

Regularität und Variabilität der n-Tilgung im  
Lëtzebuergesch („Eifeler Regel“):  
Ein unüberwachtes, induktives Lernverfahren

Magisterarbeit im Fach Computerlinguistik

**Vorgelegt von**

Johannes Kiehl  
geboren am 27. Dezember 1968 in Hamburg

**Angefertigt am**

Lehrstuhl für Computerlinguistik/  
Quantitative Linguistik  
Fach Linguistische Datenverarbeitung  
Universität Trier

**Betreuer:** Prof. Reinhard Köhler

**Beginn/Abgabe:** 23.2.2001/29.8.2001

## **Eidesstattliche Versicherung**

Ich versichere, dass ich diese Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat.

Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Trier, den 29. August 2001

## Zusammenfassung

Diese Arbeit beschäftigt sich mit der empirischen Untersuchung einer Sprachregel des Lëtzebuergischen – der sogenannten „Eifeler Regel“, die die Tilgung des auslautenden -n beim Zusammentreten bestimmter Laut- und Wortpaare beschreibt – aufgrund eines maschinenlesbaren Textcorpus.

Auf diese Weise werden zuerst die Kontextbedingungen, dann die Eigenschaften der für die Tilgung zugänglichen Lexeme selbst untersucht. Schließlich wird ein unüberwachtes, regelinduzierendes Lernverfahren vorgestellt, das ein Modell der orthographischen Form dieser Lexeme entwickelt und nach seiner Genauigkeits-/Vollständigkeitsquote optimiert.

Die Implementierung der verschiedenen Softwaremodule erfolgt in Perl. Es wird gezeigt, dass mit dem beschriebenen Verfahren eine brauchbare Klassifikation der mit -n auslautenden Lexeme erreicht werden kann, besonders im Hinblick auf reale Texte des Lëtzebuergischen.

## Aufbau der Arbeit

Nach einer kurzen Einführung in das Thema wird in Kapitel 2 ein Abriss der lëtzebuergeschen Orthographie in ihrer historischen Entwicklung gegeben und das derzeit entstehende Korrektursystem CORTINA vorgestellt.

Die beiden folgenden Kapitel widmen sich der Darstellung der Fachliteratur zur n-Tilgung im Lëtzebuergeschen (Kapitel 3) und zur corpuslinguistischen Methodik (Kapitel 4) und ziehen erste Schlussfolgerungen für die weitere Untersuchung.

In Kapitel 5 und 6 werden die Anwendungsbedingungen der n-Tilgung empirisch verifiziert. Das folgende Kapitel 7 stellt die Verfahren des maschinellen Lernens vor und diskutiert eine Auswahl der relevanten Arbeiten zur Generalisierung morphologischer Regeln.

In Kapitel 8 werden diese Techniken in einem praktischen Lernsystem eingesetzt. Die Leistung dieses Systems wird quantitativ und qualitativ bewertet. Ein Ausblick (Kapitel 9) gibt Anstöße für weiterführende Untersuchungen.

# Inhaltsverzeichnis

- 1 Einleitung 7**
  - 1.1 Motivation und Methodik 8
  - 1.2 Bemerkungen zum Corpus 8
  - 1.3 Quellenlage zur -n-Tilgung 9
- 2 Automatische Prüfung der lëtzebuergeschen Rechtschreibung 10**
  - 2.1 Standardisierung der Orthographie 10
  - 2.2 Ziel und Umfang des Projekts CORTINA 14
  - 2.3 Vollformen-Spellchecker in Client-Server-Bauweise 14
  - 2.4 Integration der n-Tilgungsregel 16
- 3 Die lëtzebuergesche n-Tilgung in historischen und neueren Darstellungen 19**
  - 3.1 Historische Darstellungen 19
  - 3.2 Moderne Darstellungen 20
  - 3.3 Die „Eifeler Regel“ in Siebenbürgen 24
  - 3.4 Zusammenfassung 26
- 4 Methodisch-technische Aspekte der Corpusanalyse 28**
  - 4.1 Corpora in der Sprachwissenschaft 28
  - 4.2 Textauswahl 29
  - 4.3 Kodierung 32
  - 4.4 Planung und Messung von Repräsentativität 39
  - 4.5 Zusammenfassung 42
- 5 Kontextbedingungen der Tilgung 43**
  - 5.1 Lautliche und lexikalische Kontexte der Tilgung 43
  - 5.2 T-Wörter: Operationalisierung der Tilgung 43
  - 5.3 Lautlicher Kontext der n-Tilgung 45
  - 5.4 Lexikalischer Kontext der n-Tilgung 48
  - 5.5 Zusammenfassung 52
- 6 n-Tilgung als Parameter des Lexems 53**
  - 6.1 Lexem und Tilgung 53

6.2	Tilgungsklassen	53
6.3	Auswertung	55
6.4	Interpretation	58
<b>7</b>	<b>Automatischer Erwerb sprachlichen Wissens</b>	<b>61</b>
7.1	Einleitung	61
7.2	Paradigmen maschinellen Lernens	61
7.3	Besprechung ausgewählter Arbeiten	63
7.4	Zusammenfassung	67
<b>8</b>	<b>Verallgemeinerung von Wortbildungsregeln aus dem Corpus</b>	<b>68</b>
8.1	Vorüberlegungen	68
8.2	Stufe 1: Regelgenerator	70
8.3	Stufe 2: Justage der Parameter	72
8.4	Evaluation der generierten Regeln	75
8.5	Zusammenfassung	77
<b>9</b>	<b>Ausblick</b>	<b>79</b>
9.1	Lexem und Tilgung	79
9.2	Schlussbemerkung	81
<b>A</b>	<b>Literaturverzeichnis</b>	<b><i>lxxxii</i></b>
<b>B</b>	<b>Anhang</b>	<b><i>lxxxvi</i></b>

# 1 Einleitung

Im Lëtzebuergeschen – einer Sprache mit ca. 370 000 aktiven Sprechern im Großherzogtum und weiteren im angrenzenden belgischen Kanton Luxembourg<sup>1</sup> – hat sich eine Regel entwickelt, nach der -n am Wortende unter bestimmten Bedingungen im Satzzusammenhang getilgt wird. Im folgenden Text, einem Dialog aus einer Boulevardkomödie, wurden die Adjektivendungen grau unterlegt, um den Tilgungsvorgang zu verdeutlichen. Hier ist zu beachten, dass im Lëtzebuergeschen (wie auch im Niederländischen) die Adjektivformen für Nominativ und Akkusativ gleich lauten, „en neien Hutt“ also der Form nach beide Kasus tragen könnte.

*Fabi:* Ech brauch en neien Hutt fir op Munneref.

*Schnicky:* Fir drësseg?

*Fabi:* Et as e speziellen amerikanesche Modell.

*Ibby:* Kalifornesch.

*Fabi:* En extra zolitten.

*Ibby:* Si hun e getest. An engem kaliforneschen Tornado.

*Fabi:* En as nêt fortgeflunn. E soutz wéi ugegoss um Kapp.

*Schnicky:* Awer fir drësseg!

*Fabi:* En as ganz aus Seid. Vu schwaarzen asiatesche Raupen.

*Ibby:* Ganz rar Déiercher. Mussen e spezielle Containeren iwwert de Pazifik.

*Fabi:* A Kalifornie gefouert gin.<sup>2</sup>

Am Beispiel zeichnen sich die Anwendungsbedingungen der n-Tilgung ab: Am Satzende und vor *H-*, *T-* und Vokal-Anlaut bleibt das Suffix *-en* intakt, vor den Konsonanten *M-*, *R-* und *C-* wird es hingegen zu *-e* getilgt. So wie hier im dramatischen Dialog verhält sich das schriftliche Lëtzebuergesche auch in allen anderen Genres und Registern.

- 1 EURYDICE (2001), *Description nationale du Luxembourg, complément à l'étude d'Eurydice „L'enseignement des langues étrangères en milieu scolaire en Europe“* (Luxembourg: Ministère de la culture, de l'enseignement supérieur et de la recherche)
- 2 Übersetzt lautet der Dialog: Ich brauch einen neuen Hut für nach Mondorf. – Für dreiBig? – Es ist ein besonderes amerikanisches Modell. – Kalifornisch. – Ein ganz stabiles. – Sie haben ihn getestet. In einem kalifornischen Tornado. – Er ist nicht weggeflogen. Er saß wie angegossen auf dem Kopf. – Aber für dreiBig! – Er ist ganz aus Seide. Von schwarzen asiatischen Raupen. – Ganz seltene Tierchen. Müssen in speziellen Containern über den Pazifik. – Nach Kalifornien gebracht werden. – Aus G. REWENIG (1997), *Summerzauber* (Echternach: Ed. Phi) (zugleich Corpus-Text Nr. 44)

## 1.1 Motivation und Methodik

Anlass und Ausgangspunkt für die vorliegende Arbeit war das Projekt „CORTINA“, das die Entwicklung einer Software zur Rechtschreibprüfung für das Lëtzebuergesche unter Berücksichtigung der Bedingungen der n-Tilgung zum Ziel hat.

Was im vorigen Abschnitt skizziert wurde, gibt etwa die „offizielle“ Definition der n-Tilgung nach den Rechtschreibreformen von 1975 und 1999, in den Kapiteln 2 und 3 im Einzelnen besprochen, wieder. Es existieren jedoch eine ganze Reihe von Ausnahmen: Da gibt es Wörter, die selten oder nie das -n verlieren; da gibt es Folgewörter, vor denen die Tilgung ausgeführt werden kann, aber nicht muss. Der Gesetzestext bleibt hier ungenau und verweist auf die Intuition der Muttersprachler: „Mir schreiwën hei ëmmer, wéi mer schwätzen.“<sup>3</sup>

In dieser Arbeit soll ein Verfahren entwickelt werden, das weitgehend selbständig die Anwendbarkeit der Regel für einzelne Wortformen erkennt. Ziel ist die automatische Annotation dieser Worteigenschaft in Wortlisten wie dem Wörterbuch des Rechtschreibkorrektursystems.

Die Methode ist corpusbasiert, das heißt, die n-Tilgung wird nicht aufgrund präskriptiver Regeln, sondern aufgrund ihrer tatsächlichen Anwendung in geschriebenen und (verschrifteten) gesprochenen lëtzebuergeschen Texten untersucht. Eine solche empirische Untersuchung liegt bislang nur in wesentlich kleinerem Umfang für das gesprochene Lëtzebuergesch vor.<sup>4</sup>

Aus den empirisch gewonnenen Daten werden schließlich mit Hilfe eines maschinellen Lernverfahrens Erkennungsregeln erzeugt, mit denen es möglich wird, das Tilgungsverhalten von zuvor unbekanntem, im Corpus nicht belegten Formen vorherzusagen.

## 1.2 Bemerkungen zum Corpus

Für die Untersuchung wurde das „LuxText“-Corpus mit ca. 1,5 Millionen laufenden Wörtern freundlicherweise vom Luxemburger Institut Grand-Ducal zur Verfügung gestellt. Es enthält etwa zu gleichen Teilen literarische und politisch-juristische Texte aus den vergangenen 10 Jahren, daneben einen kleineren Anteil von gesprochener Sprache aus Protokollen des populären Fernsehmagazins „De Nol op de Kapp“ (dt. „den Nagel auf den

3 ARRÊTÉ MINISTÉRIEL (1975), Arrêté ministériel du 10 octobre 1975 portant réforme du système officiel d'orthographe luxembourgeoise, in *Mémorial. Amtsblatt des Großherzogtums Luxemburg B* 68, S.1385

4 P. GILLES (2001), Phonologische und variationslinguistische Aspekte der -n-Tilgung ('bewegliches -n') im Lëtzebuergeschen, erscheint in K. J. Mattheier & E. Radtke (Hrsg.), *Sammelband des Symposions 'Variation in der Sprache', Heidelberg Dez. 1995* (Frankfurt: Lang) (im Druck)

Kopf“), in dem meist zahlreiche Originaltöne von Bürgern aus den Dörfern und Gemeinden Luxemburgs zu hören sind.

Die Texte im Corpus stammen entweder von Schriftstellern, oder sie wurden von Leuten erfasst, zu deren täglicher Arbeit der schreibende Umgang mit Lëtzebuergesch gehört – Schreibkräften und Parlamentsprotokollanten. Im Fall der Parlamentsmitschriften ist es sogar üblich, dass die Rohprotokolle den Parlamentariern vorgelegt und gegebenenfalls noch einmal korrigiert werden. Das Luxtext-Corpus ist also nicht eigentlich spontansprachlich. Da jedoch die n-Tilgung, die hier untersucht werden soll, im 1975 veröffentlichten Gesetz nicht im Detail geregelt ist – „wir schreiben wie wir sprechen“ –, sollten die Texte im Corpus hier durchaus aussagekräftig für das Sprachgefühl der Autoren sein.

„Luxtext“ ist ein Rohtext-Corpus, in dem keine Wortstämme oder Grundformen annotiert sind. Das bedeutet, dass getilgte und ungetilgte Formen desselben Wortes nicht zweifelsfrei einander zugeordnet werden können. In der Untersuchung wird deshalb mit dem Konstrukt des „T-Worts“ gearbeitet, das mögliche Paare von getilgten und nicht getilgten Formen zusammenfasst. Diese Operationalisierung bildet naturgemäß auch Paare, die lexikalisch völlig verschiedene Einheiten darstellen. Die T-Wörter erweisen sich dennoch als brauchbare Objekte für die Beobachtung des Wirkens der n-Tilgung.

### 1.3 Quellenlage zur -n-Tilgung

Die n-Tilgung wurde im 18. und 19. Jahrhundert besonders für den – mit dem Lëtzebuergesch wie auch mit den moselfränkischen Mundarten verwandten – Siebenbürger deutschen Dialekt untersucht, weil sie hier, wie sich herausstellte, ein wichtiges regionales Unterscheidungsmerkmal darstellt.

Anfang des 20. Jahrhunderts war die so genannte „Eifeler Regel“ offenbar unter Forschern, die sich mit moselfränkischen Mundarten beschäftigten, weithin bekannt. Peter Christa<sup>5</sup> etwa widmet ihr in seinem „Wörterbuch der Trierer Mundart“ (1927) immerhin sechs Seiten.

Die heute einflussreichste Quelle ist wohl die 1955 veröffentlichte umfangreiche Darstellung des Luxemburger Linguisten Robert Bruch<sup>6</sup>.

Erst jüngst hat der bereits zitierte Peter Gilles ein prosodisch-phonologisches Modell der n-Tilgung im Luxemburgischen vorgelegt (vgl. Kapitel 3).

5 P. CHRISTA (1927), *Wörterbuch der Trierer Mundart* (Neudruck 1969, Wiesbaden: Sändig)

6 R. BRUCH (1954), *Das Luxemburgische im westfränkischen Kreis* (Luxemburg: Paul Linden)

## 2 Automatische Prüfung der lëtzebuergeschen Rechtschreibung

### 2.1 Standardisierung der Orthographie

Die Sprache Lëtzebuergesch – anfangs noch als „Lëtzebuenger Däitsch“ oder nur „Däitsch“ apostrophiert – entwickelt sich, parallel zur politischen Emanzipation des Großherzogtums im 19. Jahrhundert, zum „zentralen Ausdruck und Bestandteil nationaler Identität“<sup>7</sup>. Bis zum Beginn des Jahrhunderts bleibt die Sprache exklusiv auf den mündlichen Gebrauch beschränkt<sup>8</sup>, emanzipiert sich aber in den Jahren zwischen 1825 und 1890 zügig als Literatursprache. In dieser Zeit werden zunächst mundartliche Gedichte, Zeitungstexte und Volksstücke veröffentlicht. Schließlich, im letzten Drittel des Jahrhunderts, entstehen die Klassiker der mundartlichen Literatur<sup>9</sup>. Gleichzeitig mit dem literarischen setzt das lexikographische Interesse am Lëtzebuergeschen ein. 1847 erscheint mit J.F. Ganglers „Lexicon der luxemburgischen Umgangssprache“ ein erstes Wörterbuch<sup>10</sup>.

#### 2.1.1 Individuelle orthographische Systeme (1854-1900)

Für die frühen Veröffentlichungen lëtzebuergescher Texte sind orthographische Vorblätter typisch, wie sie Gangler seinem Wörterbuch voranstellt. Der Autor begnügt sich mit einem knappen Vorblatt zur „angenommenen Aussprache“. Gangler orientiert sich soweit wie möglich an der deutschen und französischen Schreibung, führt aber eigene Symbole für Vokale, Diphthonge und den Laut /ʒ/ ein.

Auch Michel Rodanges Versepos „Renert“<sup>11</sup> (1872) hat ein solches Vorblatt. Rodange zeichnet (jedoch anders als Gangler) Vokale, Diphthonge und die

7 G. BERG (1993), „*Mir wëlle bleiwe, wat mir sin.*“ *Soziolinguistische und sprachtypologische Betrachtungen zur luxemburgischen Mehrsprachigkeit* (Tübingen: Max Niemeyer), S.17

8 *Ibid.*, S. 63: „Bis ins 19. Jahrhundert hinein war Lëtzebuergesch schriftlos.“

9 F. HOFFMANN (1964), *Geschichte der Luxemburger Mundartdichtung* (Luxemburg: Bourg-Bourger), S.35 ff.

10 J.-F. GANGLER (1847), *Lexicon der Luxemburger Umgangssprache (wie sie in und um Luxemburg gesprochen wird)* (Luxemburg: V. Hoffmann)

11 M. RODANGE (1872), *Renert: oder de Fuuss am Frack an a Ma'nsgrësst*, zit. nach Werkausgabe 1974, *Gesamtwierk*, Klassiker vun der Lëtzebuenger Litteratur Bd. 1, C. Meder (Hrsg.) (Luxemburg: Kripler-Muller), S. 611

Konsonanten /ʒ/ und /l:/, /n:/<sup>12</sup> aus.

Antoine Meyer ergänzt 1854 ein Vorblatt, das ursprünglich von dem Lehrer Heinrich Gloden zu einem Band Meyers mit luxemburgischen Gedichten und Fabeln verfasst worden war<sup>13</sup>, um theoretische Überlegungen und veröffentlicht es als „Regelbüchelchen vum lezeburger Orthœgraf“<sup>14</sup>. Edmond de la Fontaine (genannt Dicks), der zwischen 1855 und 1891 zahlreiche Stücke und Erzählungen auf Lëtzebuergesch veröffentlicht hat, kritisiert in seinem „Versuch über die Orthographie der Luxemburger-Deutschen Mundart“ (1855), dass Gloden & Meyer hochdeutsche Wortstämme voraussetzen und sich mit der Übernahme der deutschen Schreibung „alle orthographischen Mängel der neuern Sprachen“<sup>15</sup> aneigneten. Er selbst entwirft in dieser Arbeit zum ersten Mal ein phonetisch motiviertes System:

Die Schriftsprache gibt überhaupt nur das Wort, nicht aber seine Bedeutung wieder, d.h., mittels der Schrift werden vor allem Laute, nicht Begriffe unterschieden. Hieraus fließt nun das oberste Gesetz der Orthographie: *Das Wort soll geschrieben werden wie es ausgesprochen wird.*<sup>16</sup>

Dieser Entwurf, dessen Schreibregeln einige Zeitgenossen (z.B. Michel Lentz) übernehmen, trägt bereits einige wesentliche Züge der modernen lëtzebuergesch Orthographie, wie sie seit 1950 aus dem „Luxemburger Wörterbuch“ (LWB) hervorgegangen ist:

- Vorrang der Lautung vor der Herkunft eines Wortes
- Vokalquantität wird durch die Zahl der folgenden Konsonanten bestimmt
- Vokalqualität und die spezifisch lëtzebuergesch Diphthonge werden als Akzentzeichen wiedergegeben

### 2.1.2 Staatliche Sprachpflege und Normierung (seit 1900)

Eine Sprachpolitik im engeren Sinne zugunsten des Lëtzebuergesch innerhalb der gewachsenen luxemburgischen Mehrsprachigkeit gibt es erst

12 Silbisches [l] und [n] gehören zum Phoneminventar des Luxemburgischen, vgl. BRUCH (1954), der die Bezeichnung „Schwebelaut“ verwendet (p. 77). Gilles (1999) gibt Minimalpaare an, die die bedeutungsunterscheidende Funktion des /n:/ zeigen sollen: P. GILLES, *Dialektausgleich im Lëtzebuergesch: Zur phonetisch-phonologischen Fokussierung einer Nationalsprache* (Tübingen: Niemeyer), S.78

13 A. MEYER (1845), *Luxemburgische Gedichte und Fabeln* (Brüssel: Delavigne & Callewaert)

14 A. MEYER (1854), *Règelbüchelchen vum Lezeburger Orthœgraf, en Uress als Pro'w, d'Fraèchen aus dem Hâ, a Versen* (Lüttich : H. Dessain)

15 E. DE LA FONTAINE (1855), Versuch über die Orthographie der luxemburger deutschen Mundart, zit. nach DICKS (1982), *Gesamtwerk* Bd. 3, Klass. vun der Lëtzebuenger Litteratur, C. Hury (Hrsg.) (Luxemburg: Kripler-Muller), S.12

16 Ibd., S.12

seit etwa 20 Jahren. Weber<sup>17</sup> fasst zusammen:

The linguistic situation of Luxembourg was not, traditionally, an object of planning and political action by the country's governments and administration. It used to be taken for granted until some 20 years ago.

Doch schon seit Beginn des 20. Jahrhunderts tritt der Staat als planende Instanz in Erscheinung. Das Kultusministerium beruft 1901 eine Wörterbuchkommission, die fünf Jahre später ein „Wörterbuch der luxemburgischen Mundart“ veröffentlicht, um eine Grundlage für die wissenschaftliche Untersuchung der Mundart zu schaffen, und zugleich den fortschreitenden Einfluss der Nachbarsprachen abzuwehren<sup>18</sup>. Für dieses Werk wurde eine einheitliche Schreibung mit Kollegen aus Lothringen und Siebenbürgen<sup>19</sup> abgestimmt. Der Germanist René Engelmann entwickelt bis 1914 ein weiteres System, das als „Engelmann-Weltersche“ Rechtschreibung ab 1912 auch im neu eingeführten Unterrichtsfach „Lëtzebuergesch“ eingesetzt wird und damit staatlichen Segen erhält. Es ist, trotz zweier späterer Reformen von 1946 und 1975, vielfach noch unverändert in Gebrauch.

Die radikale Orthographiereform des Jahres 1946, entwickelt von dem Phonetiker Jean Feltes, sollte „Engelmann-Welter“ durch ein konsequent phonetisches System ersetzen. Sie hätte die lëtzebuergesche Orthographie vollständig vom deutschen Schriftbild gelöst. Gesellschaftliche Widerstände gegen die Reform verhinderten jedoch ihre Durchsetzung<sup>20</sup>. Obwohl das System offiziell bis 1975 in den Schulen gelehrt wurde, entfaltete es keine Breitenwirkung.

Bereits acht Jahre nach der Reform von 1946 entschieden sich die Autoren des 1925 begonnenen, nach dem Krieg in staatlichem Auftrag wieder aufgenommenen und 1950-75 in Einzellieferungen veröffentlichten Luxemburger Wörterbuchs (LWB), anstelle der Feltes-Orthographie die Engelmann-Weltersche Schreibung in systematisierter und modernisierter Form anzuwenden<sup>21</sup>. Mit dem Erlass von 1975<sup>22</sup> wurde die Reform von 1946

17 F. GRIN & N. WEBER (2000), Multilingualism and language policy in Luxembourg, in T.L. du Plessis & K. Deprez (Hrsg.), *Multilingualism and Government* (Pretoria: Van Schaik), S.91

18 K. K. KLEIN (1955), Das Luxemburger Wörterbuch aus siebenbürgisch-sächsischer Sicht, *Zeitschrift für Mundartforschung* 23, S.237

19 Bis in die 1930er Jahre wurde das Südsiebenbürgische als eng verwandte Mundart angesehen. Die damals von manchen Wissenschaftlern vertretene Sicht, nach der die deutschsprachigen Südsiebenbürger und die Luxemburger die gleiche Mundart aus einer „Urheimat“ in Sachsen bewahrt hätten, ist heute überholt, vgl. 3.3 unten.

20 F. HOFFMANN (1987), Pragmatik und Soziologie des Lëtzebuergesch, J.P.Goudaillier, J.P. (Hrsg.), *Aspekte des Lëtzebuergesch*, Beiträge zur Phonetik und Linguistik 55 (Hamburg: Buske), S.125

21 Luxemburger Wörterbuch (1950-1977) (Luxemburg: Linden), p. XLVI

22 ARRÊTÉ MINISTÉRIEL (1975), S.1384

offiziell außer Kraft gesetzt und die Schreibweise des LWB als verbindlich erklärt.

Sprachpolitische Maßnahmen in jüngerer Zeit waren die Erklärung des Lëtzebuergeschen zur Nationalsprache 1984 (neben dem Deutschen und Französischen als „Amtssprachen“) und die Orthographiereform des Jahres 1999<sup>23</sup>. Gegenwärtig bereitet die amtierende Mitte-Rechts-Regierung ein Gesetz vor, das die Einbürgerung in Luxemburg an den Nachweis von Grundkenntnissen des Lëtzebuergeschen binden will<sup>24</sup>.

### 2.1.3 Die lëtzebuergesche Orthographie von 1975/1999

Nach der Fertigstellung des LWB soll eine einheitliche Schreibung mit klaren und präzisen Regeln an den Schulen unterrichtet werden. Ein Erlass des Kulturministers setzt jenen von 1946 außer Kraft. Die offizielle Orthographie soll von nun an auf dem Luxemburger Wörterbuch basieren. In einem wohl von Alain Atten stammenden<sup>25</sup> Zusatz zu dem Erlass wird deren Prinzip als teils phonetisch, teils historisch-etymologisch beschrieben:

Eng Schrëft, déi sech jhust dem Gehéier no schreift, as dem Lëtzebuenger Dixonär séng nët: hallef verléisst se sech op d'Ouer, hallef op d'Gewunnecht. 'T as keng reng phonetesesch, ma och eng historesch.<sup>26</sup>

Die Schreibung der Wortstämme, auch die Groß- und Kleinschreibung orientieren sich am Deutschen. Die Schreibung integrierter Lehnwörter wie *Maschinist* und *Politik*, die sowohl im Deutschen als auch im Französischen gebräuchlich sind, „zéie mir wéinst der Endong .. dacks iwwert deen däitsche Leescht“<sup>27</sup> (d.h., sie wird der Endung wegen oft ‚über den deutschen Leisten geschlagen‘). Am Deutschen orientiert sich auch die Schreibung der Vokalquantität (ein einfacher Vokal vor einfachem Konsonanten wird, wie ein verdoppelter Vokal, lang ausgesprochen).

Die langen Konsonanten (Schwebelaute) werden nicht differenziert, phonologische Minimalpaare wie /stal/ – /stal:/ (,still' – ,Stall'), /van/ – /van:/ (,wenn' – ,Wanne') sind, von der Großschreibung abgesehen, nicht am Schriftbild zu unterscheiden.

Die Reform von 1999 bringt einzelne Vereinfachungen und Systema-

23 RÈGLEMENT GRAND-DUCAL (1999), Règlement grand-ducal du 30 juillet 1999 portant réforme du système officiel d'orthographe luxembourgeoise, in *Mémorial. Amtsblatt des Großherzogtums Luxemburg* A 112, S.2040-2048

24 R. GRAF (2001), Loi sur la nationalité: Compli-simplification, *woxx* 596 (6.7.2001), S.3

25 Atten war damals Mitglied der Wörterbuchkommission. Als Autor des (anonymen) *Annexe* nennt ihn Roth (1996): L. Roth, Zu der Schreifweis vun der Lëtzebuenger Sprooch. Internet: [http://www.eis-sprooch.lu/orthographie/eisschreifweis\\_1.asp](http://www.eis-sprooch.lu/orthographie/eisschreifweis_1.asp)

26 ARRÊTÉ MINISTÉRIEL (1975), S.1366

27 Ibd., S.1367

tisierungen; sie löst sich damit weiter von der deutschen Schreibung. So gilt etwa die Konsonantenverdopplung nach kurzem Vokal nun auch für die einsilbigen (Auxiliar-) Verben *sinn, hunn, ginn, stinn* und *dinn, dunn* als Varianten von *doen*. Die Schreibung flektierter Verben richtet sich nun systematisch nach der Schreibung der infiniten Form.

## 2.2 Ziel und Umfang des Projekts CORTINA

Im Januar 1998 richtete die luxemburgische Staatsregierung einen ständigen Sprachrat ein, den Conseil permanent de la langue luxembourgeoise (CPLL). Der CPLL koordiniert eine Reihe von sprachplanerischen Projekten:

- Planung und gezielten Ausbau des Lëtzebuergesch,
- Entwicklung einer vereinfachten Orthographie,
- Bearbeitung und Neuedition des Luxemburger Wörterbuchs (LWB) von 1950-1975
- Zusammenstellung ein- und mehrsprachiger Handwörterbücher.

Der CPLL empfahl unter anderem die Entwicklung eines Rechtschreibmoduls für gängige Textverarbeitungsprogramme, um den Status des Lëtzebuergesch als geschriebene Sprache zu stärken. Die praktische informatische Umsetzung übernahm im Auftrag die Abteilung CREDI (Cellule de recherche, d'étude et de développement en informatique) des staatlichen „Centre de Recherche Public – Gabriel Lippmann“. Das Projekt mit dem Titel „CORTINA“ (Correction orthographique informatique appliquée à la langue luxembourgeoise) wird von den Ministerien für Kultur, Hochschulwesen und Forschung sowie Erziehung, Berufsausbildung und Sport finanziert.

Im Rahmen der ersten Stufe des Projekts, zwischen Juni 2000 und Juni 2001, wurde ein Korrekturprogramm entwickelt, das anhand eines Vollformen-Wörterbuchs mit ca. 75 000 Einträgen falsch geschriebene Wörter erkennt und Alternativen vorschlägt, wobei der syntaktische und der semantische Satzzusammenhang unberücksichtigt bleiben. In einer jetzt anlaufenden Fortsetzung des Projekts soll das Wörterbuch auf 150 000 Vollformen erweitert und ein marktreifes Produkt entwickelt werden.

## 2.3 Vollformen-Spellchecker in Client-Server-Bauweise

Um den Programmieraufwand bei der Anpassung des Korrektursystems an verschiedene Betriebssysteme und Textverarbeitungsprogramme zu minimieren, entschieden sich die Entwickler für eine Client-Server-Architektur. Der Server ist portabel in Java implementiert; er stellt die Kernfunktionen des Systems bereit. Als Dialogschnittstelle zwischen Client und Server fungieren Sockets; auf dieser Schnittstelle wurde ein eigenes Protokoll spezifiziert. Für die Integration des Korrektursystems in eine beliebige Anwendung genügt der Bau eines anwendungsspezifischen Clientmoduls in einer beliebigen Programmiersprache. Voraussetzung ist lediglich, dass die Softwareumgebung der Client-Anwendung Sockets

bereitstellt.

Der Dialog in Tabelle 1 illustriert das Kommunikationsprotokoll zwischen Client (C) und Server (S).

Tabelle 1. Wort-plus-Kontext-Anfrage als Beispiel für das Client-Server-Protokoll.

---

```

C: (Initialize dialog, port 9000)
C: REQ_CHECK_CONTEXT_START           ; start word-in-context correction
C: 1, "."                             ; first word (length, string)
S: REP_NON_WORD                       ; reply: is not a word
C: 11, "Plaidéieren"                 ; second word
S: REP_NOK                             ; reply: not spelt correctly
S: 3                                   ; will send three alternatives
S: 10, "Planéieren"                  ; suggestion 1
S: 10, "Plazéieren"                  ; suggestion 2
S: 10, "Plädéieren"                  ; suggestion 3
C: REQ_CHECK_CONTEXT_END             ; end of context
C: REQ_CLOSE_CONNECTION

```

### 2.3.1 Funktionalität des Servers

Der Server beantwortet Anfragen (*requests*) zu einzelnen Wortformen; dazu greift er auf das Wörterbuch des Systems zu.

Das Serverprotokoll gibt hierfür drei Anfragen vor, die sich hinsichtlich der Größe des geforderten Kontextfensters unterscheiden: Anfragen nach

1. Einzelwort (REQ\_CHECK\_WORD),
2. Wort plus Kontext (REQ\_CHECK\_CONTEXT\_START) und
3. Wort im Kontext mit Tilgung (REQ\_CHECK\_CONTEXT\_ER\_START).

Bei der **Einzelwortanfrage** unterscheidet das System zwei Fälle:

- a. Wort ist bekannt,
- b. Wort ist unbekannt.

Im Fall b. schlägt das System als Korrektur die  $k$  ähnlichsten Einträge im Wörterbuch vor. Ähnlichkeit ist über die Distanz  $d$  definiert, es ist die Anzahl der Grapheme, die nicht zur längsten gemeinsamen Teilkette (LCS) der verglichenen Wörter gehören. Zwei Wörter gelten als „ähnlich“, wenn  $d$  minimal und dabei kleiner als eine Schwelle  $t$  ist. Der Korrekturvorschlag unterbleibt, wenn keine Einträge mit  $d < t$  gefunden werden.

Bei der **Wort-plus-Kontext-Anfrage** (vgl. Abbildung 1) muss ein weiterer Fall unterschieden werden:

- c. Unbekanntes Wort beginnt mit Majuskel und folgt auf Punkt.

In diesem Fall wird zusätzlich geprüft, ob die Form in Kleinschreibung im Wörterbuch vorhanden ist. Die Behandlung von Fall c. ist eine defensive Heuristik für die Satzende-Erkennung: Jede Großschreibung nach Punkt wird vorsichtshalber wie eine Satzgrenze behandelt.

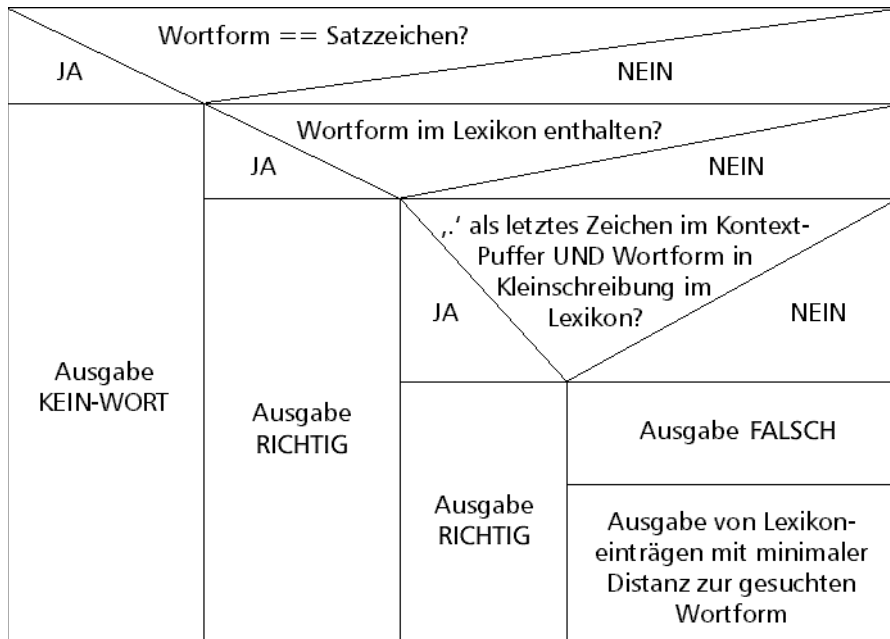


Abbildung 1. Ablaufschema der serverseitigen Bearbeitung einer Anfrage nach Wort plus Kontext

Bei der Anfrage nach **Wort im Kontext, mit Tilgung** wird zusätzlich geprüft, ob zwischen dem angefragten und dem folgenden Wort die n-Tilgungsregel berücksichtigt wurde, deren genaue Wirkungsweise das Thema der vorliegenden Arbeit ist.

### 2.3.2 Funktionalität der Clients

Im Rahmen des Projekts sind verschiedene Client-Prototypen entwickelt worden, darunter Makros für die Textverarbeitungsprogramme Word und StarWriter, sowie Online-Clients (Server-Side Java und JavaScript, Socket- und HTTP-Kommunikation).

Die Clients stellen die Benutzerschnittstelle bereit. Einige der Clients – die Word- und StarWriter-Makros und ein Online-Prototyp – verarbeiten ganze Sätze Wort für Wort. Diese Clients zerlegen (tokenisieren) den Eingabetext in einzelne Wörter und Satzzeichen und verwenden hierfür verschiedene Heuristiken.

Die ‚n‘-Tilgungsregel gilt nur innerhalb von Teilsätzen, die eine prosodische Phrase bilden, d.h., „ohne Pause“ gesprochen würden. Die Umsetzung dieser Regel betrifft also nur diejenigen Clients, die Wörter im Satzzusammenhang prüfen und korrigieren.

## 2.4 Integration der n-Tilgungsregel

Zwischen Februar und Mai 2001 veröffentlichten Luxemburger Zeitungen

und das Fernsehprogramm „De Magazin“ eine Internetadresse, unter der die verschiedenen Online-Prototypen aufgerufen werden konnten. Die Zahl der Zugriffe von außerhalb des Instituts stieg bald bis auf 2 000 Seitenaufrufe wöchentlich an.

Aus den vom System protokollierten Benutzereingaben geht zwar nicht hervor, welche der Anfragen aus Einzelwort- und welche aus Satzkorrektur-Umgebungen stammen; dennoch vermitteln sie einen ersten Eindruck davon, wie wichtig die Beherrschung der n-Tilgung durch das Korrektursystem ist.

Insgesamt wurden 239 218 Wortformen (Tokens) protokolliert, davon 28 363 verschiedene (Types). Betrachtet man daraus mittel- bis hochfrequente Formen ( $f > 20$ ) auf „-e“, so wird sofort deutlich, dass die Nichtberücksichtigung der n-Tilgungsregel zu einer hohen Quote falscher Zurückweisungen durch das Korrektursystem führt (Tabelle 2). Die Wörter in der ersten Spalte (30 Types) sind im Tilgungszusammenhang richtig, jedoch nur ungetilgt im Vollformenwörterbuch enthalten. Spalte zwei (16 Types) enthält französische und deutsche Wörter, die im Lëtzebuergesch mehr oder weniger schlecht lexikalisiert sind, Spalte drei (6) weitere unbekannte Wörter, Spalte 4 (2) systematische Fehlschreibungen (\**denen* wohl durch Interferenz mit dt. *den/denen*, ?*mee* dialektal [Luxemburg-Stadt] für *mä* „aber“).

Die Integration der Tilgungsregel in das Korrektursystem ist also, trotz des bewussten Verzichts auf die Prüfung der (Satz-) Grammatikalität, schon auf dieser Stufe des Projekts erforderlich, wenn eine Wortkorrektur aus dem Satzkontext ermöglicht werden soll.

Ziel der vorliegenden Arbeit ist es, zu prüfen, ob die dafür notwendige Eintragung der obligatorischen oder fakultativen n-Tilgung im Lexikon automatisch durchgeführt werden kann. Sinnvoll erscheint dies vor allem aus zwei Gründen: Zum Einen, um die Erweiterung des Cortina-Wörterbuchs mit nicht-annotierten Wortlisten zu beschleunigen; zum Anderen, um Benutzer, die ein lokales (Benutzer-) Wörterbuch anlegen möchten, bei der Angabe der Tilgungseigenschaft selbst eingegebener Wörter unterstützen zu können.

*Tabelle 2.* Vom System zurückgewiesene Wörter auf „-e“ (Frequenz > 20, in Klammern angegeben).

<b>n-Auslaut-Tilgung</b>	<b>franz. oder dt.</b>	<b>sonst. unbekannt</b>	<b>orthogr. Fehler</b>
Hie (37)	Je (24)	Alice (51)	dene (26)
Kuebe (122)	Navette (24)	Claude (21)	(r.: <i>deene/n</i> )
Nee (27)	Sie (36)	Josée (25)	mee (176) (r.: <i>mä</i> )
Plaze (122)	die (65)	he (37)	
alle (97)	je (28)	ne (85)	
dee (194)	le (190)	the (99)	
deene (178)	nie (32)		
dèse (45)	notre (66)		
e (1336)	que (71)		
eise (145)	rue (21)		
grousse (35)	sie (111)		
gudde (41)	texte (43)		
géinge (102)	tourne (52)		
hate (23)	une (25)		
hie (82)	ville (23)		
hire (37)	votre (48)		
jonke (23)			
kenne (23)			
könne (90)			
leschte (50)			
menge (21)			
misse (22)			
neie (34)			
nächste (27)			
nëmme (39)			
onse (24)			
schéine (30)			
soe (22)			
wie (23)			
wousste (113)			
zesumme (31)			
zweete (23)			
Äre (32)			
éischte (42)			

### 3 Die lëtzebuergesche n-Tilgung in historischen und neueren Darstellungen

Der Prozess der Tilgung von auslautendem -n gilt im Lëtzebuergeschen als sehr systematisch und kaum regionaler und individueller Variation unterworfen. Es gibt jedoch systematische Variation: Wörter, die das auslautende -n unabhängig vom Folgewort immer behalten (a., rechte Seite), und Folgewörter, vor denen getilgt werden kann, aber nicht muss (b., rechte Seite); vgl. die folgenden Paare aus dem Corpus (C48):

- a. ...duerchgezu\_ gët... – ...Eisebunn gët...  
 ...geschwë\_ geléiert... – ...dënn gemat...  
 ...virera\_ keng... – ...momentan keng...  
 ...Wäi\_ kaalgestallt... – ...Terpetäin verréiert...
- b. ...a\_ si weidergefuer... – ...an sou weider..., *auch*: ...a\_ sou weider...

Für diese Ausnahmen werden verschiedene Erklärungen genannt: So werden etwa dynamischer Akzent, die (geographische) Herkunft eines Wortes, seine diachrone Lautentwicklung oder der rheinisch-fränkische Tonakzent für die ausbleibende Tilgung verantwortlich gemacht.

In einigen anderen Sprachen und Mundarten – etwa in Eifeldialekten, im Trierer Dialekt, in südniederländischen und in Siebenbürger deutschen Dialekten – sind analoge oder verwandte Erscheinungen belegt<sup>28</sup>. Diese variieren jedoch häufig stark im Gebrauch und sind teilweise im Verschwinden begriffen, während die n-Tilgung im Luxemburgischen hochsystematisch zu sein scheint (möglicherweise unter fortschreitender Aufgabe der Ausnahmen).

In der älteren Literatur werden meist die Begriffe „Eifeler Regel“ oder, in teilweiser Vermischung mit anderen Phänomenen, „bewegliches n“ (bzw. französisch „n mobile“) verwendet. Gilles<sup>29</sup> schlug jüngst die Bezeichnung „n-Tilgung“ vor, um den phonologischen Status der Tilgungsregel zu betonen.

#### 3.1 Historische Darstellungen

Die n-Tilgung in verschiedenen west- bzw. moselfränkischen Mundarten

28 PAULIDES (1952) weist darauf hin, dass die „Eifeler Regel“ eine Entsprechung im Brabantischen hat, jedoch sind nur Flexionssilben betroffen. Ein ähnliches Beispiel präsentiert CHRISTA (1927) für den Trierer Dialekt. Vgl. J.-P. PAULIDES, *Das Luxemburgische im Spiegel des Niederländischen*, S. 73; und P. CHRISTA, S.7-12

29 GILLES (1999), S.222

wird seit dem 19. Jahrhundert in dialektkundlicher Literatur beschrieben. Eine der frühesten Darstellungen stammt wohl von dem Luxemburger Schriftsteller de la Fontaine (Dicks), der 1855 in seinem „Versuch über die Orthographie“ folgende Definition gibt:

*n* am Schlusse einer nicht betonten Endung fällt weg, wenn das darauf folgende Wort mit *b, f, g, j, k, l, m, p, r, s* und *w* anfängt ... Die Elision des *n* findet nicht statt:  
 1. Wenn die Endung volltonig ist. Z.B. *D'Kinéckin lèchelt* (Die Königin lächelt). ... 2. Wenn das darauf folgende Wort mit einem Vokal anlautet. ... 3. Wenn das darauf folgende Wort mit *d, h, n, t* und *z* anfängt.<sup>30</sup>

Während die phonetischen Tilgungsbedingungen bei Fontaine schon vollständig beschrieben sind, werden die morphologisch-lexikalischen Kriterien wohl noch nicht von ihm erfasst, wenn er hier die „Volltonigkeit“ der Auslautsilbe nennt; das gilt auch für Follmann, der 1886 zwar bereits die besondere Funktion der Hinterzungenvokale („u“, „uo“ in Follmanns Schreibung) in der Auslautsilbe erwähnt, jedoch eine zu strenge morphologische Beschränkung formuliert. Follmann vermutet – vielleicht unter dem Einfluss von Scheiners Siebenbürgischem Wörterbuch (vgl. Abschnitt 3.3 unten) –, nur bei Flexionssilben und den Stämmen von Funktionswörtern könne eine Tilgung stattfinden.

In der Literatur des frühen 20. Jahrhunderts wurde die n-Tilgung meist unter der Bezeichnung „Eifeler Regel“ behandelt. Diese Bezeichnung geht nach Capesius<sup>31</sup> auf Büschs 1888 erschienenen Aufsatz „Über den Eifeldialekt“<sup>32</sup> zurück, der offenbar besonders von Siebenbürger Mundart- und Volkskundlern gelesen wurde, die den Begriff prägten – Büsch selbst verwendet ihn nicht. In einer summarischen Darstellung, die weitere Assimilationsphänomene einschließt, spricht er vom „beweglichen n“.

## 3.2 Moderne Darstellungen

### 3.2.1 Robert Bruch (1953/1954)

Bruch behandelt in seiner zwischen 1953 und 1954 in drei Bänden veröffentlichten Arbeit zur Geschichte des Lëtzebuergeschen<sup>33,34</sup> ausführlich die n-Tilgung und ordnet ihre Erscheinungsformen lautgeschichtlich ein.

30 LA FONTAINE (1855), 19-20

31 B. CAPESIUS (1966), Die Behandlung des auslautenden n in den siebenbürgisch-sächsischen Mundarten (Die sogenannte "Eifler Regel"), *Zeitschrift für Mundartforschung* 33, S.97-126

32 TH. BÜSCH (1888), *Über den Eifeldialekt*, Programm, Malmedy (zitiert nach ibd., S.23)

33 R. BRUCH (1953), *Grundlegung einer Geschichte des Luxemburgischen*, Publications littéraires et scientifiques du ministère de l'éducation nationale (Luxemburg: Paul Linden), S.143

34 BRUCH (1954), S.25-30

Er definiert das „bewegliche -n“ (und zwar, anders als zuvor Büsch, ausdrücklich auf die Tilgung im Wort- und Morphemübergang bezogen) als

...die quantitative Nullvariante eines potenziell vorhandenen (etymologisch ererbten) Phonems, das nur noch unter bestimmten Umständen in Erscheinung tritt.<sup>35</sup>

Bruch formuliert eine „Hauptregel“ über die lautlichen Anwendungsbedingungen und die Anwendungsdomäne, und grenzt davon einzelne „Sonderfälle“ ab. Nach der Hauptregel bleibt die Tilgung (hier: Assimilation) des -n auf den Sandhi – d.h., den Ort des Zusammentreffens zweier Wörter oder Morpheme<sup>36</sup> innerhalb der Intonationsphrase – beschränkt:

In den lx. Maa. wird heutiges stamm- oder suffixauslautendes -n innerhalb einer Expirationseinheit (im Sandhi) oder in der Wortbildungsfuge vor allen Konsonanten außer h, d, t (und ts) assimiliert.<sup>37</sup>

Der Autor verweist auf verwandte Erscheinungen in der Eifel und in Brabant und ordnet das „bewegliche -n“ damit geographisch in seine Theorie von der westfränkischen Kreisbewegung (gegenüber Frings' „Rheinischem Fächer“<sup>38</sup>) ein.

Er unterscheidet acht „Sonderfälle“, in denen die n-Tilgung ganz oder teilweise unterbleibt:

1. Lehnwörter (*Jeanne, Lektoun, Maschinn, Roman*)
2. Schwebe-n (/n:/) (*Bann, Mann*)
3. historisches -nn, -nd (*dënn, dünn', blann, blind', Honn, Hunde'*)
4. historisches -ane (*Bunn, Bahn' < mhd. bane*)
5. das Präfix *on-*
6. Synkope eines intervokalischen ,g' (*Ren, Regen', Won, Wagen'*)
7. Diphthong aus historischem Langvokal (*fein < mhd. fîn, Loun < ahd. lôn*)
8. Folgendes *si, se, sech*

Damit liegt zum ersten Mal ein detaillierter Überblick über die Menge der nicht tilgenden Wörter vor, auch wenn ihre Klassifikation und die darin implizierten Erklärungsansätze noch teilweise unbefriedigend sind. Bruch selbst führt Ausnahmen für die Sonderfälle an: etwa *geschwë /geschwënn* zu 3., *hu /hunn* zu 4. Offensichtlich gibt es neben der Etymologie also zumindest weitere Faktoren, die die n-Tilgung steuern. Die Unterteilung der nicht tilgenden Formen auf *-oun* in eine Klasse „Lehnwörter“ (1.) und

35 Ibd., S.29

36 H. BUSSMANN (1990), *Lexikon der Sprachwissenschaft* (Stuttgart: Kröner), Stichwort ‚Sandhi‘

37 BRUCH (1954), S.25

38 Frings nannte so die mundartlichen Spuren des vom Südosten des Rheinlands nach Nordwesten hin abnehmenden Einflusses der Zweiten Lautverschiebung (nach

„historischer Langvokal“ (7.) erscheint willkürlich (dies kritisiert auch Gilles<sup>39</sup>).

Unter 8. zählt Bruch einzelne mit s- anlautende Folgewörter auf, vor denen die (sonst vor s- obligatorische) Tilgung fakultativ ist. Zuvor hatte schon Büsch beobachtet, „die einzige Ausnahme von der Regel ist in der 3. Pers. Plur., wenn das Pron. sie (se) nachgesetzt wird *schreiwen se*“<sup>40</sup>. Bruch nennt nun drei Pronomina, begründet dies aber nicht.

Bruchs Fassung der n-Tilgung lässt sich folgendermaßen zusammenfassen: Die Anwendbarkeit der Tilgungsregel ist vom vokalischen Nukleus der Endsilbe – hier besonders seiner diachronen Entwicklung – und dem Anlaut des Folgeworts abhängig. Darüber hinaus führt Bruch lexikalische Kriterien ein: Lehnwörter sind von der Tilgung ausgeschlossen, einige Pronomina fungieren nicht oder nur fakultativ als Tilgungskontext.

Der „Annexe“ des Orthographiegesetzes von 1975<sup>41</sup> enthält einige Bemerkungen zur „Eifeler Regel“, die auf die Bruch'schen Ausnahmen, abgesehen von 8. (vor *si, se, sech*), allerdings nicht eingehen. Die Absatzüberschrift betont den Vorrang der Sprechpraxis: „Nët geschwat an nët geschriwwen: Keen -n no der ‚Äifeler Regel‘.“ Der Hinweis „fir eenzel Silben hält se do stall, wou d'Wuert ouni -n onkloër gët“ deutet jedoch eher auf das Ziel einer Vereindeutigung im schriftlichen Ausdruck, als auf eine Übernahme der gesprochenen Praxis.

### 3.2.2 Peter Gilles (1999/2001)

Gilles<sup>42</sup> zeichnete im Rahmen seiner Dissertation zum Dialektausgleich in Luxemburg spontane und vorgegebene gesprochene Äußerungen von 26 Luxemburger Muttersprachlern auf. Aufgrund dieser Daten entwickelte er ein prosodisch-phonologisches Modell der n-Tilgung, das er jüngst in einer gesonderten Arbeit weiter ausgebaut hat<sup>43</sup>.

Gilles gelingt es, die von Bruch aufgezählten Sonderfälle in ein einheitliches Modell zu integrieren. Nach dem Vorbild von Wetzels<sup>44</sup> Analyse der Silbenstruktur des Südlimburgischen beschreibt er die n-Tilgung als Resultat einer phonologischen Silbenkontaktbeschränkung: In der prosodischen

BUSSMANN, u. diesem Stichwort).

39 GILLES (2001), S.4

40 BÜSCH (1888), S.23

41 ARRÊTÉ MINISTÉRIEL (1975), S.1384f.

42 GILLES (1999)

43 GILLES (2001)

44 W. L. WETZELS (1993), La phonologie de la flexion adnomiale dans un dialecte Limbourgeois (Pay Bas), in B. Laks & A. Riolland (Hrsg.), *Architecture des représentations phonologiques* (Paris: Editions du CNRS), S.203-231

Phrase werden extraprosodische Elemente entweder resibilifiziert (also z.B. erst mit der folgenden Silbe realisiert) oder, wenn die Silbenstrukturregeln der Sprache dies nicht zulassen, getilgt.

Für die Tilgung bzw. den Erhalt von *-n* ist nicht eine Regel verantwortlich, die in Abhängigkeit vom Folgekontext *-n* tilgt bzw. erhält, sondern es handelt sich vielmehr um einen Anpassungsprozess, der sich aus den Erfordernissen der Silbenstruktur ergibt.<sup>45</sup>

Vor den Konsonanten /d/ und /t/ wird das *-n* nicht getilgt. Gilles' Modell erklärt das damit, dass die drei Laute „...vollständige segmentelle Identität im *oral cavity*-Knoten“ aufweisen<sup>46</sup>; ihre phonologischen Merkmale sind also so ähnlich, dass /n/+/d/ bzw. /n/+/t/ sich zu einer partiellen Geminate (d.h., eine Konsonantenfolge, die mit unverändertem Vokaltrakt artikuliert wird) verbinden können:

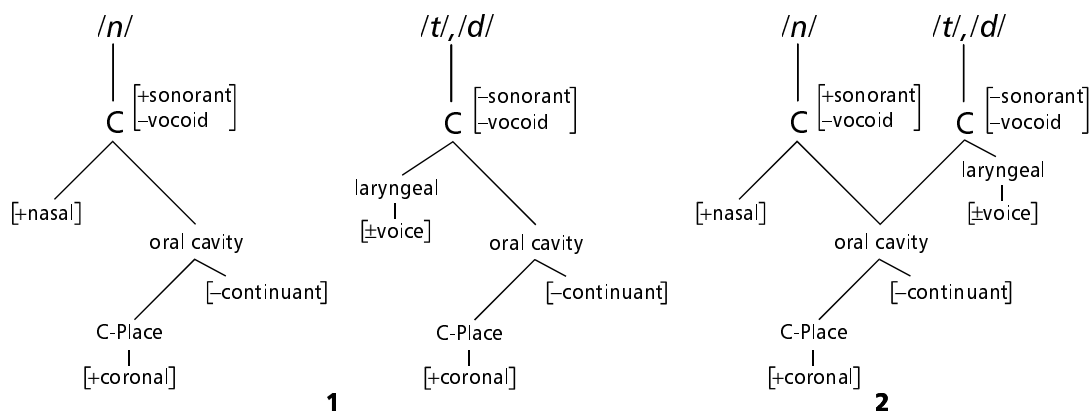


Abbildung 2. Verbindung der Merkmalsbäume zweier Konsonanten (1) zu einer ‚partiellen Geminate‘ (2) (nach Gilles, a.a.O., pp. 17/19)

Analog erklärt Gilles den *n*-Erhalt vor /h/: Der *oral-cavity*-Knoten ist bei diesem Laut nicht spezifiziert, das /h/ übernimmt die Ortsmerkmale des /n/ und beide Laute verbinden sich zu einer partiellen Geminate.

Die von Bruch genannten „Sonderfälle“ integriert Gilles in sein Silbenmodell der prosodischen Phrase:

- Die offenbar systematische Nichttilgung von Wörtern auf *-oun* und *-o:n* wird von ihm als Zusatzbedingung in die phonologische Kontaktbeschränkung aufgenommen; sie bleibt also ein – wenn auch phonologisch formulierter – unerklärter „Sonderfall“. (Die prosodische Phonologie kennt einen Regeltyp, sogenannte P1-Regeln, die auch auf grammatische Einheiten oder Gruppen von Lexemen zugreifen können.)

45 GILLES (2001), S.23

46 GILLES (2001), S.17

- Anders verhält es sich nach Gilles mit den von Bruch erwähnten Wörtern, die auf silbisches /n:/ (Schwebe-n) auslauten. Gilles zufolge führt der Schwebeton zu einer prosodisch-phonologischen Struktur, die der in Abbildung 2-2 verwandt ist: Der auf den Hochton folgende Konsonant erscheint geminatenähnlich verdoppelt. Wie schon dort entsteht eine zulässige Silbenstruktur, *ohne* dass das Auslaut-n getilgt werden muss.

Für die fakultative Tilgung vor einigen mit /z/ anlautenden Wörtern – Gilles fand hier in seinem Material die Pronomina *si/se, selwer, sech, sénger, sainen* und das Adverb *sou* – macht der Autor zwei Faktoren verantwortlich: Einmal handelt es sich um häufig enklitisch mit einem Verb verwendete Wörter, die möglicherweise in der Phonologie der Sprecher mit dem Verb zu einer untrennbaren Einheit verschmelzen. Als zweiten Grund nennt Gilles die Homorganität, d.h., die artikulationsphonetische Ähnlichkeit von /z/ mit /n/ (unterhalb des *oral-cavity*-Knotens unterscheiden sich die beiden Laute nur in einem Merkmal,  $\pm$ *continuant*). Unklar bleibt in Gilles Darstellung, warum der enklitische Gebrauch von Funktionswörtern die angenommene Silbenstrukturregel aufheben soll, während sie für Komposition und Affigierung offensichtlich gilt.

In seinem Corpus beobachtet Gilles bei den Äußerungen jüngerer Sprecher eine Tendenz zur Systematisierung der Tilgung vor /z/: Einige Sprecher tilgten vor /z/-Funktionswörtern ebenso wie vor anderen Wörtern mit /z/-Anlaut.

Zusammenfassend lässt sich Gilles' Modell der n-Tilgung als streng phonologisch beschreiben; lediglich die /z/-Funktionswörter, vor denen die n-Tilgung nur fakultativ gilt, bilden hier eine morphologisch-lexikalisch definierte Ausnahmeklasse. Einige der zuvor von Bruch genannten Ausnahmen bleiben allerdings ungeklärt (manche Wörter auf -ein, -äin; Lehnwörter).

### 3.3 Die „Eifeler Regel“ in Siebenbürgen

Das Südsiebenbürgische galt im 18. und 19. Jahrhundert als eng mit dem Lëtzebuergeschen verwandt; später setzte sich die Auffassung durch, dass eine frühe Sprachverwandschaft gemeinsam mit homologen Entwicklungsprozessen zu den augenfälligen Gemeinsamkeiten in Wortschatz und Grammatik geführt hat<sup>47</sup>. Zu diesen Gemeinsamkeiten gehört eine analoge n-Tilgungsregel, deren erste Fassung<sup>48</sup> in Binders schon 1795 erschienenem Aufsatz „Über die Sprache der Sachsen in Siebenbürgen“ an Gilles' ‚partielle Geminaten‘ denken lässt:

47 J. TOCKERT (1926), Ueber Luxemburgische Lexikographie, in *Luxemburgische Gesellschaft für Sprach- und Dialektforschung, Jahrbuch 1925*, S.30-51

48 Nach einer Angabe von CAPESIUS (1966), S.116

In den Zusammensetzungen und Verbindungen, wo der erste Theil des Wortes oder das erste Wort *n* hat, wird dieses (wie solches, was die Zusammensetzung betrifft, im Lateinischen, Griechischen, Hebräischen und anderen Sprachen vor mehreren Hauptlauten der Fall ist) in den folgenden Konsonanten, wenn er nicht *d* oder *t* ist, verwandelt; aber die Verdopplung der Konsonanten wird in der Aussprache nicht gehört.<sup>49</sup>

Arbeiten zu den siebenbürgischen Mundarten beeinflussten die luxemburgische Forschung nachhaltig. Noch 1901 wurde die Orthographie des ersten „Luxemburger Wörterbuchs“ mit Siebenbürger Wissenschaftlern abgestimmt, damit, wie Klein<sup>50</sup> später schreibt, zwei „in allem Wesentlichen gleiche“ Sprachen nicht durch verschiedene Schreibung „auf dem Papiere getrennt“ würden.

Die jüngste Formulierung der Tilgungsregel für das Südsiebenbürgische stammt von Capesius<sup>51</sup>: Die „Eifeler Regel“ wirkt sowohl im Satzzusammenhang (in „fließender Rede“) als auch in der Kompositionsfuge. Der Tilgung unterliegen Flexionsendungen und grammatische Wörter – also Präpositionen, Adverbien, Konjunktionen und Determinantien. Alle anderen stammauslautenden -n werden grundsätzlich nicht getilgt. Es gibt jedoch einzelne Gegenbeispiele, besonders in der Kompositionsfuge, z.B. [ʃti:gi:s], „Steingasse“. Diese veranlassen Scheiner zu der Vermutung, dass die Regel in einem früheren Sprachstadium allgemeinere Geltung hatte<sup>52</sup>. Ohne Ausnahme fallen die Derivationsilben für Diminutive, Adjektive und Feminina unter die Tilgungsregel. (Anders im Lëtzebuergeschen, wo viele Substantivstämme von der n-Tilgung betroffen sind, die Feminin-Derivationsendung *-in* dagegen nicht.) Der dem -n vorhergehende Vokal beeinflusst die Tilgung nicht.

Die n-Tilgung erfolgt dann, wenn das darauffolgende Wort mit einem nicht homorganen Laut beginnt; dies sind alle Konsonanten mit Ausnahme der Dentale /d/, /t/, /ts/ und des Spirans /h/. Capesius nimmt auch das /n/ in diese Reihe auf, obwohl die Frage noch nicht endgültig beantwortet werden könne, ob es sich in diesem Fall um eine Verschmelzung des auslautenden mit dem folgenden *n* handelt.

In den siebenbürgischen Darstellungen sind schon früh artikulationsphonetische Erklärungsansätze zu erkennen. So schreibt J. Roth 1887: „[n bleibt] wie es *teils seine Natur zulässt*, teils die Vermeidung des Hiatus mit

49 J. BINDER (1795), Über die Sprache der Sachsen in Siebenbürgen, *Siebenbürgische Quartalsschrift*, 4, S.386 (zit. nach CAPESIUS 1966)

50 *Korrespondenzblatt des Vereins für siebenbürgische Landeskunde* (1901), Nr. 24, S. 127 (zit. nach KLEIN 1955)

51 CAPESIUS (1966), S.97

52 A. SCHEINER (1896), Die Mundart der Siebenbürger Sachsen, *Forschungen zur deutschen Landes- und Volkskunde*, 9(2) (seitenidentischer Neudruck, 1971, Wiesbaden: Martin Sändig), S.35

Notwendigkeit fordert ... vor nachfolgenden Dentalen und Vokalen aufrecht.“<sup>53</sup> Scheiner sieht in der Tilgung das Resultat des typisch romanischen Satzakzents, die „Überwindung des Worttons durch den Satzton“<sup>54</sup>. Capesius weist diese Erklärung allerdings mit dem Hinweis auf die Wirksamkeit der „Eifeler Regel“ in der Wortfuge zurück<sup>55</sup>. Auch Capesius gibt eine artikulationsphonetische Erklärung:

Seiner Grundlage nach ist der *-n*-Ausfall zweifellos lautphysiologisch, d.h. er entspringt dem Bestreben nach Konsonantenerleichterung. Er tritt nämlich überall dort ein, wo die Artikulation des nächsten Lautes an einer anderen Stelle erfolgt als der des *-n*, und unterbleibt dort, wo die Artikulationsstelle nicht gewechselt wird. Zweitens ist er aber auch durch grammatikalische Kategorien (Sinngelänge) bedingt. Er erfolgt nur in den Flexionsendungen, in Bildungssilben und bei Indeklinabeln, wird aber vermieden, wo das *-n* zum Stamm eines Vollwortes gehört.<sup>56</sup>

Für die Ausnahmen – also diejenigen Fälle, in denen die n-Tilgung regelmäßig ausbleibt, obwohl die genannte Kontextbedingung gegeben ist – macht Capesius einerseits diachrone Faktoren, andererseits, nicht immer befriedigend, eine „morphologische Analogiewirkung“ verantwortlich. Wenn ein Wort sich historisch aus einer auf n + (Vokal, ...) auslautenden Form entwickelt habe, dann werde das übrig gebliebene Auslaut-n synchron nicht getilgt. Als „Analogiewirkung“ erklärt Capesius die seltene Tilgung des Stammaslauts von „Vollwörtern“, die von den Sprechern als Stamm + Derivationsuffix bzw. als Analogiebildung zu einem Determinans aufgefasst würden.

Capesius' Modell der „Eifeler Regel“ im Südsiebenbürgischer Deutsch lässt sich so zuspitzen: Die Anwendbarkeit der n-Tilgung hängt von morphologischen Eigenschaften eines Wortes (Flektierbarkeit, Funktion der Endsilbe) und von der Phonetik des folgenden Anlauts (Lautübergang) ab.

### 3.4 Zusammenfassung

Die n-Tilgungsregel im Lëtzebuergeschen scheint auf lexikalische und lautliche Eigenschaften sowohl des zu tilgenden Wortes selbst als auch des Folgewortes zuzugreifen. Das trifft auch auf Gilles' prosodisch-phonologisches Modell zu, das mittels besonderer Regeln auch auf grammatische Einheiten oder Gruppen von Lexemen zugreift.

Die Tilgungsregel bezieht sich im Lëtzebuergeschen stets auf den unmittelbaren Wortkontakt innerhalb einer Intonationsphrase.

Die Regelfassung der staatlichen Orthographiekommission, die 1975 und

53 J. ROTH (1887), in: *Korrespondenzblatt des Vereins für siebenbürgische Landeskunde* 10, S.94 (zit. nach CAPESIUS 1966, Hervorhebung von mir)

54 SCHEINER (1896)

55 CAPESIUS (1966), S.122

56 *Ibid.*, S.109

1999 festgeschrieben wurde, gibt die komplizierten Ausnahmen nicht an, ergänzt aber (präskriptiv), dass bei Einsilbern nur dann die Tilgung ausgeführt werden soll, wenn das Wort erkennbar bleibt.<sup>57</sup> Im von der Kommission beauftragten „Texte coordiné“, der als Erläuterung zum Gesetzestext erscheinen soll, werden weitere Einschränkungen vorgeschlagen: Bei Eigennamen, am Ende von Verszeilen und in Überschriften sollte auf die n-Tilgung verzichtet werden.<sup>58</sup>

Während die Tilgung in den Nachbarländern offenbar geschwunden ist, gibt es Anzeichen dafür, dass sie in L. – auch als abgrenzendes Merkmal einer kleinen europäischen Sprache – auf dem besten Weg ist, sich systematisch auf alle -n-Wörter auszudehnen, auch auf die hier untersuchten Ausnahmen. Dies scheinen Gilles' Daten über die Tilgung vor Funktionswörtern mit /z/-Anlaut zu belegen. Texte einiger Luxemburger Journalisten zeigen die systematische Tilgung auch von *-oun*, *-on*, *-un*<sup>59</sup>.

57 ARRÊTÉ MINISTÉRIEL (1975), S.1385

58 F. SCHANEN & J. LULLING (2001) Texte coordiné (unveröffentlicht), S.3

59 Nach mündlichem Hinweis von J. LULLING

## 4 Methodisch-technische Aspekte der Corpusanalyse

### 4.1 Corpora in der Sprachwissenschaft

Unter dem Begriff ‚Corpus‘ wird in der Sprachwissenschaft eine „endliche Menge von konkreten sprachlichen Äußerungen ... als empirische Grundlage für sprachwissenschaftliche Untersuchungen“ verstanden.<sup>60</sup> Seit den 1960er Jahren, als die Arbeit am maschinenlesbaren Brown-Corpus aufgenommen wurde, hat sich der Begriff zunehmend auf maschinell verarbeitbare Corpora verengt.<sup>61</sup>

Ein Corpus spiegelt die Sprache in ihrer konkreten Verwendung, ist also Ausdruck der Performanz in der von Chomsky eingeführten Dichotomie. Das schützt einerseits die corpuslinguistische Sprachbetrachtung davor, aufgrund unrealistischer Gewichtungen oder Überidealisierungen wichtige und aufschlussreiche Phänomene auszublenden. Immerhin ist nicht auszuschließen, dass sich vermeintliche Fehler der Sprecher als systematische, jedoch von der idealisierten, präskriptiv geprägten Grammatik nicht erfasste Regelmäßigkeiten herausstellen.<sup>62</sup> Bergenholtz/Mugdán sprechen vom „heilsamen Zwang“ zur Beschäftigung mit „unliebsamen Beispielen“.<sup>63</sup>

Andererseits muss die Corpuslinguistik sich mit den individuellen Fehlleistungen der Sprecher abfinden, jenen „grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic)“, die Chomsky ausdrücklich aus der Betrachtung durch die theoretische Linguistik ausnimmt.<sup>64</sup>

Das Corpus ist für die Linguistik nicht, wie Garside & al. ausführen, was die axiomatische Wissensbasis für die KI ist: es gestattet keine Inferenzen, sondern lediglich Häufigkeits- bzw. Wahrscheinlichkeitsaussagen. Die

60 BUSSMANN (1990), Stichwort ‚Corpus‘

61 G. LEECH (1997), Introducing corpus annotation, in R. Garside, G. Leech, A. McEnery (Hrsg.), *Corpus Annotation* (London: Addison Wesley Longman), S.1

62 G. SAMPSON (1987), Probabilistic models of analysis, in R. Garside, G. Leech, G. Sampson (Hrsg.), *The computational analysis of English* (London: Longman), S.27-28

63 H. BERGENHOLTZ & J. MUGDÁN (1989), Korpusproblematik in der Computerlinguistik: Konstruktionsprinzipien und Repräsentativität, in I. Bátori, W. Lenders, W. Putschke (Hrsg.), *Computational Linguistics* (Berlin: De Gruyter), S.143

64 N. CHOMSKY (1965), *Aspects of the theory of syntax* (Cambridge, Mass.: M. I. T. Press), S.3

besondere Stärke der corpuslinguistischen Methode liegt in ihrer großen Robustheit – der Fähigkeit, beliebigen Text zu verarbeiten. Zugleich nimmt sie aber stets ein gewisses Fehlerrisiko in Kauf.<sup>65</sup>

Linguistische Corpora unterscheiden sich hinsichtlich

- der Vielfalt (*diversity*) der in sie aufgenommenen Textsorten, Genres, Stilebenen, Autoren und Epochen,
- des Umfangs an annotierten linguistischen Beschreibungsebenen,
- der Differenziertheit (*granularity*) dieser Annotation<sup>66</sup>.

So wurden in das Brown-Corpus Stichproben aus 500 Texten und 15 verschiedenen Genres der Textsorte „gedruckte Prosa“ aufgenommen; die 1971 begonnene, halbautomatische Annotation galt morphologischen Wortkategorien und Lemmata. 77 Wortkategorien wurden unterschieden.<sup>67</sup>

Ein aktuelles, im Vergleich mit dem Brown-Corpus besonders umfangreiches Corpus ist das British National Corpus (BNC) mit Stichproben aus über 4 000 Texten (davon 20% Sprechsprache), die in etwa 100 Kategorien (Genre, Form der Veröffentlichung, demographische Angaben zu Autor oder Zielpublikum u.a.) eingeteilt wurden. Das BNC wird mit dem probabilistischen CLAWS-Tagger (65 Wortkategorien) annotiert.<sup>68</sup>

Weitere Typen von Corpora sind diachrone Corpora wie das Century-of-Prose-Corpus der Cleveland State University oder die seit 1990 entstandenen parallelen, mehrsprachigen Corpora, etwa Eurotra ET-10/63, Multext oder Crater.<sup>69</sup>

## 4.2 Textauswahl

### 4.2.1 Vorüberlegungen

Für die Zusammenstellung eines Corpus müssen qualitative und quantitative Entscheidungen getroffen werden:

- Welche Textsorten sollen repräsentiert sein?
- Wie lang soll der Zeitraum sein, aus dem die Texte stammen?
- Wie viele Texte sind erforderlich?
- Welche Länge oder Mindestlänge soll der einzelne Textausschnitt bzw. Text haben?

65 G. LEECH (1987), General introduction, in R. Garside, G. Leech, G. Sampson (Hrsg.), *The computational analysis of English* (London: Longman), S.3

66 G. LEECH (1997), S. 2-3

67 G. LEECH (1987), S.7, und (1997), S.8

68 BNC (2000), Composition of the BNC, Internet: <http://info.ox.ac.uk/bnc/what/balance.html>

69 T. McENERY, J.M. LANGÉ, M. OAKES, J. VÉRONIS (1987), The exploitation of multilingual annotated corpora for term extraction, in Garside & al., S.220

Diese Entscheidungen sind von theoretischen Vorannahmen abhängig. Das Ziel jeder corpuslinguistischen Untersuchung ist nach Engwall<sup>70</sup> „a theoretical conceptualization of the language system, the *langue*“. Die theoretische Begriffsbildung setzt jedoch Vorannahmen voraus, die ihrerseits die Zusammenstellung des sprachlichen Datenmaterials bestimmen: Welchen Einfluss haben diachrone Veränderungen auf die *langue*? Welche Äußerungssituationen und Textsorten sind zu unterscheiden? Spielen regionale oder demographische Faktoren eine Rolle?

Eine weiteres wichtiges Kriterium für die genannten Entscheidungen bildet das mit einem Corpus verbundene Forschungsinteresse. Je nachdem, auf welchen linguistischen Beschreibungsebenen die Beobachtungen interpretiert werden sollen, kann eine synchrone, dabei breit über verschiedene Textkategorien gefächerte Auswahl ebenso sinnvoll sein wie eine diachrone, die jedoch auf einige wenige Kategorien beschränkt bleibt. Je nachdem, ob der Untersuchungsgegenstand ein weit verbreitetes, leicht zu belegendes Phänomen darstellt oder eines, das nur in wenigen Texten überhaupt auftritt, schwankt die Zahl und Länge der erforderlichen Texte.

Um die Ausgewogenheit des Corpus bezüglich einzelner Texte und Autoren zu sichern, ist es üblich, gleich große Stichproben der einzelnen Texte zu erheben, indem man etwa eine feste Anzahl von Wörtern vom Beginn und vom Ende jedes Textes entnimmt. Eine Alternative zu Stichproben fester Größe besteht darin, die Messwerte bezüglich der unterschiedlich großen Stichproben zu normieren. Da viele Texteigenschaften nicht linear über den Text verteilt sind, wirft die Normierung jedoch eine Reihe ungelöster Fragen auf.<sup>71</sup>

Engwall beschreibt die Textauswahl als einen Prozess der schrittweisen Verfeinerung auf das vom Forschungsinteresse definierte Ziel hin. Biber<sup>72</sup> sieht sie als Teil eines ständigen Entwurfs- und Revisionsprozesses, dem ein Corpus unterzogen werden sollte (vgl. 4.4 unten). Grundsätzlich ist es wichtig, die Textauswahl möglichst genau aufzuschlüsseln, um die Replizierbarkeit der Experimente und die intersubjektive Interpretierbarkeit der Ergebnisse zu sichern. Erst durch genaue Spezifikation der Merkmale der einzelnen Texte lassen sich „Unterpopulationen“ bilden, um sie z.B. einer vergleichenden Untersuchung zu unterziehen.<sup>73</sup>

70 G. ENGWALL (1994), Not chance but choice: Criteria in corpus creation, in B.T. Atkins & A. Zampolli, *Computational approaches to the lexicon* (Oxford: University Press), S.50

71 Vgl. KÖHLER & GALLE (1993), Dynamische Eigenschaften von Textmaßen, in H.P. Pütz & J. Haller (Hrsg.), *Sprachtechnologie: Methoden, Werkzeuge, Perspektiven*, Sprache und Computer 13 (Hildesheim: Olms), S.3-15

72 D. BIBER (1993), Representativeness in corpus design, *Journal of literary and linguistic computing*, 8(4), S.243-257

73 ENGWALL (1994), S.78

### 4.2.2 Textauswahl im Rahmen dieser Arbeit

Das Luxtext-Corpus ist eine Textsammlung, die am staatlichen Institut Grand Ducal – Section de Linguistique et Onomastique (IDG-LEO) für die lexikographische Arbeit zusammengetragen wurde (das Institut arbeitet derzeit an einer Neufassung des LWB). Es enthält sowohl schriftliche als auch transkribierte mündliche Texte, sowohl literarische Sprache als auch Spontanäußerungen. Die Aufgliederung<sup>74</sup> nach Textsorten und Tokens zeigt Tabelle 3.

Tabelle 3. Gliederung des Corpus nach Textsorten.

Textsorte	Texte	Tokens
Fernsehreportagen	2	126 493
Parlamentsprotokolle	20 <sup>75</sup>	588 272
Literatur	26 <sup>76</sup>	762 120
Gesamt	48	1 476 885

Aus rechtlichen Gründen konnte ich das Corpus nicht als Ganzes analysieren, sondern musste ein Subcorpus mit den für diese Arbeit relevanten Tokens bilden. Dabei wurden voraussichtlich irrelevante Tokens übersprungen, aber als laufende Tokens gezählt (vgl. hierzu Abschnitt 4.3.3 unten).

Nach heutigen Maßstäben ist das Luxtext-Corpus ein kleines Corpus. Dennoch habe ich – besonders im Hinblick auf die Erweiterbarkeit der Arbeitsumgebung für Experimente an größeren Corpora – bei der Entwicklung der Corpuswerkzeuge die Spezifikation von Subpopulationen und die Entnahme von Stichproben fester Größe vorgesehen.

Jeder Text im Corpus erhielt einen Vorspann mit Entstehungsjahr, Ursprungsdatei und Autor (Tabelle 4).

Die einzelnen Corpus-Texte sind als ID-Datensatz in einer **Corpus-Indexdatei** (`efc.idx`) verzeichnet, die nach Fertigstellung und nach jeder Änderung oder Erweiterung des Corpus generiert wird.

Der Corpuszugriff erfolgt einheitlich über das hierfür entwickelte Perl-Modul `eifelExtract`. Eine **Corpus-Auswahldatei** (`efc.sel`) gibt an, welche Texte und Textausschnitte für das jeweilige Experiment sichtbar sein sollen. Um Texte auszuwählen, übernimmt der Experimentator die entsprechenden ID-Datensätze aus der Indexdatei in die Auswahldatei.

74 Der Gesamtumfang des Luxtext-Corpus beträgt ca. 1,8 Millionen Wortformen, die Corpuszusammenstellung war jedoch zum Zeitpunkt dieser Arbeit noch nicht abgeschlossen.

75 Die Parlamentsprotokolle (Protocoles de la Chambre des Députés) umfassen jeweils mehrere Einzelsitzungen, die hier nicht als Einzeltexte ausgewiesen sind.

76 25 der Texte aus dem Luxtext-Bestand sowie ein weiterer (vgl. Anhang).

Darüberhinaus kann mittels der optionalen Schlüsselwörter `sampling` und `spread` eine Stichprobe (in Tokens) und eine Verteilung dieser Stichprobe auf den jeweiligen Text (Anzahl der Unterteilungen) gewählt werden. Jede Stichprobe wird bis zum nächsterreichbaren Satzende verlängert, ihre Größe ist also größer oder gleich dem bei `sampling` angegebenen Wert. Mit `sampling = 2000`; `spread = 3` werden so beispielsweise aus jedem Text ca. 2 000 Tokens entnommen, davon je ein Drittel am Textanfang, in der Mitte und am Ende.

Das in Tabelle 5 dargestellte Beispiel zeigt, wie auf diese Weise ein Subcorpus mit je 40 000 laufenden Wortformen aus den Genres „Literatur“ und „Radiodiskussion“ ausgewählt wird.

*Tabelle 4.* Textvorspanne für einen literarischen (lit) und einen Parlamentstext (pol). Das Textgenre ist im Dateinamen kodiert.

---

```

<file name=lit.hosch.KaleKaffi.edite2.doc>      # Luxtext-Dateiname
<file wordcount=16573>                        # Tokens
<file year=2000>                              # Entstehungsdatum
<file author=Jhemp Hoscheit>                 # Autor

<file name=pol.ch.21VI2000.S39 99-00 3.doc>
<file wordcount=26156>
<file year=21.6.2000>

```

*Tabelle 5.* Corpus-Auswahldatei mit zwei Genres und je zwei Texten.<sup>77</sup>

---

```

# Selection from Index
# <Text-No>:<First-Line>:<Filename>:<Words-in-Text>:<Tokens-
  Sampled>:<Publication-Date>:<Author>
# Only first column compulsory

#### Literatur 1998
7:92315:E.lit.Pica.Hosch:91720:34815:1998:Jhemp Hoscheit
4:44291:E.lit.Kréiwénkel.Bra:48256:18530:1998:Josy Braun

#### RTL-Sendung "De No1 op de Kapp"
26:266612:P.lc.No1.1:74439:31943:unbekannt:kein
27:298557:P.lc.No1.2:52054:21691:unbekannt:kein

# Optional keys:
# Sampling = <words to sample from each text>
# Spread = <slices of sample to spread over each text>
sampling = 20000
spread = 4

```

### 4.3 Kodierung

Die Orthographie – und ihre Kodierung für den Rechner – repräsentiert den geschriebenen Text, während die Annotation eine über den Text

<sup>77</sup> Das Feld „Tokens sampled“ im ID-Datensatz gibt die Zahl der Tokens auf -Vokal oder -Vokal+n an, also die Zahl der möglichen Belege für die n-Tilgung.

hinausreichende, linguistische Interpretation darstellt.

Bei der (manuellen oder automatischen) Umkodierung schriftlicher Texte sollte möglichst viel Information aus dem Typoskript erhalten bleiben: Neben der Orthographie besonders auch die Typographie, die wichtige Schlüssel zur Textstruktur liefern kann.

Allerdings gibt es schon auf der orthographischen Ebene mehrdeutige Zeichen, die nicht nur kodiert, sondern auch (interpretierend) annotiert werden müssen. Geoffrey Leech<sup>78</sup> nennt hier

- einfache und doppelte Anführungszeichen,
- Großschreibung am Satzanfang,
- Punkt (Abkürzung, Satzende, Ordinalzahlen usw.), und
- Kursivstellung (Zitat, Emphase, u.a.).

Grundsatz jeder Kodierung und Annotation sollte sein, dass das Rohcorpus jederzeit rekonstruierbar bleibt.<sup>79</sup>

Der Begriff ‚Markup‘ wird für das Umkodieren, besonders aber für das Einfügen von Textstruktur- und typographischer Information in einen elektronischen Text benutzt.

### 4.3.1 LOB orthographic coding scheme

Die Markup-Konvention des LOB-Corpus unterscheidet zwischen Bezeichnern (*designators*) und Marken (*markers*). Bezeichner sind eindeutige Zeichenfolgen, die ein Zeichen aus der Textvorlage repräsentieren, das sich nicht im (7-Bit-) Zeichenvorrat des Corpus darstellen lässt. Marken tragen zusätzliche Information über vorhergehende oder folgende Symbole. So steht der Bezeichner \*?22 für das mathematische Wurzel-Symbol; die Marke " modifiziert einen vorhergehenden Vokal zu Vokal+Umlaut/Trema, die Marke \*1 kodiert die Textauszeichnung „kursiv“ für den folgenden Text.

Das LOB-Corpus wurde vollständig manuell erfasst. Dabei wurden bereits verschiedene Mehrdeutigkeiten der orthographischen Form aufgelöst (Tabelle 6).

Tabelle 6. Kodierung orthografisch mehrdeutiger Zeichen im LOB-Corpus.

Orthographie	Bedeutungsalternativen und Kodierung (∅ = nicht markiert)	
Großschreibung	^ Am Satzanfang	∅ sonst (z.B. Eigennamen)
Punkt	\0etc. Abkürzung	∅ Satzende
Apostroph	*' öffnendes, **' schließendes Anführungszeichen	∅ Apostroph in kontrahierten Formen
Anführungszeichen	*" öffnend, **" schließend	
Trennstrich	*- Gedankenstrich	∅ Trennstrich oder Minus

78 G. LEECH (1997), S.14

79 G. LEECH (1997), S.6

Im Corpus wird jedes Wort – von Leerzeichen begrenzte Zeichenketten; auch die Satzzeichen werden als „Wörter“ behandelt – in einer eigenen Zeile dargestellt.<sup>80</sup>

### 4.3.2 Text encoding initiative (TEI)

Die „Text encoding initiative“, 1987 von einzelnen Wissenschaftlern und Institutionen aus dem Arbeitsfeld Textarchivierung/Corpuslinguistik ins Leben gerufen, hat einen flexiblen Standard für die Kodierung und Annotation von Texten und für ihren elektronischen Austausch definiert. Im Kern zielt die TEI auf eine einheitliche Darstellung von Textinhalt und Textstruktur<sup>81</sup>. Den Zeichenvorrat für die Kodierung definiert ISO 646 (ein 7-Bit-Alphabet), die für das Markup definierten Ausdrücke sind Elemente der Metasprache SGML.<sup>82</sup>

TEI definiert eine Grundmenge (*core set*) von Texteigenschaften, die in jeder Darstellungsform eines Textes vorliegen. Die betreffenden Kodierungen und Annotationen sind für TEI-kodierte Texte obligatorisch, sollten also an jedem Text vorgenommen werden. Darüber hinaus sind zahlreiche weitere Beschreibungsmöglichkeiten vorgesehen, die sich in ihrer Funktion teilweise überschneiden. Dieser Teil der Markup-Sprache ist erweiterbar. Insgesamt definiert die aktuell veröffentlichte Fassung etwa 460 verschiedene Markup-Tags.<sup>83</sup>

Zur Grundmenge (*core*) gehört die Markierung von Absätzen, sowie auf Satzebene die Auszeichnung von Hervorhebungen, fremdsprachigen Passagen, Erläuterungen und zitierten (in Anführungszeichen eingeschlossenen) Passagen in ihren verschiedenen textfunktionalen Lesarten. Markiert werden Namen, Zahlen, Daten, Abkürzungen und Adressen, Gedichtzeilen und dramatische Dialoge.

Für Zeichen, die im Zeichenvorrat nicht zur Verfügung stehen, werden Bezeichner nach ISO 646 verwendet. Auch mehrdeutige Satzzeichen werden mit Hilfe eindeutiger Bezeichner kodiert. Die Übersicht in Tabelle 7<sup>84</sup> zeigt beispielhaft die definierten Lesarten der Zeichen ‚.‘ und ‚,‘ und ihre Kodierung.

80 B. BOOTH (1987), Text input and pre-processing: Dealing with the orthographic form of text, in: Garside & al, S.97-109

81 C.M.SPENBERG-McQUEEN & L.BURNARD (Hrsg.) (1999), TEI guidelines for electronic Text encoding and interchange (P3) (Chicago und Oxford: ACH/ACL/ALLC Text Encoding Initiative), S.10

82 Ibd., S.25

83 Ibd., S.30

84 Ibd., S.143

Tabelle 7. Kodierung einzelner Lesarten von Punkt und Komma nach TEI-Konvention.

TEI entity	Interpretation
stop.abbr	a stop used to end an abbreviation
stop.sent	a stop used to end a sentence
stop.abse	a stop used both to end an abbreviation and to end a sentence
stop.dec	a stop used as a decimal point
comma.dec	a comma used as a decimal point
midline.dec	a midline dot used as a decimal point
stop.space	a stop used as a numeric space character
comma.space	a comma used as a numeric space character

### 4.3.3 Encoding des Corpus für diese Arbeit

Das Luxtext-Corpus des 1998 aus der Wörterbuchkommission hervorgegangenen „Conseil permanent de la langue luxembourgeoise“ ist eine Sammlung von teils nach ISO 8859-1 („Windows“), teils nach Apples „Standard-Roman“-Zeichensatz kodierten Einzeltexten. In den Texten wurde fremdsprachiges Material, Ziffern und Eigennamen durch Helfer markiert, indem diese Elemente in spitze Klammern eingeschlossen wurden.

Um die Texte für die vorliegende Arbeit umzukodieren, wurde ein Perl-Script geschrieben, das anhand des im einzelnen Text vorgefundenen Zeichenvorrats automatisch zwischen Windows- und Apple-Kodierung unterschied und die Sonderzeichen in XML-Notation überführte.

Satzzeichen wurden in XML-ähnliche Bezeichner umgesetzt und, wie bei der Kodierung des LOB-Corpus, mit Leerzeichen umschlossen, sie werden also in den folgenden Verarbeitungsstufen wie Worttokens behandelt (jedoch nicht bei der Tokenzählung). Zeilenumbrüche im Original wurden ebenfalls kodiert. Tabelle 8 zeigt die bei der Kodierung verwendeten Bezeichner für Satzzeichen. Die einheitliche Benennung durch &punct... erleichtert gegenüber den analogen ISO-Bezeichnern die Darstellung als reguläre Ausdrücke, etwa für die manuelle Suche im Corpus; teilweise bedeutet sie bereits eine erste funktionale Annotation (vgl. &punctDots; für beliebig lange Ketten von Punkten, &punctMult; für emphatischen Gebrauch von Kombinationen aus ‚!‘ und ‚?‘, &punctAste;). Kontrahierte Artikel (z.B. *d’Gäärtner* für *de+Gäärtner*) bleiben im Corpus mit dem folgenden Wort verbunden; sie werden erst bei der Verarbeitung expandiert.

Tabelle 8. Verwendete Satzzeichen-Codes und ihre Entsprechung nach TEI-Konvention.

Bezeichner	Definition (EBNF)	ISO-8879-Entsprechung
&punctAste;	*{*}	&ast;, &ast;&ast; ...
&punctBr;	NUL = Zeilenumbruch	
&punctColo;	:	&colon;
&punctComm;	,	&comma;
&punctDots;	..{.}	
&punctExcl;	!	&excl;
&punctLang;	[	&lsqb;
&punctLbra;	\( <sup>85</sup>	&lpar;
&punctLquo;	(` ,)	&lsquo;
&punctMult;	(?!)(?!){?!} z.B. ?!, !?, !!!, ...	
&punctPeri;	.	&period;
&punctQues;	?	&quest;
&punctQuot;	"	&quot;
&punctRang;	]	&rsqb;
&punctRbra;	\)	&rpar;
&punctRquo;	(´ ´)	&rsquo;
&punctSemi;	;	&semi;
&punctSlas;	(\\ /)	&sol;, &bsol;

Für die Annotation fremdsprachiger Einschübe im Text definiert die TEI blockweise Start-Tag-/Ende-Tag-Auszeichnungen.

John eats a **<foreign lang=fr>croissant</foreign>** every morning.

**<mentioned lang=fr>Croissant</mentioned>** is difficult to pronounce with your mouth full.

A **<term lang=fr>croissant</term>** is a crescent-shaped piece of light, buttery, pastry that is usually eaten for breakfast, especially in France.<sup>86</sup>

Demgegenüber habe ich die im Ausgangscorpus als fremdsprachig (oder Eigenname, Ziffer) annotierten Passagen Wort für Wort ausgezeichnet. Gegenüber der blockweisen Auszeichnung in XML erleichtert dies die kontextfreie Verarbeitung von Einzelwörtern.

Ziel der Corpusuntersuchung ist die Betrachtung von Wörtern, die ein realisiertes oder möglicherweise getilgtes -n im Auslaut haben (vgl. T-Wort-Definition in Abschnitt 5.2 unten) und ihres Folgekontexts.

85 Ein vorangestelltes „\“ fungiert als Escape-Marke: Es signalisiert, dass das folgende Zeichen nicht als Metazeichen (EBNF-Symbol) zu lesen ist.

86 Vgl. SPERBERG-McQUEEN & BURNARD (1999), 124

Tabelle 9. Annotation fremdsprachiger Passagen (hier fett) und ihre Umkodierung.

Format des Luxtext-Corpus	Umkodierter Text
Well mir eis allerdings op d'Finanze wëlle konzentréieren, proposéieren ech, deen <débat d'orientation> zu engem spéideren Zäitpunkt, no der <Rentrée>, ze féieren.	Well mir eis allerdings op d'Finanze w&uml;lle konzentr&eacute;ieren &punctComm; propos&eacute;ieren ech &punctComm; deen <&punctLess;> <d&eacute;bat> <d'orientation> <&punctMore;> zu engem sp&eacute;ideren Z&auml;itpunkt &punctComm; no der <&punctLess;> <Rentr&eacute;e> <&punctMore> &punctComm; ze f&eacute;ieren &punctPeri;

Da das IGD-LEO aus rechtlichen Gründen keine zusammenhängenden Texte für eine externe Untersuchung zur Verfügung stellen konnte, wurden nur die zu betrachtenden Wörter extrahiert. Das Ausgabeformat des umkodierten Corpus besteht in einem T-Wort pro Zeile plus seinem Folgekontext von vier Worttokens (Satzzeichen zählen dabei nicht); jeder Zeile ist die laufende Tokennummer vorangestellt (auch hier zählen nur Worttokens, nicht Satzzeichen). Der Beispieltext in Tabelle 10 (folgende Seite) zeigt das Resultat der Corpus-Extraktion und -Umkodierung.

*Tabelle 10.* Ausgangsformat des Luxtext-Corpus (oben), Format des extrahierten Corpus (Mitte), teilweise rekonstruierter Ausgangstext (unten).

Ausgangstext
Här President, Dir Dammen an dir Hären, Mir hunn dat <20.> Jorhonnert verlooss, mee mir hunn et net definitiv hannert eis gelooss. Et war ee gewallegt Jorhonnert, vollgepaakt mat tragesche Momenter, zertrëppelte Mënschen an zerstéierten Dreem. Et war een déischtert Jorhonnert, awer och eent ...
Extrahiertes und umcodiertes Corpus
<file name=E.pol.Etat.nat.2000> <file wordcount=15262> 4: Dammen an dir H&auml;ren &punctComm; &punctBr; Mir 5: an dir H&auml;ren &punctComm; &punctBr; Mir hunn 7: H&auml;ren &punctComm; &punctBr; Mir hunn dat <&punctLess;> <20> 9: hunn dat <&punctLess;> <20> <&punctPeri;> <&punctMore;> Jorhonnert verlooss 14: &punctComm; mee mir hunn et net 16: hunn et net definitiv &punctBr; hannert 25: ee gewallegt Jorhonnert &punctComm; vollgepaakt mat 30: tragesche Momenter &punctComm; zertr&euml;ppelte &punctBr; M&euml;nschen an 32: &punctComm; zertr&euml;ppelte &punctBr; M&euml;nschen an zerst&eacute;ierten Dreem 33: &punctBr; M&euml;nschen an zerst&eacute;ierten Dreem &punctPeri; Et 34: an zerst&eacute;ierten Dreem &punctPeri; Et war 35: zerst&eacute;ierten Dreem &punctPeri; Et war een 39: een d&auml;ischtert Jorhonnert &punctComm; awer och ...
Rekonstruierter Ausgangstext
[?] [?] [?] Dammen an dir Hären, Mir hunn dat <20.> Jorhonnert verlooss, mee mir hunn et net definitiv hannert [?] [?] [?] [?] ee gewallegt Jorhonnert, vollgepaakt mat tragesche Momenter, zertrëppelte Mënschen an zerstéierten Dreem. Et war een déischtert Jorhonnert, awer och [?] ...

Um einzelne Beispiele im Satzkontext nachzulesen, wäre manchmal eine Rekonstruktion des Ausgangstexts wünschenswert. Die geforderte Beschränkung auf relevante Tokens erlaubt dies nur teilweise, Position und Anzahl der fehlenden Tokens bleiben jedoch sichtbar (vgl. Tabelle 10). Für die Rekonstruktion von Kontexten wurde ein Konkordanz-Werkzeug entwickelt.

## 4.4 Planung und Messung von Repräsentativität

### 4.4.1 Repräsentativität

Der Begriff der Repräsentativität eines Corpus ist eng mit dem der statistischen Stichprobe verbunden. Eine Stichprobe soll in allen wesentlichen Eigenschaften der Grundgesamtheit entsprechen, auf die sie sich bezieht.

Auf Corpora übertragen, gehören hierzu nach Biber zwingend sowohl Bedingungen der Sprachverwendung (Verwendungssituation) als auch linguistische Eigenschaften der Texte:

Representativeness refers to the extent to which a sample includes the full range of variability in a population.<sup>87</sup>

Allerdings kann ein Corpus nur bezogen auf eine mehr oder weniger künstliche Grundgesamtheit, niemals auf eine Sprache bzw. einen synchronen Sprachzustand, und nur hinsichtlich einer Zahl von Merkmalen repräsentativ sein. Bergenholtz & Mugdan plädieren deshalb dafür, besser ein „exemplarisches Corpus“ anzustreben.<sup>88</sup>

### 4.4.2 Bestimmung einer Grundgesamtheit

Um Repräsentativität für ein Corpus zu erreichen, muss also zunächst die Grundgesamtheit definiert werden – etwa auf der Grundlage einer umfassenden Bibliographie (so bei LOB). Sollen allerdings unpublizierte Texte (Sprechsprache, Briefe, e-Mails) einbezogen werden, muss der Rahmen der Stichprobe (*sampling frame*) durch theoriegeleitete Kategorisierung festgelegt werden.

Biber nennt als entscheidenden Vorzug einer Kategorisierung – also einer modellhaften Strukturierung und Hierarchisierung der Grundgesamtheit –, dass so für ein ausreichendes Gewicht der linguistisch interessanten Eigenschaften gesorgt werden kann: eine nur demographische Strukturierung der Grundgesamtheit spiegelt die Verwendungshäufigkeit bestimmter Merkmale, aber nicht ihre Wichtigkeit.

It is not necessary to have a corpus to find out that 90% of the texts in a language are linguistically similar (because they are all conversations); rather, we want to analyse the linguistic characteristics of the other 10% of the texts, since they represent the large majority of the kinds of registers and linguistic distributions in a language.<sup>89</sup>

Eine solche bewusste Gewichtung der Daten steht im Widerspruch zur

87 BIBER (1993), S.243

88 Nach BERGENHOLTZ & MUGDAN (1987), S.147, ein ursprünglich von Bungarten 1979 vorgeschlagener Begriff.

89 BIBER (1993), S.248

erklärten Absicht, die Sprachverwendung zum Gegenstand zu machen. Darauf weisen McEnery & Wilson<sup>90</sup> hin: Kategorien bedeuten immer eine Interpretation durch den, der sie bildet:

...these strata, like corpus annotation, are an act of interpretation on the part of the corpus builder because they are founded on particular ways of dividing up language into entities such as genres which it may be argued are not naturally inherent within it: different linguists may specify different genre groupings according to their theoretical perspectives on linguistic variation.

Sie sprechen sich für eine Kombination aus einem demographischen Verfahren (z.B. Wahl der Informanten aufgrund von Alter, Geschlecht, Geburtsort, Schichtzugehörigkeit) und einem theoretisch-pragmatischen Vorgehen (Festlegung „wichtiger“ Sprechakte und Äußerungssituationen) aus.

Werden demographische Parameter für die Textauswahl herangezogen, so setzt dies selbstverständlich voraus, dass ein statistischer Zusammenhang zwischen diesen Parametern und den Spracheigenschaften erwartet wird, die man beobachten möchte.<sup>91</sup>

#### 4.4.3 Auswahl der Stichproben

Stichproben, die repräsentativ für eine Grundgesamtheit sein sollen, müssen randomisiert sein, also eine Zufallsauswahl darstellen.

Bergenholtz & Mugdan<sup>92</sup> unterscheiden zwei Varianten:

- Randomisierte Stichprobe
- Geschichtete Stichprobe

Bei Texten ist neben der Zahl und der Länge der Stichproben auch der Ort der Entnahme (z.B. Kapitel) eine wichtige Größe. Zur Distribution linguistischer Merkmale innerhalb und zwischen Texten hat Biber<sup>93</sup> experimentelle Daten vorgelegt.

Während die randomisierte Stichprobe mit einem Zufallsverfahren aus der gesamten Population entnommen wird, setzt die geschichtete Stichprobe voraus, dass zuvor eine Kategorisierung der Population – beispielsweise in Textsorten, Verwendungssituationen o.ä. – vorgenommen wurde. Die Stichproben werden dann zufällig innerhalb der einzelnen Kategorien (Schichten, Strata) ausgewählt.

Die Stichprobengröße sollte so gewählt werden, dass Anzahl und Distribution der zu beobachtenden sprachlichen Einheiten möglichst

90 T. McENERY & A. WILSON (1996), *Corpus Linguistics* (Edinburgh: UP), S.65

91 BERGENHOLTZ & MUGDAN (1987), S.146

92 Ibd., S.147f.

93 BIBER (1993), S.248-255

repräsentativ für den Gesamttext sind. Biber untersuchte verschiedene lexikalische und syntaktische Eigenschaften (Zahl bestimmter Wortarten, Verbkategorien, Modi und Satztypen) und stellte fest:

- (1) Häufige Eigenschaften zeigten stabile Distribution schon über die ersten 2 000 Wörter eines Textes, während seltenere Eigenschaften größere Variabilität aufwiesen;
- (2) Eigenschaften, die *im Text* nichtlinear verteilt waren, variierten auch stark *zwischen* Texten.

Wenn das zu untersuchende Merkmal im Pilotcorpus eine Verteilung nach (2) aufweist, dann sollte der Diversität im Corpus gegenüber der Länge der einzelnen Textausschnitte der Vorzug gegeben werden.

#### 4.4.4 Messung von Repräsentativität

Biber beschreibt Corpusgestaltung als einen zyklischen Prozess, der die Stadien Gestaltung – Textsammlung – empirische Evaluation mehrfach durchlaufen sollte.

Folgende Fragen sichern dabei die Repräsentativität des Corpus bezüglich der Vielfalt an sprachlichen Erscheinungen in der Grundgesamtheit:

1. Sind die einzelnen Stichproben groß genug?
2. Genügt die Gesamtzahl an Stichproben im Corpus?
3. Reichen die Stichproben für jede Schicht (*register*), um die Distribution aller untersuchten Eigenschaften für jede einzelne Schicht zuverlässig messen zu können?

Biber stellt eine Reihe deskriptiv-statistischer, univariater Methoden vor, um diese Fragen zu beantworten:

**Stichprobengröße:** Die Stichproben – also die einzelnen Textausschnitte, die das Corpus bilden – sind erst dann lang genug gewählt, wenn die Ergebnisse unabhängiger Messungen bestimmter Eigenschaften jeweils eines Textausschnitts anhand verschiedener, gleichlanger Stichproben daraus eine hohe Korrelation zeigen. Biber teilt also jeden Text in zwei bis vier kürzere Abschnitte, misst für jeden getrennt verschiedene typische Merkmale und berechnet je Merkmal einen Korrelationskoeffizienten (*reliability coefficient*). Dieser muss positiv oder größer 0,5 sein (abhängig von der gewählten Korrelation).

**Gesamtzahl der Stichproben:** Diese genügt, wenn der Standardfehler der Mittelwerte jedes zu messenden Sprachmerkmals unter einer zuvor definierten Schwelle bleibt. Dieser „tolerierbare Fehler“ (*tolerable error*) wird getrennt für jedes zu beobachtende Sprachmerkmal abhängig von dessen Mittelwert und Standardabweichung in einem „Pilotcorpus“ festgelegt.

**Zahl der Stichproben je Schicht (*register*):** Diese genügt, wenn die Varianz der Messwerte in den einzelnen Schichten gleich ist. Für jede Schicht wird

die Standardabweichung der Messwerte jedes Merkmals berechnet und mit deren Durchschnitt normiert (Standardabweichung/Mittelwert). Die Summe dieser normierten Abweichung (*normalized deviation*) sollte zwischen den einzelnen Schichten konstant sein.

Multivariate Verfahren (Faktorenanalyse und Clusteranalyse) bieten weitere Möglichkeiten für die Analyse und Bewertung der Variationsbreite und -typen im Corpus und zwischen Registern.<sup>94</sup>

#### 4.5 Zusammenfassung

Mit dem „Luxtext“-Corpus des IGD lag ein Textcorpus des Lëtzebuergeschen vor, das Texte aus drei Genres über einen Zeitraum von ca. 10 Jahren enthält, davon 8% schriftlich erfasste gesprochene Sprache.

Es wurde ein Softwaremodul für den Corpuszugriff entwickelt, das die Entnahme von Teilstichproben fester und variabler Größe ermöglicht.

Das Corpus ist ein Rohtext-Corpus, in dem lediglich Eigennamen, Ziffern und fremdsprachige Passagen neutral markiert sind. Die folgenden Untersuchungen können daher nicht auf lexikalische Grundformen oder Kategorien zurückgreifen.

Ein Ansatz zur Messung der Repräsentativität von Corpora konnte hier nur referiert werden. Grundsätzlich stellt der geringe Umfang des Corpus – bedingt durch die begrenzte Textproduktion in lëtzebuergescher Sprache – ein Problem dar, besonders angesichts der Tatsache, dass die n-Tilgung nur in bestimmten Kontexten beobachtet werden kann.

94 BIBER (1993), S.248-255

## 5 Kontextbedingungen der Tilgung

### 5.1 Lautliche und lexikalische Kontexte der Tilgung

Wie wir in Kapitel 3 gesehen haben, unterliegt die n-Tilgungsregel im Lëtzebuergeschen bestimmten Anwendungsbedingungen, die sich einerseits auf die lautlichen und lexikalischen Eigenschaften des Wortes selbst, andererseits auf das Folgewort, seinen Anlaut oder seine grammatische Kategorie beziehen.

In diesem Kapitel soll zunächst der Einfluss des Folgeworts auf die Tilgung untersucht werden. Tabelle 11 listet noch einmal übersichtartig die von verschiedenen Autoren genannten Kontexteigenschaften auf, die nicht oder nicht zwingend zur n-Tilgung führen.

Tabelle 11. Kontexteigenschaften, die keine Tilgung auslösen

Autor	Folgewort-Anlaut	Folgewort-Lexem
Dicks	Vokale und h,n,d,t,ts <sup>95</sup>	-
Capesius		-
Bruch	Vokale und h,d,t (ts)	si, se, seltener sech (Pronomina)
Gilles		z-FUNK (si, se, sech, selwer, senger, säinen, sou) > Homorganität, <i>clitic group</i>

### 5.2 T-Wörter: Operationalisierung der Tilgung

In Kapitel 4 habe ich die Zusammensetzung und Annotation des LuxText-Corpus vorgestellt. Es wurde deutlich, dass eine automatische Analyse nur auf dem Vergleich von Zeichenketten beruhen, nicht aber auf Lemmata oder syntaktische Funktionen zurückgreifen kann. Auf dieser Beschreibungsebene lässt sich die Tilgung des auslautenden ‚n‘ formal so fassen:




1. Eine getilgte Wortform  $w_t$  hat die Form  $\{ V \mid C \} V_0^{96}$
2. Für die zugehörige, ungetilgte Form  $w_e$  gilt:  $w_e = w_t + [ n \mid nn ]$

95 Bei Dicks „z“ geschrieben (LA FONTAINE 1855)

96 Die gültigen Zeichenketten sind in EBNF notiert (BACKUS, NAUR & al (1960). Kursivgestellte Zeichen stehen hier für nichtterminale Symbole. – Vgl. P. Naur, J. W. Backus & al. (1960), Report on the algorithmic language ALGOL 60, *Communications of the ACM* 3, S.299-314

3. Einem  $V_0 = ,\ddot{e}'$  in der getilgten Form entspricht unter bestimmten Voraussetzungen ein  $V_0 = ,e'$  in der ungetilgten Form.<sup>97</sup>

Um die Kontexte eines Wortes  $w$  zu ermitteln und ihren Einfluss auf die n-Tilgung von  $w$  festzustellen, muss eine Operation definiert werden, die  $w_t$  und  $w_e$  als Realisierungen von  $w$  identifiziert, die also getilgte und ungetilgte Form als zusammengehörig erkennt. Der Tilgungsprozess ist jedoch nicht umkehrbar: Für eine Wortform, die formal der Bedingung (1.) entspricht, kann nicht immer eindeutig bestimmt werden, ob überhaupt eine Tilgung vorliegt, und, wenn ja, ob einfaches oder doppeltes ‚n‘ getilgt wurde. Ähnlich wie beim „imperfect hashing“ kommt es also bei der Zuordnung  $w_t \rightarrow w_e$  zu Kollisionen. Die möglichen Kollisionsfälle lassen sich in folgende Kategorien einteilen:

- 
 i. Mehrere  $w_t$ : Zur getilgten Form  $w_t$  existiert eine bedeutungsverschiedene, gleichlautende Form, die nicht Ergebnis einer Tilgung ist  
 Beispiel:  $si_{AUX}$  (dt.: bin/sein/sind,  $w_e = ,sinn'$ ) –  $si_{PRON}$  (dt.: sie)
- 
 ii. Flexivtilgung:  $w_t$  fällt mit einer anderen Flexionsform des selben Wortes  $w$  zusammen (Sonderfall von i.)  
 Beispiel:  $Safe_{SING}$  –  $Safe_{PL}$  (Tilgungsform zu  $w_e = ,Safen'$ )<sup>98</sup>
- 
 iii. Mehrere  $w_e$ : Zu  $w_t$  existieren zwei verschiedene, mögliche  $w_e$  auf -n und auf -nn  
 Beispiel: Zu  $w_t = ,E'$ :  $Enn$  (dt.: Ende) –  $En$  (dt.: ein)<sup>99</sup>
- iv. Orthographische Störung: Die Zuordnung scheitert aufgrund uneinheitlicher Schreibung im Corpus. Tippfehler und idiosynkratische Schreibungen einzelner Autoren fallen als extrem seltene Formen durchs Raster; jedoch gibt es systematische Fälle (z.B. ältere Schreibweise des LWB, vgl. \*hun = haben, \*sin = sein<sup>100</sup>)
- v. Homonymie durch e-Trema-Vereinfachung: Die von mir gewählte Überführung von  $\ddot{e} \rightarrow e$  in der letzten Wortsilbe

97 Die unter 3. genannte Schreibkonvention soll die erfolgte Tilgung bei französischen Plural-Nomina, deren Singularform auf -e endet, deutlich machen. Sie gehört zu den seit 1999 gültigen Schreibregeln. Ich habe mich dafür entschieden, endsilbisches  $V_0 = ,\ddot{e}'$  stets nach  $,e'$  zu transformieren, da mit einer systematischen Anwendung dieser Regel im Corpus nicht zu rechnen war.

98 Die offizielle Orthographieregelung empfiehlt die Trema-Schreibung nur für französische Lehnwörter; das Beispiel „Safe“ – als englisches Lehnwort – bliebe dann im Tilgungskontext ambig. Nach dem jüngeren „Texte coordonné“ (SCHANEN & LULLING 2001) soll das Trema allerdings immer geschrieben werden, wenn „eine Verwechslung zwischen Singular- und Pluralmorphem möglich wäre“ (3.2.3).

99 Die zugehörige Form  $w_e$  ist „En“, dies ist für eine automatische Analyse aber nicht ersichtlich.

100 Für „sinn“ überwiegt die veraltete Schreibung:  $sin$  ( $n = 4014$ ),  $sinn$  ( $n = 2230$ ).

führt vereinzelt zum Zusammenfallen von Formen. Im CORTINA-Lexikon sind allerdings nur zwei Formen betroffen: Zënn (dt. Zinn) und gënn (dt. gönne), Konflikt mit ze (hd. zu), Gen (hd. Gen).

Für die hier gewählte, automatische Analyse des lëtzebuergeschen Wortschatzes ist ein vereinfachtes Modell der Tilgung unerlässlich. Die genannten Fehlerquellen müssen in Kauf genommen, jedoch später bei der Interpretation der Ergebnisse berücksichtigt werden.

Im Folgenden werden alle Formen, die nach einer n-Tilgung ein identisches Schriftbild haben (unter Berücksichtigung von Groß-/Kleinschreibung), wie ein Type behandelt. Das so aufgefasste Type wird im folgenden „T-Wort“ genannt und mit  $w$  bezeichnet, um die inhaltliche Differenz zum lexikalischen Type deutlich zu machen; ein Lexikon aus T-Wörtern heißt „T-Lexikon“. Ein T-Wort repräsentiert all die Formen, die durch n-Tilgung auf eine gemeinsame Form abgebildet werden. Die Tilgung selbst wird durch zwei Regeln operationalisiert:

$$\begin{aligned} (-n \mid -nn) &\rightarrow \emptyset \\ -\ddot{e} &\rightarrow e \end{aligned}$$

Bei den späteren Frequenzzählungen wird die Zahl der auf -nn auslautenden Formen gesondert erhoben. So kann für jedes T-Wort die Wahrscheinlichkeit einer durchgängigen Schreibung der ungetilgten Form mit -n bzw. -nn gegenüber einer gemischten Schreibung berechnet werden.

## 5.3 Lautlicher Kontext der n-Tilgung

### 5.3.1 Fragestellung

Eine der bestimmenden Größen für die Tilgung des auslautenden -n ist der physiologische Artikulationsort des Folgelauts (Quelle z.B. Capesius oder Gilles).

In den einzelnen Darstellungen (Bruch, Capesius) schält sich eine dreiwertige Einteilung heraus: In tilgende und die Tilgung blockierende Laute sowie, als dritte Klasse, das ‚s-‘, das zu den tilgenden Lauten gerechnet wird, aber eine Reihe von Ausnahmen zulässt.

### 5.3.2 Experiment

Eine Vorerhebung soll klären, ob diese drei Folgelautklassen sich anhand des Corpus bestätigen lassen, oder ob andere Modelle – ein Lautkontinuum bzw. eine größere Zahl von Klassen – eine adäquatere Beschreibung der Situation bieten. Als Quellcorpus dient C40 (die Subcorpora und ihre Zusammensetzung sind im Anhang aufgeschlüsselt).

Das Vorgehen wird in drei Schritten beschrieben: Zunächst wird ein Lexikon aller Tilgungspaare  $w_i$  zusammengestellt. Anschließend wird eine Tabelle

aller Folgegrapheme  $g_j$  (unter Vernachlässigung von Groß-/Kleinschreibung) dieser Tilgungspaare angelegt. Schließlich wird für jedes Folgegraphem berechnet, wie häufig die  $w_i$  im Text vor diesem Graphem als getilgte Form  $w_{ti}$  bzw. als ungetilgte Form  $w_{ei}$  realisiert werden.

Tabelle 12. Frequenz und Tilgungsdruck der Folgegrapheme  $g_j$  (Corpus: C40, Sample:100%).

$g_j$	Absolute Häufigkeit	$\theta_j$
p	9 055	1
y	3*	1
q	222	1
k	11 574	0,99
m	18 160	0,99
w	10 051	0,99
b	10 252	0,99
g	14 868	0,99
j	2 021	0,99
c	1 877	0,98
f	7 050	0,98
l	6 153	0,98
r	4 176	0,98
v	7 568	0,96
s	21 936	0,89
&foreign;	13 051	0,73
x	5	0,6
h	15 565	0,15
n	8 412	0,14
u	2 496	0,12
z	6 316	0,11
ë	1 428	0,09
à	23	0,09
o	7 274	0,08
a	20 837	0,08
e	25 462	0,07
d	48 762	0,07
ö	57	0,07
ä	459	0,06
i	2 809	0,06
t	2 554	0,06
é	494	0,05
ü	5	0
œ	1*	0
î	12	0

### 5.3.3 Auswertung

Für jedes Folgegraphem  $g_j$  wurde der Tilgungsdruck<sup>101</sup> auf den Vorgänger als

101 Ich habe mich für die Bezeichnungen „Tilgungsdruck“ und „Tilgungsbereitschaft“ (Kap. 6) entschieden, um der Markiertheit der getilgten Formen gerecht zu werden – das isolierte Lexem ist ungetilgt, erst bestimmte Kontextfaktoren führen zur Tilgung. Allerdings wurde in den Experimenten durchweg die Zahl der *nicht getilgten* Formen als Bezugsgröße verwendet. So wurde es notwendig, die Quotienten über das Inverse des Anteils der nicht getilgten Formen zu definieren. Es gilt:  $|w_t|/|w| = (1 - |w_e|/|w|)$ .

## Quotient

$$\theta_j = 1 - \frac{\text{Anzahl der nicht getilgten Vorgänger } w_{ei}}{\text{Anzahl der Vorgänger } w_i}$$

berechnet:

Tabelle 12 (vorige Seite) zeigt den ermittelten Quotienten zu jedem  $g_j$  (\*Grapheme mit weniger als 5 Belegen werden später nicht berücksichtigt).

Die Rangdarstellung in Abbildung 3 soll die Clusterung (Klassenbildung) um einzelne Quotienten herum verdeutlichen. Die Ordinate zeigt den Tilgungsdruck  $\theta$  (in Schritten zu 1/100), auf der Abszisse sind die Summen der absoluten Häufigkeiten aller  $g_j$  zu einem gegebenen  $\theta$  abgetragen.

Die clusterartigen Verdichtungen um einzelne  $\theta$  legen die Annahme zweier scharf abgegrenzter Kategorien nahe:

1. Tilgende Kontexte ( $0,96 \leq \theta \leq 1$ ): Dies sind die Konsonanten außer h,n,d,t,z. v ( $\theta = 0,96$ ) steht etwas abgesetzt am rechten Rand. Dies geht jedoch im wesentlichen auf das Konto zweier hochfrequenter Kollokationen.<sup>102</sup>
2. Nicht tilgende Kontexte ( $0 \leq \theta \leq 0,15$ ): Vokale sowie h,n,d,t und z. Die Kategorie wirkt homogen.

Zwischen den Kategorien stehen drei isolierte Laute/Grapheme, nämlich s ( $\theta = 0,89$ ), x ( $\theta = 0,6$ ) und das Pseudo-Graphem ‚&foreign;‘ ( $\theta = 0,73$ ), das für die Markierung „fremdsprachig/Ziffer“ steht. Das (hier sehr selten auftretende) x hat in allen Fällen einen Lautwert nahe bei [z], vgl. „xénophobe“.

und umgekehrt.

102 81% der n-Erhalts-Fälle vor ‚v‘ (237 von 290) betreffen die Präposition „vun/vum“ (analytischer Genitiv im L.). 37% der Fälle gehen allein auf das Konto der Kollokationen „Enn vun“ und „Enn vum“ (108 Belege). „Enn“ (dt.: Ende) erscheint jedoch nur durch eine Zuordnungskollision (Typ iii. vgl. 5.2 oben) als ungetilgte Realisation des Artikels „E/En“. Diese Verzerrung erklärt den Rechtsversatz des ‚v‘.

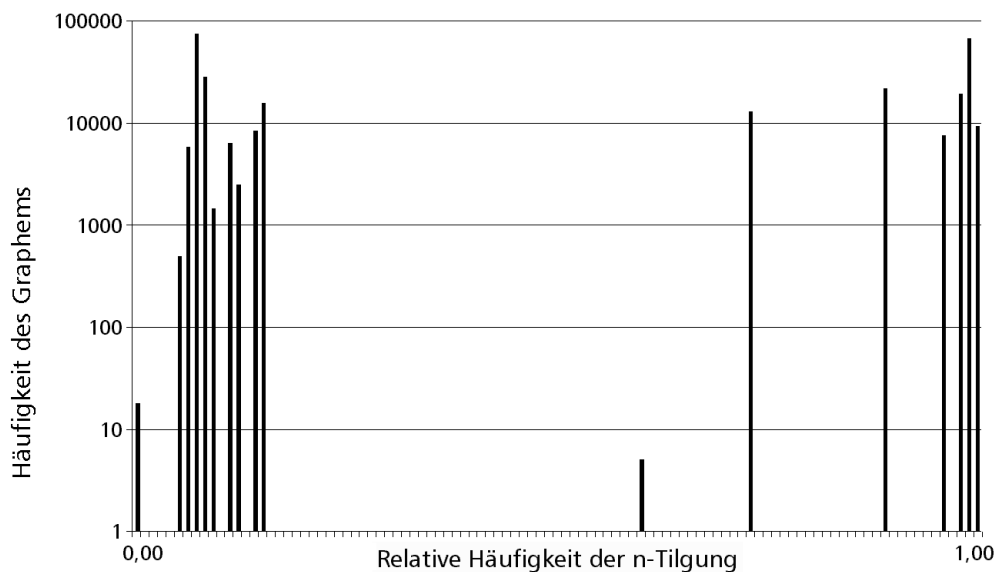


Abbildung 3. Frequenzen der Folgegrapheme gegen relative Häufigkeit der n-Tilgung („Tilgungsdruck“)

### 5.3.4 Interpretation

Mit der Untersuchung des „Tilgungsdrucks“ einzelner Grapheme sollte zunächst die Rechtfertigung der in grammatischen und lexikographischen Arbeiten angegebenen Tilgungskontexte überprüft werden. Das dabei angelegte Raster ist zugegebenermaßen grob: Die Überlagerung des beobachteten Zusammenhangs durch äußere Effekte – den Einfluss fester Syntagmen oder besonders häufiger Abfolgen bestimmter Wortarten, nichtzufällige Verzerrungen wie etwa die systematische Ungleichverteilung von Singular-/Pluralformen – ist sehr wahrscheinlich. Umso klarer treten die Kategorien des lautlichen „Tilgungskontexts“ und „Erhaltungskontexts“ hervor. Die Sonderstellung des ‚s‘ zeichnet sich bereits ab.

## 5.4 Lexikalischer Kontext der n-Tilgung

### 5.4.1 Fragestellung

Robert Bruch<sup>103</sup> beschreibt die n-Tilgung vor s- als obligatorisch, mit Ausnahme der Pronomina *si* und *se*, seltener *sech*, wo sie fakultativ sei. Der Gesetzestext von 1975 ergänzt seine Aufzählung um weitere Wörter:

Virum s- kann den -n jhust bei *si*, *se*, *säin*, *sech*, *séng*, *sou* a beim sonneren s virun *du*, *de* stoëbleiwen: *Wann si kommen od. wa si kommen; wann s de këns od. wa s de këns.*<sup>104</sup>

Peter Gilles fand in seinen Feldversuchen, was er allgemeiner als freie

103 BRUCH (1954)

104 ARRÊTÉ MINISTÉRIEL (1975), S.1384

postlexikalische Varianz nach mit stimmhaftem ‚s‘ anlautenden Funktionswörtern zusammenfasst, nämlich

... die Pronomen *si/se* ‚sie‘, *selwer* ‚selbst‘, *sech* ‚sich‘, *sénger* ‚seiner‘, *sainen* ‚seinen‘ oder das Adverb *sou* ‚so‘. ... Dagegen wird vor ‚echten‘ Lexemen mit z-Anlaut (*sätzen* ‚sitzen‘, *siven* ‚sieben‘ usw.) -n ausnahmslos getilgt.<sup>105</sup>

Offenbar gibt es also nicht nur lautliche, sondern auch lexikalische Ausnahmekontexte für die Tilgungsregel, es herrscht jedoch Uneinigkeit darüber, welches diese Kontexte sind.

Die Frage nach Folgewörtern mit geringem Tilgungsdruck soll in diesem Abschnitt so allgemein wie möglich gestellt werden: Gibt es Folgewort-Lexeme, die – über den Einfluss ihres Anlauts hinaus – signifikanten Einfluss auf die n-Tilgung eines Wortes haben?

#### 5.4.2 Experiment

Im Abschnitt 5.3.3 habe ich den Tilgungsdruck  $\theta$  eines Graphems als Durchschnitt der Tilgungsquote  $(1 - |w_e|/|w|)$  aller links davon auftretenden T-Wörter definiert. Im folgenden Experiment geht es um den Einfluss, den ein Folgegraphem  $g_j$  auf die Tilgungsquote eines *einzelnen* T-Wortes ausübt; ich schreibe

$\theta_{j,w}$  für den Tilgungsdruck von  $g_j$  bezüglich  $w$ ,

$|w|_x$  für die Zahl der Belege von  $w$  vor (dem Lexem oder Graphem)  $x$ ,

$\gg$  für die Relation „konfident größer“.

Sei  $S_{j,w}$  die Menge aller mit  $g_j$  anlautenden Folgewörter zu  $w$  im Corpus. Gesucht sind nun alle Wortpaare aus  $w$  und einem Folgewort  $s \in S_j$ , für die gilt:

$$\theta_{j,w} \gg \left( 1 - \frac{|w_e|_s}{|w|_s} \right) \quad (1)$$

Die Berechnung erfolgt anhand des gewählten Teilcorpus (C40, Sample: 100%) und geht in drei Schritten vor sich:

1. Es wird eine Liste der Folgewörter aller möglichen T-Wörter<sup>106</sup> zusammengestellt,
2. die Tilgungsquoten<sup>107</sup> je T-Wort und Folgegraphem werden in einer Tabelle aufgeschlüsselt, um schließlich
3. die Tilgungsquote jedes Wortpaars  $(w,s)$  zu berechnen und mit der

<sup>105</sup> GILLES (2001), S.26 ff.

<sup>106</sup> Bedingung: Auslaut auf Vokal oder ‚n‘/‚nn‘.

<sup>107</sup> Berechnung wiederum als Erhaltsquoten.

Tilgungsquote von *w* vor dem Anlaut von *s* zu vergleichen.

Die Zahl der Belege für die einzelnen Wortpaare schwankt stark. Um die verschieden großen Stichproben dennoch vergleichen zu können, wird deshalb jeweils die untere bzw. obere Schwelle eines Konfidenzintervalls anstelle der gemessenen Quoten eingesetzt; so kann gesagt werden: mit einer Wahrscheinlichkeit von *k* Prozent hätte die Vergleichsoperation bei einer anderen Stichprobe dasselbe Resultat.

Das Konfidenzintervall  $c/2$  für eine normalverteilte Proportion  $p$  berechnet sich aus:

$$\frac{c}{2} = z \cdot \sqrt{\frac{p \cdot (1-p)}{N}} ; z = 1,645$$

Damit lässt sich für die Bedingung (1) genauer schreiben:

$$g_{j,w} - \frac{c_{g_{j,w}}}{2} > g_{s,w} + \frac{c_{g_{s,w}}}{2} ; g_{s,w} = \left(1 - \frac{|w_e|_s}{|w|_s}\right) \quad (1')$$

Aus den gegebenen Daten und mit den Konfidenzbedingungen ( $z = 1,645$ ; mindestens 3 Belege je Wortpaar) konnten 95 Wortpaare identifiziert werden, die die Bedingung (1') erfüllten.

### 5.4.3 Auswertung

Als Ergebnis liegt eine Menge von (T-Wort, Folgewort)-Paaren vor, zwischen denen die *n*-Tilgung häufiger als erwartet ausbleibt. Die bei diesen Paaren besonders niedrige Tilgungsquote kann – wenigstens ist dies vorstellbar – auf zwei verschiedene Ursachen hinweisen:

- a. eine Eigentümlichkeit bestimmter Kollokationen,
- b. eine spezifische Eigenschaft des Folgeworts.

So unterbleibt (a.) in Kollokationen wie „en bloc, en faveur, en masse, en plus, en ville, en vue“<sup>108</sup> regelmäßig die Tilgung aus dem unmittelbar nachvollziehbaren Grund, dass es sich hier nicht um den lätzebuergischen Artikel *en* – der regelmäßig getilgt wird –, sondern um die gleich geschriebene französische Präposition handelt.

Dagegen stehen (b.) Lexeme, die allgemein, unabhängig vom vorangehenden T-Wort, einen niedrigeren Tilgungsdruck als ihre Anlautklasse ausüben.

Die um ambige T-Wörter bereinigte Resultatliste<sup>109</sup> umfasst 37 Wortpaare, in

108 Jede der genannten Kollokationen ist mehrfach im luxemburgischen Satzzusammenhang belegt (C48).

109 Acht Paare, deren erstes Wort eines aus der Liste „En/Enn“, „Ma/Mann“, „Se/Sënn“, „si/sinn“, oder „So/Sonn“ war, wurden nicht berücksichtigt. In allen Fällen handelt es

denen 14 verschiedene Folgewörter vorkommen. Abbildung 4 zeigt die 14 Folgewörter absteigend nach dem Grad ihrer Abweichung von der jeweiligen Anlautklasse geordnet.

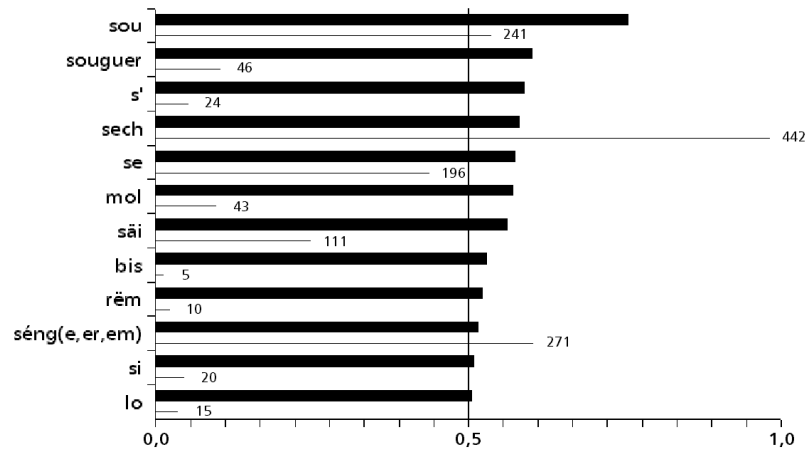


Abbildung 4. Lexeme, deren Tilgungsdruck konfident unter dem ihrer Anlautklasse liegt (Maß:  $\theta_g/(\theta_s+\theta_g)$ ;  $> 0,5 \Rightarrow$  niedrigerer Tilgungsdruck). Die kleinen Zahlen geben die Anzahl der Tokenpaare an, die als Belege gezählt wurden

Bei den vorliegenden Corpusdaten waren nicht tilgende Kollokationen (a.) meist zu selten, um den Sprung über die Konfidenzschwelle zu schaffen; bei den hier nicht berücksichtigten Paaren, die unter der Schwelle blieben, stehen einige Fälle von Homonymie wie die zuvor erwähnten Kollokationen mit der Präposition *en*.

#### 5.4.4 Interpretation

Nach dem Gesetzestext kann die n-Tilgung vor den Wörtern *si/se* ‚sie‘, *säin/séng* ‚sein/e‘, *sech* ‚sich‘ und *sou* ‚so‘ unterbleiben. Nach Gilles handelt es sich um „häufig klitisierte Elemente“, d.h., dass die Sprecher sie als morphologische Einheit mit dem vorhergehenden Wort auffassen. Diese Erklärung ist wenig befriedigend, da auch Präfixe getilgt werden.

Die als Ergebnis der Suche nach lexikalischen Kontexten mit niedrigerem Tilgungsdruck vorliegende Liste enthält jede der im Gesetzestext genannten „Ausnahme“-Formen, darüber hinaus jedoch noch 4 weitere: Dies sind die Wörter *mol*, *bis*, *rëm* und *lo*. Tabelle 13 stellt diese Wörter neben die Paare, bei denen die niedrige Tilgungsquote festgestellt wurde, und zeigt einige Textbeispiele.

sich um Kollisionen von Lexemen, die bedeutungsverschieden sind und unterschiedliches Tilgungsverhalten zeigen.

Tabelle 13. Nicht mit s- anlautende Lexeme *g* mit niedrigem Tilgungsdruck

<i>g</i>	Paare ( <i>g,w</i> )	Beispiele (aus C48)
<i>mol</i> (dt. ‚mal‘)	wa(nn) mol, ee(n) mol, e(n) mol, da(nn) mol	
<i>bis</i>	u(n) bis	vun haut <i>u bis</i> 2015 Vum fréien Hierscht <i>un bis</i> Mëtt Abrëll Stréch vum Pillem <i>un bis</i> erof bei de Fouss
<i>rëm</i> (dt. ‚wieder‘)	scho(n) rëm	
<i>lo</i> (dt. ‚da‘, ‚jetzt‘)	a(n) lo	kee Schnellzuch, <i>a lo</i> déing hie mol <i>a lo</i> hun ech just virstellen <i>an lo</i> as mä Papp ginn, <i>an lo</i> hunn déi Saachen

Hier fällt besonders ins Auge, dass drei der Formen jeweils auch eine betonte (Lang-) Form besitzen, die mit unbetontem /ə/ beginnt,

*emol – mol, erëm – rëm, elo – lo,*

und über die das Règlement 1999 ausführt: „Et muss een sech bewosst sinn, datt sou ënnerschiddlech Wuertkonstruktiounen entstinn: ... *an elo – a lo ... kënns d’ërem? – kënns de rëm?*“<sup>110</sup> Das Règlement sieht also gerade im Fall *a(n) lo* die Tilgung vor, während die Daten aus dem Corpus eher darauf hinweisen, dass die Sprecher vor dem „verschluckten“ Schwa häufig auf die Tilgung verzichten.

Für das Wortpaar *u(n) bis* gab es zwar nur 5 Belege, davon zeigten aber immerhin drei die Nichttilgung von *un* vor *bis*, während *un* sonst vor *b*-ausnahmslos tilgte.

## 5.5 Zusammenfassung

In diesem Kapitel wurden die Kontexte – lautlich und lexikalisch – untersucht, die die n-Tilgung auslösen. Die Ergebnisse gestatten es, in den folgenden Experimenten einen lautlicher Tilgungskontext (Konsonant außer *h, d, n, t, ts*) und einen Erhaltungskontext (komplementär zum Tilgungskontext, außer *s*) vor auszusetzen. *s*, vor dem die Tilgung durchweg seltener stattfindet, wird als „neutraler“ Kontext behandelt.

Die Liste der Wörter, deren Tilgungsdruck unter dem ihrer Anlautklasse lag, kommt im Rahmen der Regelinduktion und -justage (Kapitel 8) als Liste „ungewöhnlicher“ Kontexte, die bei der Bewertung der generierten Regeln ausgeschlossen bleiben, noch einmal zum Einsatz.

## 6 n-Tilgung als Parameter des Lexems

### 6.1 Lexem und Tilgung

Im vorigen Kapitel wurden Zusammenhänge zwischen der Tilgung des auslautenden -n und dem Folgewort beschrieben. Nun soll näher untersucht werden, inwieweit die Tilgung vom Lexem selbst abhängt.

Frühere grammatische Beschreibungen des Lëtzebuergeschen erwähnen einen solchen Zusammenhang: Nicht alle auf -n auslautenden Wörter dürfen getilgt werden, bei manchen ist die Tilgung möglich, aber nicht zwingend. Es sind verschiedene Vermutungen darüber angestellt worden, welche Worteigenschaften die Tilgung beeinflussen. Genannt werden phonetische, phonologische, morphologische und etymologische Kriterien (vgl. Kapitel 3 oben).

Ziel der Untersuchung wird sein, eine Klassifikation der T-Wörter  $w$  in „sichere Tilger“, „sichere Erhalter“ und eine „Problemklasse“ zu erreichen, die dann genauere Betrachtung verdient.

### 6.2 Tilgungsklassen

In diesem Experiment soll der Zusammenhang zwischen den einzelnen Lexemen und der n-Tilgung betrachtet werden. Dieser Zusammenhang, falls er existiert, muss vom Einfluss des Folgekontexts getrennt behandelt werden.

Dies lässt sich erreichen, indem man für jedes T-Wort  $w$  und hier für jeden einzelnen Folgekontext  $K_i$  die Zahl der nicht getilgten Realisierungen  $|w_e|$  zur Zahl der Realisierungen insgesamt  $|w|$  in Beziehung setzt. Ergebnis ist ein Profilvektor  $P$ , der das Verhalten von  $w$  in den betrachteten Folgekontexten beschreibt<sup>111</sup>:

$$P_w = \left[ \frac{|w_e|_{K_1}}{|w|_{K_1}}, \frac{|w_e|_{K_2}}{|w|_{K_2}}, \dots, \frac{|w_e|_{K_n}}{|w|_{K_n}} \right]$$

<sup>111</sup> Schreibkonvention:  $|x|_K$  bezeichnet die absolute Häufigkeit des Lexems  $x$ , im Kontext  $K$ , im untersuchten Corpus.

Aufgrund der Erkenntnisse aus dem vorigen Kapitel sollen die Folgekontexte nicht individuell, sondern zusammengefasst zu vier Gruppen untersucht werden, wie es Tabelle 14 zeigt. Angesichts der oben (Kapitel 5) festgestellten starken Homogenität und der guten Klassentrennung scheint das eine zulässige Vereinfachung zu sein.

Tabelle 14. Zusammenfassung der Folgegrapheme in vier Kontextklassen.

Kontextklasse	Bedingung für Folgegraphem
Erhaltungskontext	Vokale und h, n, d, t, z
Tilgungskontext	Konsonanten außer h, n, d, t, z
unbekannter Kontext	s, Sonderzeichen, Ziffern/Fremdsprache (&foreign;)
Nullkontext	Clausen-Ende <sup>112</sup>

Ein Beispiel aus dem T-Lexikon zu C40 soll dieses Vorgehen illustrieren. Abbildung 5 zeigt den relativen n-Erhalt  $|w_e|/|w|$  der Wörter auf -unn, bezogen auf zwei Kontextklassen; die beiden Achsen der Darstellung repräsentieren diese Kontexte: Erhaltungskontext (Ordinate) und Tilgungskontext (Abszisse). Die Größe der Kreise ist proportional zur Zahl der unterschiedlichen Lexeme (als T-Wörter), die aufgrund übereinstimmender Tilgungseigenschaften auf einen gemeinsamen Punkt abgebildet erscheinen.

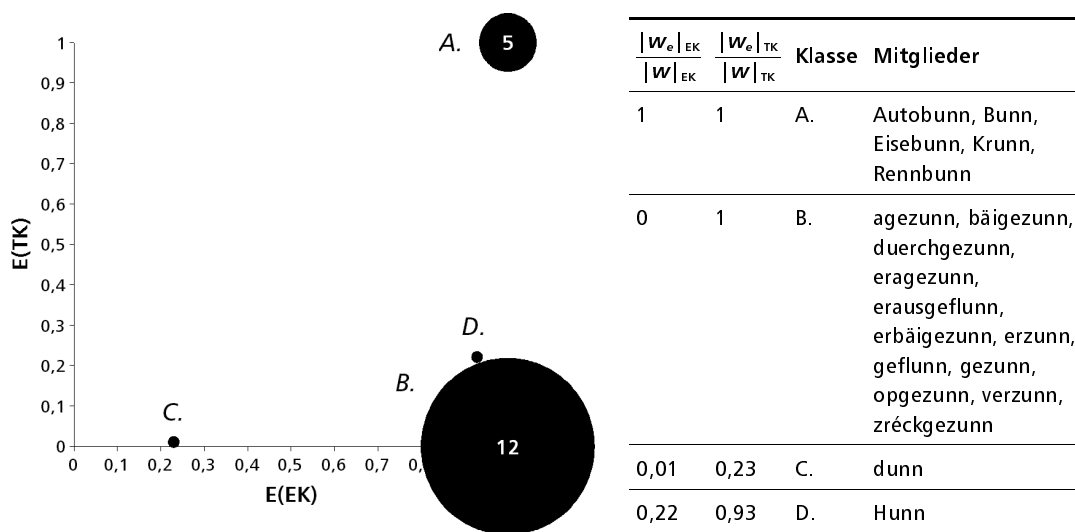


Abbildung 5. Lexeme auf -unn bilden eine Erhalter- (A.) und eine Tilgerklasse (B.)

Im gezeigten Beispiel widerstehen von den Wörtern auf -unn 5 der Tilgung (Gruppe A.), 12 folgen der Tilgungsregel (Gruppe B.). Gruppe A. enthält Komposita zu Bunn (dt.: Bahn), Gruppe B. Partizipformen der Verben zéien und fléien (dt.: ziehen, fliehen) bzw. ihrer Partikelkomposita. Zwei weitere Wörter stehen abseits dieser beiden Klassen. Diese Separierung ist in beiden

112 Eines der Zeichen [,:;!?!...] (vgl. Kap. 4)

Fällen ein Hinweis auf Kollisionen (Homographien innerhalb der T-Wortklasse). Im ersten Fall kollidieren drei Types miteinander:  $du_{\text{PRON}}$ ,  $du/dunn$  (dt.: dann) und die Alternativform  $dunn$  (zu  $doen$ , dt.: tun). Im zweiten Fall sind es die beiden Types  $Hu/Hunn$  (dt.:  $habe_{1.SG}$ ,  $haben_{1./3.PL}$ ) und  $Hunn$  (dt. Hahn).

Wie ich es hier an einer kleinen Auswahl von T-Wörtern gezeigt habe, wurde das Tilgungsverhalten sämtlicher T-Wörter aus C40<sup>113</sup> quantitativ erfasst.

Die Häufigkeit von Erhalt und Tilgung wurden in absoluten Zahlen erfasst, um Statistiken über Wörter mit stark unterschiedlichen Vorkommenshäufigkeiten berechnen und miteinander vergleichbar machen zu können.

Als zusätzliche Maßzahl wurde die absolute Häufigkeit der nicht getilgten, auf -nn auslautenden Formen erfasst. Mit ihr können T-Wörter identifiziert werden, bei denen möglicherweise eine Kollision eines mit einfachem mit einem mit doppeltem n auslautenden Lexem vorliegt (vgl. 5.2 oben).

Aufgrund des so zusammengestellten Frequenzlexikons können nun Klassen von Lexemen gebildet werden, die das auslautende ‚n‘ regelmäßig tilgen bzw. regelmäßig erhalten.

Wie am Beispiel zu sehen war, müssen allerdings Schwellenwerte festgelegt werden, um die klaren (disjunkten) Klassen von den uneindeutigen Fällen trennen zu können. Hierfür werden, zunächst adhoc, Schwellenwerte als Zugehörigkeitskriterien festgelegt. In einem späteren Stadium werden die Schwellenwerte systematisch variiert.

Dabei kommen zwei Maße für statistische Konfidenz zum Einsatz, um die einzelnen, in der absoluten Häufigkeit stark differierenden Lexeme vergleichbar zu machen.

Im folgenden wird für die so bestimmten Tilgungsklassen kurz geschrieben:

- $\omega_T$  für Lexeme, die das auslautende -n regelmäßig in den entsprechenden Kontexten tilgen, und
- $\omega_N$  für Lexeme die das auslautende -n regelmäßig nicht tilgen.

### 6.3 Auswertung

Das aus C40 (Sample: 100%) extrahierte T-Wörter-Lexikon umfasst 21 894 unterschiedliche Wörter (Types) in 469 032 Realisierungen (Tokens). Um die n-Tilgung als Proportion berechnen zu können, wurde die Bedingung gestellt, dass mindestens je ein Beleg für eine ungetilgte Form im Erhalt- oder im Tilgungskontext vorliegt. Opfer dieser Auswahl sind in erster Linie die Tokens mit extrem niedriger Frequenz (*hapax legomena*). 346 371

113 ...sofern sie mindestens einmal im Erhaltungs- oder Tilgungskontext vorgefunden wurden.

T-Wort-Tokens (73,8%) erfüllten diese Bedingung, jedoch nur noch 8 936 T-Wort-Types (40,8%).

Tabelle 15. Anzahl der in T-Wort-Tokens in den einzelnen Kontextklassen

Kontextklasse	Zahl der Belege	Anteil der $w_e$ in %
Erhaltungskontext	144 378	92,2%
Tilgungskontext	104 928	5,1%
unbekannter Kontext	40 736	32,2%
Nullkontext	56 329	96,2%

Tokens/Types: 346371/8936

Die dritte Spalte der Tabelle 15 zeigt, trotz der nur groben, ersten Auswahlbedingung, bereits deutlich den Einfluss der Tilgungsregel auf die T-Wörter: Über 90% Erhalt im Erhaltungskontext und am (Teil-) Satzende, gegen 5% im Tilgungskontext. Sie bestätigt so die gewählte Operationalisierung der „T-Wörter“.

Über Schwellenwerte für den Anteil der nichtgetilgten Formen  $w_e$  im Erhaltungs- und im Tilgungskontext wurden statistisch „gute“ Beispiele für die Klassen der regelmäßigen „Tilger“  $\omega_T$  und der „Nie-Tilger“  $\omega_N$  bestimmt. Als Schwellenwerte habe ich zunächst gewählt:

	$ w_e / W _{\text{Erhaltungskontext}}$	$ w_e / W _{\text{Tilgungskontext}}$
$\omega_T$	> 0,5	< 0,5
$\omega_N$	> 0,5	> 0,5

$\omega_T$  sind demnach T-Wörter, die im Erhaltungskontext überwiegend nichtgetilgt, im Tilgungskontext aber überwiegend getilgt erscheinen.  $\omega_N$  tilgen in beiden Kontexten überwiegend nicht.

Die Zahl der Belege je Type schwankt und ist oft generell niedrig. Ersetzt man jedoch die zu vergleichenden Zahlenverhältnisse durch die jeweils ungünstigsten Werte innerhalb ihres Konfidenzintervalls, so kann man (mit einer bestimmten Wahrscheinlichkeit) annehmen, dass das Verhältnis durch das Hinzuziehen weiterer Belege nicht ungünstiger wird als der angenommene Wert.

Zu jeder Proportion wurde ein binomiales Konfidenzintervall berechnet ( $z=1,645$ , Konfidenz 90%, mind. 3) und für die Größerbedingung die untere, für die Kleinerbedingung die obere Konfidenzschwelle (*confidence limit*) anstelle der errechneten Proportion eingesetzt.

Unter diesen Voraussetzungen ließen sich 863 T-Wort-Types als  $\omega_T$  und 101 Types als  $\omega_N$  klassifizieren.

Das bedeutet nicht, dass es sich bei den übrigen knapp 8 000 Types nicht auch um regulär tilgende bzw. niemals tilgende Lexeme handelt. Ursache für diese rigide Auswahl ist vielmehr das angesetzte Konfidenzkriterium, das zusammen mit der gewählten Schwelle alle Formen mit weniger als 5

Belegen je Kontextklasse ausschließt. Eine erste Sichtung dieser, keiner Klasse zugeordneten Wörter (vgl. Tabelle 2 im Anhang B) erlaubt folgende Einteilung:

a. Zu wenige Belege ( $ w _{\text{Erhaltung-KT}}$ oder $ w _{\text{Tilgung-KT}} < 5$ )	7 891 Types
b. Singular/Plural-Kollision, Bsp.: Disque +n, Debatte +n)	40 Types
c. Nominativ/Akkusativ-Kollision, Bsp.: large +n)	2 Types
d. Orthographische Fehler, individuelle Schreibweisen	12 Types
e. Kollision nicht verwandter Lexeme (Bsp.: zu +nn, Wee +n)	18 Types

f. Zweifelsfälle 9 Types  
Die Lexeme, die die Klassen der  $\omega_T$  und  $\omega_N$  bilden, zeigen folgende Verteilung auf die verschiedenen Kontexte:

$\omega_T$		
Kontextklasse	Zahl der Belege	Anteil der $w_e$ in %
Erhaltungskontext	115 157	98,3%
Tilgungskontext	84 637	0,8%
unbekannter Kontext	35 527	31,0%
Nullkontext	37 879	98,9%

Tokens/Types: 273 200/863

$\omega_N$		
Kontextklasse	Zahl der Belege	Anteil der $w_e$ in %
Erhaltungskontext	2 825	97,7%
Tilgungskontext	2 730	93,1%
unbekannter Kontext	839	95,2%
Nullkontext	1 385	98,8%

Tokens/Types: 7 779/101

Die  $\omega_T$ -Klasse umfasst gegenüber der  $\omega_N$ -Klasse offenbar Wörter mit höherer mittlerer Frequenz. Deutlicher wird dies, wenn man den Mittelwert der Lexemfrequenzen beider Klassen berechnet:

	Frequenz	
	Mittelwert	Varianz
$\omega_T$	316,6	456,6
$\omega_N$	77,0	64,4

Warum sind die „Ausnahmen“ der Tilgungsregel niederfrequenter? Für Peter Gilles ist die Tonkontur HLH für die unterbleibende Tilgung verantwortlich. Jedoch müsse ein Wort einen gewissen Grad an dynamischer Betonung aufweisen (*stress*), um diese Tonkontur tragen zu können. Hochfrequente Wörter und Wortsegmente tragen häufig keinen *stress*: „Daher bleiben Funktionswörter und hochfrequente Verben mit gleicher segmenteller

Weitere Zusammenhänge zwischen der Zugehörigkeit zu einer der beiden Klassen und bestimmten Worteigenschaften, etwa Großschreibung oder Artikulationsort der Endsilbe, werden in Kapitel 9 angesprochen.

## 6.4 Interpretation

Im letzten Abschnitt wurde eine Klassifikation des T-Lexikons nach dem Tilgungsverhalten der einzelnen Wörter in verschiedenen phonetischen bzw. graphemischen Kontexten beschrieben.

Diese Klassifikation ist realisierbar und liefert, trotz materialbedingter und systematischer Einschränkungen, befriedigende Ergebnisse.

Allerdings ist die Ausbeute an statistisch „guten Beispielen“ klein. Wollte man die auf Vokal und -n endenden Wörter aus dem vorliegenden Corpus anhand der beiden erzeugten Klassen als  $\omega_T$  und  $\omega_N$  annotieren, so könnte man noch 60% der laufenden Wortformen (Tokens), aber nur 4,4% der unterschiedlichen Lexeme (Types) annotieren.

Im direkten Vergleich mit dem CORTINA-Vollformenwörterbuch<sup>115</sup> fällt die Relation noch ungünstiger aus: Hier können nur 846 (3,1%) der 26 547 auf -n auslautenden Formen annotiert werden.

Die Tilgungseigenschaften der Wörterbucheinträge direkt aus dem Corpus zu bestimmen, ist offenbar nicht möglich. Als mögliche Lösungsstrategien bieten sich an:

- Vergrößerung des Corpus: Die obigen Zahlen lassen einen Umfang von 20 Millionen Wortformen nötig erscheinen (Faktor 40), eine für das geschriebene Lëtzebuergesche zur Zeit nicht erreichbare Zahl
- Einbeziehung der seltenen Formen: Durch Herabsetzung der Schwellen, Zusammenführen ähnlicher oder gleich auslautender Formen
- Verallgemeinerung: Indem man von den vorhandenen „guten Beispielen“ auf das Verhalten anderer Wörter schließt

Im Folgenden soll eine Kombination aus den beiden letzten Ansätzen verfolgt werden. Für den dritten Ansatz spricht besonders, dass er die morphologische und lexikalische Produktivität des Lëtzebuergeschen berücksichtigt. Schon das existierende CORTINA-Wörterbuch verzeichnet zahlreiche morphologisch mögliche Flexionsformen der einzelnen Wörter, ohne Rücksicht darauf, ob diese auch praktisch verwendet werden, ob sie also in einem Corpus des Lëtzebuergeschen – wie groß er auch sei – belegbar wären. Inkongruenzen zwischen Wörterbuch und Corpus sind also in beiden Richtungen zu erwarten: Vom Wörterbuch aus, weil es aus systematischen Gründen wenig gebrauchte Derivationen enthält; vom

<sup>115</sup> Wortbestand vom Mai 2001: 75 400 Vollformen

Corpus aus, weil es die lexikalische Produktivität der Sprachbenutzer widerspiegelt.

Ein Vergleich der Flexions- und Derivationsparadigmata des Verbs „sichen“ (dt.: suchen) in den verschiedenen Wortlisten veranschaulicht das (Tabelle 16). Die erste Spalte zeigt die im CORTINA-Wörterbuch verzeichneten Formen (auf -n), die zweite die im Corpus C40 vertretenen Belege, die dritte schließlich die beiden Formen, die die Frequenzschwelle zur  $\omega_T$ -Klasse ( $z=1,64$ , Schwellenwerte 0,5/0,5) überwinden konnten.

Tabelle 16. Beispiel für den Formenreichtum des Wörterbuchs

Cortina-WB	Corpus (C40)	$\omega_T$ (0,5/0,5)
aussichen	aussichen	
auszesichen		
besichen	besichen	
duerchsichen	duerchsichen	
eraussichen	eraussichen	
erauszesichen	erauszesichen	
erbäisichen		
nosichen		
	nozesichen	
ofsichen		
ofzesichen	ofzesichen	
opsichen	opsichen	
opzesichen		
	raussichen	
sichen	sichen	sicheN
unzesichen	unzesichen	
usichen	usichen	
versichen	versichen	versicheN
zesummenzesichen		
zesummesichen	zesummesichen	
ënnersichen	ënnersichen	

Nicht zufällig handelt es sich hierbei um das Simplex „sichen“ und das nicht abtrennbare „versichen“: Infinitivformen von Verben stehen – bedingt durch die (dem Deutschen vergleichbare) Wortstellung im Lëtzebuergesch – häufig am Ende eines (Teil-) Satzes, damit aber im Nullkontext. Bei Verben ohne abtrennbare Präfixe lauten jedoch Infinitiv- und bestimmte flektierte Präsensformen häufig gleich<sup>116</sup>. Mit ihren Realisationen als finite Verben liefern „sichen“ und „versichen“ also überdurchschnittlich viele Beispiele im Erhaltungs- und im Tilgungskontext.

116 Auch bei „sichen“: „se fannen, wat se sichen“; „Ech sichen d'Mënschen“, sot de klenge Prënz“; „mir sichen no engem System“ (Beispiele aus C40)

Das weitere Vorgehen wird darin bestehen, das Tilgungsverhalten gut belegter Lexeme wie „sichen“, „versichen“ jeweils auf eine gesamte Wortgruppe hin zu verallgemeinern. Zugleich können die Hapax legomena aus dem Corpus zur Überprüfung dieses Vorgehens herangezogen werden.

## 7 Automatischer Erwerb sprachlichen Wissens

### 7.1 Einleitung

Die beiden klassischen Ansätze des maschinellen Lernens stammen aus den 1960er und 1970er Jahren: Rosenblatts ‚Perceptron‘ mit seinem an biologischen Vorbildern orientierten, konnektionistischen Ansatz und Winstons ‚Klötzchenwelt‘-Lernsystem, das gegebene Beispiele in ein möglichst einfaches semantisches Modell einordnete. Neuere Anstöße kamen aus dem Bereich der Künstlichen Intelligenz (KI) (eine historische Übersicht bietet Kodratoff<sup>117</sup>). Im folgenden Abschnitt wird eine Systematisierung der verschiedenen Ansätze skizziert.

Brill & Marcus<sup>118</sup> erinnern daran, dass das unüberwachte Lernen grammatischen Wissens daneben auch Wurzeln in der strukturalistisch geprägten Sprachwissenschaft hat. Nach R. S. Wells' Aufsatz „Immediate Constituents“ von 1947 modellierten sie Methoden zur automatischen Analyse von lexikalischen Distributionen. In Abschnitt 7.3 werden verschiedene Arbeiten zum automatischen Wissenserwerb aus Corpora – darunter auch Arbeiten Brills – vorgestellt und in den skizzierten Rahmen eingeordnet.

### 7.2 Paradigmen maschinellen Lernens

Briscoe & Caelli<sup>119</sup> unterscheiden anhand der Form des Wissenserwerbs und seiner systeminternen Repräsentation vier Paradigmen lernender Systeme:

- Symbolisch-empirische Systeme (*symbolic empirical learners*): Der Wissenserwerb erfolgt induktiv durch die Abstraktion von Beispielen;
- Erklärungsbasierte (analytische) Systeme (*explanation based learners*): Das System lernt, indem es neues Wissen deduktiv aus gegebenem Hintergrundwissen ableitet, oft in Form eines mathematischen Beweises;
- Genetische, nach dem Vorbild der evolutionären Auslese modellierte Systeme;
- Konnektionistische, an der parallelen Verarbeitung des Nervensystems

117 Y. KODRATOFF (1988), *Introduction to machine learning* (London: Pitman)

118 E. BRILL & M. MARCUS (1992), Tagging an unfamiliar text with minimal human supervision, *Proc. of the AAAI Fall Symposium*

119 G. BRISCOE & T. CAELLI (1996), *Symbolic machine learning*, Bd. 1 von *A compendium of machine learning* (Norwood: Ablex), S.7

orientierte Systeme.

Die Vertreter der ersten beiden Paradigmen zählen zu den sogenannten symbolischen Systemen.

Eine feinere Taxonomie aufgrund der eingesetzten Lernstrategien stellt Michalski<sup>120</sup> auf:

- Eingebautes Wissen (*direct implanting*): Diese Systeme lernen nicht im eigentlichen Sinn, sondern erhalten ihr gesamtes Wissen bereits bei der Konstruktion, z.B. in Form einer Datenbank, die möglicherweise in einem gewissen Rahmen mathematische oder statistische Schlussfolgerungen ziehen kann;
- Lernen nach Anleitung (*L. from instruction*): Das System wählt eingegebenes Wissen aus und überführt es in eine effiziente interne Repräsentation;
- Lernen durch Deduktion (*L. by deduction*): Hier zieht das System Schlüsse aus vorhandenem Hintergrundwissen (und gegebenenfalls neu dargebotenem Wissen). Es erwirbt so logisch äquivalentes oder spezifischeres Wissen. Erklärungsbasierte Systeme gehören zu dieser Klasse;
- Lernen aufgrund von Analogien (*L. by analogy*): Das System muss vorhandenes Wissen auf einen neuen Anwendungsbereich ausweiten;
- Lernen durch Beispiele (*L. from examples*): Solche Systeme lernen induktiv (vgl. oben symbolisch-empirische Systeme). Ein „Lehrer“ (bzw. Lehrer-Prozess) stellt die Beispiele für das zu lernende Konzept bewusst, systematisch oder zufällig zur Verfügung;
- Beobachtung und Entdeckung (*L. by observation and discovery*): Auch diese Klasse lernender Systeme arbeitet induktiv. Die Systeme müssen den Beobachtungsraum jedoch selbständig gliedern. Michalski unterscheidet passive Beobachtung (das System sucht eine Struktur in einer gegebenen Menge von Beobachtungen) und aktives Experimentieren (das System führt selbst Änderungen in einer gegebenen Umgebung herbei und beobachtet diese).

Für sprachwissenschaftliche Aufgaben sind – neben zahlreichen Systemen des induktiven Typs – in den letzten Jahren auch Analogien lernende Systeme in den Blickpunkt gerückt, eine Einführung gibt Daelemans<sup>121</sup>.

120 R. S. MICHALSKI (1987), Learning strategies and automated knowledge acquisition: An overview, in L. Bolc (Hrsg.), *Computational models of learning* (Heidelberg/New York: Springer), S.1-19

121 W. DAELEMANNNS (1999), Introduction, JETAI special issue on memory-based language processing, *JETAI* 11(3), S.287-292

### 7.3 Besprechung ausgewählter Arbeiten

Die Bestimmung des Tilgungsverhaltens lëtzebuergescher Wörter anhand der Graphemabfolge entspricht der Entdeckung eines Zusammenhangs zwischen der Form eines Objekts und seiner Kategorie. Ein vergleichbares Problem ist die Zuweisung von Wortarten (-Tags) an unbekannte Wortformen aufgrund ihrer morphologischen Struktur. Im Folgenden werden einige aktuelle Arbeiten besprochen, die sich diesem Problem von der Seite des maschinellen Lernens her nähern.

#### 7.3.1 Brill: Induktion von Regeln zur Wortartenbestimmung

Brill<sup>122,123</sup> entwarf einen Tagger, der aus einem annotierten Corpus in einem iterativen Lernprozess Regeln für die Wortartzuweisung induziert. Brills Verfahren, das er unter dem Namen „Transformation based error-driven learning“ (TBEDL) bekannt gemacht hat, wurde für die verschiedensten Fragestellungen adaptiert, so etwa die Desambiguierung der syntaktischen Funktion von Präpositionalphrasen<sup>124</sup> oder die automatische Informationsextraktion<sup>125</sup>.

Das System vergleicht den Ausgangszustand, ein mit einem kleinen Tag-Lexikon und einfachen Heuristiken vorgetaggetes Corpus, mit dem Zielzustand, dem selben Corpus mit richtig zugewiesenen Tags („*the truth*“). Dabei entwickelt das System eine Folge von Transformationsregeln, die die Annotation des Ausgangscorpus möglichst fehlerfrei in die des Zielcorpus überführen.

Jede Regel besteht aus einer Transformation („ersetze Tag *P* durch Tag *Q*“) und einer sie auslösenden Umgebung. Für jedes falsch zugewiesene Tag instantiiert das System eine vorgegebene Menge von Regelschemata als Regelhypothesen. Bei jeder Iteration des Lernvorgangs wird ermittelt, welche Regelhypothese die größtmögliche Zahl falsch zugewiesener Tags korrigiert; diese wird beibehalten.

Brills Lernverfahren lässt sich als überwachtes, induktives Lernverfahren einordnen (Brill entwickelte später auch eine unüberwachte Version, die ohne ein manuell getaggetes Trainingscorpus auskommt<sup>126</sup>). Der Prozess der

122 E. BRILL (1992), A simple rule-based part of speech tagger, in *Proc. of the third conference on applied natural language processing*

123 E. BRILL (1994), A report of recent progress in transformation-based error-driven learning, *Proc. of the 12. ARPA workshop on human language technology*, S.722-727

124 E. BRILL & P. RESNIK (1994), A rule-based approach to prepositional phrase attachment disambiguation, in *Proc. of the 15. COLING international conference*, S.998-1004

125 M.E. CALIFF (1998), Relational learning techniques for natural language information extraction (Dissertation), *Technical Report AI 98-276* (Austin: University of Texas)

126 E. BRILL & M. POP (1999), Unsupervised learning of disambiguation rules for part of

Regelfindung ähnelt dem *Decision-trees*-Verfahren; durch die vorgegebenen Regelschemata ist der Entscheidungsraum jedoch vorgegeben.<sup>127</sup>

Während des Lernprozesses wird in jedem Durchlauf der Nettobeitrag jeder Regelhypothese zur Gesamt-Aufklärungsquote gemessen. Dies ermöglicht die Fokussierung des Lernprozesses auf die erfolgversprechendsten Regeln (Schutz vor *overtraining*).<sup>128</sup> Dehaspe & Forrier<sup>129</sup> zeigen, dass sich die Effizienz von Brills Lernverfahren steigern lässt, wenn die Regelschemata nach ihrer logischen Allgemeinheit sortiert werden.

Ein interessanter Aspekt im Rahmen der vorliegenden Arbeit ist Brills Strategie bei der Behandlung unbekannter Wortformen. Hier liegt der Schwerpunkt nicht auf der Reduzierung an sich bekannter Lesarten anhand des Satzkontexts, sondern auf der Erstzuweisung einer plausiblen Wortkategorie aufgrund von morphologischen Eigenschaften der Form selbst. Die erste Taggerversion (1992) wies unbekanntem Wörtern als Ausgangsannotation die Tags *Nomen* oder *Eigennamen* zu, es stellte sich jedoch heraus, dass gerade diese Wörter nach Abschluss der Lernphase eine besonders hohe Fehlerrate aufwiesen. Brill setzt deshalb 1994 ein eigenes TBEDL-Lernsystem speziell für das Problem der Erstzuweisung von Tags an unbekannte Formen ein. Dieses System sucht nach Zusammenhängen zwischen Wortart und morphologischer Form: Kriterien sind häufige Affixe, Stamm-Ähnlichkeit (operationalisiert als Identität zweier Formen nach Tilgung oder Einfügung eines Teilstrings) und unmittelbar benachbarte Lexeme („unbekannte Wörter rechts der Form X erhalten das Tag Y“). Morphologische Prozesse wie Umlautung oder Assimilation lassen sich in den vorgegebenen Regelschemata jedoch nicht beschreiben. Hier setzt eine Arbeit von Andrei Mikheev an, die im folgenden Abschnitt besprochen wird.

### 7.3.2 Mikheev, Daille: Induktion morphologischer Regeln

Das von Mikheev<sup>130</sup> beschriebene Verfahren leitet Regeln aus einem allgemeinen Lexikon mit Wortartangaben ab und gewichtet sie nach der Frequenz der betreffenden Wörter in einem nicht annotierten Corpus (Mikheev nutzte hierfür das Brown-Corpus).

speech tagging, in: K. Church & al. (Hrsg.), *Natural language processing using very large corpora* (Dordrecht: Kluwer), S.27-42

127 L.A. RAMSHAW & M.P. MARCUS (1994), Exploring the statistical derivation of transformational rule sequences for part-of-speech tagging, in *The balancing act: Proc. of the ACL workshop on combining symbolic and statistical approaches*, S.93

128 *Ibid.*

129 L. DEHASPE & M. FORRIER (1999), Transformation-based learning meets frequent pattern discovery, *Proc. of the Language logic and learning workshop*, S.40-52

130 A. MIKHEEV (1997), Automatic rule induction for unknown word guessing, *Computational Linguistics* 23(3), S.405-423

Ähnlich wie die morphologische Komponente bei Brill, so generiert auch Mikheev Regeln, die bestimmte Affixe abtrennen und den verbleibenden „Stämmen“ entweder direkt eine Wortart zuweisen oder sie mit einem Lexikoneintrag in Beziehung setzen.

So können Regeln beispielsweise lauten: „Wenn das unbekannte Wort  $W$  in  $x+y$  zerlegt werden kann und die Form  $x+z$  im Lexikon eine Wortartmarkierung ( $p$  oder  $q$ ) hat, so weise  $W$  die Wortart  $r$  zu.“

Im dargestellten Experiment reduziert Mikheev die Menge der möglichen Regelschemata auf 3 allgemeine Endungsschemata und 3 Schemata, die auf Wörterbucheinträge zurückgreifen (Tabelle 17).

Tabelle 17. Endungsschemata ohne (obere drei), mit Wörterbuchwissen (untere drei).

Regelschema	Erläuterung	Beispiel
Ending <sup>-</sup>	Endungsregeln für Formen mit Bindestrich	evil- <u>doer</u> → NOMEN
Ending <sup>c</sup>	Endungsregeln für groß geschriebene Formen	Kash <u>mir</u> → EIGENNAME
Ending <sup>*</sup>	Endungsregeln für sonstige Formen	daring → ADJ
Suffix <sup>0</sup>	Durch Abtrennen der Endung entsteht ein Wort, das mit Wortart $p$ im Wörterbuch steht	book <u>ed</u> = book <sub>v</sub> + <u>ed</u> → (VFİN ∨ PART)
Suffix <sup>1</sup>	Die Ersetzung der Endung durch die Kette $z$ erzeugt ein Wort mit Wortart $p$	tr <u>ied</u> = try <sub>v</sub> - <u>y</u> + <u>ied</u> → (VFİN ∨ PART)
Prefix	Durch Abtrennen des Präfixes entsteht ein Wort mit Wortart $p$	<u>un</u> screw = <u>un</u> +screw <sub>v</sub> → VINF

Der Suchraum reduziert sich weiter durch strikte Längenbeschränkungen für die einzelnen Segmente (abzutrennende „Affixe“: bis 5, verbleibende „Stämme“: mindestens 3, einzufügende Segmente: höchstens 1 Graphem).

So unterscheidet sich Mikheevs Ansatz von dem Brills in vier Punkten:

- Der Algorithmus generiert Regeln gleicher Komplexität. Brills Algorithmus dagegen erzeugt eine geordnete Liste von interagierenden Regeln, die als komplexe Produktionen gelesen werden können (eine Regel kann vorhergehende Transformationen überschreiben). Damit sind Mikheevs Regeln weniger mächtig, jedoch leichter lesbar
- Regeln, die auf das Lexikon verweisen, können Beschränkungen (*constraints*) für die Wortarten der Einträge formulieren (bei Brill genügt stets das Vorhandensein eines Eintrags)
- Regeln können morphologische Substitution modellieren, wie etwa die Beziehung *try* ~ *tries*
- Die Regelanwendung weist einer Form, wenn nötig, die Menge aller möglichen Tags zu, während Brill stets das wahrscheinlichste Tag auswählt.

Daille<sup>131</sup> entwickelte nach dem Vorbild des Mikheevschen Algorithmus ein Verfahren, um Derivationsregeln für französische Adjektive (z.B. *gaz* ~ *gazeux*, *higiène* ~ *higiénique*) zu finden.

### 7.3.3 Gaussier: Unsupervised learning of derivational morphology

Gaussier<sup>132</sup> entwirft ein probabilistisches Modell der lexikalischen Derivation: Die Wahrscheinlichkeit dafür, dass ein Wort eine morphologische Variante eines anderen ist, wird aus der Wahrscheinlichkeit der orthographischen Transformation der einen Form in die andere abgeleitet. Das Verfahren zielt auf die automatische Segmentierung von lexikalischen Stämmen (*stemming*) und die Entdeckung lexikalischer Ableitungsprozesse. Die Lernbeispiele stammen aus einem Lexikon mit Wortarten-Tags. Das Verfahren lässt sich mit Michalski als ‚Lernen durch (passive) Beobachtung‘ klassifizieren.

Es beruht auf einem vereinfachten Derivationsmodell: Wörter werden dann als verwandt betrachtet, wenn sie aus einem „Stamm“ mit einer bestimmten Mindestlänge und einem Paar von „Affixen“ („Pseudosuffix-Paar“, ohne Beschränkung der Länge) gebildet werden können.

Der Grad der Verwandtheit wird probabilistisch bestimmt. Zwei Wörter sind umso näher „verwandt“, je mehr Wortpaare mit denselben Wortart-Tags und demselben, abtrennbaren Pseudosuffix-Paar im Lexikon vorkommen; die Wahrscheinlichkeit des Suffixpaars wird also zum Ähnlichkeitsmaß. So erscheint etwa das Wortpaar [*admiration*<sub>N</sub>, *admirer*<sub>V</sub>] als ähnlicher in Gaussiers Sinn als das Paar [*demonstrateur*<sub>A</sub>, *demonstration*<sub>N</sub>]; dies aufgrund der häufigeren Kookkurrenz des Suffixpaars [-*ion*+N, -*er*+V] mit jeweils gleichen „Stämmen“.

Aufgrund der so operationalisierten Ähnlichkeit wird eine Clusteroperation nach der Maximummethode durchgeführt, die Wortpaare zu Wortgruppen (sogenannten „relationalen Familien“) zusammenfasst.

Der letzte Verarbeitungsschritt besteht darin, innerhalb der Cluster paarweise alle möglichen Segmentierungen und Affigierungen, die jeweils eine Form in die andere überführen könnten, zu berechnen und nach dem Aufwand (Einfügungen und Löschungen) zu gewichten. Die Wortpaare und Gewichte errichten einen minimalen Spannbaum (*minimum spanning tree*). In diesem Prozess führt jede Einfügung eines zusätzlichen Tochterknotens zur Reduktion der Menge der möglichen Segmentierungen beim Mutterknoten:

131 B. DAILLE (2000), Morphological rule induction for terminology acquisition, *Proc. of the 18. COLING international conference*, S.215-221

132 E. GAUSSIER (1999), Unsupervised learning of derivational morphology from inflectional lexicons, in *Proc. of the ACL'99 workshop on unsupervised methods in natural language processing*

1. produire[produit] (produi/re,produ/ire,prod/uire,...)
2. produire[produit production] (produ/ire,prod/uire,...)

Die so entstehenden Ableitungsbäume interpretiert Gaussier als Derivationsparadigmata mit Ableitungsregeln. Die Iteration der Verarbeitungsschritte zwischen Clusterung und Berechnung der minimalen Bäume führt zu einer schrittweisen Reduktion der Regeln, bis ein stabiler Zustand erreicht ist.

## 7.4 Zusammenfassung

In Kapitel 6 trat das Problem auf, dass die Lexikoneinträge des CORTINA-Wörterbuchs nur zu einem kleinen Teil (3%) in ausreichender Zahl und in den für die Tilgungsregel relevanten Kontexten aus dem Corpus belegt werden konnten. Allerdings gilt dies, mehr oder weniger stark, für jedes aus einem Corpus zusammengestellte Lexikon – immer gibt es einen „langen Schwanz“ (Manning & Schütze<sup>133</sup>) extrem seltener Formen.

Ein möglicher Ausweg besteht darin, seltene Formen, die sich im Hinblick auf die zu beobachtende Eigenschaft gleich verhalten, zu einem Objekt zusammenzufassen. Dies tut auch der linguistisch geschulte Beobachter, wenn er verschiedene Flexionsformen zu einem Lemma zusammenfasst oder Komposita mit gleichem Grundwort als Klasse behandelt.

Für den Rechner stellt sich dabei ein Klassifizierungsproblem, das der Zuordnung lexikalischer Kategorien (vgl. Brill) oder der Erkennung von Wortfamilien (vgl. Daille, Gaussier) ähnelt.

Es wurden verschiedene Anwendungen maschinellen Lernens auf solche Fragestellungen besprochen. Brill generierte eine Menge von möglichen Regeln, die die Morphologie eines Wortes seiner lexikalischen Kategorie zuordnen, und wählte jeweils die Lösung mit der höchsten absoluten Aufklärungsquote; Mikheev entwickelte besonders das morphologische Instrumentarium dieses Verfahrens weiter; Gaussier integrierte das Konzept von Derivationsparadigmata, indem er Cluster von Regelableitungen bildete, die häufig gemeinsam auftraten.

Im folgenden Kapitel soll ein induktives Lernverfahren mit Elementen aus den vorgestellten Arbeiten auf das Problem der lätzebuergeschen  $n$ -Tilgung zugeschnitten werden.

133 C. MANNING & H. SCHÜTZE (1999), *Foundations of statistical natural language processing* (Cambridge, Mass.: MIT Press), S.23

## 8 Verallgemeinerung von Wortbildungsregeln aus dem Corpus

### 8.1 Vorüberlegungen

Die Anwendbarkeit der n-Tilgung auf ein Wort wird offenbar vom Kontext und von Eigenschaften des Wortes selbst gesteuert (vgl. Kap. 5 und 6). Von diesen beiden Einflussfaktoren schienen die Kontextbedingungen vergleichsweise einfach fassbar: Abgesehen von Kontexten auf s-, teilten sich die Folgekontexte klar in zwei disjunkte Klassen auf. Lediglich für eine kleine Gruppe von Folgesilben schienen sich die Klassengrenzen zu verwischen (vgl. Kap. 5).

Der vom Wort selbst ausgehende Einfluss auf die Tilgung erwies sich als komplizierter. Von zwei oberflächlich ähnlichen Wörtern konnte ohne weiteres eines im Tilgungskontext das auslautende -n verlieren, das andere aber nicht. Dennoch ist die sogenannte Eifeler Regel für Luxemburg durchweg als systematisch beschrieben worden. Die hier vorgelegten Ergebnisse widersprechen dem nicht: Der überwiegende Teil der im Corpus ausreichend belegten Lexeme gehört entweder der einen oder der anderen Klasse an, Zweifelsfälle sind meist auf Kollisionen (vgl. 5.2) zurückzuführen.

Der begrenzte Umfang der verfügbaren Corpora für das Lëtzebuergesche gab den Anstoß dafür, ein Verfahren einzusetzen, das aus den vorhandenen Belegen auf allgemeinere Regeln für den Zusammenhang zwischen Lexem und Tilgung schließt. Ein solches Verfahren wird im Folgenden vorgestellt. Es beruht auf der Annahme, dass ein hinreichend starker Zusammenhang zwischen der Graphemstruktur des einzelnen Wortes und seinen Tilgungseigenschaften besteht. Dabei ist zu berücksichtigen, dass in der Schreibung nicht alle phonologischen Charakteristika repräsentiert sind, insbesondere nicht suprasegmentale Tonkonturen, und dass der von manchen Autoren vermutete Einfluss von Wortbedeutung, Herkunft oder Wortart durch eine solche Analyse nicht erfasst werden kann (vgl. Kapitel 3).

Wenn es gelingt, Regeln zu finden, die aus der Schreibung des einzelnen Wortes eine treffsichere Vorhersage über sein Tilgungsverhalten machen, dann bietet eine solche Lösung zwei Vorteile: zum Einen lassen sich Regeln über Zeichenketten leicht in einen effizienten Stringerkenner überführen, der beispielsweise die Benutzer des Korrektursystems beim Anlegen eigener Wörterbücher unterstützen könnte. Zum Anderen bleibt die Regelform – im Gegensatz zu anderen Formen der Wissensrepräsentation, etwa Markovketten oder neuronale Netze – lesbar und interpretierbar, sie kann leicht ergänzt werden und bietet darüber hinaus Einblicke in die Wirkungs-

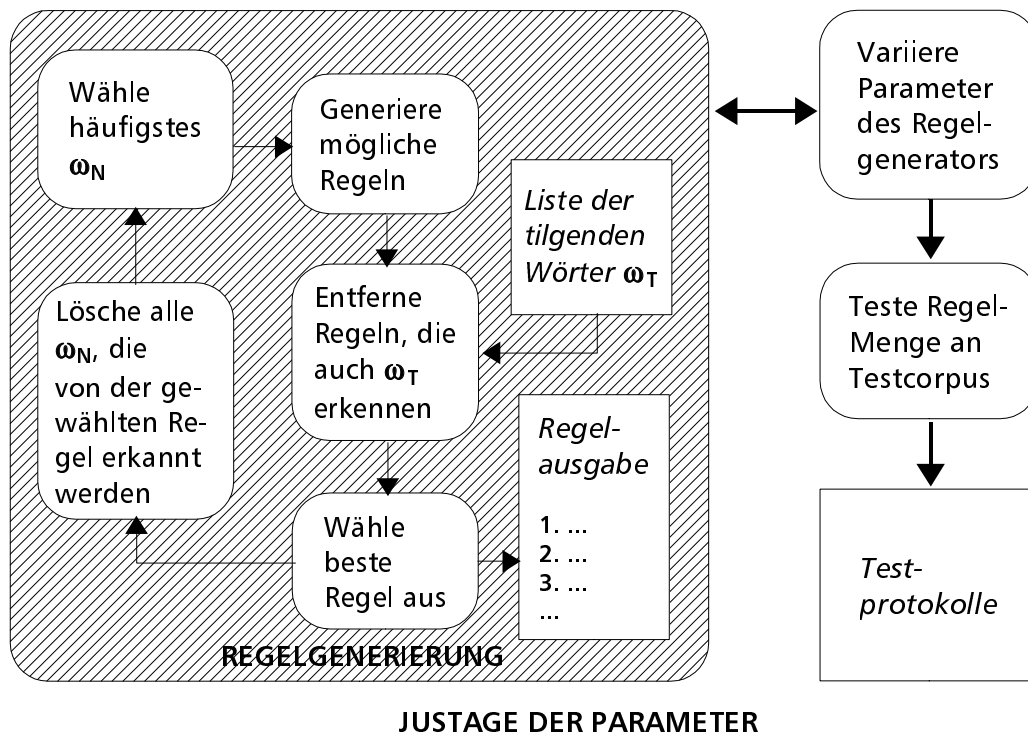


Abbildung 6. Blockdiagramm des Lernsystems mit Regelgenerator und Justageschleife

weise der n-Tilgung.

Das hier vorgestellte symbolisch-empirische Lernsystem<sup>134</sup> generiert solche Regeln aufgrund des Vergleichs einer Liste mit „guten“ (d.h., repräsentativen) Beispielen mit einem Textcorpus. Die Regeln formulieren Bedingungen über Graphemfolgen am Wortende und definieren eine Menge von nicht tilgenden Wortformen. Die Anlage des Lernsystems ist zweistufig: Es besteht aus einem Regelgenerator, der die dem System vorgelegten Beispiele durch eine Menge von Regeln möglichst effizient und erschöpfend „erklärt“, und aus einem Kontroll- und Justagesystem, das die Parameter des Generators systematisch variiert und die erzeugten Regelsätze an einem Testcorpus überprüft (vgl. Abbildung 6). Weder Lern- noch Testcorpus sind annotiert, das Lernverfahren ist also als vollständig unüberwacht einzuordnen.

In seinem Aufbau ähnelt das System Brills „Transformation based, error-driven learning“: Wie dieses misst es laufend den erreichten Nettogewinn und gibt zu jedem Zeitpunkt der jeweils „gewinnträchtigsten“ Regel den Vorzug. Die Gewinnung der Lernbeispiele aufgrund einer statistischen Klassifizierung der Belege aus dem Corpus ähnelt dagegen dem Vorgehen von Gaussier (vgl. Kapitel 7).

134 Vgl. BRISCOE & CAELLI (1996), S. 7

## 8.2 Stufe 1: Regelgenerator

Der Regelgenerator erzeugt Erkennungsregeln (*pattern matching rules*) für nichttilgende Lexeme ( $\omega_N$ ). Dies geschieht durch Verallgemeinerung von Beispielen und Gegenbeispielen aus dem T-Lexikon, die mit Angaben zur Frequenz im Corpus, zum Tilgungsverhalten und zum jeweiligen Anteil an auf -nn auslautenden Realisierungen annotiert sind.

### 8.2.1 Parameter und Regelschemata

Der Generator erhält als Parameter:

- Kriterien für die Auswahl der „guten Beispiele“, d.h., von regelmäßig tilgenden ( $\omega_T$ ) und regelmäßig nichttilgenden ( $\omega_N$ ) Wörtern aus dem T-Lexikon. Diese Kriterien sind Konfidenzniveau, Stichprobenverteilung (normal oder binomial) und Schwellenwerte (vgl. Kapitel 6)
- Beschränkungen über die zulässigen Regelschemata (*templates*).

Die Regelschemata bestehen aus zwei Grundtypen – einer Identitäts- und eine Teilsting-Regel –, die mit zwei weiteren Bedingungen kombinierbar sind, nämlich „Form endet mit -nn“<sup>135</sup> und „Form beginnt mit einer Majuskel“<sup>136</sup>. Aus diesen Kombinationsmöglichkeiten ergeben sich acht verschiedene Regelschemata (Tabelle 18).

Tabelle 18. Vorgegebene Regelschemata.

Schema	Erkannte Muster
gleich+N (<Arg>)	xxxn, xxxnn
gleich+nn (<Arg>)	xxxnn
gleich+Maj+N (<Arg>)	Xxxn, xxxnn
gleich+Maj+nn (<Arg>)	Xxxxnn
endet-mit+N (<Arg>)	-xxxn, -xxxnn
endet-mit+nn (<Arg>)	-xxxnn
endet-mit+Maj+N (<Arg>)	X...xxxn, X...xxxnn
endet-mit+Maj+nn (<Arg>)	X...xxxnn

Der Regelgenerator wird mit einer Parameterliste wie der folgenden aufgerufen:

$$\{S_{\omega_N}=0,5; S_{\omega_T}=0,5; C/(BIN,90\%); T=(I,T,nn)\}$$

In diesem Beispiel werden die Schwellen für die Tilgungsquoten der  $\omega_N$  und

135 Die Eigenschaft „Form endet auf -nn“ wurde mit folgender Bedingung bestimmt: (Formen auf -nn / (Formen auf -n + Formen auf -nn)) – e > t. e ist hier der Standardfehler des Quotienten multipliziert mit z=1,645. Für die Schwelle t wurde aufgrund einer Voruntersuchung der Wert 0,65 gewählt.

136 Die Majuskelregeln wurden im Experiment nicht realisiert.

der  $\omega_T$  jeweils auf 0,5 festgelegt (d.h., die  $\omega_N$  müssen in beiden Kontexten eine Tilgungsquote  $< 0,5$  haben; die  $\omega_T$  müssen im Erhaltungskontext eine Tilgungsquote  $< 0,5$ , im Tilgungskontext eine Quote  $> 0,5$  haben, vgl. Kapitel 6.3); beim Vergleich der Tilgungsquoten mit den Schwellenwerten kommt ein 90%iges Konfidenzintervall unter Annahme einer binomialen Stichprobenverteilung zum Einsatz; aus den Regelschemata werden Identitäts-, Teilstring- und nn-Endungsregeln zugelassen.

### 8.2.2 Regelgenerierung

Zu Beginn der Regelgenerierung bildet das Programm zwei Untermengen aus dem T-Lexikon: Eine Menge nicht tilgender Lexeme  $\omega_N$  als durch die Regeln zu erklärende Beispiele, und eine Menge tilgender Lexeme  $\omega_T$  als Gegenbeispiele. Aufgrund der positiven Beispiele werden die Schemata parametrisiert. Die Strategie des Algorithmus kann gierig (*greedy*) genannt werden, denn aus der Menge der zu einem gegebenen Zeitpunkt möglichen Parametrisierungen wird jeweils die am wenigsten spezifische ausgewählt. Die erzeugten Regeln können als Erkennungsregeln für die eingegebenen  $\omega_N$  verstanden werden.

Die Generierung selbst verläuft zyklisch in fünf Schritten:

1. Wähle aus der  $\omega_N$ -Liste das Wort mit der höchsten Frequenz aus
2. Bilde sämtliche Regeln, die dieses Wort erkennen würden
3. Verwirf alle Regeln, die zugleich eines der Wörter aus der  $\omega_T$ -Liste erkennen
4. Bewerte jede Regel anhand der Summe der Frequenzen der von ihr erkannten  $\omega_N$ ; gib die Regel mit der höchsten Bewertung aus
5. Lösche alle Wörter aus der  $\omega_N$ -Liste, die von der höchstbewerteten Regel erkannt wurden

Weiter bei 1., bis die  $\omega_N$ -Liste erschöpft ist.

Das folgende Beispiel zeigt die Menge der Regelhypothesen, die aufgrund des  $\omega_N$  „*Ämstänn*“ generiert werden (ohne Majuskel-Regeln):

```
gleich+nn (Ämstä)
gleich+N (Ämstä)
endet-mit+nn (Ämstä)
endet-mit+N (Ämstä)
endet-mit+nn (mstä)
endet-mit+N (mstä)
endet-mit+nn (stä)
endet-mit+N (stä)
endet-mit+nn (tä)
endet-mit+N (tä)
endet-mit+nn (ä)
endet-mit+N (ä)
```

### 8.2.3 Verfahrensalternativen

Aus der Menge der Regelhypothesen wird die allgemeinste (am wenigsten spezifische) Regel ausgewählt, die noch keinen Widerspruch zu den Gegenbeispielen  $\omega_T$  ergibt. Gegenüber diesem strengen Vorgehen lässt sich auch eine fehlertolerante Variante des Verfahrens denken: So müsste die Erkennung einzelner  $\omega_T$  durch eine Regel nicht automatisch zum Ausschluss dieser Regel führen, sondern würde beispielsweise ihre Bewertung mindern.

## 8.3 Stufe 2: Justage der Parameter

Die Aufrufparameter des im vorigen Abschnitt beschriebenen Regelgenerators  $\{S_{\omega_N}; S_{\omega_T}; C; T\}$  legen fest, wie die „guten Beispiele“ für das Regellernen bestimmt und welche Regeltypen dabei erzeugt werden sollen.

In der zweiten Lernphase, der Justage, werden diese Parameter systematisch variiert. Jeder Programmlauf mit einer neuen Parameterkonstellation erzeugt einen eigenen Regelsatz; jeder Regelsatz wird an den auf -n, -nn oder Vokal auslautenden Wörtern aus dem Testcorpus überprüft. Die Prüfbedingungen lauten: Wie viele Fälle von ausgebliebener Tilgung wurden vom Regelsatz richtig vorausgesagt, wie viele falsch, und wie viele gar nicht?

### 8.3.1 Testcorpus und Aufrufparameter

Als Testcorpus dient ein vom Lerncorpus unabhängiges Corpus (C8). In ihm ist, wie schon im Lerncorpus, die n-Tilgung *nicht* annotiert – ein Wort, das auf Vokal auslautet, kann also sowohl Ergebnis einer Tilgung als auch eine vom Kontext unbeeinflusste Form sein. Beim Testen kann deshalb nicht die absolute Fehlerzahl, sondern nur die relative Zahl im Vergleich der einzelnen Regelsätze gemessen werden.

Über einen Aufrufparameter kann das Justageprogramm veranlasst werden, eine Liste von Folgewörtern mit abweichendem Tilgungsdruck (vgl. Kapitel 5.4) einzulesen, die dann bei der Fehlerzählung nicht gewertet werden.

### 8.3.2 Ausgabe des Justagesystems

Die Treffer-/Fehlerwerte der einzelnen Regelsätze werden in einer Logdatei gesammelt und dazu herangezogen, den geeignetsten Regelsatz für den gewünschten Zweck – zunächst also für die Annotation des CORTINA-Wörterbuchs in Bezug auf die n-Tilgung – zu ermitteln.

Ein häufig eingesetztes Maß für die Bewertung mustererkennender Algorithmen ist die Genauigkeits-/Vollständigkeitsquote<sup>137</sup> (precision/recall),

137 Deutsche Termini folgen R. HENZLER (1992), *Information und Dokumentation* (Berlin: Springer), S.165

so auch in den zitierten Arbeiten zur Regelgeneralisierung von Mikheev<sup>138</sup> und Daille<sup>139</sup>. Die Vollständigkeitsquote drückt das Zahlenverhältnis zwischen erkannten und übersehenen Mustern aus (hits:misses), die Genauigkeitsquote jenes zwischen erkannten und irrtümlich erkannten (hits: false hits).

Zur Bewertung der Regelsätze wurden diese Maße getrennt für Tokens und Types berechnet. Hierfür mussten in der Testphase für jedes T-Wort  $w$  (aus dem T-Lexikon  $W$  des Testcorpus), in seinen Realisierungen  $w_e$  (ungetilgt) und  $w_t$  (getilgt), die Treffer  $h$ , die irrtümlichen Treffer  $f$  und die übersehenen Fälle  $m$  gezählt werden:

	$w_e$ im Tilgungskontext	$w_t$ im Tilgungskontext
$w$ von einer Regel erkannt	$h_w$	$f_w$
$w$ von keiner Regel erkannt	$m_w$	–

Die Genauigkeits-/Vollständigkeitsquoten (als *prec* und *recl* geschrieben) werden als Quotienten der Summen der Treffer- und Fehlerzahlen berechnet. Die auf Tokens bezogenen Maße werden aus den tatsächlichen Häufigkeiten berechnet, die auf Types bezogenen dagegen berücksichtigen nur je eine Treffer-/Fehlersituation pro Type.

$$\begin{aligned}
 prec_{TOK} &= \frac{\sum_{w \in W} h_w}{\sum_w f_w + h_w} \\
 recl_{TOK} &= \frac{\sum_w h_w}{\sum_w m_w + h_w} \\
 prec_{TYP} &= \frac{\sum_w \delta(h_w)}{\sum_w \delta(f_w) + \delta(h_w)} ; \delta(x) = \begin{cases} 0: & x=0 \\ 1: & x>0 \end{cases} \\
 recl_{TYP} &= \frac{\sum_w \delta(h_w)}{\sum_w \delta(m_w) + \delta(h_w)}
 \end{aligned}$$

Bei der Bewertung der verglichenen Regelsätze wurden die Tokenmaße herangezogen; ihre Summe musste maximal sein, das heißt, ein schlechterer Genauigkeitswert konnte durch einen besseren Vollständigkeitswert ausgeglichen werden. Die Betrachtung der Tokenmaße anstelle der Typemaße bevorzugt Regeln, die besonders häufige Lexeme erkennen; dies scheint durch die geplante Anwendung gerechtfertigt: Die Akkuratheit eines mit diesen Regeln annotierten Wöretbuchs wird für häufig gebrauchte Wörter höher sein als für seltene.

### 8.3.3 Bestimmung eines optimalen Regelsatzes

Ein Beispiel für die Regelbewertung zeigen die Tabellen 19 und 20. Im hier gezeigten Testlauf waren alle Regeltemplates außer Majuskelregeln zugelassen, die Schwellenwerte für die  $\omega_T$  und die  $\omega_N$  wurden unter Verwendung eines binomialen Konfidenzintervalls (90%,  $z = 1,645$ ) zwischen 0,35 und 0,95 in Schritten von 0,1 variiert. Zum Vergleich wurden zwei Klassen mit minimalen Zugehörigkeitsbedingungen einbezogen: Hier genügten drei Belege für jedes Wort, sofern jeweils kein Beleg für „falsches“ Tilgungsverhalten vorlag; in der Auswertung sind diese Klassen mit *wk* (weak) anstelle des Schwellenwerts gekennzeichnet.<sup>140</sup>

T-Wörter, auf die ein Wort mit besonders niedrigem Tilgungsdruck folgte (Liste aus Kapitel 5, Abschnitt 4), wurden bei der Zählung nicht berücksichtigt.

Tabelle 19. Genauigkeits- (linke Tabelle) und Vollständigkeitsquoten (rechts) für verschiedene Parametrisierungen des Regelgenerators, bezogen auf Tokens.

$S_{\omega_T} \backslash S_{\omega_N}$		$S_{\omega_N}$								$S_{\omega_T}$							
		wk	0.35	0.45	0.55	0.65	0.75	0.85	0.95	wk	0.35	0.45	0.55	0.65	0.75	0.85	0.95
wk		0,74	<b>0,94</b>	<b>0,94</b>	<b>0,99</b>	<b>0,99</b>	<b>0,99</b>	1	1	0,63	<b>0,67</b>	<b>0,66</b>	<b>0,63</b>	<b>0,63</b>	<b>0,61</b>	0,54	0,5
0.35		0,98	0,99	0,99	0,98	0,98	0,98	1	1	0,57	0,56	0,53	0,47	0,47	0,41	0,29	0,1
0.45		0,95	0,94	0,94	0,99	0,99	0,99	1	1	0,6	0,62	0,6	0,56	0,55	0,54	0,47	0,3
0.55		0,93	0,94	0,94	0,99	0,99	0,99	1	1	0,62	0,64	0,64	0,6	0,6	0,59	0,52	0,47
0.65		0,91	0,94	0,94	0,99	0,99	0,99	1	1	0,62	0,64	0,64	0,6	0,6	0,59	0,52	0,47
0.75		0,7	<b>0,94</b>	<b>0,94</b>	<b>0,99</b>	<b>0,99</b>	<b>0,99</b>	1	1	0,64	<b>0,67</b>	<b>0,66</b>	<b>0,63</b>	<b>0,63</b>	<b>0,61</b>	0,54	0,5
0.85		0,68	0,91	0,91	0,96	0,96	0,96	1	1	0,64	0,67	0,67	0,63	0,63	0,62	0,54	0,5
0.95		0,43	0,89	0,89	0,96	0,96	0,96	1	1	0,67	0,67	0,67	0,63	0,63	0,62	0,54	0,5

CI:(BIN.90%): T:(I.T.nn)

Tabelle 20. Genauigkeits-/Vollständigkeitsquoten bezogen auf Types; die unterstrichenen Felder kennzeichnen optimale Regelsätze (vgl. im Text).

$S_{\omega_T} \backslash S_{\omega_N}$		$S_{\omega_N}$								$S_{\omega_T}$							
		wk	0.35	0.45	0.55	0.65	0.75	0.85	0.95	wk	0.35	0.45	0.55	0.65	0.75	0.85	0.95
wk		<u>0,98</u>	<u>0,98</u>	<u>0,98</u>	<u>0,99</u>	<u>0,99</u>	0,99	1	1	<u>0,53</u>	<u>0,5</u>	<u>0,5</u>	<u>0,47</u>	<u>0,47</u>	0,45	0,43	0,41
0.35		0,99	0,98	0,98	0,98	0,98	0,98	1	1	0,45	0,37	0,34	0,28	0,27	0,2	0,13	0,03
0.45		<u>0,97</u>	0,98	0,98	0,99	0,99	0,99	1	1	<u>0,5</u>	0,43	0,41	0,38	0,37	0,36	0,33	0,23
0.55		<u>0,95</u>	<u>0,98</u>	0,98	0,99	0,99	0,99	1	1	<u>0,52</u>	<u>0,47</u>	0,46	0,43	0,43	0,42	0,39	0,37
0.65		<u>0,94</u>	0,97	0,97	0,99	0,99	0,99	1	1	<u>0,52</u>	0,47	0,46	0,43	0,43	0,42	0,39	0,37
0.75		<u>0,92</u>	<u>0,98</u>	<u>0,98</u>	<u>0,99</u>	<u>0,99</u>	0,99	1	1	<u>0,54</u>	<u>0,5</u>	<u>0,5</u>	<u>0,47</u>	<u>0,47</u>	0,45	0,43	0,41
0.85		0,87	<u>0,96</u>	<u>0,95</u>	0,96	0,96	0,97	1	1	0,55	<u>0,51</u>	<u>0,5</u>	0,48	0,48	0,46	0,43	0,41
0.95		0,69	<u>0,94</u>	<u>0,94</u>	0,96	0,96	0,97	1	1	0,6	<u>0,51</u>	<u>0,51</u>	0,48	0,48	0,46	0,43	0,41

CI:(BIN.90%): T:(I.T.nn)

140 Die genaue Zugehörigkeitsbedingung für minimale „sichere Tilger“  $\omega_T$  war disjunktiv formuliert: Es genügten entweder drei Belege für Erhalt im Erhaltskontext oder Tilgung im Tilgungskontext bzw. kein Beleg für die umgekehrten Fälle, oder es mussten 90%-konfident über 95% der Belege in den „zulässigen“ und weniger als 10% in den „unzulässigen“ Kontexten vorliegen. Eine analoge disjunktive Bedingung galt für die minimalen „sicheren Nicht-Tilger“  $\omega_N$ .

Jedes Tabellenfeld steht für einen einzelnen Regelsatz, der aus den  $\omega_T$  und  $\omega_N$  zu den am Tabellenrand angegebenen Schwellenwerten (zeilenweise:  $\omega_T$ -Schwellen, spaltenweise:  $\omega_N$ -Schwellen) generalisiert wurde. Die jeweils linken Tabellen zeigen die Genauigkeits-, die rechten die Vollständigkeitsquoten. Die durch Unterstreichung hervorgehobenen Felder repräsentieren Regelsätze, für die die Summe beider Quoten maximal ist.

Das Ergebnis ist noch nicht eindeutig; mehrere Regelsätze erreichen (annähernd) gleiche Werte. Zieht man jedoch die nach Types berechneten Quoten zur Vereindeutigung heran, so bleibt genau ein „optimaler“ Regelsatz übrig: Es ist der Regelsatz ( $S_{\omega_T}=0,75$ ;  $S_{\omega_N}=0,35$ ), für den die Summe  $\sum_{\text{Tok,Typ}} \text{prec} + \text{recl} = 3,09$  maximal ist.

## 8.4 Evaluation der generierten Regeln

Das Cortina-Wörterbuch (Stand vom April 2001, mit ca. 75 000 Einträgen, davon ca. 17 000 mit -n-Auslaut) wurde versuchsweise mit einem maschinell erzeugten Regelsatz annotiert. Die Regeln definieren Nicht-Tilger  $\omega_N$ , das heißt, bei den von keiner Regel erfassten Einträgen sollte es sich um Tilger  $\omega_T$  handeln. Ein Teil dieser Wörter wurde anschließend von Jérôme Lulling, dem sprachwissenschaftlichen Mitarbeiter des Projekts CORTINA und Luxemburger Muttersprachler, überprüft. Die Korrekturen wurden in einen neuen Regelsatz überführt, so dass die seither im Projekt eingesetzten Regeln als Produkt eines teilüberwachten Lernverfahrens angesehen werden müssen. Dennoch ist es aufschlussreich, einige der so zu Tage getretenen Schwachstellen der vollständig unüberwacht gelernten Regeln zu analysieren.

Bei der manuellen Überprüfung erhielten 513 zuvor nicht annotierte Einträge ein Tag „n“ für „tilgt nicht“. Hiervon waren 60 Ortsnamen, die nach dem vorliegenden Entwurf<sup>141</sup> zum „texte coordonné“ – einer Ergänzung und Erläuterung des Gesetzestextes von 1999 – grundsätzlich nicht getilgt werden sollen.<sup>142</sup>

### 8.4.1 Movierte Feminina

Die größte Gruppe unter den übrigen Formen sind 132 movierte Feminina, wie z.B.

*Chefin, Biologin, Gesellin, Ierfpränzessin, Asiatin, Fabrikantin, Éirepresidentin, Haaptkonkurrentin, Atheistin, Hygiène spezialistin, Fundamentalistin, Feministin, Duerchschnëttstouristin,*

141 SCHANEN & LULLING (2001)

142 Dies gilt unabhängig von der Phonologie der betreffenden Namen. Das Problem der Ortsnamen wird nun durch eine Ausnahmeliste nach BRAUN (1999) gelöst: *Eis Sprooch richte gschreiwten*, 3. Auflage (Bartreng: Rapidpress)

### *Friemepolizistin*<sup>143</sup>

Das Corpus enthält nur wenige Belege für movierte Feminina überhaupt. Eine Zählung der Formen auf *-istin* illustriert dies. Im Corpus (C48) sind nur 10 Types belegt, davon kein einziger mit mehr als zwei Tokens.<sup>144</sup> Da die „schwächste“ Bedingung, die eine Form *w* erfüllen musste, um einer  $\omega_N$ -Partition anzugehören,  $|w| > 3$  lautete (d.h., für  $S_{\omega_N} = ,wk'$ ), war keine der *-istin*-Bildungen in den beim Lernen eingesetzten Partitionen vertreten.

#### 8.4.2 Phonetisch nicht voll integrierte Wörter

Eine weitere größere Gruppe bilden, mit 40 Formen, phonetisch nicht voll integrierte französische Lehnwörter mit nasaliertem Vokal / $\tilde{a}$ /, / $\tilde{e}$ /, / $\tilde{i}$ /, / $\tilde{o}$ / im Auslaut (z.B. *Carcan*, *Bassin*, *Moyen*, *Assurancëjargon*). Hier muss der gewählte Ansatz, der auf der Verallgemeinerung von geschriebenen Endungen ohne phonetische Annotation beruht, scheitern. Den nasalierten Formen auf *-en* stehen zahlreiche ähnliche, mit dem regelmäßig tilgenden lëtzebuergeschen Pluralmorphem *-en* gebildete Formen gegenüber, vgl. etwa die Paare *Doyen*<sub>SG</sub> – *Gameboyen*<sub>PL</sub>, *Soutien*<sub>SG</sub> – *Soucien*<sub>PL</sub> (zu *Souci* ‚Sorge‘).

Nicht tilgbar sind auch manche englischen Lehnwörter: Hier waren dies die Formen *Fan*, *Slogan*, *Screen*, *Login*.

#### 8.4.3 Vorderer Vokal + *-nn*

Bei den auf vorderen Vokal+nn auslautenden Formen, die vom System falsch klassifiziert wurden, sind zwei Gruppen zu unterscheiden: Imperativformen (22 Formen) und Nomina auf *-ënn* (ebenfalls 22 Formen).

Imperative<sup>145</sup> wie *fann* (0), *erfann* (1), *bënn* (0), *verdënn* (0), *erkenn* (0), *verblenn* (0), *penn* (0), *verbrenn* (4) sind im Corpus praktisch nicht vorhanden (die Beispiele wurden willkürlich ausgewählt, in Klammern ist jeweils die Zahl der Belege vermerkt).

Die Erkennung der Regelmäßigkeit der Nichttilgung der Wörter auf *-ënn* – eine kurze Sichtung des Corpus lässt vermuten, dass Formen auf *-ë* durchweg auf Plurale wie *Frontaliëren*, *Cassiëren* zurückgehen; die einzige

143 Eine „gefensterte“ Auswahl: Jede 10. Form aus der nach Endungen sortierten Liste.

144 Die männlichen Ausgangsformen zeigen erheblich höhere Frequenz: *Polizistin* ( $f=2$ ) – *Polizist* ( $f=56$ ); *Pianistin* (1) – *Pianist* (3); *Galeristin* (1) – *Galerist* (3); *Artistin* (1) – *Artist* (28); *Automobilistin* (2) – *Automobilist* (19). Dies gilt umso mehr, als die männlichen Formen im Corpus – der ja eine Zusammenstellung von auf *-n/-V* auslautenden Formen und ihrer Kontexte ist – nicht vollständig erhalten sind. – Nicht belegbar sind dagegen die Ausgangsformen zu *Floralistin*, *Parachutistin*, *Receptionistin*, *Feministin* und *Terroristin* (jedoch 8 mal die Pluralform: *Terroriste/Terroristen*).

145 Die lëtzebuergesche Form zu dt. *ich find'*, *ich bind'* lautet dagegen stets auf *-en*: *ech fannen*, *ech bënnen*.

Ausnahme bildet offenbar *geschwënn*, das zu *geschwë* getilgt werden kann – wird dem System offenbar durch die gewählte Operationalisierung der T-Wörter erschwert: *ë* in der letzten Silbe wurde wie *e* behandelt, um die Beziehung zwischen Tilgungspaaren wie *Avenuen* ~ *Avenuë* erfassen zu können. Das System „übersah“ hier vor allem Komposita zu *-grënn* (,Gründe‘, 7), *-sënn* (,Sünde‘, ,Sinn‘, 4), *-frënn* (,Freunde‘, 4) und *-hënn* (,Hunde‘, ,Hähne‘, 2).

#### 8.4.4 Weitere nicht erkannte Formen

Nicht erkannt wurden im übrigen

- 36 Formen auf *-o:n* und *-onn*: Komposita zu *-won* (,Wagen‘), *-zon* (,Zone‘); Partizipia auf *-bonn* und *-wonn* (bzw. *-schwonn*),
- 24 Formen auf *-e:n*, darunter Komposita zu *-ween* (,Wagen‘) und *-reen* (,Regen‘) (von Bruch<sup>146</sup> historisch mit der Synkope eines intervokalischen *-g-* erklärt), aber auch *-steen* (,Stein‘), *-zeen* (,Szene‘) und *Zireen* (,Sirene‘). Die Tilgung der Komposita zu *-steen* ist im Corpus allerdings sehr uneinheitlich, dort finden sich „*Grafsteen müssen*“, „*Meilesteen gesat*“ neben „*Bordstee gerannt*“ und „*Schläifstee geschlaff*“.<sup>147</sup> Das CORTINA-Wörterbuch verzeichnet immerhin *Steebléck* und *Steeplatten*,
- 19 Formen auf *-i:n*, z.B. *Protein*, *Hermelin*, *Termin*, *Kantin*, *Festungsruin*, *Buslawin*,
- 17 Formen auf *-éin*: das Verb *begéin* und Komposita zu *-léin* (,Löhne‘), *-spéin* (,Späne‘), *-tréin* (,Träne‘) und *-téin* (,Töne‘),
- 13 Formen auf *-a:n*, darunter *Dekan*, *Fasan*, *Cellophan* und Komposita zu *-organ*,
- 5 Substantiva auf *-unn*: Komposita zu *-dunn* (,Balken‘) und *-hunn* (,Hahn‘).

### 8.5 Zusammenfassung

Mit dem in diesem Kapitel vorgestellten Regellernverfahren konnten gute Erkennungsraten für das Testcorpus C8 erreicht werden: Der ermittelte „beste“ Regelsatz erreichte eine korrekte Klassifikation für 97,2% der auf *-n* endenden Wortformen des Testcorpus.

Das Schwellenpaar, das die Trainingsdaten für diesen Regelsatz definiert, zeichnet sich durch eine hohe Schwelle für die  $\omega_T$  und eine niedrige für die

146 BRUCH (1954), S.27

147 Eine gesonderte Durchsicht der Belege zu *Steen* ‚Stein‘ (Grundform und Komposita) zeigt freie Variation, teilweise sogar in Texten des selben Autors. Ein Zusammenhang der Tilgungsquote mit Kasus oder Numerus ist nicht zu erkennen.

$\omega_N$  aus. Damit hat das System sich sozusagen für eine bestimmte Lernstrategie „entschieden“: Da die Regeln aufgrund der  $\omega_N$  generiert, aufgrund der  $\omega_T$  verworfen werden, bedeutet diese Schwellenkonfiguration, dass zunächst zu viele, zu detaillierte Regeln erzeugt und dann an einer kleinen, zuverlässigen Liste von Gegenbeispielen überprüft werden.

Die hier noch nicht realisierten „Majuskelregeln“ versprechen eine Steigerung der Trefferquote, da ein relativ starker Zusammenhang zwischen den Tilgungseigenschaften und der Großschreibung eines Wortes zu bestehen scheint (vgl. Kapitel 9).

Bei der Evaluation am Wörterbuch zeigt sich das Verfahren stark bei häufigen, dagegen schwächer bei seltenen Wörtern. Das Verfahren ist also insofern robust, als es in realen Texten vergleichsweise wenig Fehler macht. Im konkreten Fall – der Annotation eines Wörterbuchs mit seinem typischen, hohen Anteil an seltenen Formen – ist dieses Verhalten weniger günstig.

Als falsch hat es sich erwiesen, den Unterschied zwischen *ë* und *e* in der Auslautsilbe zu vernachlässigen. Hier muss eine andere Lösung gefunden werden – vielleicht analog zur Behandlung von *-n/-nn*, die beim T-Wort zwar fehlen, aber eine numerische „Spur“ in der Spalte „-nn-Wahrscheinlichkeit“ hinterlassen.

## 9 Ausblick

### 9.1 Lexem und Tilgung

Anhand der nun vorliegenden Daten können weitere Untersuchungen zur schriftlichen Anwendung der „Eifeler Regel“ durchgeführt werden. Einige Zusammenhänge zwischen Eigenschaften und Tilgung eines Wortes sollen hier kurz angerissen werden.

Offenbar besteht ein starker Zusammenhang zwischen der Wortart und der Tilgungsklasse, wenn man Großschreibung eines Wortes als Indiz für die Wortart Substantiv<sup>148</sup> gelten lässt:

*Tabelle 21.* Anteil großgeschriebener Formen im T-Lexikon gegenüber den Tilgungsklassen  $\omega_T$  (tilgt regulär) und  $\omega_N$  (tilgt nicht)

	T-Lexikon zu C40		$\omega_T$		$\omega_N$	
Tokens	346371		267756		7779	
Types	8936		859		101	
Caps (Token)	66302	19,1%	34476	12,9%	7454	95,8%
Caps (Types)	5224	58,5%	377	44,0%	93	92,2%

Nie-Tilger sind also meistens (93 von 101 Types) Substantive! Dieser Zusammenhang ist auch dann noch auffallend, wenn man berücksichtigt, dass Substantive als typische „Inhaltswörter“ einen hohen Anteil an den niederfrequenten Lexemen haben, von denen wir gesehen haben, dass sie ohnehin eher zum n-Erhalt neigen.

*Tabelle 22.* Anteil von Formen mit wahrscheinlicher -nn-Schreibung

	T-Lexikon		$\omega_T$		$\omega_N$	
Tokens	346371		267756		7779	
-nn (Token)	21962	6,3%	18477	6,9%	2165	27,8%

Die Doppel-,n'-Schreibung ist offenbar ein Indiz für eine phonologische Struktur, die die ,n'-Tilgung blockiert (Tabelle 22): Der Anteil der Formen mit doppeltem ,n'<sup>149</sup> ist bei den  $\omega_N$  fast viermal so hoch wie bei den  $\omega_T$ .

Einen weiteren Hinweis auf phonologische Faktoren ergibt der Vergleich der Vokale in den Endsilben (Tabelle 23). Bei den regulär tilgenden  $\omega_T$  herrschen

148 Großschreibung am Satzanfang ist im T-Lexikon eigens ausgezeichnet und hier nicht eingerechnet.

149 Formen mit einer hohen Wahrscheinlichkeit der Doppel-,n'-Schreibung, vgl. 5.2 oben



## 9.2 Schlussbemerkung

Auf den Nicht-Muttersprachler, der die richtige Handhabung der n-Tilgung im Lëtzebuergesch erlernen möchte, wirkt die Vielfalt und Verschiedenartigkeit der Ausnahmen verwirrend. Aber auch Muttersprachlern fällt es meiner Beobachtung nach meist schwer, die Kriterien für Tilgung oder Erhalt in Form von Regeln anzugeben. Nach einzelnen Wörtern gefragt – was im Verlauf der Entstehung dieser Arbeit öfter vorkam –, probierten viele Sprecher zuerst halblaut das Wort im Satzzusammenhang aus, bevor sie es in die eine oder andere Klasse einordneten.

Auf diese Intuition kann der Lernende natürlich nicht zurückgreifen. Der hier beschrittene Weg der Reduktion und Abstraktion von Beispielen aus einem Textcorpus könnte dabei helfen, gewissermaßen Faustregeln für die n-Tilgung zu entwickeln, die das Lehren und Lernen von Lëtzebuergesch als Fremdsprache erleichtern.

Der beschriebene „Sprechtest“ der Muttersprachler passt zu der (bei Gilles, aber auch implizit bei Bruch geäußerten) These, dass Länge- und Akzentverhältnisse in der Auslautsilbe eine entscheidende Rolle für die Anwendbarkeit der Tilgung spielen. Dennoch können die Bedingungen der „Eifeler Regel“ für die Schriftsprache wohl nicht vollständig aus der Sprechweise abgeleitet werden. Hier spielen offenbar auch Kriterien wie die Wahrung der Erkennbarkeit des Schriftbildes – so beim Verzicht auf Tilgung bei Eigen- und Ortsnamen oder bei kurzen Worten wie „flunn“, „zunn“ und „Wäin“ – eine Rolle. Darüber hinaus glaube ich, Belege für eine Verselbständigung der orthographischen Form gefunden zu haben wie die in Kapitel 8 erwähnte Reihe „Reen“ – „Zireen“ – „Steen“, von der nach dem Bruchschens Modell nur das erste Wort (wegen der durch die -g--Synkope bewirkten Akzentverhältnisse) ein „Nie-Tilger“ sein sollte. Wenn diese Beobachtung zutrifft, dann deutet sie, gemeinsam mit der von Gilles und anderen beobachteten Tendenz zur Vereinfachung und Verallgemeinerung der Tilgung, auf einen Entwicklungsprozess hin: Möglicherweise beeinflusst die zunehmende Schriftlichkeit der lëtzebuergesch Sprache schon heute die Form und Gültigkeit der „Eifeler Regel“.

## A Literaturverzeichnis

- Arrêté Ministériel (1975). Arrêté ministériel du 10 octobre 1975 portant réforme du système officiel d'orthographe luxembourgeoise. In: *Mémorial. Amtsblatt des Großherzogtums Luxemburg B* 68, S.1366-1390
- Berg, G. (1993). „Mir wëlle bleiwe, wat mir sin.“ *Soziolinguistische und sprachtypologische Betrachtungen zur luxemburgischen Mehrsprachigkeit*. Tübingen: Niemeyer
- Bergenholtz, H. & J. Mugdan (1989). Korpusproblematik in der Computerlinguistik: Konstruktionsprinzipien und Repräsentativität. In I.S. Batori, W. Lenders, & W. Putschke (Hrsg.), *Computational Linguistics*. Berlin: De Gruyter, S.141-149
- Biber, D. (1993). Representativeness in corpus design. *J. of literary and linguistic computing* 8(4), S.243-257
- BNC (2000). Composition of the BNC. Internet: <http://info.ox.ac.uk/bnc/what/balance.html>
- Booth, B. (1987). Text input and pre-processing: Dealing with the orthographic form of text. In R. Garside, G. Leech & A. McEnery (Hrsg.), *Corpus Annotation*. London: Longman Addison-Wesley, S.97-109
- Braun, J. (1999). *Eis Sprooch richteg schreiwen*. 3. Auflage. Bartreng: Rapidpress
- Brill, E. & M. Marcus (1992). Tagging an unfamiliar text with minimal human supervision. *Proc. of the 1992 AAAI Fall Symposium*
- Brill, E. & M. Pop (1999). Unsupervised learning of disambiguation rules for part of speech tagging. In K. Church & al. (Hrsg.), *Natural language processing using very large corpora*. Dordrecht: Kluwer, S.27-42
- Brill, E. & P. Resnik (1994). A rule-based approach to prepositional phrase attachment disambiguation. *Proc. of the 15. COLING international conference*, S.998-1004
- Brill, E. (1992). A simple rule-based part of speech tagger. *Proc. of the third conference on applied natural language processing*
- (1994). A report of recent progress in transformation-based error-driven learning. *Proc. of the 12. ARPA workshop on human language technology*, S.722-727
- Briscoe, G. & T. Caelli (1996). *Symbolic machine learning*. Band 1 von *A compendium of machine learning* (2 Bände). Norwood: Ablex
- Bruch, R. (1953). *Grundlegung einer Geschichte des Luxemburgischen*. Luxemburg: Paul Linden (Reihe: Publications littéraires et scientifiques du ministère de l'éducation nationale)
- (1954). *Das Luxemburgische im westfränkischen Kreis*. Luxemburg: Paul Linden
- (1955). Précis populaire de grammaire Luxembourgeoise. Neudruck 1973 in L.Senniger (Hrsg.), *Beiträge zur lux. Sprach- und Volkskunde X* (3., durchgesehene Auflage). Luxemburg: Institut Grand-Ducal
- Bußmann, H. (1990). *Lexikon der Sprachwissenschaft*. Stuttgart: Kröner
- Califf, M.E. (1998). Relational learning techniques for natural language information extraction (Dissertation). *Technical Report AI 98-276*. Austin: University of Texas
- Capesius, B. (1966). Die Behandlung des auslautenden n in den siebenbürgisch-sächsischen Mundarten (Die sogenannte „Eifler Regel“). *Zeitschrift für Mundartforschung* 33, S.97-126

- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Mass.: MIT Press
- Christa, P. (1927). *Wörterbuch der Trierer Mundart. Mit Sprachgesetzen derselben und Sprachproben in Prosa und Poesie*. Seitenidentischer Neudruck 1989, Wiesbaden: Sändig
- Church, K.W. & R.L. Mercer (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics* 19(1), S.1-24
- Daelemanns, W. (1999). Introduction. JETAI special issue on memory-based language processing. *J. of empirical and theoretical Artificial Intelligence* 11(3), S.287-292
- Daille, B. (2000). Morphological rule induction for terminology acquisition. *Proc. of the 18. COLING international conference*, S.215-221
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics* 19(1), S. 61-73
- Dehaspe, L. & M. Forrier (1999). Transformation-based learning meets frequent pattern discovery. *Proc. of the language logic and learning workshop*, S.40-52
- Engwall, G. (1994). Not chance but choice: Criteria in corpus creation. In B.T. Atkins & A. Zampolli (Hrsg.), *Computational approaches to the lexicon*. Oxford: University Press, S.49-82
- EURYDICE (2001). Description nationale du Luxembourg, complément à l'étude d'Eurydice „L'enseignement des langues étrangères en milieu scolaire en Europe“. Luxemburg: Ministère de la culture, de l'enseignement supérieur et de la recherche
- Gangler, J.-F. (1847). *Lexicon der Luxemburger Umgangssprache (wie sie in und um Luxemburg gesprochen wird)*. Seitenidentischer Neudruck 1973, Wiesbaden: Sändig
- Gaussier, E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. *Proc. of the ACL'99 workshop on unsupervised methods in natural language processing*
- Gilles, P. (1999). *Dialektausgleich im Lëtzebuergeschen: Zur phonetisch-phonologischen Fokussierung einer Nationalsprache*. Tübingen: Niemeyer
- (2001). Phonologische und variationslinguistische Aspekte der -n-Tilgung (,bewegliches -n') im Lëtzebuergeschen. Erscheint in K. J. Mattheier & E. Radtke (Hrsg.), *Sammelband des Symposiums ‚Variation in der Sprache‘*, Heidelberg Dez. 1995. Frankfurt: Lang (im Druck)
- Graf, R. (2001). Loi sur la nationalité: Compli-simplification. In: *woxx* 596 (6.7.2001), S.3
- Grin, F. & N. Weber (2000). Multilingualism and language policy in Luxembourg. In T.L. du Plessis & K. Deprez (Hrsg.), *Multilingualism and Government*. Pretoria: Van Schaik, S.82-91
- Henzler, R. (1992). *Information und Dokumentation*. Berlin: Springer
- Hoffmann, F. (1964/1967). *Geschichte der Luxemburger Mundartdichtung* (2 Bände). Luxemburg: Bourg-Bourger
- (1987). Pragmatik und Soziologie des Lëtzebuergeschen. In J.P. Goudaillier (Hrsg.), *Aspekte des Lëtzebuergeschen*, Beiträge zur Phonetik und Linguistik 55. Hamburg: Buske
- Klein, K.K. (1955). Das Luxemburger Wörterbuch aus siebenbürgisch-sächsischer Sicht. *Zeitschrift für Mundartforschung* 23, S.237
- Kodratoff, Y. (1988). *Introduction to machine learning*. London: Pitman
- Köhler, R. & M. Galle (1993). Dynamische Eigenschaften von Textmaßen. In H.P. Pütz & J. Haller (Hrsg.), *Sprachtechnologie: Methoden, Werkzeuge, Perspektiven*, Sprache und Computer 13. Hildesheim: Olms, S.3-15
- Kroemmer, M.-Th., J. Hansen, & L. Roth (2000). *Klengen Dictionnaire „Franséisch*

- Lëtzebuergesch*". Luxemburg: Actioun Lëtzebuergesch/Imprimerie St. Paul (auch unter <http://www.eis-sprooch.lu>)
- La Fontaine, Edmond de (1855). Versuch über die Orthographie der luxemburger deutschen Mundart.  
Benutzt wurde die Werkausgabe 1982, C. Hury (Hrsg.), *Dicks Gesamtwierk* (III), (Luxemburg: Kripler-Muller)
- Leech, G. (1987). General introduction. In R. Garside, G. Leech, & G. Sampson (Hrsg.), *The computational analysis of English*. London: Longman, S.1-15
- (1997). Introducing corpus annotation. In R. Garside, R. Leech, & T. McEnery (Hrsg.), *Corpus Annotation. Linguistic Information from computer text corpora*. London & New York: Longman, S.1-18
- Luxemburger Wörterbuch (1950-1977). Luxemburg: Linden
- Manning, C. & H. Schütze (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press
- McEnery, T. & A. Wilson (1996). *Corpus Linguistics*. Edinburgh: University Press
- McEnery, T., J.-M. Langé, M. Oakes, & J. Véronis (1987). The exploitation of multilingual annotated corpora for term extraction. In R. Garside, G. Leech & A. McEnery (Hrsg.), *Corpus Annotation*. London: Longman Addison-Wesley, S.220-230
- Meyer, Antoine (1845). *Luxemburgische Gedichte und Fabeln. Nebst einer grammatischen Einleitung und einer Wörtererklärung der dem Dialekt mehr oder weniger eigenartigen Ausdrücken von Gloden*. Brüssel: Delavigne & Callewaert
- (1854). *Règelbüchelchen vum Lezeburger Orthœgraf, en Uress als Pro'w, d'Fraèchen aus dem Hâ, a Versen*. Lüttich: H. Dessain
- Michalski, R. S. (1987). Learning strategies and automated knowledge acquisition: An overview. In L. Bolc (Hrsg.), *Computational models of learning*. Heidelberg/New York: Springer, S.1-19
- Mikheev, A. (1997). Automatic rule induction for unknown word guessing. *Computational Linguistics* 23(3), S.405-423
- Naur, P., J.W. Backus, F.L. Bauer, & al. (1960). Report on the algorithmic language ALGOL 60. *Communications of the ACM* 3, S.299-314
- Paulides, J.-P. (1952). Das Luxemburgische im Spiegel des Niederländischen. Vergleichende Studien zur luxemburgischen und niederländischen Lautlehre unter Berücksichtigung der deutschen. In *Annuaire de la section de linguistique, ethnologie et onomastique* (Luxemburg: Linden/Institut Grand-Ducal), S.51-104
- Ramshaw, L.A. & M.P. Marcus (1994). Exploring the statistical derivation of transformational rule sequences for part-of-speech tagging. *The balancing act: Proc. of the ACL workshop on combining symbolic and statistical approaches*, S.93
- Règlement Grand-Ducal (1999). Règlement grand-ducal du 30 juillet 1999 portant réforme du système officiel d'orthographe luxembourgeoise. In: *Mémorial. Amtsblatt des Großherzogtums Luxemburg* A 112, S.2040-2048
- Rewenig, G. (1997). *Summerzauber. En turbulent Kaméidistéck an zwee Akten fräi no Motiver aus „La trilogia della villeggiatura" vum Carlo Goldoni*. Echternach: Ed. Phi
- Rodange, M. (1872). Renert, oder de Fuuss am Frack an a Ma'nsgrésst.  
Benutzt wurde die Werkausgabe 1974, C. Meder (Hrsg.), *Gesamtwierk*, Reihe Klassiker vun der Lëtzebuenger Litteratur (Luxemburg: Kripler-Muller)
- Sampson, G. (1987). Probabilistic models of analysis. In R. Garside, G. Leech, & G. Sampson (Hrsg.), *The computational analysis of English*. London: Longman, S.27-28

- Schanen, F. & J. Lulling (2001). Texte coordonné (unveröffentlicht)
- Scheiner, A. (1896). *Die Mundart der Siebenbürger Sachsen*. Forschungen zur deutschen Landes- und Volkskunde 9, 2 Bde. Stuttgart: Engelhorn (Band 2 liegt vor als seitenidentischer Neudruck, 1971, Wiesbaden: Martin Sändig)
- Sperberg-McQueen, C.M. & Lou Burnard (Hrsg.) (1999). *TEI Guidelines for Electronic Text Encoding and Interchange (P3)*. (3., revidierte Fassung.) Chicago und Oxford: ACH/ACL/ALLC Text Encoding Initiative (im Internet erhältlich per Listserver TEI-L unter uicvm.uic.edu)
- Tockert, J. (1926). Ueber Luxemburgische Lexikographie. In: Luxemburgische Sprachgesellschaft (Gesellschaft für Sprach- und Dialektforschung), *Jahrbuch* 1925, S.30-51.
- Wetzels, W.L. (1993). La phonologie de la flexion adnomiale dans un dialecte Limbourgeois (Pays-Bas). In B. Laks & A. Riolland (Hrsg.), *Architecture des représentations phonologiques*. Paris: Editions du CNRS, S.203-231

## ***B* Anhang**

*Table B1. Corpus-Indexdatei (efc.idx) mit sämtlichen Texten des verwendeten Corpus. Angaben über Tokenzahl, Jahr und Autor (Felder 4,6,7). „Tokens sampled“ gibt die Zahl der mit Vokal oder -n auslautenden Tokens an*

---

```

# Index to Corpus
# <Text-No>:<First-Line>:<Filename>:<Words-in-Text>:
  <Tokens-Sampled>:<Publication-Date>:<Author>
1:1:E.lit.Atlantik.Re:82867:31974:1985:Guy Rewenig
2:31979:E.lit.Eisefresser.Re:8977:2889:1994:Guy Rewenig
3:34872:E.lit.Geeschter.Hosch:24798:9415:1997:Jhemp Hoscheit
4:44291:E.lit.Kréiwénkel.Bra:48256:18530:1998:Josy Braun
5:62825:E.lit.Palazzo.Re:19502:7393:1998:Guy Rewenig
6:70222:E.lit.Pazifik.Re:58630:22089:1998:Guy Rewenig
7:92315:E.lit.Pica.Hosch:91720:34815:1998:Jhemp Hoscheit
8:127134:E.lit.Porto.Bra:47733:18112:1997:Josy Braun
9:145250:E.lit.Prenz.Bra:15629:6387:1994:Josy Braun
10:151641:E.lit.Rick.Kart:7853:3179:1986:Rene Kartheiser
11:154824:E.lit.Spunten.ap.Scha:11065:4417:1999:Raymond Schaack
12:159245:E.lit.Spunten.arm.Scha:9758:3883:1999:Raymond Schaack
13:163132:E.lit.Spunten.aus.cours.Sch:10283:3851:1999:Raymond Schaack
14:166987:E.lit.Spunten.eischt.Scha:6613:2496:1999:Raymond Schaack
15:169487:E.lit.Spunten.inter.Scha:4098:1647:1999:Raymond Schaack
16:171138:E.lit.Spunten.parais.1.Scha:10777:4194:1999:Raymond Schaack
17:175336:E.lit.Spunten.parais.2.Scha:12462:4741:1999:Raymond Schaack
18:180081:E.lit.Spunten.parais.3.Scha:13176:5087:1999:Raymond Schaack
19:185172:E.lit.Spunten.parais.4.Scha:10547:3989:1999:Raymond Schaack
20:189165:E.lit.Spunten.stage.Scha:11776:4563:1999:Raymond Schaack
21:193732:E.lit.Steps.Scha:99955:38158:1995:Raymond Schaack
22:231894:E.lit.Waasser.Haus:53465:20202:1998:Georges Hausemer
23:252100:E.pol.Dec.gvt.99:8171:3801::
24:255903:E.pol.Etat.nat.2000:14152:6304::
25:262209:E.pol.Etat.nat.99:10702:4401::
26:266612:P.lc.No1.1:74439:31943::
27:298557:P.lc.No1.2:52054:21691::
28:320250:P.pol.Cr.chamb.13X98:6698:2470::
29:322722:P.pol.Cr.chamb.14X98:14832:5683::
30:328407:P.pol.Cr.chamb.15II2000:30931:12947::
31:341356:P.pol.Cr.chamb.15X98:28897:11601::
32:352959:P.pol.Cr.chamb.16II2000:38033:15695::
33:368656:P.pol.Cr.chamb.16XI199&15II2000:41888:17539::
34:386197:P.pol.Cr.chamb.17II2000:34075:13754::
35:399953:P.pol.Cr.chamb.21III2000:43994:16487::
36:416442:P.pol.Cr.chamb.22III2000:39776:15622::
37:432066:P.pol.Cr.chamb.24XI98:25344:9766::
38:441834:P.pol.Cr.chamb.25XI98:20806:8288::
39:450124:P.pol.Cr.chamb.26XI98:23292:9721::
40:459847:P.pol.Cr.chamb.27X98:21854:9308::
41:469157:lit.HellegMuechtCorrige:55720:23564:2000:Fernand le Chartreux
42:492725:lit.hosch.Johannes-1-3.doc:7614:3105:unbekannt:Jhemp Hoscheit
43:495834:lit.hosch.KaleKaffi.edite2.doc:16573:6306:2000:Jhemp Hoscheit
44:502144:lit.re.SUMMERZAUBER 2.doc:22273:7182:1997:Guy Rewenig
45:509330:pol.ch.21VI2000.S39 99-00 3.doc:26156:9451:21.6.2000:
46:518784:pol.ch.22III2000.CHD2000-26-27-:78912:29615:22.3.2000:
47:548402:pol.ch.25V2000.S37-38DELPHINE.d:46775:16814:25.5.2000:
48:565219:pol.ch.27VI2000.S40 99-00 2.doc:32984:11777:27.6.2000:
# Summary: <text-class>,<No-of-members>,<Total-No-of-words>,
  <Total-tokens-sampled>
# 1c,2,126493,53634
# lit,26,762120,292168
# pol,20,588272,231044
# Summary done.

```

---

Corpus-Auswahl C40: Zeilen 1-40, C8: Zeilen 41-48, C48: Zeilen 1-48

**Tabelle B2.** Zu Kapitel 6.3: Liste der T-Wörter, die aufgrund untypischer Distribution in den Erhalts- und Tilgungskontexten (EK und TK) nicht klassifiziert wurden.

Die Spalten E/EK, T/EK usw. enthalten die jeweilige Zahl der Erhalts- bzw. Tilgungsrealisierungen im angegebenen Kontext. Die Angaben in der Spalte Problemklasse stehen für: B – Sing.-Pl.-Kollision; C – Nom.-Akk.-Kollision; D – Orthographie; E – Nicht verwandte Lexeme; ? – Zweifelsfälle, uneinheitliche Handhabung der Tilgung im Corpus.

T-Wort	Freq.	E/EK	T/EK	E/TK	T/TK	Problemklasse
ausgesiN	92	28	0	3	5	?
BouN	21	0	6	5	5	?
ExameN	103	37	0	18	27	?
FührerschäiN	23	6	0	5	5	?
ItalieN	17	8	0	2	4	?
ReeN	54	21	0	5	3	?
SchäiN	20	12	0	2	5	?
SpuenieN	26	10	0	4	2	?
SteeN	80	25	6	12	17	?
administrativeN	24	8	4	0	7	B
AmbassadeN	29	3	10	1	5	B
AssembléeN	24	2	7	0	15	B
AssuranceN	45	2	14	0	18	B
ChanceN	139	16	35	1	33	B
ClienteN	103	38	9	12	16	B
CommerceN	51	1	29	0	6	B
CoureurN	18	6	0	1	6	B
DebatteN	18	4	3	0	6	B
DemandeN	68	2	28	0	26	B
DemokratieN	38	2	14	0	8	B
DeponieN	25	6	3	0	10	B
DisqueN	17	2	5	0	6	B
EnquêteN	41	3	12	0	15	B
EnveloppeN	41	4	10	1	18	B
EpicerieN	41	3	16	0	10	B
ExerciceN	18	2	6	0	8	B
FailliteN	67	11	12	0	23	B
GarageN	61	4	12	0	15	B
GarantieN	27	1	6	0	9	B
GareN	67	1	23	0	27	B
GesteN	32	3	9	1	10	B
GlaceN	52	2	14	1	24	B
KategorieN	26	3	5	0	11	B
LycéeN	89	1	40	1	31	B
MadameN	309	0	30	1	29	B
MessageN	26	1	10	0	8	B
MinistèreN	110	4	64	0	20	B
NoexamenN	30	7	0	7	8	B
PartieN	68	3	21	0	31	B
PassageN	20	1	10	0	6	B
PropriétaireN	47	5	14	0	14	B
ReprocheN	20	4	2	0	7	B
SchiN	28	12	0	5	1	B
SiteN	50	3	16	0	18	B

T-Wort	Freq.	E/EK	T/EK	E/TK	T/TK	Problemklasse
StageN	18	2	7	0	7	B
ToiletteN	35	2	8	0	12	B
VisiteN	21	3	7	0	6	B
VitesseN	32	2	18	0	7	B
ZoneN	33	2	10	0	15	B
EltreN	40	21	0	2	7	C
largeN	14	1	5	0	6	C
ännerstëtzeN	77	10	0	1	6	D
danzeN	38	9	0	1	6	D
deeN	6828	3334	0	8	2488	D
GrompereN	22	8	0	1	6	D
maacheN	2121	303	0	7	245	D
mengeN	1125	309	0	1	36	D
neeN	83	5	1	1	6	D
rappeN	44	13	0	1	5	D
ruffeN	84	33	0	1	6	D
WeencheN	16	6	0	1	5	D
zuN	2907	1	1151	1	874	D
DecheN	32	12	0	2	8	D?
DeN-&unsafeCap;	5109	1018	3	2	1651	E
duN	1651	177	600	5	623	E, s.o. (Du)
DuN-&unsafeCap;	755	104	165	0	422	E, s.o. (Du)
SiN-&unsafeCap;	1566	11	672	0	639	E, s.o. (Si)
DoN	138	0	41	2	74	E Don=Doen (tun), Do (Adv.)
DuN	252	7	82	0	134	E Dunn=elo (dann), Du (Pron)
EeN	57	8	10	0	20	E Een (Artikel), Ee (Ei)
noN	1946	13	947	1	560	E Frz. non, no (nach)
HuN	30	13	1	2	7	E Hunn (1.Hahn, 2.Haben)
MaN-&unsafeCap;	449	4	302	0	97	E Mann, Ma (aber)
maN	850	18	528	6	223	E mann (wenig), man (Pron.), man=maachen, ma (aber)
MoN	65	3	10	1	27	E Monn (Mund), frz. Mon
doN	4561	13	1739	0	1541	E s.o. (Do)
seN	3847	17	1805	2	1309	E sen/senn/sënn = sinn (Verb), se (Pronomen)
SiN	120	10	41	0	45	E Sinn (Verb), Si (Pronomen)
SoN-&unsafeCap;	49	1	18	1	16	E Sonn (Sonne), So (Partikel)
TéiN	30	3	6	9	4	E Téin (Töne), Téi (Tee)
WeeN	502	4	116	6	237	E Ween (Pl. zu Won, dt.: Wagen), Wee (Weg)

*Tabelle B3.* (Zu Kapitel 8.) Höchstbewerteter Regelsatz ( $S_{\omega_T} = 0,75$ ;  $S_{\omega_N} = 0,35$ ). Die Angabe ADH/DIS steht für die Zahl der Lernbeispiele  $\omega_T$  und  $\omega_N$ . Spalte 2 zeigt die Regelbewertung während der Generierung (Token-Zahl der von dieser Regel erkannten Beispiele).

---

# RB Levels: 0.75/0.35 ADH/DIS: 346/162	
endet-mit+N Ma	731
endet-mit+N E	574
endet-mit+N ou	3904
endet-mit+N Se	269
endet-mit+N ä	507
endet-mit+N to	276
endet-mit+N io	310
endet-mit+N So	147
endet-mit+N ai	144
endet-mit+N ro	228
endet-mit+N ta	134
endet-mit+N Wo	96
endet-mit+N bu	116
endet-mit+N la	108
endet-mit+N Gre	70
endet-mit+N fo	65
endet-mit+N h	61
endet-mit+N Fre	58
endet-mit+N chi	127
endet-mit+N zi	72
endet-mit+N Li	45
endet-mit+N Ke	43
endet-mit+N täi	39
endet-mit+N ço	35
endet-mit+N r	57
endet-mit+N ri	52
endet-mit+N ko	29
endet-mit+N Kru	29
endet-mit+N lo	41
endet-mit+N ma	26
endet-mit+N Ho	26
endet-mit+N pho	25
endet-mit+N di	25
endet-mit+N bi	24
endet-mit+N Pa	22
endet-mit+N He	21
endet-mit+N hwe	20
endet-mit+N co	19
endet-mit+N ewa	19
endet-mit+N ü	18
endet-mit+N cra	17
endet-mit+N Ka	16
endet-mit+N Mi	15
endet-mit+N Gewe	15
endet-mit+N pa	14
gleich+N i	14
endet-mit+N oi	11
endet-mit+N Bu	11
endet-mit+N Léi	10
endet-mit+N lka	10
endet-mit+N dwa	9
endet-mit+N Zee	8

