Noah Bubenhofer, Sandra Hansen-Morath, Marek Konopka

Korpusbasierte Exploration der Variation der nominalen Genitivmarkierung

Abstract: The variation of the strong genitive marker of the singular noun has been treated by diverse accounts. Still there is a consensus that it is to a large extent systematic but can be approached appropriately only if many heterogeneous factors are taken into account. Over thirty variables influencing this variation have been proposed. However, it is actually unclear how effective they can be, and above all, how they interact. In this paper, the potential influencing variables are evaluated statistically in a machine learning approach and modelled in decision trees in order to predict the genitive marking variants. Working with decision trees based exclusively on statistically significant data enables us to determine what combination of factors is decisive in the choice of a marking variant of a given noun. Consequently the variation factors can be assessed with respect to their explanatory power for corpus data and put in a hierarchized order.

DOI 10.1515/zgl-2014-0024

Dr. Noah Bubenhofer: Dresden Center for Digital Linguistics, Technische Universität Dresden, Helmholtzstr. 10, D-01062 Dresden, E-Mail: noah.bubenhofer@tu-dresden.de
Sandra Hansen-Morath: Institut für Deutsche Sprache, Abteilung Grammatik, R5, 6–13, D-68161 Mannheim, E-Mail: hansen@ids-mannheim.de
Dr. Marek Konopka: Institut für Deutsche Sprache, Abteilung Grammatik, R5, 6–13, D-68161 Mannheim, E-Mail: konopka@ids-mannheim.de

1 Einleitung

Die Variation der starken Markierung des Nomens im Genitiv Singular (im Weiteren: ‚Markierungsvariation') ist in ihrem formalen Kern gut greifbar: Die Nomina können vor allem mithilfe der Endungen *-es* oder *-s* markiert werden, sie können aber auch unmarkiert bleiben, und alle diese Möglichkeiten können uns sogar bei ein und demselben Nomen begegnen:

(1) Kelly Makin führte die Regie des **Films**. (St. Galler Tagblatt, 04.11.1999)
(2) Natürlich nennen wir Ihren Namen als Quelle im Abspann des **Filmes**. (Hannoversche Allgemeine, 22.03.2008)
(3) Vin Diesel (40) liegt am Set des **Film** „Babylon A.D." weit hinter dem Drehplan. (Braunschweiger Zeitung, 24.07.2007)

Wenige Nomentypen können sich der Variation entziehen. Die übrigen schwanken zwischen mindestens zwei Varianten, und es stellt sich die Frage, was über die Variantenwahl entscheidet. Die Antworten darauf fallen sehr heterogen aus. Nichtsdestoweniger scheint Einigkeit über die weitgehende Systematizität der Markierungsvariation zu herrschen. Die vielen verschiedenen Postulate von Einflussfaktoren aus der bisherigen Forschung gehen in diese Untersuchung als Hypothesen ein, die korpuslinguistisch zu evaluieren sind. Im Folgenden gilt es somit:

– zu überprüfen, inwiefern die bisher postulierten Einflussfaktoren zur Modellierung der Markierungsvariation als konsistentes System beitragen können,
– weitere Hypothesen zu Einflussfaktoren, ihrer Gewichtung und ihrem Zusammenwirken zu generieren.

In diesem Beitrag[1] wird vorgeschlagen, bei der Auffindung relevanter Einflussfaktoren der Markierungsvariation von einem maschinellen Lernverfahren Gebrauch zu machen – das Ziel ist also ein datengeleitet gewonnenes probabilistisches Modell der Genitivvariation. Um daraus einen linguistischen Nutzen zu ziehen und das komplexe Zusammenspiel verschiedener Faktoren interpretierbar

1 Die vorliegende Studie wurde im Rahmen des Projekts „Korpusgrammatik" (vgl. http://www.ids-mannheim.de/gra/projekte/korpusgrammatik.html) am Institut für Deutsche Sprache in Mannheim (IDS) durchgeführt. Sie stellt die erste Untersuchungsphase einer größer angelegten Pilotstudie zur Markierungsvariation dar (vgl. Fuß/Konopka 2014).

zu machen, wird das statistische Modell als Entscheidungsbaum visualisiert. Die korpuslinguistische Grundlage besteht aus dem Deutschen Referenzkorpus (DeReKo), das in der hier benutzten Version (Institut für Deutsche Sprache 2011a) gemäß Connexor[2]-Analyse ca. 4,3 Milliarden Token[3] und 250 Millionen Sätze umfasste. Für diese Studie sind aus dem Korpus eigens Genitivformen maschinell annotiert und potenzielle Einflussfaktoren klassifiziert worden.

Der Beitrag beschreibt das methodische Vorgehen und präsentiert auch die ersten linguistischen Folgerungen: Zunächst werden die in der Forschung postulierten Faktoren diskutiert und ausgewählte Darstellungen der Markierungswahl in sog. Entscheidungsbäume „übersetzt", um angelegte Abhängigkeiten zwischen möglichen Einflussfaktoren transparenter zu machen (Abschnitt 2). Danach werden das Inventar und die Distribution der Genitivmarkierungen einerseits und der Genitivnomina andererseits vorgestellt (Abschnitt 3). Anschließend wird die mit Rücksicht auf die Distributionsvorgaben und postulierten Faktoren konzipierte Extraktion der Genitivnomina aus den morphosyntaktisch annotierten Korpora beschrieben. In einem weiteren Schritt wird präsentiert, wie für die Extraktionsdaten im Prozess maschinellen Lernens Entscheidungsbäume erzeugt werden, mit denen das Zusammenwirken von Einflussfaktoren visualisiert und statistisch fundiert beurteilt werden kann (Abschnitt 4). Daran schließen sich Überlegungen an, wie die automatisch erzeugten Entscheidungsbäume linguistisch zu interpretieren sind (Abschnitt 5). Der Beitrag endet mit einem Fazit einschließlich einer ersten Beurteilung der angewandten Methodik aus statistischer und linguistischer Sicht (Abschnitt 6).

2 Zum linguistischen Forschungsstand

Die Forschung zur Variation der starken Genitivmarkierung ist sehr umfangreich. Eine exhaustive Besprechung des linguistischen Forschungsstands ist nicht Aufgabe dieses Beitrags, zumal eine ausführlichere Darstellung in Fuß/Konopka (2014) zu finden ist. Es ist in diesem Kontext auch auf umfassende Behandlungen der Markierungsvariation zu verweisen wie Fehringer (2011), Szczepaniak (2010), Duden (2009), Duden (2007), Pfeffer/Morrison (1979, 1984) und Appel (1941). Im Folgenden wird nur eine allgemeine Charakterisierung der Gesamtmenge postulierter Einflussfaktoren der Markierungsvariation (Ab-

2 Eines der Tagging-Werkzeuge, mit denen DeReKo morphosyntaktisch annotiert ist (vgl. http://www.connexor.com).

3 Connexor-Token sind sowohl Wörter als auch Satzzeichen.

schnitt 2.1) sowie eine exemplarische Darstellung der Auffassungen zu Hierarchien und Abhängigkeiten unter diesen Faktoren (Abschnitt 2.2) vorgenommen. Der Fokus liegt damit auf Variationsfaktoren des entkontextualisierten Nomens; denkbar wäre auch der Einbezug von Kontextbedingungen wie Varietät, Register etc. Die im folgenden referierte Forschung ignoriert solche Faktoren, in unser Modell sind jedoch die Faktoren Publikationsdatum, Land, Register, Domäne und Region mit eingeflossen (vgl. Tabelle 2).

2.1 Variationsfaktoren

Um die Ausgangslage unserer Untersuchung zu skizzieren, werden in diesem Abschnitt vier umfassendere Darstellungen als „Quellen“ herangezogen: die Dudengrammatik (Duden 2009, 195ff.), der Zweifelsfälle-Duden (Duden 2007, v. a. 369ff.) sowie zwei neuere ausdrücklich korpusbezogene Studien von Szczepaniak (2010) und Fehringer (2011). Alle diese Quellen streben ähnlich wie unsere Studie eine mehr oder weniger ganzheitliche Beschreibung der Markierungsvariation an. Das heißt zum einen, dass sie ausführlich den Grundwortschatz behandeln und nicht nur Sonderwortschatzbereiche fokussieren wie Eigennamen oder Fremdwörter, und zum anderen, dass sie mehrere Markierungsvarianten, darunter die Hauptvarianten *-s* und *-es*, berücksichtigen und sich nicht einseitig auf eine Variante bzw. ein Phänomen wie z. B. die Weglassung der Markierung beschränken. Sie werden hier als repräsentativ für den Wissensstand über die Variationsfaktoren zu Beginn unserer Untersuchung herangezogen. Dieser erweckt, wie im Folgenden gezeigt wird, nicht unbedingt einen homogenen Eindruck.

Die in den vier Quellen thematisierten Faktoren der Genitivmarkierungsvariation sind im Wesentlichen in Tabelle 1 enthalten. Die meisten Faktoren wurden in mehreren Quellen gefunden (siehe dritte Spalte). Ihre Anordnung in Tabelle 1 ist unabhängig von den Quellen und dient nur der Übersichtlichkeit.

Tab. 1. Mutmaßliche Faktoren der Genitivmarkierungsvariation

Faktorenbereich/ Faktor (nummeriert)	Ausprägungen (A), bzw. bei binären Faktoren beispielhafte Realisierungen der positiven Ausprägung (R)[4]	Quelle[5]
Auslaut		
1. Letztlauttyp	A: Konsonant, Vokal, Diphthong	D4, D9, F
bei konsonantischem Auslaut		
2. *s*-Auslaut	R: *-s, -ss, -ß, -tz, -z, -x, -ce, -ts, -chs*	D4, D9, F, S
3. *sch*-Auslaut	R: *-sch*, *-ch*	D4, D9
4. Wortausgang /st/	R: *-st, -zt*	D4, D9
5. Sonoritätshierarchie	A: Liquid, Nasal, Frikativ, Affrikate, Plosiv	F, S
6. Konsonantenhäufung	R: *Herd*	D4, D9, F, S
Endreim/Endsilbe (mit konsonantischem Auslaut)		
7. Vokallänge	A: lang, Diphthong, kurz	D4, D9, F, S
8. spezielle(r) Endreim Endsilbe	R: *-en, -em, -el, -ler, -ner, -end*	D4, D9, S
9. spezielles Suffix	R: *-ig, -ich, -ing, -ling, -chen, -lein* etc.	D4, D9, F
Betonung		
10. Endsilbenbetonung	R: *Vertrag*	D4, D9, F, S
Wortbildung		
11. Silbenanzahl	A: 1, 2, 3 etc.	D4, D9, F, S
12. Komplexität	A: Simplex, Präfigierung, Suffigierung, Kompositum	D9, F, S
bei Präfigierung		
13. Präfixbetonung	R: *Vortrag*	S
bei Suffigierung		
14. Suffixbetonung (Nebenakzent)	R: *Reichtum*	S

4 Mit Realisierungen sind hier Erscheinungsformen einer Faktorausprägung gemeint. So hat z. B. der Faktor ‚*s*-Auslaut' zwei Ausprägungen, auf die man mit ‚ja' oder ‚nein' bzw. numerisch mit ‚1' und ‚0' referieren kann. Die erste, positive Ausprägung des Faktors hat in Texten wiederum mehrere Realisierungen. Deren Beispiele werden in Tabelle 1 aufgeführt und mit „R" gekennzeichnet.

5 Erklärung der Abkürzungen: D4 = Duden (2009), D9 = Duden (2007), F = Fehringer (2011), S = Szczepaniak (2010). Die aufgeführten Quellen thematisieren nicht immer alle Ausprägungen des Faktors bzw. alle Realisierungen einer Ausprägung.

Faktorenbereich/ Faktor (nummeriert)	Ausprägungen (A), bzw. bei binären Faktoren beispielhafte Realisierungen der positiven Ausprägung (R)	Quelle
bei Kompositum		
15. Fuge	A: *-s, -es* und andere Fugen	D9
16. semantische Transparenz	A: transparent, opak	F
17. Lexikalisierung	A: stark, niedrig	S
18. Präferenz des Grundworts	A: *-s, -es* etc.	F
Nomenklasse:		
19. Fremdwort	R: *Hit*	D4, D9, S
20. Eigenname (allgemein)	R: *Rhein*	D4, D9, F
21. Personenname	R: *Werther*	D4, D9, F
22. geografischer Name	R: *Irak*	D4, D9, F
23. Fachwort	R.: *Biedermeier*	D9
24. Titel o. Ä.	R.: *Kaiser*	D9
25. Abkürzung (v. a. Akronym)	R.: *AKW*	D4, F
26. Konversion	R.: *Jemand*	D4, D9
27. Appellativum aus Eigenname, Monatsname, Produktbezeichnung o. Ä.	R.: *Opel*	D4, D9
28. starkes Nomen auf *-en*	R: *Rahmen*	D4, D9
29. Genus	A: Maskulin, Neutrum	S
Frequenz		
30. Häufigkeit des Nomens	A: hoch, niedrig	D4, F, S
31. Quotient Häufigkeit des komplexen Wortes / Häufigkeit des Grundworts	A: (z. B.) ≥ 1, < 1	F
Syntax		
31. Vorhandensein eines Artikels auf *-s*	R: *des*	D4
33. Position des Genitivattributs	A: Voranstellung, Nachstellung	D4, D9
34. Vorangehen eines Titels	R: [*Kaiser*] *Karl*	D9
35. mehrgliedriger Eigenname	R: *Walther von der Vogelweide*	D9

Faktorenbereich/ Faktor (nummeriert)	Ausprägungen (A), bzw. bei binären Faktoren beispielhafte Realisierungen der positiven Ausprägung (R)	Quelle
36. feste Verbindung	R: *Auditorium maximum*	D4
37. Paarformel	R: *Grund und Boden*	D4
38. formelhafte Wendung	R: *Manns* [*genug*]	D9

Auffällig ist die hohe Anzahl der Faktoren. Keiner führt alleine zur Ausnahmslosigkeit einer Markierungsvariante. Wohl aber haben einige wenige Kombinationen von Faktoren nahezu ausschließlich eine Variante zur Folge. So führen der *s*-Auslaut (Faktor 2) beim Grundwortschatz (also beim Nichtvorliegen der positiven Ausprägung von Faktor 19–27) regulär nur zur Endung -(*s*)*es* (z. B. *Hauses*, *Erlebnisses*) und der Endreim mit einem Schwa (Faktor 8) – ebenfalls beim Grundwortschatz – zur Endung -*s* (z. B. *Segels*). Für andere Faktoren bzw. Faktorenkombinationen werden nur Tendenzen postuliert, die im Übergewicht einer bestimmten Endung resultieren. So sollen beim Grundwortschatz z. B. eine Konsonantengruppe am Nomenende (Faktor 6) zur Bevorzugung von -*es* (z. B. *Herdes*, *Volkes*), ein Auslaut auf Nasal oder Liquid (Faktor 5) aber zur Bevorzugung von -*s* (z. B. *Qualms*, *Kerls*) führen. Wie komplexere Kombinationen von Faktoren wirken bzw. wie die Faktoren untereinander zu hierarchisieren sind, bleibt erst einmal unklar.

Die Darstellungen fallen in den einzelnen Quellen teilweise unterschiedlich aus. Die Dudengrammatik gewichtet besonders stark lautliche Faktoren und die Zugehörigkeit zu einer bestimmten Nomenklasse und gibt bei der Endungswahl teilweise getrennte Regeln für Nomina des Grundwortschatzes und für Fremdwörter an. Unter den Faktoren, die zur Weglassung der Genitivendung führen können, rücken dann morphologische und syntaktische Bedingungen in den Vordergrund sowie Wortschatzbereiche wie Eigennamen, Wörter auf -*en*, Paarformeln, mehrteilige feste Verbindungen oder Farbbezeichnungen. Der Zweifelsfälle-Duden betont wiederum, dass die Variation zwischen -*s* und -*es* nur teilweise geregelt ist, und strukturiert die diesbezügliche Darstellung nach festem und schwankendem Gebrauch der Endungen. Probleme, die mit der Weglassung der Endung verbunden sind, werden im Kontext einschlägiger Nomenklassen beschrieben wie Personennamen, geografische Namen, Monatsnamen, Fachwörter, Fremdwörter. Derartige Nomenklassen werden bei Szczepaniak und Fehringer lediglich gestreift, denn beide fokussieren die Variation zwischen den häufigsten Markierungen, -*s* und -*es*, und diese ist für den Grundwortschatz charakteristisch. Obwohl beide Autorinnen viele Faktoren teilen, legen sie in ihren Studien deutlich verschiedene Schwerpunkte. Szczepaniak präferiert stark die Parameter

des „phonologischen Wortes“, zu denen vor allem der Wortbildungstyp sowie die Faktoren ‚Silbenanzahl‘, ‚Betonung‘, ‚Sonoritätshierarchie‘[6] und ‚Konsonantenhäufung‘ gehören. Solche morphologischen und phonologischen Faktoren zieht zwar auch Fehringer heran, sie betont aber in erster Linie die Rolle des Faktors ‚(Token-)Frequenz‘. Während Szczepaniaks Grundidee ist, dass die Wahl von *-s* desto wahrscheinlicher wird, je phonologisch komplexer das Nomen ist, argumentiert Fehringer, dass die Markierungswahl bei Simplizia durch die Frequenz des Nomens determiniert ist und dass diese auch für komplexe Wörter, die auf diesem Nomen als Grundwort aufbauen, ausschlaggebend ist, – sofern sie semantisch transparent erscheinen. Entscheidend sei, dass die bekannten phonologischen Faktoren nur im Bereich frequenter Nomina wirksam seien. Welche Faktoren wiederum die Markierungswahl bei den seltenen Nomina beeinflussen, sei unklar – jedenfalls zeigten seltene Nomina keine deutlichen Präferenzen für eine Variante.

Außer den oben genannten grammatischen Faktoren und der Frequenz des Genitivnomens werden in den Quellen sporadisch rhythmische und stilistische Parameter angesprochen (Duden 2007, 371; Pfeffer/Morison 1984, 18), jedoch nicht präzisiert. Darüber hinaus beschäftigt sich Szczepaniak (2010) mit der historischen Entwicklung der Genitivendungen, die auf eine Abnahme von *-es* und eine Zunahme von *-s* hinausläuft. Weitere denkbare situations- oder sprecherabhängige Parameter werden in den Quellen nicht fokussiert.

2.2 Faktorendarstellungen als Entscheidungsbäume

Nicht zuletzt, weil die Darstellungen in der Fachliteratur unterschiedlich strukturiert sind, entsteht oft der Eindruck, dass die Faktoren jeweils anders hierarchisiert werden. Um angelegte Abhängigkeiten zwischen Faktoren transparenter und die verschiedenen Darstellungen miteinander vergleichbarer zu machen, wurden die Darstellungen in binäre Entscheidungsbäume „übersetzt“. Im Folgenden werden exemplarisch zwei Teilbäume präsentiert, die den (stärker varianten) Bereich der auf Konsonanten endenden Nomina des Grundwortschatzes betreffen.[7] In diesem Bereich zeigen die Darstellungen übrigens die deutlichsten Unterschiede.

6 Szczepaniak spricht eigentlich von einer Hierarchie konsonantischer Stärke („consonantal strength“), die als Konversion der Sonoritätshierarchie aufgefasst werden kann.

7 Ausgeschlossen bleiben hier Nomina auf *s*-Laut, *sch*-Laut, die *st*-Gruppe, die sehr stark zu *-es* tendieren, sowie Nomina mit speziellen Endreimen bzw. Suffixen, die fast ausnahmslos *-s* zu sich nehmen. Für eine ausführlichere Diskussion der Forschungsbeiträge, vgl. Fuß/Konopka 2014.

Ein relativ allgemeines Bild ergibt sich aus den Ausführungen der Dudengrammatik (siehe Abbildung 1), wo dazu passend der Hinweis erscheint: „Feinere Regeln lassen sich teilweise nur schwer geben“ (Duden 2009, 198).

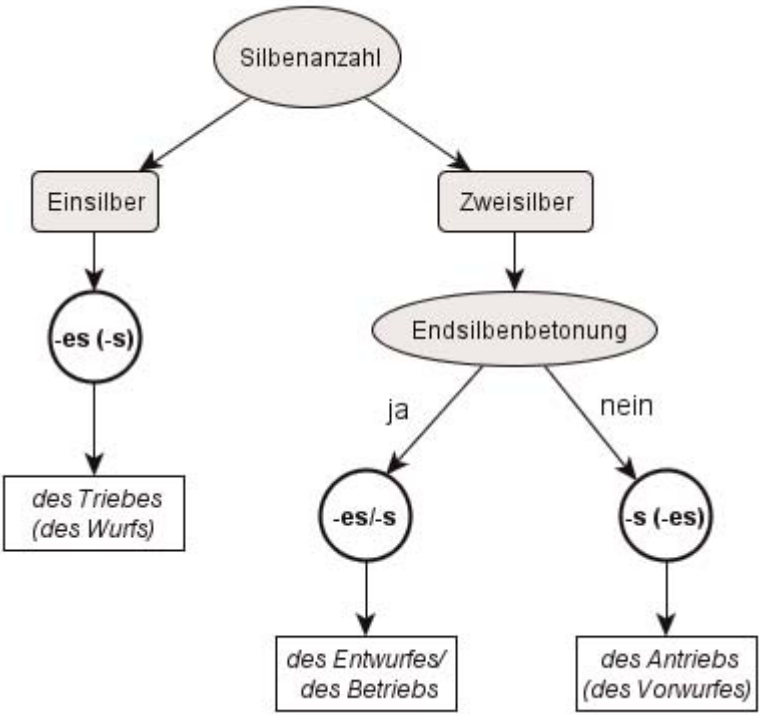

Abb. 1: Entscheidungsbaum: Markierungsvariation für den Bereich der auf Konsonanten endenden Nomina des Grundwortschatzes nach Duden (2007)

Mit der Darstellung der Dudengrammatik sind die Ausführungen Fehringers (2011) prinzipiell kompatibel. Fehringer stellt die Silbenanzahl und die Betonung an den Ausgangspunkt ihrer Überlegungen. Dann aber führt sie neue Faktoren ein. Sie geht auch viel weiter, was die Detailliertheit der Festlegungen angeht (vgl. Abbildung 2).

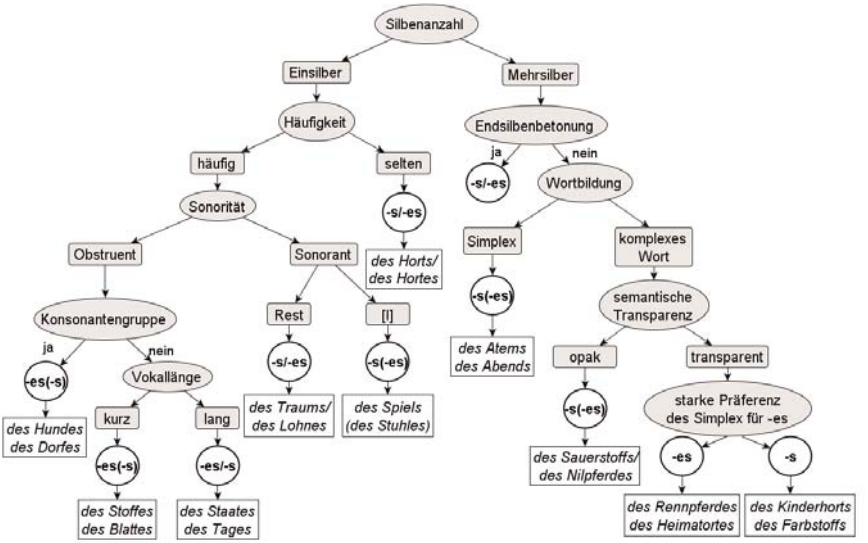

Abb. 2: Entscheidungsbaum: Markierungsvariation für den Bereich der auf Konsonanten endenden Nomina des Grundwortschatzes nach Fehringer (2011)

Ihre Darstellung enthält in Bezug auf Einsilber zwei Besonderheiten:

- Zum zentralen Faktor wird die Frequenz der Genitivformen eines Nomens erhoben.
- Es wird die Ansicht Pfeffers/Morrisons (1979, 1984) vertreten, dass kurze Vokale vor Konsonanten am Nomenende eher zur Endung *-es* führen als lange Vokale in der gleichen Position.[8]

Bei Mehrsilbern behandelt Fehringer nur komplexe Wörter ohne Endsilbenbetonung. Hier setzt sie vor allem auf ‚semantische Transparenz', die den Einfluss des Grundwortes auf das komplexe Wort ermöglichen und so u. U. die bei Mehrsilbern allgemein wirksame Tendenz zu *-s* überschreiben könne.

Zu betonen bleibt: Während die Dudengrammatik bei allgemeineren Regeln bleibt, geht Fehringer ins Detail. Sie gewichtet dabei besonders stark den noch nicht etablierten Faktor ‚Frequenz' und bringt für komplexe Wörter den Faktor ‚semantische Transparenz' ins Spiel.

8 Die umgekehrte Position vertritt Szczepaniak (2010, 116f.). Ansonsten wird die Vokallänge vor abschließendem Konsonanten kaum thematisiert.

2.3 Zwischenfazit Analyse

Ein allgemein anerkanntes Modell der Variation der Genitivmarkierung existiert höchstens in sehr groben Zügen. Die Darstellungen in der vorgestellten Literatur haben zwar Kernfaktoren wie ‚Silbenanzahl' oder ‚Betonung' gemeinsam, aber bereits diese können in das Variationsmodell spezifisch eingebaut werden, denn die Darstellungen reichen von einer vorsichtigen Postulierung weniger Parameter (Duden 2009) bis hin zur Erörterung zahlreicher Einflussfaktoren, die im Variationsmodell auch an mehreren Stellen wirksam werden können. Unser Wissensdefizit kann durch folgende Fragen umrissen werden:

- Wie relevant sind die einzelnen Faktoren?
- Kann die große Menge an heterogenen Faktoren ein konsistentes Modell der Markierungsvariation ergeben?
- Welche Hierarchien sind zwischen den einzelnen Faktoren der Markierungsvariation festzustellen?

Was einzelne Faktoren angeht, so ist besonders interessant, ob sich die von Fehringer postulierte Schlüsselposition des Faktors ‚Frequenz' bestätigen lässt. Außer der Frage nach seiner Relevanz muss also geklärt werden,

- ob bzw. wie ‚Frequenz' in einem konsistenten Variationsmodell mit andersartigen Faktoren zusammengeführt werden kann.

Um bei der Suche nach Antworten auf diese Fragen voranzukommen, bedarf es einer Extraktion der Genitive aus dem Korpus. Dabei sind gefragt einerseits die bestmögliche Präzision bei möglichst vollständiger Erfassung der relevanten Genitivfälle und andererseits eine genaue Erfassung von Ausprägungen der postulierten Faktoren, um deren Wirkung quantifizierbar und adäquat darstellbar zu machen. Hierzu ist es nötig, die Distribution der verschiedenen Genitivnomina zu analysieren und das Ergebnis in die Extraktion zu integrieren.

3 Genitivmarkierung und Genitivnomen – Inventar und Distribution

Will man Nomina im Genitiv Singular in einem sehr umfangreichen Korpus zuverlässig erfassen, kann man dies in vertretbarer Zeit nur mithilfe eines automatisierten Suchverfahrens erreichen. Hier werden die notwendigen linguistischen Vorannahmen diskutiert, auf denen das in Abschnitt 4 zu beschreibende Extraktionsverfahren basiert.

Die kanonischen Formen der starken Genitivmarkierung sind die Endungen *-es* und *-s*, und ihre eigentlichen Domänen werden gebildet durch stark flektierende Maskulina und Neutra (z. B. [*eines*] *Haus**es***, [*des*] *Vater**s***)[9] sowie artikellose Eigennamen (z. B. [*die Figur*] *David**s***, *Marta**s*** [*Mann*], [*außerhalb*] *Europa**s***).[10] Die Genitivmarkierung *-s* erscheint außerdem regulär an gemischt flektierenden Maskulina und Neutra, und zwar in Kombination mit der schwachen Markierung *-(e)n* (z. B. [*des*] *Friede**ns,*** [*des*] *Herz**ens***). Sie kann zuweilen auch bei Nomina beobachtet werden, die standardmäßig schwach flektieren (z. B. [*des*] *Bär**s***, [*des*] *Bären**s*** statt [*des*] *Bären*).

Die starke Genitivmarkierung hat spezifisch schriftliche Ausprägungen wie in *des Ergebnis**ses***, *Hingis**'*** [*Halbfinal*] oder *Gino'**s*** [*Kunstcafé*], und die Reihe der Genitivnomenvarianten wird erst dann vollständig, wenn man auch das gänzliche Unterbleiben einer Markierung mit berücksichtigt wie in [*des kleinen*] *Peter*, [*des alten*] *Rom*, [*des*] *Barock* oder (nicht normgerecht) [*eines*] *Garten*. Alles in allem gibt es also für im Genitiv stehende Nomina in der geschriebenen Sprache acht Markierungsmöglichkeiten: *-es*, *-s*, *-ens*, *-ns*, *-ses*, [nach *s*, *x* u. ä.] *-'*, *-'s*, *-Ø*.[11]

Damit liegen die allgemeinen Rahmenbedingungen für die Extraktion der Genitivnomina aus dem Korpus vor. Dass diese Vorgaben für die Extraktion aber alleine noch nicht ausreichend sind, liegt auf der Hand: Nomina, die bestimmten Deklinationstypen bzw. den Eigennamen zuzurechnen sind, in einem Korpus mithilfe eines automatischen Werkzeugs auffindbar zu machen, also ohne direkten Rückgriff auf unser Sprach- und Weltwissen, ist keine triviale Aufgabe. In unserem Fall kommt erschwerend hinzu, dass die Nomina in genau nur einem Kasus stehen dürfen, dem Genitiv. Nicht zuletzt, weil der Xerox-Tagger[12] im DeReKo die bei einer Wortform möglichen Kasus nicht disambiguiert (hierzu Abschnitt 4.1), muss die Taggeranalyse bei der Extraktion durch eine zusätzliche, über die Wortgrenzen hinausgehende Prüfung von maschinell erkennbaren Textmerkmalen ergänzt werden, die typischerweise das Vorkommen eines Genitivs

9 Wenn nicht anders vermerkt, stammen alle in diesem Beitrag aufgeführten Beispiele und Belege für Genitive aus DeReKo.

10 Werden solche Eigennamen im Maskulinum und Neutrum in bestimmten Konstruktionen mit einem Artikel gebraucht, tendieren sie stark zur Endungslosigkeit (vgl. Duden 2009, 201), z. B.: *des jungen Beethoven*.

11 Die Endungen *-ens, -ns* können als Kombinationen aus einem Nicht-Nominativ-Kennzeichen *-en*/*-n* und einem Genitivkennzeichen *-s* analysiert werden. Hier werden sie aber als Ganzes betrachtet, weil bei der automatischen Suche nach Genitivnomina im Korpus die Kandidaten mit Lemmata in der Nominativform abgeglichen werden.

12 Die anderen Tagger, mit denen DeReKo morphosyntaktisch annotiert ist, unterschieden keine Kasus. Vgl. Abschnitt 4.1.

begleiten. Solche oberflächennahen Merkmale stehen im Fokus bei der folgenden Erörterung der Distribution der Genitivnomina.

3.1 Die Genitivphrase

Der erste Blick soll der Umgebung des Genitivnomens gelten. Da die Genitivmarkierung am Nomen prinzipiell fehlen kann, wird das Genitivnomen selbst erst im zweiten Zugriff herangezogen.

Dem Genitivnomen wird meist ein Artikel im Genitiv (im Weiteren: ‚Genitivartikel') oder ein entsprechend flektiertes Adjektiv[13] vorangestellt. Anstelle von Letzterem kann in allen Positionen eine Ortsnamenableitung auf *-er* (z. B. *Berliner*) erscheinen (was nicht mehr gesondert vermerkt wird). Es entstehen dabei folgende Nominalphrasentypen, die hier als Genitivphrasen bezeichnet werden:

> **Genitivphrase I: Genitivartikel** + **Genitivnomen** (*des Landes*, *unseres Teamkapitäns*, *des Barock*)
> **Genitivphrase II: Adjektiv auf *-en*** + **Genitivnomen** ([*Akt*] *rohen Mutwillens*, [*Tonnen*] *Berliner Mülls*, [*wegen*] *leichten Unwohlseins*)

Im zweiten Fall ist die Markierung des Nomens als Genitiv dringend geboten, da ansonsten die Nominalphrase nicht als Genitiv erkennbar ist.

Der Genitivartikel kann auch weiter links vom Genitiv entfernt stehen. In dem Fall muss ein Adjektiv auf *-en* direkt vor dem Genitivnomen erscheinen (aber nicht unbedingt direkt auf den Genitivartikel folgen). Die beiden bisherigen Genitivnomenbegleiter werden miteinander kombiniert:

> **Genitivphrase III: Genitivartikel** + **[... +] Adjektiv auf *-en*** + **Genitivnomen** (*eines anderen Vorfalls*, *des* [*am Wochenende*] *beerdigten Kapitäns*)

Schließlich kann das Nomen auch allein die Genitivphrase bilden, d. h. ohne einen Genitivartikel und ohne ein entsprechend flektiertes Adjektiv erscheinen, was bekanntermaßen typisch für Eigennamen ist. Zumindest in der Schrift ist hier eine Genitivmarkierung notwendig, um den Kasus erkennbar zu machen:

> **Genitivphrase IV**: (selbständiges) **Genitivnomen** (*Bambergs* [*Altstadt*], *Shakespeare's*[14] [*Alterswerk*], *Walter Gross'* [*lyrische Sendung*], [*Gnade*] *Gottes*])

---

13 Im Text u. a. auch als Zahl mit Punkt für ein Ordinalzahl-Adjektiv, z. B. *des 18. Jahrhunderts*. Auch adjektivisch flektierte Pronominaladjektive (dazu Wiese 2009, 167) wie *alle*, *wenige* oder *andere* sind hier mit gemeint.

14 In den Daten vertretene Variante zu *Shakespeares*.

Die oben genannten Möglichkeiten des Genitivphrasenaufbaus können wie folgt zusammengefasst werden:

Genitivphrasen I – IV: [Genitivartikel +] [[... +] Adjektiv auf *-en* +] Genitivnomen

3.2 Die übergeordnete Konstruktion

Syntaktisch gesehen können alle Genitivphrasen von einer Präposition, einem Adjektiv oder einem Verb regiert wie auch adverbial oder adnominal verwendet werden. In der adnominalen – offenkundig häufigsten – Verwendung ist die Genitivphrase Teil einer umfassenderen Nominalphrase, in der sie nach ihrem nominalen Bezugselement oder davor steht. Bei der Nachstellung der **adnominalen Genitivphrase** folgt diese typischerweise direkt[15] auf das Bezugsnomen:

(1) **Nomen + Genitivphrase** ([*der*] *Ruf des Landes*, [*ein*] *Akt königlichen Mutwillens*, [*eine*] *Verwandte des am Wochenende beerdigten Kapitäns*, [*die*] *Gnade Gottes*)

Bei der **Voranstellung** der Genitivphrase folgt auf diese direkt ein anderes Nomen oder dessen Attribut, das ein stark flektiertes Adjektiv ist oder mit einem solchen Adjektiv endet – die Genitivphrase schließt dabei einen Artikel zum folgenden Nomen aus:

(2) **Genitivphrase + Nomen** (*Bambergs Altstadt, Shakespeare's Alterswerk*)

(3) **Genitivphrase + [... +] flektiertes Adjektiv + Nomen** (*Walter Gross' lyrische Sendung, Chaplins* [*ewig*] *scheiternder Tramp*)

Die Distributionen (1)–(3) könnten insbesondere beim Auffinden von selbständigen Genitivnomina während der Extraktion behilflich sein, denn in solchen Fällen fehlen die für die umfangreicheren Genitivphrasen charakteristischen Hinweise auf den Genitiv – der Genitivartikel und das Adjektiv auf *-en*. Selbständige Genitivnomina (meist Eigennamen) sind ihrerseits typisch für die Distributionen (2) und (3).

Mit der adnominalen Verwendung in ihrer Häufigkeit vergleichbar scheint nur die Verwendung des Genitivnomens mit einer regierenden **Präposition** zu

15 Dies ist zwar meistens der Fall, aber natürlich keine notwendige Bedingung, vgl. z. B. elliptische Konstruktionen wie in *Die andere Seite der Einzelkämpfer-Medaille ist freilich **die des** teilweise schon jahrzehntelangen Vertrauensverhältnisses im Team* (Burgenländische Volkszeitung, 28.05.2008, S. 31; „Das ist der Vorteil eines einzigen Arztes").

sein. In dieser Verwendung wie auch in der Verwendung mit einem regierenden **Adjektiv** kann das regierende Element vor oder nach der Genitivphrase stehen: Geht es der Genitivphrase voran, steht es direkt davor, folgt es ihr nach, steht es typischerweise direkt[16] nach dem Genitivnomen.

(4) **Genitivpräposition/Genitivadjektiv + Genitivphrase** (*während des ganzen Mittelalters*, *ledig Gottes* [*und aller Dinge*])

(5) **Genitivphrase + Genitivpräposition/Genitivadjektiv** (*eines politischen Strafverfahrens wegen*, *eines Kapitaldelikts schuldig*)

Wird die Genitivphrase von einem **Verb** regiert, kann dieses ebenfalls vor der Genitivphrase oder danach stehen. In beiden Stellungen ist es öfter nicht adjazent zu der Genitivphrase bzw. dem Genitivnomen:

(6) **Genitivverb + [... +] Genitivphrase:** ([*Er*] *gedachte Melitta's* [*wie einer Toten*], [*Seine ... Leitung*] *beraubt* [*Haydns Musik*] *ihres Charmes*)

(7) **Genitivphrase + [... +] Genitivverb:** ([...] *eines natürlichen Todes stirbt*, [*Nachdem das IOC sechs weitere Sportler*] *des Dopings* [*während der Olympischen Spiele 2008*] *überführt hatte*)

Wird die Genitivphrase von einem adjektivisch flektierten **Partizip** eines Genitivverbs regiert, kann dieses Partizip nur nach der Genitivphrase stehen:

(8) **Genitivphrase + [... +] Genitivpartizip**: [*Der*] *des Dopings überführte* [*Österreicher*]

Die Genitivpräpositionen, -verben, -und -adjektive erweitern somit die Reihe der Indizien für das Vorkommen eines Genitivs. Allerdings sind sie als solche Indizien nur bedingt von Nutzen. Nicht nur dass sie sowohl nach links als auch nach rechts regieren können, sie regieren oft auch andere Kasus bzw. komplexere Konstruktionen, sodass im Umfeld kein von ihnen regiertes Genitivnomen vorkommen muss. Darüber hinaus kann – insbesondere bei diskontinuierlichen Bildungen mit einem Genitivverb – zwischen dem den Genitiv regierenden Element und dem Genitivnomen eine größere Lücke entstehen, deren Füllung kaum maschinell kontrollierbar erscheint.

Die noch verbliebene, **adverbiale Verwendung** scheint schließlich außerhalb der Genitivphrase keine maschinell verwertbaren Anhaltspunkte für das Vorliegen eines Genitivs zu bieten: Die Genitivphrase vom Typ I oder II (siehe 3.1) modifiziert

16 Relativ seltene Ausnahmen bilden vor allem Fälle, in denen direkt nach der Genitivphrase ein das Adjektiv modifizierendes Element steht, wie in *des Obersteiner „Weihnachtsmarktes" sehr überdrüssig* (Rhein-Zeitung, 15.12.2006; „Angebot war wunderbar").

den gesamten Satz bzw. die gesamte Verbalphrase und setzt in diesen keine einfach systematisierbaren Ausdrücke voraus:

(9) [*Das „Lärmkorsett" ist*] *meines Wissens* [*bis heute eingehalten worden.*]

(10) [*Sie müssen sich*] *schweren Herzens* [*eingestehen, dass die Fahrt über den Julier bei dieser Schnee- und Wetterlage kaum zu bewältigen ist.*]

Bei den adverbialen Genitivphrasen handelt es sich heutzutage bekanntlich nur noch um feste Wortverbindungen (vgl. Duden 2009, 821). Folglich kommen hier auch nur wenige Nomenlexeme infrage.

Zum Schluss sei noch angemerkt, dass es sowohl in der übergeordneten Konstruktion als auch in der Genitivphrase zu oberflächensyntaktischen Abweichungen von den oben konstatierten Regularitäten kommen kann, und zwar vor allem bei Ellipsen, Koordinationen und mehrgliedrigen Phrasen (z. B. *Drohungen wie die des Ministers*; *des langjährigen Landes- und Bundesministers*; *des Herrn Professor Gottscheds*). Sie legen der automatischen Genitivextraktion die eine oder andere Beschränkung auf.

3.3 Nomenklassen

In den vorangegangenen Abschnitten wurde bereits deutlich: Es gibt Korrelationen zwischen bestimmten Nomenklassen und bestimmten Distributionen des Genitivnomens und der Genitivmarkierung. Unter den Nomenklassen fallen zunächst **Eigennamen** auf, die charakteristisch für Genitivphrase IV sind (z. B. *Bambergs* [*Altstadt*], vgl. oben 3.1). Sie wird von selbständigen Nomina gebildet, die verstärkt auf die Genitivmarkierung angewiesen sind. Ein besonderes Gewicht bei der Suche nach dem Genitivnomen kommt hier also der Markierung und den Indizien außerhalb der Genitivphrase (vgl. oben 3.2) zu. Sind Eigennamen hingegen Teil der Genitivphrase eines anderen Typs, bleibt der Genitiv am Nomen häufig unmarkiert (z. B. *des* [*historischen*] *Palästina*). In diesem Fall kann es sich bekanntlich nur um Genitivphrasen vom Typ I und III handeln, die einen Artikel enthalten (vgl. oben 3.1). Das Gewicht der Genitiverkennung verschiebt sich in diesen Phrasen auf andere Elemente als das Nomen – vor allem auf den Artikel.

Eine Tendenz zur Weglassung der Genitivmarkierung bei Genitivphrasen vom Typ I und III ist außer bei Eigennamen vielfach auch bei **Fremdwörtern**, **Abkürzungen**, **Konversionen**, **Neologismen** u. Ä. beobachtet worden. Für die optimale Untersuchung der Markierungsvariation sind somit der „Grundwortschatz"[17] und

17 Diese Bezeichnung wurde aus Duden (2007, 196) übernommen.

der „Sonderwortschatz“ zu unterscheiden. Der Grundwortschatz umfasst stark und gemischt flektierende Maskulina und Neutra sowie typischerweise schwach flektierende Nomina, sofern sie gelegentlich eine starke Genitivendung aufweisen. Der Sonderwortschatz setzt sich aus verschiedenen Nomenklassen zusammen wie Eigennamen, Fremdwörtern, Abkürzungen, Konversionen, Neologismen u. Ä., die wiederum jeweils spezifisches Markierungsverhalten zeigen können. Problematisch für die vorliegende Studie erscheint, dass sich der heterogene Sonderwortschatz mithilfe von maschinell greifbaren, formalen Kriterien schwer und in jedem Fall bei Weitem nicht vollständig erfassen und klassifizieren lässt.

3.4 Zwischenfazit Datenextraktion

Bei der Genitivmarkierung muss von acht Varianten (einschließlich der Weglassung der Markierung) ausgegangen werden. Bei der Suche nach Genitivnomina im Korpus sind verschiedene Nomenklassen zu berücksichtigen, bei denen von spezifischen Präferenzen in Bezug auf Distribution und Markierungswahl auszugehen ist.

Die wichtigsten Hinweise auf ein Genitivnomen liegen innerhalb der Genitivphrase mit dem Genitivartikel[18] und einer realisierten Genitivmarkierung vor. Fehlt eines dieser beiden Indizien, rücken einerseits die Zugehörigkeit des Nomens zu einer bestimmten Sonderwortschatzklasse, welche unter gegebenen Umständen die Weglassung des Genitivartikels oder der Genitivmarkierung lizenzieren kann, und andererseits die Gestalt der übergeordneten Konstruktion in den Vordergrund. So ist die Sonderwortschatzklasse ‚Eigenname‘ typisch für selbständige Genitivnomina ohne Genitivartikel und damit auch für adnominale Genitivphrasen, die dem übergeordneten Nomen vorangehen. Im Umkehrschluss könnten also die spezifische Distribution in der übergeordneten Konstruktion (hier: Stellung vor einem anderen, artikellosen Nomen) und die Zugehörigkeit zu einer bestimmten Nomenklasse (hier: ‚Eigenname‘) als Hinweise benutzt werden, die das Fehlen des ansonsten so ausschlaggebenden Artikels bei der Genitivdiagnose aufwiegen könnten.

Sieht man von adnominalen Genitivphrasen ab, so kann die übergeordnete Konstruktion – in dem Fall ein Satz bzw. eine Verbalphrase – nur dadurch Hinweise auf ein Genitivnomen geben, dass sie Ausdrücke (Präpositionen, Ad-

18 Für eine Liste der Genitivartikel, die bei der in Abschnitt 4 beschriebenen Extraktion verwendet wurden, vgl. http://hypermedia.ids-mannheim.de/call/public/korpus.genitivdb.

jektive, Verben) enthält, die den Genitiv regieren können.[19] Allerdings ist der Wert dieser Hinweise nicht allzu hoch einzustufen, da die genannten Ausdrücke oft verschiedene Kasus bzw. Konstruktionen regieren können und ihre Stellung in Bezug auf die Genitivphrase nur schlecht systematisierbar ist.

4 Genitivextraktion und Baummodellierung

Im Folgenden wird gezeigt, wie die mutmaßlichen Variationsfaktoren statistisch ausgewertet werden, um valide Aussagen über die Wahl der Genitivmarkierung machen zu können. Dabei wird überprüft, ob die Wahl der Genitivmarkierung überhaupt an bestimmte Regeln gebunden ist und falls bestimmte Prinzipien nachzuweisen sind, welche Kombinationen der möglichen Einflussfaktoren ausschlaggebend sind.[20] Mit dieser Methodik werden zwei Ziele verfolgt: zum einen sollen Erkenntnisse über die Wahl der Genitivendung gewonnen werden, zum anderen soll die Frage beantwortet werden, ob es mit maschinellen Lernverfahren gelingt, linguistisch sinnvolle Regeln über die Wahl der Genitivmarkierung abzuleiten. Hierzu wird zunächst ein Trainingskorpus bestehend aus nominalen Genitiven erstellt, in dem die verschiedenen Einflussfaktoren maschinell annotiert werden. Basierend auf diesem Korpus werden explorative Methoden des maschinellen Lernens verwendet und ein Modell erzeugt, welches die Genitivmarkierungen in Abhängigkeit verschiedener Einflussfaktoren mit einer hohen Trefferquote vorhersagt und sich in einem Entscheidungsbaum darstellen lässt. Der Baum ermöglicht es, bestehende Hypothesen zu evaluieren sowie neue Hypothesen zu generieren, die in weiteren Studien mithilfe inferenzstatistischer Verfahren getestet werden können.[21]

Wir verorten unser Vorgehen im Paradigma der datengeleiteten Korpusanalysen: Eine Reihe von in der Literatur postulierten Einflussfaktoren wird anhand großer Datenmengen überprüft, allerdings nicht so, dass jeder Faktor isoliert an der Datenrealität getestet wird, sondern indem der datengeleitete Algorithmus des maschinellen Lernens in einem iterativen Prozess aus den in Frage kommenden Faktoren ein Modell erstellt, das die Wahl der Genitivendung mit einem *möglichst* einfachen System *möglichst* weniger Faktoren *möglichst* gut

19 Für die Listen der Genitivpräpositionen, -adjektive und -verben, die bei der in Abschnitt 4 beschriebenen Extraktion zum Einsatz kamen, vgl. http://hypermedia.ids-mannheim.de/db/liesmich.txt.

20 Vgl. hierzu auch Bubenhofer et al. (2013), die mithilfe maschineller Lernverfahren Fugenelemente in nominalen Komposita vorhergesagt haben.

21 Vgl. Fuß/Konopka (2014).

vorausgesagt werden kann. Das Modell gibt jedoch nicht nur die relevanten Faktoren zurück, sondern zeigt auch an, welchen Anteil der Fälle jede modellierte Kombination von Faktoren korrekt voraussagen kann. Ein solches probabilistisches Modell hilft, die Gültigkeit der linguistischen Regeln, die aus dem Modell abgeleitet werden können, einzuschätzen und nennt gleichzeitig die im Korpus belegten Ausnahmen.

4.1 Extraktion

Die Datengrundlage für die Analysen bildet eine Datenbank von Belegen für Genitivnomina mit dazugehörigen verschiedenartigen Metainformationen. Die Genitivnomina wurden maschinell aus dem morphosyntaktisch annotierten DeReKo (Institut für Deutsche Sprache 2011a) extrahiert, welches belletristische, wissenschaftliche und populärwissenschaftliche Texte sowie eine große Zahl von Zeitungstexten enthält. Das Korpus wurde u. a. mit dem TreeTagger (Schmid 1994) und der ‚Xerox FST Linguistic Suite[22] annotiert. Nur Letztere bietet eine vollständige morphologische Analyse, bei der Kasus-Informationen mit ausgegeben werden. Allerdings werden zu einer Wortform alle möglichen Kasus ausgegeben und nicht weiter disambiguiert, sodass man sich nicht allein auf die Kasus-Angabe ‚Genitiv' verlassen kann.[23]

Um die im Untersuchungsinteresse stehenden Genitive sicher aus den Daten extrahieren zu können, reicht deshalb die Berücksichtigung der Kasus-Information des Xerox-Taggers nicht aus. Aus diesem Grund wurde ein regelbasierter Ansatz[24] gewählt, um auf der Grundlage der grammatischen Xerox-Annotation Genitive zu extrahieren. Da die Lemmatisierung durch den Xerox-Tagger ebenso nicht disambiguiert, wurde zusätzlich die zuverlässigere Lemmatisierung durch den TreeTagger berücksichtigt.

Die Regeln wurden auf der Basis des Distributionsverhaltens von Genitiven (siehe Abschnitt 3) definiert und beziehen sich einerseits auf den potenziellen Genitiv und andererseits auf den Kontext des potenziellen Genitivs. Berücksichtigt wurden dabei insbesondere folgende Informationen:

22 Vgl. http://open.xerox.com/Services/fst-nlp-tools (Juli 2014).

23 Vgl. die Online-Testfrage unter http://open.xerox.com/Services/fst-nlp-tools/Consume/176 (Juli 2014).

24 Hierfür diente ein Perl-Script (Wall et al. 2001).

Das potenzielle Genitivnomen betreffend:
- Morphologische Analyse des Xerox-Taggers bezüglich Numerus, Genus und Kasus
- Genitivmarkierung (identifiziert durch Differenz zwischen Wortform und TreeTagger-Lemmatisierung)

Den Kontext betreffend:
- Genitivartikel
- Adjektiv auf *-en*, numerische Ordinalzahlen oder Ortsnamenableitung auf *-er* vor dem potenziellen Genitivnomen
- Vor- oder nachgestellte Genitiv-Präpositionen

Insgesamt wurden 18 Regeln implementiert und je nach Distributionsregel Punkte (sowie Strafpunkte)[25] vergeben. Die Punkte wurden anschließend aufsummiert: Je höher die Punktzahl für einen Beleg ist, desto wahrscheinlicher handelt es sich dabei um einen Genitiv. Berücksichtigt für die Analyse wurden alle Belege, die eine Punktzahl >= 2 aufweisen.[26]

Um die Genitivextraktion zu optimieren, wurden sechs Extraktionsdurchgänge durchgeführt. Jeder Durchgang der Genitivextraktion wurde an einem Gold-Standard gemessen. Hierzu wurden 1000 Sätze zufällig aus dem DeReKo-Teilkorpus ‚Mannheimer Korpus 2' (mk2) extrahiert, das aus verschiedenen Textsorten besteht (Romane, Zeitungen, Zeitschriften, Sachtexte etc.) und aus diesem Grund relativ ausgewogen ist.[27] Aus diesem Teilkorpus wurden alle als Neutra, starke Maskulina oder schwache Maskulina mit Genitivmarkierung klassifizierten Wörter extrahiert (3456 Fälle), wobei dafür die entsprechende Annotation des Xerox-Taggers und zur Absicherung der Flexionsklasse die CELEX-Datenbank (vgl. 4.2) verwendet worden sind. In einem nächsten Schritt erfolgte eine manuelle Durchsicht und Klassifizierung der Belege. Folgende relevanten Informationen wurden kodiert: der Genitiv, das Genus, der Numerus und die Zugehörigkeit zur starken Flexionsklasse. Die Fälle wurden von mindes-

25 Es wurde beispielsweise ein Strafpunkt vergeben, wenn der Tagger das betreffende Nomen nicht als Genitiv klassifiziert hat.

26 Der Schwellenwert wurde durch optimierende Annäherungen an den Gold-Standard experimentell festgelegt.

27 Vgl. http://www.ids-mannheim.de/cosmas2/projekt/referenz/virtuell1.html?sigle=mk2&archiv=W (30. Juli 2014).

tens zwei Personen parallel codiert und Abweichungen manuell entschieden.[28] Die endgültige Liste enthält 393 Belege für starke Genitive (oder schwache Genitive mit Genitivmarkierung) im Singular.

Der Extraktionsalgorithmus wurde fortlaufend am Gold-Standard getestet und daran optimiert. Durch den Abgleich des sechsten Durchgangs der Genitivextraktion mit dem Goldstandard wurden die 3456 zu evaluierenden Fälle wie folgt klassifiziert:

- richtig positiv: 291 Fälle
- richtig negativ: 3147 Fälle
- falsch positiv: 10 Fälle
- falsch negativ: 8 Fälle

Daraus ergeben sich folgende Evaluationswerte:

- Präzision: 0,967
- Ausbeute: 0,973
- F: 0.97[29]

97 % der Fälle wurden demnach korrekt klassifiziert.

4.2 Anreicherung mit linguistischen Daten und Metainformationen

Zu den extrahierten Belegen wurde eine Reihe von weiteren Informationen erhoben, die unter anderem bei den unterschiedlichen Hypothesen zu Einflussfaktoren der Markierungsvariation eine Rolle spielen (vgl. Abschnitt 2). Neben morphosyntaktischen Informationen, die durch einen Abgleich mit den annotierten Korpusdaten gewonnen werden konnten, waren dies zusätzlich phonologische, prosodische und semantische Informationen sowie extralinguistische Metadaten. Eine Reihe von morphologischen und phonologisch-prosodischen Eigenschaften konnte unter Nutzung der CELEX-Datenbank ermittelt werden (Baayen u. a. 1995). Der deutsche Teil von CELEX enthält 51.728 Grundformen

28 Insgesamt klassifizierte der Xerox-Tagger 9158 Nomen im mk2-Korpus. Neben den 3456 starken Maskulina und Neutra, die von mindestens zwei Personen durchgesehen worden sind, wurden die restlichen Fälle von mindestens einer Person auf fehl-annotierte Genitive durchgesehen.

29 Die Trefferquote gibt den Anteil der korrekt gefundenen Fälle gemessen an allen möglichen korrekten Fällen an. Die Präzision gibt den Anteil der korrekt gefundenen Fälle gemessen an allen gefundenen Fällen an. Das F-Maß verrechnet Precision und Recall gleichgewichtet zum harmonischen Mittelwert: F = 2 * (Präzision * Recall) / (Präzision + Recall).

(bei Verben ist dies die Infinitivform und bei Nomen die Nominativ-Singular-Form) und 365.530 flektierte Wortformen, die aus mehreren deutschsprachigen Korpora stammen.[30] Zusätzlich wurden die Daten mithilfe von Sonderwortschatzlisten im Hinblick auf verschiedene lexikalische Eigenschaften kategorisiert.[31] Dabei wurden herangezogen: in anderen IDS-Projekten entstandene bzw. begründete Listen zu Abkürzungen,[32] Eigennamen, Fremdwörtern,[33] Neologismen[34] und eine selbst erstellte Liste zu Personentiteln (z. B. *Bundeskanzler* oder *Generalsekretär*). Es wurden mehrere Extraktionen durchgeführt, die fortlaufend durch eine manuelle Stichprobenkontrolle und verschiedene Fehlerbehebungen optimiert wurden. Für die sechste und letzte Extraktion der Daten wurden zusätzlich selbst erstellte Listen zu Konversionen (z. B. *Jemand* oder *Zuwenig*) und Stilbezeichnungen (z. B. *Jugendstil* oder *Barock*) verwendet. Beim Einsatz aller Listen ging es darum, im Korpus möglichst viele zuverlässige Vertreter der betreffenden Wortschatzbereiche zu finden. Eine vollständige Erfassung dieser Wortschatzbereiche ist nicht möglich und so konnte sie auch nicht angestrebt werden.[35]

Durch die Anreicherung der Daten mithilfe der CELEX-Datenbank, der verschiedenen Listen und des Extraktionsskripts stand zu jedem Genitivbeleg eine Menge an Informationen zur Verfügung. Darüber hinaus wurden die Belege mit außersprachlichen Metadaten[36] angereichert, die im Rahmen des Projekts „Korpusgrammatik“[37] erhoben wurden. Die Kategorisierung nach verschiedenen Metadaten, wie z. B. „Medium“ oder „Register“ erlaubt es u. a, die Belege dahingehend zu klassifizieren, ob sie im Rahmen des konzeptionellen Kontinuums in eher „mündlichen“ oder „schriftlichen“ Texten vorkommen. Die Metadaten wer-

30 Zwar stellt die Datenbank dlexDB (vgl. Heister et al. 2011) ein umfangreicheres und neueres Instrument dar, allerdings sind für die vorliegende Studie vor allem die phonologisch-prosodischen und morphologischen Eigenschaften der betreffenden Lemmata interessant, die nur durch die CELEX-Datenbank ermittelt werden konnten.

31 Zur Motivation vgl. 3.3.

32 Die Abkürzungsliste wurde projektübergreifend am IDS erstellt. Die Liste hat ihren Ursprung in der COSMAS-I-Toolbox zur Konvertierung von Fremdtexten in das IDS-Korpusformat. Anschließend wurde diese Liste im Rahmen des Projektes elexiko (vgl. http://www.ids-mannheim.de/lexik/elexiko.html (26.05.2014)) gepflegt und ergänzt.

33 Diese Listen wurden im Rahmen des IDS-Projektes elexiko erstellt (vgl. http://www.ids-mannheim.de/lexik/elexiko.html (26.05.2014)).

34 Diese Liste ist Bestandteil des IDS-Projekts „Lexikalische Innovationen“ (Leitung: Doris Steffens).

35 Genauere Informationen zum Einsatz der meisten Listen in Fuß/Konopka (2014).

36 Sie betreffen Parameter der Quellentexte wie ‚Medium‘, ‚Register‘, ‚Domäne‘, ‚Jahr‘, ‚Land‘ und ‚Region‘ vgl. Bubenhofer/Konopka/Schneider 2013.

37 Vgl. http://www1.ids-mannheim.de/gra/projekte/korpusgrammatik.html (26.05.2014).

den als mögliche Einflussfaktoren in die Baummodellierung aufgenommen. Auf diese Weise kann überprüft werden, ob ein Wandel im Beriech der Genitivmarkierung in konzeptionell mündlichen Texten initiiert wird.

Die extrahierten Genitivnomina werden in der Genitiv-Datenbank (*GenitivDB*)[38] zur Verfügung gestellt, die die Grundlage für die in diesem Beitrag beschriebene Baummodellierung bildete. In Tabelle 2 sind die wichtigsten Informationen aufgeführt, die neben der Information über die Art der Genitivmarkierung in der Genitivdatenbank zu jedem Beleg enthalten sind. Die vollständige Liste der über 80 Angaben morphologischer, lexikalischer, prosodischer, phonologischer und extralinguistischer Art findet sich unter http://hypermedia.ids-mannheim.de/db/liesmich.txt.

Tab. 2: Zusatzinformationen zu den Belegen in der Genitivdatenbank (Auswahl)

Kategorie (Spaltenname in der *GenitivDB*)	**Information (alle binären Variablen sind in der Datenbank mit ‚0' (= negativer Wert) und ‚1' (= positiver Wert) kodiert)**
Lemma	Grundform (Lexem)
MorphGen	Wahrscheinlichkeit für ein Genitivnomen nach Xerox
Mask	Handelt es sich beim Genitivnomen um ein Maskulinum?
Neut	Handelt es sich beim Genitivnomen um ein Neutrum?
Wk	Handelt es sich um ein schwach zu flektierendes Nomen? (= *weak*)
Art	Vorhandensein und Position eines Artikels
AdjEN	Steht ein Adjektiv auf *-en* adjazent davor?
PropN	Ist das Genitivnomen ein Eigenname? (= *proper noun*)
Fremdw	Handelt es sich beim Genitivnomen um ein Fremdwort?
Abk	Handelt es sich beim Genitivnomen um eine Abkürzung?
Neo	Handelt es sich beim Genitivnomen um einen Neologismus?
Stil	Handelt es sich beim Genitivnomen um eine Stilbezeichnung?
Konversion	Handelt es sich beim Genitivnomen um eine Konversion?
Zeitausdruck	Handelt es sich beim Genitivnomen um einen Zeitausdruck?
Titel	Handelt es sich beim Genitivnomen um einen Titel?
NNPrae	Steht vor dem Genitivnomen ein anderes Nomen?
PropNPrae	Steht vor dem Genitivnomen ein Eigenname?
TitelPrae	Steht vor dem Genitivnomen ein Titel?
Kompositum	Ist das Genitivnomen ein Kompositum?

38 Die aktuelle Version der Genitiv-DB ist auf folgender Seite einsehbar: http://hypermedia.ids-mannheim.de/call/public/korpus.genitivdb (26.05.2014)

Kategorie (Spaltenname in der *GenitivDB*)	Information (alle binären Variablen sind in der Datenbank mit ‚0' (= negativer Wert) und ‚1' (= positiver Wert) kodiert)
Fuge	Fuge des Kompositums (falls vorhanden)
HK[39]	Häufigkeitsklasse des Lemmas
HKZG	Häufigkeitsklasse des Zweitglieds beim Kompositum
HKQuot	Quotient aus: HK Kompositum/HK Zweitglied bei Komposita
CELEX	Ist das Nomen oder sein Zweitglied in CELEX berücksichtigt?
AnzSilb	Anzahl der Silben (laut CELEX)
LetztlautDISC	Phonetische Umschrift des letzten Lautes des Lexems im DISC-Format (laut CELEX)
Letztlauttyp	Ist der letzte Laut ein Vokal oder ein Konsonant? (laut CELEX)
Letztlautart	Artikulationsart des Auslautkonsonanten (nasal, liquid etc.) – falls vorhanden (laut CELEX)
LetztreimDISC	Phonetische Umschrift des Reims der letzten Silbe im DISC-Format (laut CELEX)
letztsilbbetont	Ist die letzte Silbe des Genetivnomens betont? (laut CELEX)
vorletztsilbbetont	Ist die vorletzte Silbe des Genetivnomens betont? (laut CELEX)
SilbbetontDist	Entfernung der betonten Silbe von der letzten Silbe
PreOrtho	Orthografische Umschrift des Präfixes – falls vorhanden (laut CELEX)
SuffOrtho	Orthografische Umschrift des Suffixes – falls vorhanden (laut CELEX)
MorphStat	Morphologischer Status (laut CELEX) (Ausprägungen: morphologisch komplex, Konversion, monomorphemisch, contracted form bzw. lexikalisierte Flexion, irrelevante und unbestimmte Morphologie)
Sepa	Ist das Nomen von einem trennbaren Verb abgeleitet? (laut CELEX)
Year	Jahresangabe zum Text, aus dem das Genitivnomen stammt
Country	Landesangabe zum Text, aus dem das Genitivnomen stammt (Ausprägungen: Österreich, Schweiz, Deutschland, Deutschland-Ost, Deutschland-West)

39 Die Genitivnomina wurden mit der DeReWo-Grundformliste vom Dezember 2011 (v-ww-bll-250000g-2011-12-31-0.1, vgl. http://www.ids-mannheim.de/kl/projekte/methoden/derewo.html) abgeglichen. Darin „hat eine Grundform die Häufigkeitsklasse N, wenn die häufigste Form etwa 2^N-mal häufiger vorkommt als diese Form. Für die Grundformenliste ist der Eintrag mit der höchsten Frequenz 'der,die,das' mit f('der,die,das') = 373.738.420 […]." (Dokumentation, S. 6); folglich: Je höher die Häufigkeitsklasse, desto seltener das Wort.

Kategorie (Spaltenname in der *GenitivDB*)	Information (alle binären Variablen sind in der Datenbank mit ‚0' (= negativer Wert) und ‚1' (= positiver Wert) kodiert)
Register	Registerangabe zum Text, aus dem das Genitivnomen stammt (Ausprägungen: Pressetexte, Gebrauchstexte, Literarische Texte)
Domain	Thematische Domäne des Textes, aus dem das Genitivnomen stammt (Ausprägungen: Fiktion, Kultur/Unterhaltung, Mensch/Natur, Politik/Wirtschaft/Gesellschaft, Technik/Wissenschaft)
Region	Regionangabe zum Text, aus dem das Genitivnomen stammt (Ausprägungen: überregional, Herkunft unbekannt, Nordwest, Nordost, Mittelwest, Mittelost, Mittelsüd, Südwest (einschließlich Schweiz), Südost (einschließlich Österreich))
Prob	Wahrscheinlichkeit für ein Genitivnomen (vgl. Abschnitt 4.1)

Unser Ziel war es, die Wahl der Genitivmarkierung durch Parameter wie die in Tabelle 2 aufgelisteten so genau wie möglich vorherzusagen. Hierzu bedienten wir uns einer Methode des maschinellen Lernens.

4.3 Modellierung eines Entscheidungsbaums

Durch die Modellierung eines Entscheidungsbaumes ist die explorative Vorhersage und das Aufdecken von Regeln für das Verhalten einer bestimmten Variable (hier: der Genitivmarkierung) in Abhängigkeit von verschiedenen Faktoren möglich. Es handelt sich dabei um einen gerichteten, azyklischen Graphen, der aus einem Wurzelknoten an der Spitze und beliebig vielen inneren Knoten sowie mindestens zwei Blättern besteht. Die Klassifikation erfolgt vom Wurzelknoten abwärts über innere Knoten, bis ein Blatt erreicht wird, welches die Zielinformation, d. h. eine Ausprägung der im Untersuchungsfokus stehenden Variable, enthält (hier: z. B. die Genitivmarkierung *-es*). Jeder Knoten repräsentiert ein so genanntes Attribut, d. h. einen möglichen Einflussfaktor. Je nach Ausprägung des Einflussfaktors folgt man einem unterschiedlichen Zweig, bis man zu einem Blatt gelangt. Die einzelnen Blätter enthalten die jeweiligen Klassifikationen der Variable, die es zu erklären gilt; in der vorliegenden Studie bestehen diese aus den Genitivmarkierungsvarianten *-s*, *-es*, *-ses*, *-ens*, *-ns*, [nach *s*, *x* etc.] -', *-'s*, Ø (vgl. Abschnitt 3).

Entscheidungsbäume werden mit Verfahren des maschinellen Lernens auf der Basis von Trainingsdaten erstellt. Um die Wahl der Genitivmarkierungen vorherzusagen, wurde im Rahmen dieser Studie der Algorithmus C4.5 (vgl. Quinlan 1993), der in der Software WEKA (vgl. Witten/Frank 2005) implementiert ist, angewendet. Der Algorithmus testet jeden Faktor daraufhin, ob er die Daten-

menge in Gruppen aufteilt, die in sich so wenig Varianz wie möglich aufweisen. Der Faktor, der die Varianz in einer Gruppe am besten erklärt, wird ausgewählt und der Trainingsdatensatz in Teilmengen aufgeteilt. Das Maß für die Aufteilung ist die Kullback-Leibler-Divergenz (vgl. Kullback/Leibler 1951), die auf der Berechnung der relativen Entropie basiert. Für jede weitere Teilmenge werden die Faktoren mithilfe des Maßes bewertet und je nach Bewertung der Faktoren in weitere Teilmengen aufgeteilt. Dieser Prozess wird wiederholt, bis eine Teilmenge keine Varianz mehr aufweist, d. h. lediglich nur noch Fälle einer Klasse enthält oder eine vorgegebene minimale Anzahl von Fällen pro Blatt erreicht ist.

Im Vergleich zu anderen Algorithmen des maschinellen Lernens muss bei C4.5 keine binäre Aufteilung erfolgen. Für die Analyse der Genitivmarkierungen bietet der Algorithmus den Vorteil, dass die Anzahl der Verzweigungen beliebig ist. Denn die meisten potentiellen Einflussfaktoren sowie die Zielvariable (= Genitivmarkierung), weisen keine binären, sondern eine beliebige Anzahl an Merkmalsausprägungen auf. Ein weiterer Vorteil des Algorithmus C4.5 besteht im Umgang mit fehlenden Attributwerten. Im vorliegenden Datenkorpus kommt es vor, dass durch die Listenvergleiche nicht für alle Lemmata vollständige Informationen vorliegen. Durch den gewählten Algorithmus werden unbekannte Werte bei der Berechnung ignoriert.

Für die Analyse der nominalen Genitivmarkierung wurden nach jeder Extraktion mehrere Bäume trainiert, die sich zum einen durch statistische Einstellungen und zum anderen durch die Auswahl möglicher Einflussfaktoren unterscheiden. In diesem Beitrag wird der letzte dieser trainierten Entscheidungsbäume, der nach der sechsten Extraktion entstanden ist, vorgestellt und interpretiert. Punktuell wird zum Vergleich der letzte Entscheidungsbaum herangezogen, der nach der fünften Extraktion entstanden ist (für Visualisierungen beider Bäume siehe http://hypermedia.ids-mannheim.de/call/public/korpus.ansicht?v_id=5032).

4.4 Statistische Evaluation der Entscheidungsbäume

Die statistische Evaluation erfolgte auf der Basis einer Berechnung mit zehnfacher Kreuzvalidierung und einem Stoppkriterium von mindestens 2000 Fällen pro Blatt. Der letzte Entscheidungsbaum nach der fünften Extraktion der Daten wurde mit einem 50 %igen Konfidenzintervall berechnet.[40] Nach seiner Durchsicht und

40 Für einen Vergleich verschiedener Konfidenzintervalle bei der Berechnung von Entscheidungsbäumen vgl. Drazin/Montag (http://wwww.samdrazin.com/classes/een548/project2report.pdf, 26.05.2014).

Interpretation erfolgte eine sechste, optimierte Extraktion. In der letzten der danach erfolgenden Baummodellierungen wurden zur besseren linguistischen Interpretierbarkeit verschiedene Einflussfaktoren zusammengelegt (vgl. dazu Abschnitt 5). Darüber hinaus konnten vor der Modellierung einige der noch verbliebenen Extraktionsfehler manuell behoben werden. Um den Baum zu verkleinern und somit die linguistische Interpretierbarkeit zu vereinfachen, wurde außerdem das Konfidenzintervall auf 95 % erhöht. Die statistische Evaluation dieses Baumes wird im Folgenden erläutert.

Durch die Berechnung des letzten Baumes aus Extraktion 6 wurden 2.861.871 Instanzen (als Genitivtoken) klassifiziert. Insgesamt wurden 22 mögliche Einflussfaktoren bei der Baummodellierung berücksichtigt (vgl. Abschnitt 5). Der berechnete und visualisierte Baum besteht im Ganzen aus 1.559 Blättern, was eine Reduktion der Komplexität im Vergleich zu davor modellierten Bäumen bedeutet. So hatte der letzte Baum aus Extraktion 5, der mit 33 möglichen Einflussfaktoren modelliert wurde, noch 8.167 Blätter (bei 3.547.402 klassifizierten Instanzen).

Tabelle 3 zeigt die Evaluationsmaße (vgl. Fußnote 29) für die Vorhersage der einzelnen Endungen im Rahmen des letzten Entscheidungsbaumes aus Extraktion 6.

Tab. 3: Evaluationsmaße für den letzten Entscheidungsbaum aus Extraktion 6

Genauigkeit	Trefferquote	F-Maß	Markierungsvariante
0	0	0	*-’s*
0.92	0.58	0.71	*-Ø*
0.99	0.99	0.99	*-ens*
0.90	0.88	0.89	*-es*
0.86	1	0.92	*-ns*
0.91	0.94	0.92	*-s*
0	0	0	*-’*
1	0.99	1	*-ses*
0.91	**0.91**	**0.91**	**GESAMT**

Das durchschnittliche F-Maß beträgt 0.91. 91 % der möglichen Markierungsvarianten werden dabei korrekt vorhergesagt bei einem Recall von 0.91. Das F-Maß der Endungen *-s*, *-ens*, *-ns* und *-ses* liegt über 0.90. Die relativ häufig vorkommende *es*-Endung wird in 90 % der Vorkommen korrekt zugeordnet (bei einer Trefferquote von 88 %). Die Vorhersagekraft für die Nullmarkierung (Ø) stieg im Vergleich zur Berechnung des letzten Baumes aus Extraktion 5 von 64 % auf ca. 71 %. Der Grund dafür liegt nicht etwa in einem signifikanten Anstieg der Trefferquote, sondern in einer steigenden Genauigkeit (von 86 % auf 92 %). Die

niedrigfrequenten Markierungen mit finalem Apostroph (wie in *Hingis'* [*Halbfinal*]) und apostrophierten *s* (wie in *Reni's* [*Frischmarkt*]) werden nicht erkannt (die Trefferquote, die Genauigkeit und das F-Maß liegen bei 0).

Um genauere Aussagen über die Verteilung der vorhergesagten Endungen zu treffen, betrachten wir in Tabelle 4 die Konfusionsmatrix im Hinblick auf die Verteilung der vorhergesagten Genitivendungen im Vergleich zu den tatsächlichen Markierungen (Extraktion 6).

Tab. 4: Konfusionsmatrix vorhergesagte vs. tatsächliche Genitivendungen im letzten Entscheidungsbaum aus Extraktion 6

a	b	c	d	e	f	g	h	<– classified as
0	13	0	80	2	214	0	0	**a = *-'s***
0	41850	71	7325	851	22494	0	0	**b = -Ø**
0	0	11489	78	0	48	0	0	**c = *-ens***
0	2172	0	975503	9	128734	0	0	**d = *-es***
0	0	0	0	11998	0	0	0	**e = *-ns***
0	1458	0	104679	1114	1540169	0	0	**f = *-s***
0	5	0	32	0	121	0	0	**g = -'**
0	46	0	0	0	0	0	7330	**h = *-ses***

Was die falschen Vorhersagen angeht, so wird die Nullmarkierung (Ø) oft als *s*-Endung vorausgesagt (in ca. 31 % der Fälle). Die Korrektheit steigt dabei im Vergleich zum letzten Baum aus der fünften Extraktion um ca. 6 %. Auch die relativ gut klassifizierte *es*-Endung wird – wenn falsch vorhergesagt – meist mit der *s*-Endung verwechselt; insgesamt wird die Markierung *-es* in 11 % der Fälle falsch klassifiziert. Die Klassifizierung der *s*-Endung wird etwas schlechter als in vorangegangenen Modellierungen: Wenn diese Markierung falsch erkannt wird, dann wird sie meistens als *es*-Endung klassifiziert (in ca. 6 % der Fälle). Die Markierung *-ns* wird (wie schon in vorangegangenen Analysen) zu 100 % korrekt vorausgesagt. Die Endung *-ens* bleibt mit ca. einem Prozent an falschen Klassifizierungen sehr stabil. Wenn diese falsch vorausgesagt wird, dann konkurriert sie mit den Endungen *-s* und *-es*. Die Vorhersage der *ses*-Endung verbessert sich gegenüber vorangegangenen Modellierungen: Wurden in der letzten Modellierung aus Extraktion 5 noch 2 % aller Fälle falsch klassifiziert, sind es im Rahmen dieser Baummodellierung nur noch 0,6 %. Im Fall einer falschen Vorhersage, wird die Nullmarkierung (Ø) angenommen.

5 Linguistische Auswertung und Optimierung der Entscheidungsbäume

Die Auswahl der Attribute bzw. potenzieller Einflussfaktoren, die in die Baummodellierungen eingingen, die jedem der sechs Extraktionsdurchgänge folgten, wurde von Modellierung zu Modellierung variiert. Manche Attribute wurden modifiziert. Dabei ging es darum, einerseits die Vorhersagewerte zu steigern bzw. ihr hohes Niveau aufrechtzuerhalten und andererseits den Baum nicht allzu komplex ausfallen zu lassen bzw. seine linguistische Interpretierbarkeit zu sichern. Mitunter machten die Entscheidungsbäume die Extraktionsfehler transparent und trugen zu verbesserten Folgeextraktionen bei. Im Folgenden wird wie schon in Abschnitt 4 auf die letzte Baummodellierung aus Extraktion 6 näher eingegangen. Die letzte Baummodellierung aus Extraktion 5 wird punktuell zum Vergleich herangezogen, um die Weiterentwicklung der Exploration zu veranschaulichen. Tabelle 5 präsentiert die jeweils benutzten Attribute, und zwar geordnet nach der Häufigkeit der dazugehörigen Knoten in der Baumvisualisierung.

Tab. 5: Attribute der Baummodellierung und ihre Häufigkeit in der Baumvisualisierung[41] (für Visualisierungen siehe http://hypermedia.ids-mannheim.de/call/public/korpus.ansicht?v_id=5032)

Baum aus Extraktion 5 Attribut/Einflussfaktor	Anzahl Knoten	Baum aus Extraktion 6 Attribut/Einflussfaktor	Anzahl Knoten
HK	24	HK	29
LetztreimDISC	14	Vokallänge	13
MorphStat	9	Jahr	10
ArtDist	7	Region	10
LetztlautDISC	6	Genus	9
Mask	6	Eigenname	8
PropN	6	KonsGruppe	7
Kompositum	5	Präfix	7
Wortart	5	Fuge	6
Year	5	Letztlaut	5
AnzSilb	4	SilbbetontDist	5

41 Zu Attributen bzw. Faktoren vgl. Tabelle 2 sowie http://hypermedia.ids-mannheim.de/db/liesmich.txt.

Baum aus Extraktion 5 Attribut/Einflussfaktor	Anzahl Knoten	Baum aus Extraktion 6 Attribut/Einflussfaktor	Anzahl Knoten
Country	4	Fremdwort	5
Fremdw	3	KonsArt	5
Fuge	3	HKZG	4
Domain	2	letztsilbbetont	4
HKZG	2	Suffix	3
Letztlautart	2	Silbenanzahl	2
letztsilbbetont	2	Komplexität	2
Region	2	Letztlauttyp	1
Neut	1		
SilbbetontDist	1		
Abk	0		
Medium	0		
Neo	0		
NNPrae	0		
PropNPrae	0		
Register	0		
Sepa	0		
SuffOrtho	0		
Titel	0		
TitelPrae	0		
Wk	0		
Zeitausdruck	0		

Die Metainformationen, die in die letzte Baummodellierung nach Extraktion 5 eingingen (vgl. den linken Teil von Tabelle 5), spiegelten die in der bisherigen Forschung postulierten Einflussfaktoren der Markierungsvariation nur teilweise wider. Den Forschungsfaktoren entsprachen direkt z. B. die Informationen, ob es sich beim Genitivnomen um ein Lexem des Sonderwortschatzes handelt (vgl. *PropN*, *Fremdw*, *Zeitausdruck*, *Abk*, *Neo*), ob mit dem Genitivnomen ein Kompositum vorliegt (*Kompositum*), oder auch die Genusinformationen (*Mask*, *Neut*). Wir hofften, uns den Unterschieden, die in der Forschung zwischen häufigen und seltenen Genitivformen gemacht wurden, mithilfe der Häufig-

keitsklassen *HK* (je höher die Häufigkeitsklasse, desto seltener das Wort, vgl. Fußnote 39)[42] und der abgeleiteten Attribute *HKZG* und *HKQuot* gut annähern zu können. Viele typische CELEX-Angaben wie *SuffOrtho*, *MorphStat* oder *LetztreimDISC* konnten unbearbeitet allerdings kaum mit den Forschungsfaktoren zusammengeführt werden.

Vor der Extraktion 6 wurden daher einige Änderungen vorgenommen. Die umfangreichsten betrafen den komplexen Faktor ‚Sonderwortschatz'. Da sich in bisherigen Extraktionsergebnissen bei den verschiedenen Sonderwortschatzklassen deutlich eine Affinität zur Endung *-s* bzw. zur Nullmarkierung abzeichnete, wurden die bei der Sonderwortschatzerfassung verwendeten Listen stark erweitert und an mehreren Stellen berichtigt, um die Bedeutung des Sonderwortschatzes als Einflussfaktor angemessener berücksichtigen zu können bzw. um später bei der Baummodellierung möglichst selten andere Attribute an Stellen zu aktivieren, an denen eigentlich der Sonderwortschatzstatus entscheidend war.

Auch in den auf die Extraktion 6 folgenden Baummodellierungen wurden Änderungen durchgeführt. Ziel war es, durch Attributumformulierung und das Weglassen von unbedeutenderen Attributen einerseits zu besser interpretierbaren Einflussfaktoren im Entscheidungsbaum zu gelangen und andererseits diejenigen Attribute, deren Relevanz sich in den vergangenen Modellierungen abzeichnete, genauer zu prüfen.

Im Bereich der Umformulierung von Attributen war die größte Neuerung, dass basierend auf verschiedenen Gruppierungen der sehr zahlreichen Ausprägungen des *LetztreimDISC*-Attributs drei neue, weit weniger komplexe Attribute eingeführt wurden, die *LetztreimDISC* ersetzten:

- *KonsGruppe* (abschließende Konsonantengruppe) mit zwei Ausprägungen (positiv, z. B.: *LetztreimDISC = ‚ɛrk'* wie in *Triebwerk* und negativ, z. B.: *LetztreimDISC = ‚at'* wie in *Staat*),
- *Vokallänge* (vor abschließenden Konsonanten) mit den Ausprägungen *lang, Diphthong, kurz, Schwa*,
- *Vokalhöhe* (vor abschließenden Konsonanten) mit den Ausprägungen *tief, mittel, hoch, Diphthong*.

Eine weitere wichtige Änderung war die Kombination von *Letztlautart* und bestimmten Ausprägungen von *LetztlautDISC* sowie *LetztreimDISC* zu einem neuen

42 Mit Häufigkeitsklassen werden nicht Genitivtoken klassifiziert, sondern alle Token, die einer Grundform (einem Lemma) entsprechen (vgl. DeReWo- Grundformliste vom Dezember 2011 (v-ww-bll-250000g-2011-12-31-0.1, vgl. http://www.ids-mannheim.de/kl/projekte/methoden/derewo.html, 26.05.2014).

Merkmal *KonsArt* (Art des Auslautkonsonanten) mit den Ausprägungen *s-Laut*, *sch*-Laut, *st*-Gruppe sowie – für die übrigen Fälle – den aus *Letztlautart* übernommenen Ausprägungen *L(iquid), N(asal), F(rikativ), A(ffrikate), P(losiv)*.[43]

Der hier zu beschreibenden Baummodellierung gingen nach der Extraktion 6 noch vier voraus. Sie wurden u. a. dazu genutzt, um Sicherheit hinsichtlich der Attribute zu erbringen, die eliminiert werden konnten, weil ihre Bedeutung für die bisherigen Modellierungen offensichtlich gering war oder weil sie fehleranfällig waren bzw. sich linguistisch nicht sinnvoll interpretieren ließen. Diesen Auswertungen ist auch das erst kürzlich eingeführte Attribut *Vokalhöhe* zum Opfer gefallen sowie Attribute, die schwächer belegte Sonderwortschatzklassen betrafen (*Abk, Neo, Titel, Zeitausdruck*). Letztere wiesen trotz jetzt besserer Erfassung immer noch zu wenige Token auf, als dass sie in der Baumvisualisierung sichtbar geworden wären. Außerdem fanden Zusammenfassungen und – der Verständlichkeit halber – Umbenennungen von Attributen statt wie *Mask + Neut -> Genus* mit den Ausprägungen *Mask* und *Neut*; *PreCode + SuffCode + Kompositum + MorphStat -> Komplexität* mit den Ausprägungen *Mono*, *Präf*, *Suff* und *Kompositum*; *Year -> Jahr*; *PropN -> Eigenname*; *LetztlautDISC -> Letztlaut*; *Fremdw -> Fremdwort*; *PreOrtho -> Präfix*; *SuffOrtho -> Suffix*.

Der hohe Anteil der in der Baummodellierung korrekt vorhergesagten Instanzen[44] legt nahe, dass die Markierungsvariation grundsätzlich systematisierbar ist.[45] Die Visualisierung des letzten Baumes nach der Extraktion 6 ist unter http://hypermedia.ids-mannheim.de/call/public/korpus.ansicht?v_id=5032 einzusehen. Als Hinweise auf die Relevanz der einzelnen in die Baummodellierung eingegangenen Attribute können im Folgenden dienen:

- die Häufigkeit ihres Auftretens in der Baumvisualisierung,
- die Baumposition der das Attribut repräsentierenden Knoten,
- die Systematisierbarkeit der Korrelation der Attributausprägungen mit der Markierungswahl.

Nicht möglich ist hier aus Platzgründen die Besprechung aller in die Modellierung eingegangenen Attribute im Hinblick auf alle drei Punkte. Die folgende Darstellung versucht daher nur die wichtigsten Feststellungen zusammenzufassen.

43 Zusätzlich wurde in die Modellierung auch das Attribut *Letztlauttyp* wieder aufgenommen, wobei basierend auf CELEX-Kategorien *Letztlauttyp* und *LetztreimDISC* die Ausprägungen *V(okal)*, *C(onsonant)* und *D(iphthong)* unterschieden wurden.

44 Ca. 91 %, vgl. oben Tabelle 3.

45 Vgl. dazu auch eine Analyse mit einer anderen Attributauswahl in Schneider (2014, Abschnitt 4).

Die **Häufigkeit des Auftretens von Attributknoten** in der Baumvisualisierung kann dem rechten Teil von Tabelle 5 weiter oben entnommen werden. Sie steht dafür, wie oft ein Attribut in der Visualisierung für eine Entscheidung gesorgt hat. Auffällig und einigermaßen überraschend ist die große Häufigkeit des Knotens *HK* (*Häufigkeitsklasse* mit 29 Vorkommen). Sie weist darauf hin, dass ‚Lexemfrequenz' ein wichtiger Faktor der Markierungswahl sein könnte, wenn auch die große Knotenhäufigkeit zum Teil dadurch bedingt sein kann, dass es für *HK* viele Ausprägungen gibt.[46] Dabei spielte in den Ausführungen Fehringers (2011) die Tokenfrequenz eine zentrale Rolle – ansonsten blieb die Frequenz aber in der bisherigen Forschung weitgehend unbeachtet. Die Häufigkeit des Knotens *HK* ist jetzt noch höher als in der letzten Visualisierung aus Extraktion 5 (vgl. den linken Teil von Tabelle 5), sodass der Einflussfaktor ‚Lexemfrequenz' sich deutlicher von anderen Faktoren absetzt. Es folgen die aus dem aufgelösten Attribut *LetztreimDISC* neu hervorgegangene Variable *Vokallänge* sowie die gleich frequenten extralinguistischen Variablen *Jahr* und *Region*. Was die übrigen neu formulierten Attribute angeht, so findet sich *KonsGruppe* in der oberen Hälfte der Rangliste, *KonsArt* dagegen in der unteren Hälfte, wobei selbst letzterer Faktor immer noch beachtliche fünf Knotenvorkommen aufweist. Hohe Positionen nehmen noch *Genus* und *Eigenname* (vormals *PropN*) ein. Hingegen nimmt das in der Forschung als Einflussfaktor allgemein anerkannte Attribut *Silbenanzahl* (vormals *AnzSilb*) mit zwei Knotenvorkommen eine Position gegen Ende der Liste ein, was allein aber seine Bedeutung noch nicht entscheidend relativieren kann (vgl. weiter unten). Die Anzahl der in die Baummodellierung aufgenommenen Attribute ist im Vergleich zur Modellierung anhand Extraktion 5 von 33 auf 19 gesunken. Alle 19 Attribute finden sich jetzt auch als durch Knoten repräsentierte Einflussfaktoren in der Baumvisualisierung wieder (vgl. oben Tabelle 5). Das Modell wurde gegenüber dem Modell aus Extraktion 5 insgesamt stringenter und überschaubarer.

Was die **Position der Attributknoten im Baum** angeht, so werden hier die obersten drei Ebenen genauer betrachtet (vgl. Abbildung 3). Den Wurzelknoten nimmt *HK* (*Häufigkeitsklasse*) ein. Darunter erscheinen Knoten, denen *KonsArt*, *Silbenanzahl*, *Vokallänge*, *Fremdwort* und noch einmal *KonsArt* zugeordnet sind, alles Attribute, die direkt Faktoren entsprechen, die in der Forschung postuliert werden (allerdings ‚Vokallänge' und ‚Konsonantenart' nur selten besonders gewichtet). Die an den Knoten vorgenommenen Datenaufteilungen erscheinen linguistisch weitgehend plausibel. Für *HK*, *Silbenanzahl*, *Vokallänge* und *Fremdwort* sind im Weiteren keine Einschränkungen auf einen bestimmten Bereich des

46 In unseren Daten sind 22 Häufigkeitsklassen vertreten.

Baums (den Anfang oder den terminalen Bereich) festzustellen. Hingegen tendiert *KonsArt* eher zur Position vor terminalen, die konkrete Markierungswahl repräsentierenden Knoten.

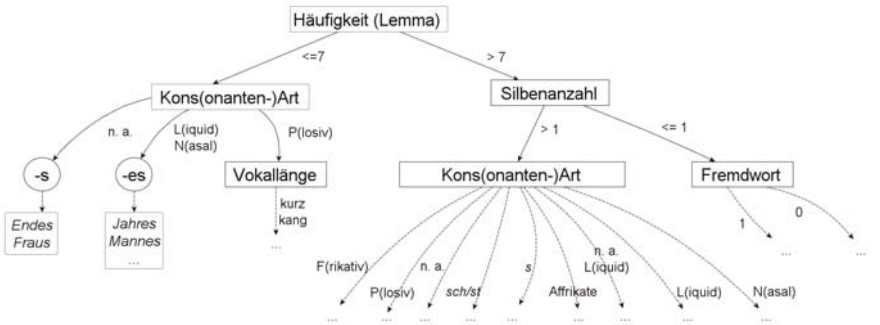

Abb. 3. Die obersten drei Ebenen der Visualisierung des letzten Baums aus Extraktion 6

Betrachtet man Attribute, deren Knoten im Baum relativ häufig auftreten, aber nicht in den obersten drei Ebenen erscheinen, so fällt es auf, dass auch *Jahr*, *Region*, *Genus*, *KonsGruppe* und *Präfix* zum terminalen Bereich tendieren.

Die **Korrelationen zwischen Attributausprägung und Markierungswahl** erscheinen teilweise systematisierbar. Die sich abzeichnenden Haupttendenzen,[47] die evtl. auch schon in dem nach Extraktion 5 modellierten Baum zu beobachten waren und oft den Forschungspostulaten entsprechen, sind in Tabelle 6 zusammengestellt.

Tab. 6: Attributausprägungen mit sich abzeichnenden Markierungswahltendenzen im letzten Baum aus Extraktion 6

Attribut	Beschreibung der Ausprägungen	Markierungswahl – Tendenz
HK (*Häufigkeitsklasse*)	niedrigere Werte	-> *es*
Vokallänge	Schwa-Silbe (*-en, -el* etc.)	-> *s*
	Diphthong[48]	-> *s*

47 Gemeint ist hier ein merkliches Ansteigen der Wahrscheinlichkeit für eine bestimmte Markierungsvariante beim Vorliegen der Faktorausprägung.

48 Bei Nomen auf Schwa-Silbe ist mit der Ausnahmslosigkeit der *-s*-Endung zu rechnen. Die Diphthonge vor abschließenden Konsonanten zeigen nur eine leichte Tendenz. Eine solche lässt sich weder bei restlichen kurzen Vokalen noch bei langen Vokalen vor abschließenden Konsonanten beobachten.

Attribut	Beschreibung der Ausprägungen	Markierungswahl – Tendenz
Jahr	spätere Texte[49]	-> *s*
Region	Südost	-> *es*
Eigenname	positiv	-> *s*
KonsGruppe	positiv	-> es
Fremdwort	positiv	-> *s*, NULL
KonsArt	*s*-Laut	-> *es*, NULL
	st-Gruppe	-> *es*
	Affrikate	-> *es*
	Plosiv	-> *es*
	Nasal	-> *s*
HKZG	niedrigere Werte	-> *es*
letztsilbbetont	positiv	-> *es*
Suffix	*-nis*	-> *ses*
	andere Suffixe	-> *s*, NULL
Silbenanzahl	Einsilber	-> *es*
Letztlauttyp	V(okal), D(iphthong)	-> *s*

Keine nennenswerten Tendenzen konnten während der Inspektion der Baumvisualisierung für Ausprägungen der Attribute *Genus*, *Präfix*, *Fuge*, *SilbbetontDist*, *Suffix* und *Komplexität* festgestellt werden.

Durch die Baumvisualisierung hat sich vor allem bei den Faktoren ‚Lexemhäufigkeit', ‚Silbenanzahl', ‚Art des Auslauts' (Attribute *KonsGruppe, KonsArt, Letztlaut*) und ‚Zugehörigkeit zum Sonderwortschatz' (Attribute *Fremdwort* und *Eigenname*) der Verdacht auf ihre Bedeutung verdichtet. Außerdem kam die starke Vermutung auf, dass die Vokallänge vor konsonantischem Auslaut und auch die außersprachlichen Dimensionen ‚Zeit' (Attribut *Jahr*) und ‚Raum' (Attribut *Region*) eine wichtige Rolle spielen können. Die Berücksichtigung des Sonderwortschatzes ist viel konsistenter ausgefallen als in den Modellierungen vor Extraktion 6, wobei teilweise weiterbestehende Schwächen v. a. bei der Erfassung von Eigennamen nicht verschwiegen werden dürfen. Im Modell ist schließlich nur ein Attribut verblieben (*Letztlaut* – phonetische Umschrift des letzten Lautes), das aufgrund eines zu niedrigen Abstraktionsgrades kaum zusammenhängend generalisierbar ist. Das gewisse Einzelausprägungen des Aus-

49 So zeigen etwa einsilbige Nomen auf die Affrikate <*pf*> in den Texten aus der Zeit nach 2002 mehrheitlich die Endung *-s* (wie in *Zweikampfs*), während sie in früheren Texten eher zu *-es* (wie in *Zweikampfes*) tendieren.

lauts doch zu klaren Tendenzen führen können, bleibt dabei unbenommen. Einige Tendenzen werden im letzten Baum aus Extraktion 6 schon durch andere Attribute als *Letztlaut* abgedeckt (z. B. *KonsArt*: *sch*-Laut -> es), andere bleiben tatsächlich nur mithilfe von *Letztlaut* aufspürbar (z. B. unter Liquiden tendiert /l/ stärker zu *-s* als /ɪ/).

6 Diskussion und Fazit

In der explorativen Phase der korpusbasierten Untersuchungen zur Variation der nominalen Genitivmarkierung wurden sechs Datenextraktionen durchgeführt. Nach jedem Extraktionsdurchgang wurden mithilfe maschinellen Lernens mehrere Entscheidungsbäume zur Markierungswahl modelliert, wobei die Anzahl und Art der die Entscheidungen bedingenden Attribute, d. h. der potentiellen Einflussfaktoren, variiert wurde. Durch die Variation der Attribute war es möglich, relevante von irrelevanten Variablen zu unterscheiden, um schließlich einen Baum zu berechnen, der die Variation der Daten bestmöglich aufklärt. Durch den letzten Baum nach der sechsten Extraktion wurden 2.861.871 Instanzen klassifiziert. Bei der Berechnung wurden insgesamt 19 mögliche Einflussfaktoren berücksichtigt, durch die ca. 91 % der Daten korrekt vorhergesagt werden konnten (bei einem Recall von 0.91). Im Zuge der Interpretation vorangegangenen Baummodellierungen konnten einige Faktoren als Einflussgrößen ausgeschlossen werden. Darüber hinaus hat sich herausgestellt, dass aus linguistischer Perspektive einige Faktoren zusammenzulegen sind. Außerdem nahm die Anzahl der Blätter drastisch ab, was die linguistische Interpretierbarkeit stark vereinfachte.

Durch die Interpretation mehrerer Baummodellierungen war es möglich, einen detaillierten Einblick in die Wirkungsweisen und die Hierarchisierungen der Faktoren zu bekommen. Darüber hinaus konnten potentielle Fehlerquellen aufgedeckt und korrigiert werden. Die Komplexität der Bäume erschwerte allerdings das Aufdecken und Visualisieren von linguistischen Regularitäten. Diese wurden schließlich durch die Vereinfachungen und Zusammenlegungen im letzten Baum besser darstellbar und interpretierbar.

Somit erwiesen sich die maschinellen Extraktionen und auf maschinellem Lernen basierenden Modellierungen von Entscheidungsbäumen als geeignete Methoden, um die Exploration der Variation im Bereich der kanonischen Markierungen *-es* und *-s* sowie der abgeleiteten standardisierten, gesprochen wie geschrieben expliziten Markierungen *-ses*, *-ens*, *-ns* voranzutreiben. Umgekehrt zeigten sie deutliche Schwächen bei der Erfassung und Klassifikation der

nicht-kanonischen, weniger standardisierten und zum Teil auch weit selteneren Markierungsvarianten *-'s, -'* und Ø.

Die Modellierung von Entscheidungsbäumen erwies sich auch als eine geeignete Methode, um Einflussfaktoren mit relativ großer Abdeckung,[50] z. B. im Bereich Sonderwortschatz ‚Fremdwort' oder ‚Eigenname', zu einem System zusammenzubauen. Wenig bis keinen Aufschluss gab sie umgekehrt über Einflussfaktoren mit einer verhältnismäßig kleinen Abdeckung, z. B. im gleichen Bereich ‚Neologismus' oder ‚Konversion'. Insgesamt sind die verwendeten Methoden offenkundig nicht dazu da, die Durchschlagskraft eines Faktors (bzw. seine Effektgröße)[51] genauer abzuschätzen, und auch nicht, um seine Effektstärke zu kalkulieren. Diese Fragestellungen müssen in den Untersuchungen, die der hier dokumentierten explorativen Phase folgen, eigens angegangen werden.[52]

Mit der verwendeten Methodik ließ sich wiederum gut prüfen, (1) ob sich die Variation in den Daten überhaupt als ein konsistentes System modellieren lässt. Diese Methodik konnte Aufschluss darüber geben, (2) ob die in der Forschung postulierten Faktoren ein solches konsistentes System ergeben bzw. wie viel Variation man mit einem bestimmten Ensemble an Faktoren erklären kann. Dabei wurde deutlich, (3) welche der postulierten Faktoren (bzw. der in die Baummodellierung eingehenden Attribute) sich in einem solchen System tatsächlich als Einflussfaktoren wiederfinden und welche Position sie darin einnehmen. Auf diese Weise konnten (4) einerseits bisherige Hypothesen über die Relevanz von Faktoren geprüft und andererseits neue Hypothesen über Einzelfaktoren und Faktorenhierarchien herausgearbeitet werden.

Ad (1): Die in diesem Beitrag beschriebenen Analysen legen nahe, dass die Markierungsvariation durch hohe Systematizität gekennzeichnet ist. Sie lässt sich im Wesentlichen tatsächlich als ein konsistentes System modellieren, und das bereits mithilfe von linguistischen und extralinguistischen Faktoren, die – wie die unbearbeiteten CELEX-Kategorien – nicht eigens für unseren Zweck angepasst wurden (Baum aus Extraktion 5).[53]

50 D. h. Faktoren, die für relativ große Teile der Datenbasis einschlägig sind.

51 Gemessen am Anteil des einschlägigen Materials, den eine Faktorausprägung in zu erwartender Weise beeinflusst.

52 Vgl. Fuß/Konopka 2014, wo auf die Durchschlagskraft von Faktoren mithilfe relativer Häufigkeiten der Ausprägungen geschlossen wird und als Maße für die Effektstärke die logarithmierte Odds Ratio sowie der Phi-Koeffizient verwendet werden. Darüber hinaus wird das Zusammenwirken der Einflussfaktoren hypothesengeleitet durch die Berechnung logistischer Regressionen untersucht.

53 Zu einem ähnlichen Ergebnis kommt Schneider (2014) mit einem modifizierten Satz an Faktoren/Attributen.

Ad (2): Interessanterweise führt eine weitgehende Anpassung der in die Baummodellierung eingehenden Attribute an die in der Forschung postulierten Faktoren (Baum aus Extraktion 6) nicht zu einer wesentlichen Verbesserung der Aufklärungsrate des Modells (gemessen etwa am Anteil der korrekt vorhergesagten Instanzen). Dennoch kann auf der Basis solcher Faktoren ein relativ konsistentes Modell entstehen. Das Ensemble aus in der Forschung postulierten Faktoren, das bei der Modellierung des letzten Baums zum Einsatz kommt, könnte aber offensichtlich noch optimiert werden.

Ad (3): Von den in die Modellierung des letzten Baums eingegangenen Faktoren finden sich alle in der entsprechenden Visualisierung wieder, was den Hypothesen über ihre prinzipielle Relevanz zumindest nicht widerspricht. Dabei ragt durch die Frequenz und die zentrale Positionierung der entsprechenden Knoten die ‚Häufigkeitsklasse des Lemmas' (Attribut *HK*) heraus, die auf den Faktor ‚Häufigkeit' im Allgemeinen verweist, der in der einschlägigen Forschung noch nicht richtig etabliert ist. Außerdem ist an den Baumvisualisierungen die Relevanz der komplexen Variationsfaktoren ‚Silbenanzahl', ‚Sonderwortschatzzugehörigkeit' (Attribute *Fremdwort* und *Eigenname*) und ‚Wortausgang' (Attribute *KonsGruppe, KonsArt*) ablesbar, wozu gemeinschaftlich und ausbalanciert die Frequenz einschlägiger Knoten, deren Position im Baum und die Systematisierbarkeit der Korrelation von Faktorausprägungen und Markierungswahl beitragen. Die Wirkung der genannten linguistischen Faktoren scheint in unseren Daten noch durch die außersprachlichen Parameter *Zeit* und *Raum* modulierbar, zumindest in der Weise, dass prinzipiell Genitivnomina in Texten aus jüngerer Zeit stärker als sonst zu *-s* und Genitivnomina in Texten aus dem Südosten stärker als sonst zu *-es* tendieren. Eine schlechte Systematisierbarkeit der Korrelation von Faktorausprägungen und Markierungswahl spricht übrigens gegen die Postulierung von Faktoren, die direkt den Attributen *Vokallänge*, *Genus*, *SilbbetontDist*, *Fuge*, *Präfix*, *Suffix* und *Komplexität* entsprechen. Außerdem ist das Attribut *Letztlaut* klarerweise überspezifiziert.[54] Vor diesem Hintergrund scheinen vorsichtigere, auf wenige Einflussfaktoren beschränkte Gesamtdarstellungen der Markierungsvariation, unsere Datenlage besser wiederzugeben.

Ad (4): Wie bereits erwähnt, scheinen die Faktoren ‚Häufigkeitsklasse', ‚Silbenanzahl' und ‚Sonderwortschatzzugehörigkeit' (hier bezogen auf alle Sonderwortschatzklassen, also außer ‚Fremdwort' und ‚Eigenname' auch auf ‚Neologismus', ‚Konversion' etc.) eine wichtige Rolle zu spielen. Insbesondere im Bereich des Grundwortschatzes scheinen zusätzlich der Wortausgang und die zeitliche und räumliche Einordnung relevant zu werden. Es zeichnet sich also folgendes Bild ab:

54 Natürlich sind einige wenige Ausprägungen von ‚Letztlaut' in Kombination mit anderen Faktoren systematisierbar, z. B.: *Letztlaut* = ‚*s*' führt bei heimischen Appellativen zu *-(s)es*.

Entgegen der landläufigen Meinung ist die Variation der Genitivmarkierung gar nicht so frei. Bei der Mehrheit der Token ist die Markierungsentscheidung bereits mehr oder weniger vorgegeben (vgl. Abbildung 4).

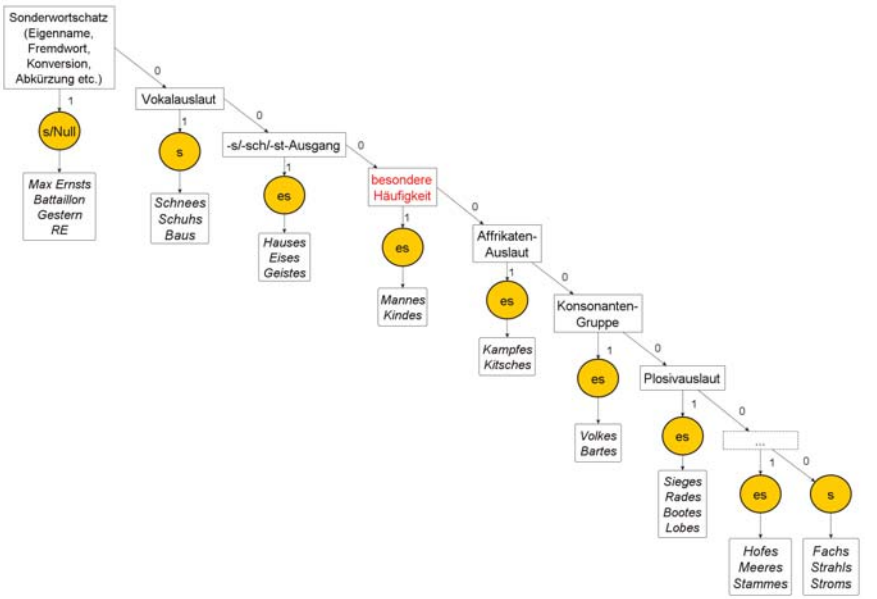

Abb. 4: Visualisierung einer komplexen Hypothese zur Markierungsvariation bei Einsilbern

Der vereinfachte Baum in Abbildung 4, der die Situation bei Einsilbern illustrieren soll, lässt sich folgendermaßen interpretieren: Der linke Bereich, der mit dem rot markierten Faktor ‚besondere Häufigkeit' endet und eine große Menge an Token involviert, ist deutlich weniger variant als der Bereich im rechten Teil der Abbildung. Die Variation in beiden Bereichen hat auch einen unterschiedlichen Charakter. Etwas stärker ausgeprägt sind im linken Bereich nur die Schwankungen zwischen der *s*-Endung und der Nullmarkierung beim Sonderwortschatz. Dagegen ist der rechte Bereich durch eine deutliche Variation zwischen der *es*-Endung und der *s*-Endung geprägt und die in der Abbildung vorgenommenen Zuordnungen sind nur als leichte Präferenzen zu verstehen, die u. U. bei Mehrsilbern schon viel schwächer durchschlagen. Hinzu kommt, dass in neuerer Zeit die Markierungsentscheidung für *-es* im Allgemeinen immer zögerlicher fällt, diese latente Verschiebung des Entscheidungsgewichts zugunsten von *-s* im Südosten womöglich aber etwas ausgebremst wird.

Abbildung 4 und das dazu Gesagte sind nur als eine sehr komplexe Hypothese zu verstehen, die die Organisation des Gesamtsystems der Markierungsvariation beschreibt. Diese Hypothese muss in ihren Bestandteilen noch präzisiert wer-

den. Dies wird im weiteren Verlauf unserer Untersuchungen zur Variation der Genitivmarkierung angestrebt.[55]

Literatur

Appel, Elsbeth (1941): Vom Fehlen des genitiv-s. (=Arbeiten zur Entwicklungspsychologie 21) München: Beck.

Baayen, R. Harald/Piepenbrock, Richard/Gulikers, Leon (1995): The CELEX Lexical Database (CD-ROM). Philadelphia: Linguistic Data Consortium.

Bubenhofer, Noah/Konopka, Marek/Schneider, Roman et al. (2013): Präliminarien einer Korpusgrammatik. (=Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, CLIP 4.) Tübingen: Narr.

Bubenhofer, Noah/Brinkmann, Caren/Hein, Katrin (2013): Maschinelles Lernen zur Vorhersage von Fugenelementen in nominalen Komposita. In: Bubenhofer, Noah/Konopka, Marek/Schneider, Roman et al.: Präliminarien einer Korpusgrammatik. (=Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, CLIP 4.) Tübingen: Narr, S. 183–227.

Drazin, Sam/Montag, Matt: Decision Tree Analysis using Weka. Machine Learning-Project II, University of Miami. http://wwww.samdrazin.com/classes/een548/project2report.pdf, zuletzt eingesehen am 30.06.2014.

Duden (2007): Band 9. Richtiges und gutes Deutsch. Wörterbuch der sprachlichen Zweifelsfälle. 6. vollst. überarb. Aufl. Mannheim/Leipzig/Wien/Zürich: Dudenverlag.

Duden (2009): Band 4. Die Grammatik: Unentbehrlich für richtiges Deutsch. 8. überarb. Auflage. Mannheim: Dudenverlag.

Fehringer, Carol (2011): Allomorphy in the German genitive. A paradigmatic account. In: Zeitschrift für Germanistische Linguistik 39/1, S. 90–112

Fuß, Eric/Konopka, Marek (2014): Variation der starken Genitivmarkierung. Manuskript. Mannheim: Institut für Deutsche Sprache.

Heister, Julian/Würzner, Kay-Michael/Bubenzer, Johannes/Pohl, Edmund/Hanneforth, Thomas/Geyken, Alexander/Kliegl, Reinhold (2011): dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. Psychologische Rundschau 62(1), 10–20.

Institut für Deutsche Sprache (2011a): Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2011-I (Release vom 29.03.2011). Mannheim: Institut für Deutsche Sprache. http://www.ids-mannheim.de/DeReKo, zuletzt eingesehen am 30.06.2014.

Kullback, Solomon/Leibler, Richard Arthur (1951): On information and sufficiency. Annals of Mathematical Statistics 22/1, S. 79–86.

Pfeffer, J. Alan/Morrison, Scott E. (1979): The genitive singular with *-s* and/or *-es* in spoken and written German. IRAL 17, pp. 303–311.

Pfeffer, J. Alan/Morrison, Scott E. (1984): The genetive singular with *-s* and/or *-es* in spoken and written German). In: Pfeffer, J. Alan (Ed.): Studies in Descriptive German Grammar. Heidelberg: Groos, pp. 9–18 (Abdruck v. Pfeffer/Morrison (1979).

Quinlan, J. Ross (1993): C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann.

55 Zu weiterführenden Analysen vgl. Fuß/Konopka 2014.

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing. Manchester, UK.

Schneider, Roman (2014): GenitivDB — a Corpus-Generated Database for German Genitive Classification. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik.

Szczepaniak, Renata (2010): *Während des Flug(e)s/des Ausflug(e)s*? German Short and Long Genitive Endings between Norm and Variation. In: Lenz, Alexandra N./Plewnia, Albrecht (eds.): Grammar between Norm and Variation. Frankfurt am Main: Peter Lang, S. 103–126.

Wall, Larry/Christiansen, Tom/Orwant, Jon/Schwartz, Randal (2001): *Programmieren mit Perl*. 2. Aufl. Köln: O'Reilly.

Wiese, Bernd (2009): Variation in der Flexionsmorphologie: Starke und schwache Adjektivflexion nach Pronominaladjektiven. In: Konopka, Marek/Strecker, Bruno (Hgg.): Deutsche Grammatik – Regeln, Normen, Sprachgebrauch. (= Institut für Deutsche Sprache – Jahrbuch 2008). Berlin, New York: de Gruyter, S. 166–194.

Witten, Ian H./Frank, Eibe (2005): Data Mining: Practical Machine Learning Tools and Techniques. 2. Aufl. San Francisco.