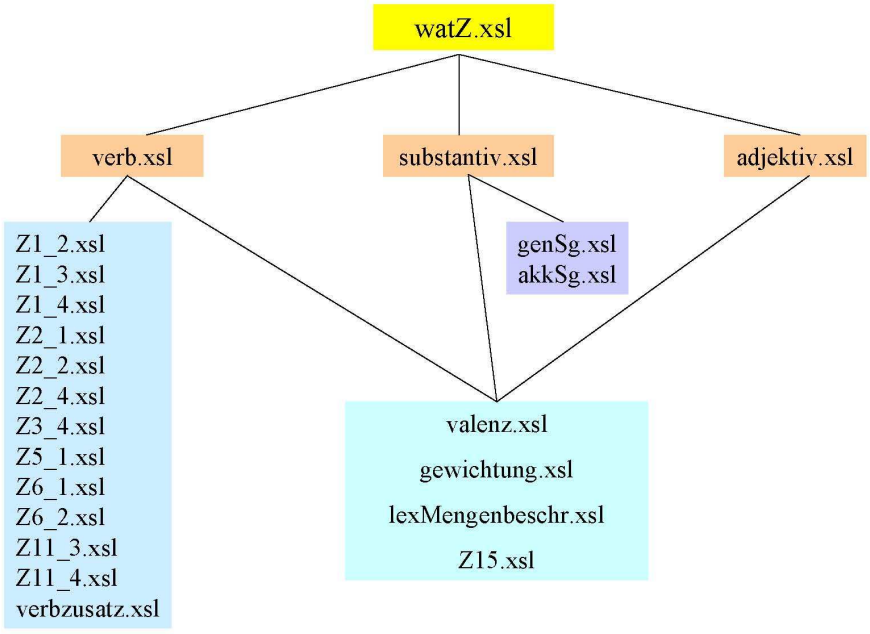

Abbildung 2.5: Schema der Stylesheets für Hauptwortklassen

3 Evaluation

Bei der Evaluation wurden 100 Einträge, die ein Zufallsgenerator ausgewählt hat, unter unterschiedlichen Aspekten ausgewertet. Das Evaluationskorpus findet sich im Verzeichnis „eval/“ in der Datei „01ramdomEntries.xml“.

3.1 XML-Daten vs. Originalversion

Hier wurde Folgendes überprüft:

1. Ist das WDG-Original in den WAT-Daten vollständig vorhanden oder sind Daten verloren gegangen. -> Innerhalb des Testkorpus waren alle Originaldaten vorhanden.

2. Stimmen die Veränderungen in Bezug auf Groß- und Kleinschreibung. -> Bei den berücksichtigten Wortklassen (Adjektive, Nomen, Verben) ist die Groß- und Kleinschreibung richtig.

3. Enthält das Element mostProb tatsächlich das Lemma mit der höchsten Trefferanzahl. -> In den Testdaten ja.

4. Enthält das Element wlRest alle Lemmavarianten außer dem Original. -> In den Testdaten ja.

5. Ist das Lemma mit der höchsten Trefferzahl tatsächlich das Richtige. -> In einem Fall traf das nicht zu, da sich die Rechtschreibregeln seit der Erstellung von WDG geändert haben. In dem Eintrag *Ausfluss* hatte die Lemmavariante *Ausfluß* die höchste Trefferanzahl. Das liegt daran, dass in den IDS-Korpora die Texte mit alter Rechtschreibung noch überwiegen.

6. Wie oft ist eine selbst generierte Lemmavariante die wahrscheinlichste. -> In den 100 Einträgen gab es insgesamt 136 Wl1- und Wl2-Elemente (da Wl2 nicht immer vorhanden sein muss). Von diesen Einträgen hatten 21 Elemente eine neu generierte Lemmavariante als die Wahrscheinlichste.

7. Auffällig im Element author war die starke Häufung des Autornamens bzw. der Quelle. 82 von den 100 Einträgen hatten name=SDW verzeichnet. Was name=SDW bedeutet ist unbekannt.

8. Der einzige Fehler wurde in den Q-Zeilen beobachtet. Es handelt sich um einen systematischen Formfehler, der in allen Q-Zeilen auftritt. Statt:
   <Q id=1><s1 info=0></s1><s2 jwk=21></s2><s3 jstw=0></s3> <s4 jbed=9></s4></Q>
   heißt es:
   <Q id=1><s1 info=0></s1><s2 info=jwk=21></s2><s3 info=jstw=0></s3><s4 info=jbed=9></s4></Q>
   Dieser Fehler muss noch behoben werden, da das Skript sonst nicht mit der DTD übereinstimmt.

9. Die Testdaten enthielten 69 Einträge mit 0-9 Treffern und 13 Einträge mit 10-20 Treffern. Es hatten also nur 18 Einträge mehr als 20 Treffer.

3.2 Stylesheets vs. WAT-Dokumentation

Die Information wird entsprechend der WDG-Dokumentation und den Stylesheets vollständig und richtig ausgelesen. Dies hat der wiederholte Vergleich verschiedener Dokumentationen über WDG mit den Stylesheets und ebenso die Richtigkeit der Stichproben aus dem Evaluationskorpus bewiesen.

3.3 WAT vs. linguistische Realität

Bei der Überprüfung der Grammatikalität der Daten fielen folgende Fehler auf:

- Die Länge des abtrennbaren Präfixes bei Verben ist in der Buchstabenanzahl des Präfixes angegeben. Augenscheinlich wurde hier teils bei 0 und teile bei 1 angefangen zu zählen. Wenn man von einer bei 0 beginnenden Zählung ausgeht, gab es in dem Evaluationskorpus folgende Fehler: *durchgeschleust* - *4, *austüftel*- *2, *verschwenk* - *3, etc.

- Bei vielen Substantiven fehlt die Geschlechtsangabe der Feminina und die Angabe über den Nominativ Plural, wenn er mit *-en* gebildet wird, z.B. *Ausbreitung*. Beide Informationen sind mit 0 belegt und befinden sich in der Zeile Z1. Möglicherweise ist diese Zeile bei kompletter Nullbelegung entfernt worden. Da die Kombination der beiden Angaben sehr häufig vorkommt, tritt der Fehler leider auch sehr oft auf. Die Fehlerquote ließe sich verbessern, wenn die Z1-Zeile bei endungslosem Genitiv, Dativ und Akkusativ eingefügt werden könnte - Informationen, die in Z2 vermerkt sind. Da diese Angaben aber ebenfalls jeweils mit 0 kodiert sind, fehlt auch diese Zeile bei den beobachteten Feminina. Alle anderen Z-Zeilen geben keinen Aufschluss auf das Genus. Ob die Fehleranzahl bei defaultmäßigem Einfügen der beiden Zeilen verringert

wird, müsste überprüft werden. Das Evaluationskorpus würde zumindest keine neuen Fehler aufweisen, wenn die nicht vorhandenen Z1- und Z2-Zeilen an jeder Stelle mit 0 kodiert eingefügt werden würden. Eine andere Möglichkeit wäre, die femininen Ableitungssuffixe wie *-ung, -keit, -heit*, etc. in den Lemmata zu suchen und ihnen bei Bedarf die entsprechenden Zeilen zuzuschreiben. Somit wären zumindest die femininen Ableitungen korrigiert.

- Ist die weibliche Genusangabe vorhanden, die Zeile Z2 mit den Kasusangaben aber nicht, so sollte diese Zeile eingefügt werden, da Feminina in allen Kasus endungslos sind und Endungslosigkeit mit 0 kodiert ist.

- Das gleiche Problem besteht bei den Infinitivstämmen, z.B. *beschwips*. Hier sollte die Zeile Z2 mit 0 0 0 0 eingefügt werden, weil in ihr die Infinitivendung mit 0 = *-en* kodiert ist. Der Einschub der Z-Zeilen ist noch nicht erfolgt.

- Die mögliche Kompositabildung ist in Z2/3 mit 0 kodiert. Die Richtigkeit dieser Angabe ist oft zweifelhaft, z.B. *feinsinnig, durchgeschleust, geknallt, garagiert*. Die Kompositabildbarkeit dieser Lemmata müsste in einem Korpus überprüft werden, um genaueren Aufschluss zu erhalten.

- In Z11 bei Verben werden zwei Angaben zum Gebrauch des Partizips gemacht. Auch hier gibt es Null-Kodierung, die fehlerträchtig ist, z.B. *nachzudrucken* - *Partizip nicht steigerungsfähig. Dieser Fehler konnte innerhalb der Stylesheets behoben werden, indem durch eine if-Abfrage die Information nur bei Partizipstämmen ausgewertet wird und sollte damit nicht mehr vorkommen.

- Der Nominativ Plural ist bei Substantiven in Z1/3 mit 0 = *-en* kodiert, wodurch wiederum Fehler zustande kamen, z.B. *Dandy* - *Nominativ Plural *-en*. Diese Information ist aber nur für Substantivstämme relevant, die sowohl Singular als auch Plural bilden können. Deshalb konnte auch dieses Problem durch eine if-Abfrage in den Stylesheets beseitigt werden.

- In Z2/2 der Substantive ist bei Dativendung insbesondere die Pluralendung gelegentlich falsch, z.B. *Froschblut*, *Rutil*, *Handumdrehen* - *Plural: *-n* (den Adlern). Dies konnte durch eine dreifache if-Abfrage nach Stammtyp gelöst werden.

- Teils gibt es Angaben, deren Informationsgehalt nicht zu entschlüssen ist, z.B. Z5/3 bei Substantiven: lexikalische Mengenbeschränkung *Wort aus SDW*, Z15/1 bei Substantiven: Fachgebietsmarkierung *TRANSIT*, et.al.

Es zeigt sich, dass alle Einträge mit definierter 0 sehr fehleranfällig sind. Dies ist ein durchgehender Fehler, der mit dem Design der Ursprungsdaten zu tun hat. Teils

lässt sich die 0-Belegung durch andere Informationen gut rekonstruieren, wie oben bereits notiert wurde. Informationen, die aber nicht in Abhängigkeit zu anderen linguistischen Angaben stehen, sind für eine Korrektur nicht greifbar.

Auswertung:

- **fehlerfreie Einträge:** 68 davon fehlten aber in 14 Fällen grundlegende morphosyntaktische Informationen. Bei diesen Fällen handelte es sich mit einer Ausnahme (*schnellfüßig*) um feminine Substantive, die Z1 und/oder Z2 mit 0 0 0 0 kodiert hätten haben müssen.

- **fehlerhafte Einträge:** 32 davon:
  - 8 Einträge mit falsch angegebener Präfixlänge.
  - 7 Einträge waren an einer Stelle in den Z-Zeilen mit einer Zahl versehen, die nicht belegt ist. Letztendlich sind das nicht interpretierbare Stellen, die keinen Fehlerwert haben, da sie nicht stören.
  - 6 Einträge mit falsch angegebenem Dativ.

4 Erweiterung

Da die Evaluation und eine nochmalige Datenenanalyse einige Mängel aufgewiesen haben, wurden weitere Korrekturen vorgenommen, die hier aufgelistet sind:

Verb:

- Eine neue **Z1**-Zeile (Z1 0 0 0 0) wurde eingefügt, wenn der Eintrag ein Infinitiv-, Präteritum- oder Präsensstamm ist und die Z1-Zeile fehlte. Damit wird die Bildung der Präsensformen und die Länge des abtrennbaren Präfixes angegeben.

- Eine neue **Z2**-Zeile (Z2 0 0 n n) wurde eingefügt, wenn der Eintrag ein Präteritum- oder Infinitivstamm ist und die Z2-Zeile fehlte. Damit wird die Bildung des Präteritums und die Endung des Infinitivs angegeben. Die Angabe *n* bedeutet, dass diese Stellen nicht interpretiert werden sollen. (Diese Angabe könnte man auch für andere nicht definierte bzw. sinnvoll belegte Stellen verwenden.)

- Eine neue **Z3**-Zeile (Z3 0 0 0 0) wurde eingefügt, wenn diese Zeile fehlte. Damit wird der persönliche/unpersönliche Gebrauch, die Reflexivität und Perfektbildung angegeben.

- Eine neue **Z5**-Zeile (Z5 0 0 0 0) wurde eingefügt, wenn diese Zeile fehlte. Damit wird angegeben, dass beim attributiven Gebrauch das Partizip II passivisch ist.

- Eine neue **Z12**-Zeile (Z12 0 0 0 0) wurde eingefügt, wenn diese Zeile fehlte. Damit wird angegeben, dass der Stamm zum Paradigma eines Vollverbs gehört.

Substantiv:

- Eine neue **Z1**- und **Z2**-Zeile (Z1 0 0 0 0, Z2 0 0 0 0) wurden bei folgenden femininen Ableitungssuffixen eingefügt, sofern diese Zeilen nicht vorhanden waren: *-ANZ, -ENS, -ENZ, -IK, -ION, -HEIT, -KEIT, -LEI, -REI, -SCHAFT, -TAET, -UNG*. Mit Z1-Zeile wird die feminine Genusmarkierung erschlossen und die Pluralendung mit *-en* angegeben. Mit Z2-Zeile werden

Angaben zu Kasus Singular gemacht (Genitiv, Dativ, Akkusativ ohne Endung) und der Stamm als Singular-/Pluralstamm markiert.

- Eine neue **Z6**-Zeile wurde, wenn nicht vorhanden, eingefügt, bei Lemmata auf *-ION, -HEIT, -KEIT, -SCHAFT, -TAET, -TUM, -UNG* mit folgender Belegung: (Z6 0 0 16 0). Damit wird die Fuge bei Wortzusammensetzungen mit *s* markiert, z.B. *Armut**s**zeugnis*.
  Endet ein feminines Substantiv auf -e wird (Z6 0 0 2 0) eingefügt. Damit wird die Fuge bei Wortzusammensetzungen mit *en* markiert, z.B. *Sonne**n**brandt*.

Adjektiv:

- Eine neue **Z1**-Zeile (Z1 0 0 0 0) wurde eingefügt, wenn diese Zeile fehlte. Damit wird die semantische Markierung des Adjektivs angezeigt.
- Eine neue **Z3**-Zeile (Z3 0 0 0 0) wurde eingefügt, wenn diese Zeile fehlte. Damit wird Flektierbarkeit angezeigt.
- Eine neue **Z4**-Zeile (Z4 0 0 0 0) wurde eingefügt, wenn diese Zeile fehlte. Damit wird der attributive Gebrauch angezeigt.