III. Empirische Grundlagen und methodische Voraussetzungen[1]

Im vorangegangenen zweiten, der Forschungsgeschichte gewidmeten Kapitel sind durchgehend Angaben zur empirischen Grundlage sowie zu den methodologischen Voraussetzungen der besprochenen Studien gemacht worden. Da beides die Ergebnisse einer jeden Untersuchung hinsichtlich ihres praktischen Erkenntniswertes sowie ihrer theoretischen Reichweite entscheidend bestimmten, soll dieser Themenbereich in diesem Kapitel für die vorliegende Arbeit ausführlich behandelt werden.

Ausgehend von der Existenz einer Wirklichkeit, die dem Menschen als solche grundsätzlich zugänglich ist, soll hier nach einer empirischen Grundlage und nach einer entsprechenden Methode für deren Auswertung nach dem allseits anerkannten Prinzip gesucht werden, daß Sprachfähigkeit (fr. *langage*) und Sprachsysteme (fr. *langue*) sich dem/r Wissenschaftler/in hauptsächlich durch den Sprachgebrauch (fr. *parole*) erschließen. Die Frage nach den empirischen Grundlagen und nach der entsprechenden Methode für deren Auswertung soll nun in den vier Schritten von (1) beantwortet werden (Abschnitte 1-4):

(1) Fragestellungen zur Datenbasis der Untersuchung
   a. *Gewinnung*: Was für empirische Daten gibt es und wie können und sollten sie gewonnen werden?
   b. *Aufbereitung*: Wie kann und sollte die Datenbasis aufbereitet werden?
   c. *Abfrage*: Welche Informationen können und sollten aus der aufbereiteten Datenbasis gewonnen werden?
   d. *Analyse*: Wie kann und sollte die gewonnene Information analysiert werden?

1. Zur Gewinnung der Datenbasis

Der erste Schritt der Untersuchung besteht also darin, empirisches Datenmaterial zusammenzutragen. Hier stellt sich die erste Frage: Was für empirische Daten gibt es und wie können und sollten sie gewonnen werden? Um diese Frage zu beantworten, muß zuerst eine Typologie der in der Forschung benutzten Daten aufgestellt werden. Dies geschieht in (2):

(2) Typologie der empirischen sprachlichen Daten (nach López Morales 1994; Moreno Cabrera 2002[2]: 56-57 und Sarmiento 1995: 91-100)
   a. «Interne» Daten (aus Eigenbeobachtung bzw. Introspektion)

---

[1] An dieser Stelle bedanke ich mich recht herzlich bei Herrn Prof. Dr. Guillermo Rojo (Real Academia Española de la Lengua & Universidade de Santiago de Compostela) und bei der von ihm geleiteten Forschungsgruppe «Grupo de Sintaxe do Español», insbesondere bei Frau Prof. Dr. Belén López Meirama, für ihr Entgegenkommen, für viele nützliche Informationen, für die Erlaubnis, das hier vorgestellte Corpus ARTHUS und die dazugehörige Datenbank BDS zu benutzen, sowie für ihre ständige Hilfe bei meinen beiden Forschungsaufenthalten an der Universidade de Santiago de Compostela (Februar 1999 und September 2000) und darüber hinaus.

   i. Daten aus der Kompetenz des Forschers
   ii. Daten aus der Kompetenz von Testpersonen
b. «Externe» Daten (aus Fremdbeobachtung)
   i. Zweckgesteuertes Sammeln von Daten mit vorherigem Wissen der Informanten (Interview-Corpora)
   ii. Zweckentfremdendes Sammeln von Daten ohne vorheriges Wissen der Informanten (Text-Corpora)

Die wichtigste Trennungslinie in dieser Klassifikation verläuft zwischen «internen» (2a) und «externen» (2b) Daten (vgl. Sarmiento 1995: 91-100). Chomsky (1986: 15-50) unterschied die *internalisierte Sprache* (engl. *Internalized Language*, *I-Language*) von der *externalisierten Sprache* (engl. *Externalized Language*, *E-Language*). Die erste ist die mentale Repräsentation des sprachlichen Wissens über die Einzelsprache, welche die universale Grammatik und gewisse sprachspezifisch gesetzte Parameter einschließt. Demgegenüber umfaßt die letzte den Prozeß und das Ergebnis der Anwendung von internalisierter Sprache in der Kommunikationssituation. Beide Datentypen werden in unserer Untersuchung als Ausdrücke sprachlicher Wirklichkeit anerkannt. Damit setzen wir uns also über die Diskussion zwischen «Generativismus» und «Funktionalismus» bewußt hinweg, inwiefern internalisierte bzw. externalisierte Sprache jeweils gültige Grundlagen für sprachwissenschaftliche Untersuchungen darstellen. Der «Generativismus» benutzt interne Daten und begründet dies mit der inhärenten Fehlerhaftigkeit externer Daten, während der «Funktionalismus» externe Daten bevorzugt. Im Endeffekt geht es den ersten eher um die Kognition und den zweiten eher um die Kommunikation, wobei es wünschenswert erscheint, eine umfassende Sicht anzustreben, d. h. beides zu verbinden, anstatt beides gegeneinander auszuspielen (vgl. aber die ausführliche Diskussion von Mensching 2005).

Interne Daten stammen also aus der «muttersprachlichen» Kenntnis einer Sprache (engl. *competence*) und werden durch bewußte «*Selbstbeobachtung*» bzw. «*Introspektion*» gewonnen. Der Wissenschaftler kann entweder sich selber (2ai) oder andere (2aii) als Informanten nehmen. Insbesondere im zweiten Fall kann zwischen einer eher passiven und einer eher aktiven Beteiligung des Informanten an der Datengewinnung unterschieden werden. Bei der passiven Untersuchung bekommt der Informant Fragebögen, auf dem Sätze stehen, zu denen er (Grammatikalitäts-)Urteile abgeben soll.[2] Bei der aktiven Untersuchung löst der Informant Aufgaben aus Testblättern, bei denen er Sätze produzieren soll, die gewisse, schon vorgegebene Elemente enthalten sollen. Interne Daten erlauben es somit, die Trennungslinie zwischen grammatikalischen und ungrammatikalischen Sätzen ziemlich genau und nachvollziehbar zu bestimmen (vgl. zu diesem Fragenkomplex Schütze 1996). Allerdings kann die ganze Bandbreite grammatikalisch möglicher Äußerungen in einer Sprache von einzelnen Informanten nicht gedeckt werden. Außerdem sind Aussagen, die Unterschiede im Gebrauch grammatikalisch möglicher Varianten betreffen, eher als subjektiv zu bewerten und daher systemlinguistisch kaum verwertbar, da es häufig zu unklaren oder einander widersprechenden Urteilen kommt. Aus diesen Gründen sind solche Daten für Untersuchungen zum Variationsraum einer Variablen denkbar ungeeignet und spielen in dieser Arbeit nur am Rande eine Rolle. Es dürfte trotzdem selbstverständlich sein, daß es

[2] Vgl. z.B. die empirische Studie und die methodologischen Überlegungen von Liceras (1994) hinsichtlich der Subjektnachstellung.

kaum möglich ist, die Kompetenz des Verfassers (2ai) auszuschalten, und daß diese immer wieder die Untersuchung bewußt oder unbewußt beeinflussen wird.

> Man geht tatsächlich immer von Texten aus, auch wenn man den Eindruck hat, man befrage mittels Introspektion die eigene Kompetenz, d. h. die nicht aktualisierte Fähigkeit, Äußerungen nach den Regeln einer bestimmten Sprache hervorzubringen. Auch wenn man den Eindruck hat, man tue bei der Beschreibung nichts anderes, als das eigene Wissen explizit zu machen, geht man doch von Texten aus, die das Produkt einer Art von «innerem» Sprechen sind, man analysiert Texte, die im Vollzug dieses «inneren Sprechens» bereits nach den Regeln hervorgebracht worden sind, die man beschreiben möchte. (Coseriu 1994[3]/1980: 39)

Externe Daten stammen aus der Umsetzung der «muttersprachlichen» Kenntnisse einer Sprache in konkreten Äußerungen im Rahmen echter Kommunikationssituationen (engl. *performance*) und werden durch «*Fremdbeobachtung*» gewonnen. Dabei darf der Wissenschaftler sich selbst *per definitionem* nicht «beobachten»; die Daten müssen entsprechend von anderen Personen stammen, die als Informanten dienen, wobei die Kommunikationssituation auch bei ihnen bewußte «Introspektion» möglichst ausschließen sollte, was nur im Idealfall gelingen dürfte. Auf der einen Seite können künstliche «Informationssituationen» hergestellt werden, bei denen die Informanten mehr oder minder im Klaren sind, daß sie Informationen für eine Untersuchung liefern (2bi). In diesem Fall hat es der Wissenschaftler in der Hand, die «Informationssituation» zu steuern oder nicht. Auf der anderen Seite können Daten aus Kommunikationssituationen gewonnen werden, die ursprünglich nicht als Quelle von sprachwissenschaftlich verwertbaren Informationen gedacht waren (2bii). Dabei ist dem Wissenschaftler naturgemäß jede Möglichkeit der Steuerung verwehrt.

Die Arbeit mit Informanten wird häufig von der Soziolinguistik genutzt, weil sie es erlaubt, relevante Auskünfte zu Person und Sprachbewußtsein der Informanten zu sammeln. Beim «zweckgesteuerten Sammeln» kann es auf der einen Seite je nach Ziel der Untersuchung sinnvoll und notwendig sein, die «Informationssituation» in eine gewisse Richtung zu lenken, um gewisse Daten von den Informanten spontan zu erhalten, die sonst nicht vorkommen würden. Auf der anderen Seite ist es nicht auszuschließen, daß die «Künstlichkeit» der «Informationssituation» zu «gekünstelten» bzw. «künstlichen» Äußerungen führen kann (López Morales 1994: 75-84). Dagegen vermeidet das «zweckentfremdende Sammeln» teilweise das Problem der «Künstlichkeit», doch können sich auch mangelnde Beeinflussungsmöglichkeiten negativ auf die anschließende Aufbereitung und Verwertung des gesammelten Materials auswirken. Außerdem sind die Möglichkeiten, Auskünfte über die Informanten und über die Kommunikationssituation zu erhalten, sehr beschränkt. Einziges Kriterium für die Wahl zwischen den beiden Methoden ist Ziel und Zweck der Untersuchung. Wie weiter unten ausgeführt wird, werden die empirischen Daten, die die Grundlage dieser Untersuchung bilden, aus beiden Typen gewonnen, denn der Bereich der Syntax gilt zurecht oder zu unrecht als der am wenigsten durch diese Unterschiede belastete (vgl. aber die Diskussion in Martín Butragueño 1997). Externe Daten erlauben zwar keinerlei Aussagen über ungrammatikalische Äußerungen (vgl. aber De Kock 2005), doch erlauben sie es, die breite Palette von Varianten einer Variable in ihrem kommunikativen Kontext nuanciert zu beschreiben und deren Gebrauch zu erklären. Darum eignet sich diese Datenart besonders gut, um Untersuchungen zur Variabilität im Sprachgebrauch wie die hier vorliegende durchzuführen (vgl. Cowart 1997).

2. Zur Aufbereitung der Datenbasis

Der zweite Schritt der Untersuchung besteht in der Aufbereitung des nach den genannten Methoden zusammengetragenen empirischen Datenmaterials. Die Entwicklung neuer Werkzeuge und Methoden im Bereich der Gewinnung, Aufbereitung und Analyse von empirischem Datenmaterial, die unter dem Stichwort *Corpuslinguistik* läuft (vgl. z.B. McEnery/Wilson 2001[2]; Habert/Nazarenko/Salem 1997; Biber/Conrad/Reppen 1998; Kennedy 1998), bringt es mit sich, vorzugsweise auf sogenannte *Corpora* von Sprache zurückzugreifen. Es gibt mittlerweile Bestrebungen, die Kriterien für Definition, Beschreibung und Einteilung von Corpora europaweit zu vereinheitlichen. Ausdruck davon sind die von John Sinclair (1996) entworfenen *Preliminary Recommendations on Corpus Typology*, welche die Politik von EAGLES (***E**xpert **A**dvisory **G**roup on **L**anguage **E**ngineering **S**tandards*, Pisa) bestimmen sollen. Sie werden im Folgenden kurz vorgestellt und bilden den Hintergrund für die Vorstellung des in dieser Untersuchung benutzten Corpus.

2.1 Zur Corpusdefinition und den spanischen Corpora

Der Begriff «Corpus» wird nach EAGLES folgendermaßen definiert und gegen andere Typen von Material(an)sammlungen abgegrenzt:

(3) Sammlungen empirischen Datenmaterials: Definitionen (Sinclair 1996; vgl. auch Marcos Marín 1994: 84; Sánchez 1995: 8-14)
   a. A *corpus* is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.
   b. Words such as *collection* and *archive* refer to sets of texts that do not need to be selected, or do not need to be ordered or the selection and/or ordering do not need to be on linguistic criteria.

Entscheidend ist also der Zweck der Materialsammlung: Es geht darum, einen «repräsentativen» Querschnitt (Stichprobe) einer Einzelsprache, wohl im Sinne der externalisierten Sprache (Gesamtpopulation), zu erhalten. Dieser Zweck entscheidet darüber, ob überhaupt eine Auswahl und eine Aufbereitung stattfinden sollen, und wenn ja, darüber, was für Kriterien dazu benutzt werden. Dies bedeutet zugleich den Verzicht auf den Gebrauch zufälliger Ansammlungen von Materialien bzw. schon bestehender Materialsammlungen, die unter anderen Kriterien als sprachwissenschaftlichen ausgewählt und aufbereitet wurden (vgl. kritisch Rieger 1979 und z.B. für das Englische Biber 1993 und für das Spanische Alvar/Corpas 1994 und Moreno Fernández 2005a). Wenn ein Corpus auf einem elektronischen Datenträger unter Benutzung international anerkannter Codierungsstandards zugänglich gemacht wird, handelt es sich um einen «*maschinenlesbaren Corpus*»: «A *computer corpus* is a corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance.» (Sinclair 1996).

Die Corpuslinguistik in der heute bekannten Form hat erst relativ spät in die spanische Sprachwissenschaft Einzug gehalten. Im Bereich der Gewinnung von Datenmaterial hat die spanische Sprachwissenschaft allerdings relativ früh eine Spitzenstellung bezogen. Schon

1964 stellte Juan M. Lope Blanch (Universidad Nacional Autónoma de México) das *Proyecto de estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica* vor (kurz HABLA CULTA, vgl. Lope Blanch 1967, 1986 sowie VV. AA. 1971-1973 und Samper 1995). Dieses Projekt hat im Laufe von knapp dreißig Jahren (1964-1993) Materialien zur gesprochenen Sprache der Gebildeten in elf Städten Spaniens und Lateinamerikas hervorgebracht (Samper/Hernández/Troya (Hrsgg.) 1998). Diese sind von der Arbeitsgruppe um José Antonio Samper Padilla (Universidad de Las Palmas de Gran Canaria) zunächst auf elektronische Datenträger gebracht worden und später auch Teil des spanischen Referenzcorpus CREA (s. u.) geworden (vgl. Pino/Sánchez 1999). Obwohl diese Materialien noch nicht erschöpfend von der neueren spanischen Sprachwissenschaft benutzt worden sind, sind immer wieder Stimmen laut geworden, welche die mangelnde Einheitlichkeit in der Vorgehensweise (vgl. z.B. Koch/Oesterreicher 1990) sowie das mangelnde Bewußtsein für soziolinguistische Fragestellungen (vgl. z.B. Cortés 1994: 60-64) kritisiert haben. Deshalb hat ALFAL (***A**sociación de **L**ingüística y **F**ilología de **A**mérica **L**atina*) 1993 das Nachfolgeprojekt PRESEEA (***Pr**oyecto para el **E**studio **S**ociolingüístico del **E**spañol de **E**spaña y de **A**mérica*) in die Wege geleitet (Moreno Fernández 1993, 1996, 2005b), von dem zahlreiche Teilergebnisse vorliegen[3].

Die ersten maschinenlesbaren Corpora für das Spanische stammen aus den 70er und 80er Jahren (vgl. Sánchez 1995: 18; De Kock (Hrsg.) 2001). An der Katholieke Universiteit Leuven erstellten Josse De Kock und seine Mitarbeiter zwei kleine Corpora aus literarischer Sachprosa von spanischen und lateinamerikanischen Autoren des 20. Jahrhunderts (De Kock/Verdonk/Gómez Molina 1991; De Kock/Gómez Molina/García Mouton/Delbecque 1992; jeweils 100.000 Wörter). An der Göteborg Universitet schufen Per Rosengren und David Mighetto (vgl. De Kock (Hrsg.) 2001) zwei weitere Corpora aus Zeitungssprache (PE77: *Banco de Datos de **P**rensa **E**spañola 1977*, ca. 1,9 Millionen Wörter) und literarischer Sprache (ONE71: *Banco de Datos de **O**nce **N**ovelas **E**spañolas 1951-19**71***, jeweils ca. 1,0 Millionen Wörter)[4]. An El Colegio de México erstellte die Forschungsgruppe um Luis Fernando Lara ein Referenzcorpus des mexikanischen Spanisch *Corpus del español mexicano contemporáneo* (1921-1974) von ca. 2 Millionen Wörtern, das die Grundlage für zwei Wörterbücher (Lara 1987, 1996), aber auch für grammatikalische Studien gewesen ist (vgl. z.B. Knauer 1998).

Im Laufe der 90er Jahre sind zahlreiche maschinenlesbare Corpora entstanden. An dieser Stelle sollen nur einige wichtige Corpora unmarkierter Sprache (s. u.)[5] stichwortartig vorgestellt werden[6] (vgl. auch De Kock (Hrsg.) 2001).

---

[3] Vgl. aber das Informationsportal *Linguas.net* (o. J.) auf http://www.linguas.net [04.2005].

[4] Vgl. das Informationsportal von Mighetto u. a. (1998-2001) auf http://spraakbanken.gu.se/lb/konk/rom2/ [04.2005].

[5] Mittlerweile gibt es mindestens zwei Einrichtungen, die maschinenlesbare Corpora (auch für das Spanische) sammeln und interessierten Forschern gegen Entgelt zur Verfügung stellen: LDC (***L**inguistic **D**ata **C**onsortium*) von der University of Pennsylvania in Philadelphia 1992 und ELRA (***E**uropean **L**anguage **R**esources **A**ssociation*), mit Unterstützung der EU 1995 in Luxemburg gegründet. Diese Corpora sind zwar unter sprachwissenschaftlichen Gesichtspunkten erstellt worden, aber sie dienen in erster Linie computerlinguistischen Zwecken.

[6] Viele Corpus-Projekte werden in erster Linie für den internen Gebrauch bestimmter Forschungsgruppen erstellt und sind ständigen Veränderungen unterworfen. Die entsprechenden Informatio-

(4) Überblick über maschinenlesbare Corpora des Spanischen (nach Marcos Marín 1994; Sánchez 1995; Instituto «Cervantes» 1997, 2005; Stand 2001)
   a. **C**orpus **O**ral de **R**eferencia del **E**spañol **C**ontemporáneo COREC (1992) von der Forschungsgruppe um Francisco A. Marcos Marín (Universidad Autónoma de Madrid), in CREA eingearbeitet (Marcos Marín 1994: 115-142; Moreno Sandoval 2002).
   b. **C**orpus del **E**spañol de la **R**epública de **C**hile CERC (1992) von Forschungsgruppen aus der Universidad de Chile und der Universidad Católica de Chile (Marcos Marín 1994: 148-155; Moreno Sandoval 2002) unter der Leitung von Francisco A. Marcos Marín.
   c. **C**orpus del **E**spañol de la **R**epública **A**rgentina CORA (1992) von Forschungsgruppen aus der Universidad de Buenos Aires und der Universidad Católica Argentina (Marcos Marín 1994: 143-147; Moreno Sandoval 2002) unter der Leitung von Francisco A. Marcos Marín.
   d. Corpus lingüístico del español contemporáneo CUMBRE (1995) von der Forschungsgruppe um Aquilino Sánchez (Universidad de Murcia) im Auftrag des Verlages SGEL, Madrid (Sánchez/Sarmiento/Cantos/Simón 1995).
   e. Base de datos informatizada de la lengua española LEXESP (1998) von der Forschungsgruppe um Núria Sebastián von der Universitat de Barcelona (Instituto «Cervantes» 1997).
   f. **A**rchivo de **T**extos **H**ispánicos de la **U**niversidade de **S**antiago de Compostela ARTHUS (2000) von der Forschungsgruppe um Guillermo Rojo an der Universidade de Santiago de Compostela (Instituto «Cervantes» 1997)
   g. **C**orpus de **R**eferencia del **E**spañol **A**ctual CREA (2000) von der Forschungsgruppe um Guillermo Rojo am Instituto de Lexicografía der Real Academia Española de la Lengua (Instituto «Cervantes» 1997).

Diese Corpora entsprechen der gegebenen Definition eines maschinenlesbaren Corpus. Nur eine genaue Beschreibung und Einteilung dieser Corpora nach festgelegten Kriterien macht es möglich, das für diese Untersuchung geeignete Corpus zu finden. Dies geschieht im nächsten Unterabschnitt.

2.2 Zur Beschreibung und Einteilung von Corpora

Corpora können durch folgende formale Angaben (35) näher beschrieben werden, wobei die in (ai, bi, ci) vorgestellte Möglichkeit die in der Sprachwissenschaft übliche Form ist, es sei denn der Untersuchungsgegenstand erfordert die andere.

(5) Sammlungen empirischen Datenmaterials: Beschreibung (aus Sinclair 1996)
   a. *Quantität*: «groß»
      i. Geschlossene Corpora
      ii. Offene Corpora
   b. *Qualität*: «authentisch»
      i. Corpora unmarkierter Sprache
      ii. Corpora markierter Sprache

---

nen sind spärlich und gelangen nicht oder nur indirekt an die Öffentlichkeit (z.B. durch kurze Hinweise in den eher seltenen Überblicken zur Corpuslinguistik (s. Haupttext), in den Selbstdarstellungen der Forschungsgruppen etwa im Internet oder durch Quellenvermerke in den wissenschaftlichen Arbeiten, die auf diesen Corpus-Projekten basieren).

c. *Einfachheit*: «einfacher Text» (*plain text*)
   i. Corpora ohne Format und ohne linguistische Annotation
   ii. Corpora mit Format und ohne linguistische Annotation
   iii. Corpora mit Format und mit linguistischen Annotation

Wichtig ist erstens, daß Corpora «groß» sind, denn je größer das Corpus ist, desto größer die Zahl verschiedener Phänomene (*types*) und die Zahl der Belege jedes einzelnen Phänomens (*token*). In dem Maße, wie die Speicherkapazitäten und die Schnelligkeit der Computer erhöht worden sind, sind auch die Anforderungen an maschinenlesbare Corpora gestiegen. War am Anfang die Anzahl von einer Million Wörtern das Übliche (Sinclair 1996), so geht der Trend heute von geschlossenen Corpora von über 100 Millionen Wörtern hin zu offenen Corpora, die regelmäßig um eine festgelegte Zahl von Wörtern ergänzt werden. Letztere «wachsen mit der Zeit» und erlauben die Beobachtung der Entwicklung der Sprache in «Echtzeit» (s. u. *monitor corpora*). Für den Zweck dieser Untersuchung genügt ein kleines geschlossenes Corpus. Einerseits sind die zu untersuchenden *types* (Permutationen bzw. Kombinationen von Elementen) so beschränkt in der Zahl, daß ein größeres Corpus wohl nicht dazu führen wird, bisher unbekannte Varianten zu entdecken. Andererseits würde eine allzu große Anzahl von *tokens* es nicht mehr erlauben, eine angemessene qualitative Analyse durchzuführen.

Wichtig ist zweitens, daß Corpora «authentische» Daten enthalten, damit sie die sprachliche Wirklichkeit möglichst getreu wiedergeben. Allerdings ist die sprachliche Wirklichkeit so komplex, daß es notwendig wird, zwischen unmarkierten und markierten Kommunikationssituationen zu unterscheiden. Bei den Letzteren handelt es sich um besondere Situationen, weil Informanten, Thematik oder Aufnahmebedingungen nach bestimmten Kriterien gewählt worden sind, die nicht mehr als repräsentativ für das Ganze gelten können (z.B. Kindersprache, Fachsprache, Experimentssituationen). Auch hier soll unser Corpus möglichst unmarkierte Kommunikationssituationen wiedergeben, wobei weder der Gebrauch weit verbreiteter (diatopischer, diastratischer, diaphasischer und diamedialer) Sprachvarietäten noch die Aufnahme der wichtigsten Textsorten als markiert gelten, sondern im Gegenteil als Ausdruck einer Streuung, die der vorzufindenden sprachlichen Wirklichkeit in ihrer Vielfalt an kommunikativen Bedürfnissen gerecht werden will.

Wichtig ist drittens, daß das Corpus so wenig wie möglich manipuliert wird, damit es für möglichst viele und verschiedene Forschungszwecke benutzt werden kann. Corpora können als einfacher Text (engl. *plain text*) in einem der üblichen Zeichencodes (z.B. ASCII) angeboten werden; üblich ist aber geworden, möglichst viele Informationen zu den im Corpus vorhandenen Texten einzuarbeiten und dabei, nach den Richtlinien der TEI (***T**ext **E**ncoding **I**nitiative*) und ähnlicher Organisationen, «Sprachen» wie SGML (***S**tandard **G**eneralized **M**arkup **L**anguage*) und ihre Abwandlungen zu benutzen. Darüber hinaus ist es möglich, die Texte mit sprachwissenschaftlichen Informationen anzureichern, wie z.B. Angaben zu Wortarten (engl. *parts-of-speech tagging*) oder Lemmata. Diese letzte Möglichkeit ist das Ziel einiger der vorgestellten Corpora, die tatsächlich in vielen Fällen noch nicht weit vorangekommen sind. In unserem Fall heißt es, daß das Corpus möglichst nur für unseren Forschungszweck manipuliert worden ist und ansonsten unverändert bleibt. Das von uns gewählte Corpus, das weiter unten vorgestellt wird, entspricht besonders gut dieser Forderung, weil es Text und Aufbereitung voneinander trennt.

Mit der Zeit haben sich unterschiedliche Typen von Corpora etabliert, die verschiedenen Untersuchungszielen gerecht werden sollen. Die Einteilungskriterien in (6) gehen von den erschlossenen Texten aus und sind: Repräsentativität (a), Geschlossenheit (b), Medium (c) und Vollständigkeit (d).

(6) Sammlungen empirischen Datenmaterials: Typologie (Sinclair 1996)
   a. *Reference corpus* vs. sublanguage/special corpus (Re – Su): «A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials.»
   b. *Monitor corpus* vs. constant size corpus (Mo – Co): «a large and constantly moving [corpus]».
   c. *Spoken corpus* vs. written corpus (Sp – Wr): Corpus gesprochener Sprache, wobei diese folgendermaßen definiert wird: «[...] any language whose original presentation is in oral form – that is, the speakers involved behave in oral mode».
   d. *Samples corpus* vs. text corpus (Sa – Te): «Samples are small, in relation to texts such as newspapers, books and radio programmes, and of a constant size [...].»

Nach den vorgestellten Beschreibungskategorien und Einteilungskriterien können die genannten maschinenlesbaren Corpora des Spanischen folgendermaßen charakterisiert werden.

(7) Beschreibung und Einteilung der vorgestellten maschinenlesbaren Corpora des Spanischen (Sinclair 1996; Instituto «Cervantes» 1997)

Corpus	Typ	Zugang	Nutzung	Mio.	Code	Annotation
COREC	Re, Co, Sp, Sa	FTP	nicht kommerziell	1,1	SGML	keine
CERC	Re, Co, Wr, Te	FTP	nicht kommerziell	2,0	SGML	keine
CERA	Re, Co, Wr, Te	FTP	nicht kommerziell	2,0	SGML	keine
CUMBRE	Re, Co, Sp/Wr, Sa	beschränkt	kommerziell	8,0	COCOA	W, Sy, Sm
LEXESP	Re, Co, Wr, Sa	?	*public domain*	5,5	?	MS, Sy
ARTHUS	Re, Co, Sp/Wr, Te	Web-KWIC	nicht kommerziell	3,0	COCOA	W, M, Sy, Sm
CREA	Re, Mo, Sp/Wr, Te	Web-KWIC	nicht kommerziell	100,0	SGML	W, M, Sy

Abkürzungen (s. auch (6)): KWIC = *Key Word in Context*; M = morphologisch; MS = morphosyntaktisch; W = Wortarten; Sm = semantisch; Sy = syntaktisch.

Für die üblichen sprachwissenschaftlichen Untersuchungen –und auch für die vorliegende– muß auf Repräsentativität im Sinne von Streuung geachtet werden (6a) und insbesondere darauf, daß beide Realisierungsformen der Sprache (gesprochen und geschrieben) gleichermaßen präsent sind (6c). Externe Sprache darf auf keine der beiden Realisierungsformen reduziert werden, denn beide Formen lösen auf ihre je eigene Weise das Problem der Kommunikation (vgl. Koch/Oesterreicher 1990). Außerdem ist es wichtig, ganze Texte und nicht nur Textpassagen zu haben (6d), damit auch die Ebene des Textes ggf. in die Untersuchung einbezogen werden kann. Somit bleiben für diese Arbeit nur noch ARTHUS und CREA übrig. Da aber CREA erst vor kurzem fertig wurde, bisher keine syntaktische Information bereithält und zudem trotz Internetzugang eine sehr restriktive Zugangspolitik zu den

Original-Materialien pflegt, konnte die Wahl nur auf ARTHUS fallen.[7] Im nächsten Abschnitt soll die Wahl von ARTHUS allerdings auch positiv gerechtfertigt werden; es wird außerdem zu zeigen sein, daß dieses Corpus sehr gut aufbereitet ist und einmalige Abfragemöglichkeiten für die Untersuchung der Satzgliedstellung bietet.

2.3 Zur Aufbereitung von Corpora und zum Corpus der Untersuchung

Für die Untersuchung der Anordnung bestimmter Satzglieder, die bestimmte Beziehungen unterhalten, genügt nicht ein maschinenlesbares Corpus im üblichen Format, ja nicht einmal ein Corpus, das mit Informationen zu den Wortarten oder gar zur hierarchischen Struktur des Satzes angereichert worden ist, denn die relationale Struktur des Satzes wird im Spanischen in den wenigsten Fällen in der Morphosyntax eindeutig sichtbar. Folge davon ist die strukturelle Ambiguität der meisten Satzglieder, die ein automatisiertes «*parsing*» sehr schwierig machen.[8] Die zur Zeit einzig mögliche Lösung ist die manuelle Erstellung einer Datenbank, die die notwendigen Informationen zur relationellen Struktur des Satzes enthält (Rojo 2001; vgl. auch García-Miguel 1994; van Halteren 1997). Eine solche Datenbank habe ich für eine frühere Untersuchung selber erstellt (vgl. Bellosta von Colbe 1994); für die vorliegende Untersuchung konnte ich auf den zeitgenössischen synchronen Teil von ARTHUS und auf die darauf aufbauende Datenbank BDS zurückgreifen (***Ba****se de* ***D****atos* ***S****intácticos del* ***E****spañol* ***A****ctual*, zeitweise auch BADSEA genannt). Sie steht einzigartig in der spanischen Sprachwissenschaft da und ist wohl auch einmalig im Vergleich zu anderen Philologien. In (8) werden die wichtigsten Informationen zu ARTHUS/BDS zusammengefaßt.

(8) Informationen zu ARTHUS und BDS (nach Grupo de Sintaxe do Español 1996, 1999; Rojo 2001)
- a. Forschungsgruppe:
  - i. *Leitung*: Guillermo Rojo
  - ii. *Mitarbeiter* (1998): Francisco García Gondar, José María García-Miguel, Belén López Meirama, María del Carmen Losada, Inmaculada Mas, María José Rodríguez Espiñeira, Victoria Vázquez Rozas
  - iii. *Korrektur* (1998): Conchita Álvarez Lebredo, Pilar Alvariño, Cristina Blanco, Fernando Castro, Susana Comesaña, Adelaida Gil, Fátima Gayoso, Eva Muñiz, Marta Rebolledo, María Paula Santalla, Susana Sotelo, David Vázquez Martínez.
  - iv. *Unterstützung*: 1988-1991: Xunta de Galicia (XUGA 82710088); 1991-1994: Ministerio de Educación y Ciencia (PB90-0376); 1997-1999: Xunta de Galicia (XUGA 20402B97).
- b. ARTHUS/BDS
  - i. Berichtszeitraum: ca. 1980-1990

---

[7] Meine Anfragen bezüglich CUMBRE bei Aquilino Sánchez und beim Verlag SGEL blieben trotz anfänglichen Entgegenkommens letztlich erfolglos.

[8] Nach den Informationen von Paul R. Bowden sind schon Versuche fürs Englische, Französische und Portugiesische unternommen worden, automatisch «Subjekt» und «Objekt» in Corpora zu ermitteln, die lediglich mit Informationen zu den Wortarten versehen worden waren (Mitteilung an die Diskussionsliste CORPORA, 10.05.1999).

ii. Form: Format: ASCII; Kodierung: COCOA (***CO****unt and* ***CO****ncordance on* ***A****tlas*)
iii. Größe: 34 Texte; 1.449.005 Wörter; 161.662 Sätze; 63 Felder für verschiedene Angaben; 3.554 verschiedene Verben; 113 verschiedene Valenzkonfigurationen.
iv. Zusammenstellung (Angaben: s. Literaturverzeichnis):

Medium	Textsorte	Spanien	Amerika	Gesamt	Frequenzen	
Geschrieben	Narrativik	385.661	153.245	538.906	37,19%	81,15%
	Essay	168.511	89.207	257.718	17,78%	
	Theater	212.507	0	212.507	14,66%	
	Zeitung	166.804	0	166.804	11,51%	
Gesprochen		207.948	65.122	273.070	18,85%	
Gesamt		1.141.431	307.574	1.449.005		
Frequenz		78,77%	21,23%			

Nach Rojo (1993) wurden die Texte zuerst mit Hilfe eines Scanners und des OCR-Programms *Readstar 3/6* (Inovatic) elektronisch erfaßt bzw. lagen schon in maschinenlesbarer Form vor (dies gilt für SEVILLA, 1VOZ, 2VOZ, 3VOZ). Danach wurden die Texte korrigiert und wie die Papierversion formatiert. Dies erlaubt, die Beispiele jederzeit an der Papierversion des Corpus zu überprüfen. Mit Hilfe des Konkordanzprogramms *Micro-OCP* (Oxford Computing Service) wurden alphabetische Listen des Corpus erstellt, die es ermöglichten, dasselbe Verb manuell einheitlich zu bearbeiten. Danach wurde das Corpus in Datenbankformat konvertiert, die manuelle Analyse aller Sätze nach 63 verschiedenen Parameter durchgeführt und die Ergebnisse dieser Analyse in eine relationale Datenbank gebracht (ursprünglich in *dBase IV 2.0*, später *FoxPro 2.0* von Borland, beide mittlerweile etwas veraltet). Um ein einheitliches Vorgehen der Forschungsgruppe zu gewährleisten, wurden Richtlinien für die manuelle Analyse festgelegt, die mehrmals revidiert wurden (zuletzt Grupo de Sintaxe do Español 1996). Außerdem wurde eigens ein Programm namens *ADI_ESQ* entwickelt, um aus den Feldern die (groben und feinen) Valenzkonfigurationen eines jeden Verbs sowie die Verben einer jeden (groben und feinen) Valenzkonfiguration zu gewinnen. Nach der Fertigstellung von BDS (1988-1994) wurde sie komplett manuell korrigiert (1997-1999) und als Internet-Datenbank zugänglich gemacht (2001)[9]. Geplant ist nunmehr, sie mit semantischer Information anzureichern. An dem Projekt, ein Frequenzwörterbuch ***Dic****cionario de* ***Frec****uencias Verbales* (DICFREC) und ein Valenzwörterbuch ***Di****ccionario de* ***C****onstrucciones* ***V****erbales del* ***E****spañol* ***A****ctual* (DICVEA) des Spanischen zu entwickeln (Grupo de Sintaxe do Español 2001; Rojo 1992, 1994) ist wohl nicht festgehalten worden. Aber schon während der Erstellung und Korrektur der Datenbank entstanden richtungsweisende Forschungsarbeiten, deren empirische Basis ARTHUS/BDS ist. Für die vorliegende Arbeit sind insbesondere Vázquez Rozas (1995) und López Meirama (1997) wichtig; aber auch García-Miguel (1995b), Rodríguez Espiñeira (1990) und Cabeza (1997) sind zu nennen.

Die Struktur der Datenbank ist in den folgenden abfragbaren Feldern zu ersehen:

[9] Unmittelbar vor Abschluß dieser Arbeit (2001) wurde eine provisorische Internet-Version von BDS fertiggestellt, zu der ich zu Testzwecken Zugang erhielt. Die aktuelle Internet-Version ist unter <http://www.bds.usc.es/> zu finden (Grupo de Sintaxe do Español 2001).

(9) BDS: Übersicht über die Felder

Allgemeines	Satz		Subjekt	CDIR
Obra	Tipo	Número/persona	Sujeto	CDIR
Referencia	Función	Predicado complejo	Unidad	Clítico
Tipo	Voz	N° argumento	Animación	Marca
Procedencia	Polaridad	Orden	Determinación	Unidad
Verbo	Modalidad		Número	Animación
[Acepción]	Perífrasis.			Determinación
[Subacepción]	Forma verbal			Número
Observaciones	Forma verbal dominante			

CIND	CPR1	CPR2	CAG	PVO
CIND	CPR1	CPR2	CAG	PVO
Clítico	Tipo	Tipo	Marca	Unidad
Clítico 2	Marca	Marca	Unidad	Marca
Unidad	Unidad	Unidad	Animación	Determinación
Animación	Animación	Animación	Determinación	Número
Determinación	Determinación	Determinación	Número	Referente
Número	Número	Número		

Abkürzungen: CAG = complemento agente (Agens eines Passivsatzes); CDIR = complemento directo (direktes Objekt); CIND = complemento indirecto (indirektes Objekt), CPR = complemento preposicional (Präpositionalobjekt); PVO = predicativo (Prädikativ):

Die Datenbank wurde mit qualitativen Daten, nämlich Angaben zu den *nominalen Kategorien* gespeist. Solche Informationen sind naturgemäß nicht quantitativ meßbar; sie können nur auf der absoluten, diskreten Skala der natürlichen Zahlen gezählt werden (Frequenzberechnung). Allerdings erlaubt die Frequenzberechnung verschiedener nominaler Kategorien auch die Kreuzklassifikation mehrerer Varianten, die zu verschiedenen Variablen gehören (vgl. z.B. Lowry 1998-2005a). Diese Möglichkeit soll ausgenutzt werden, und dazu gibt die Datenbank wertvolle Hilfestellung. Sie gibt quantitative und qualitative Informationen, nämlich einerseits Frequenzberechnungen zu einem oder mehreren Feldern, andererseits alle *tokens* (Sätze), die sich in dem Feld einer Kategorie oder in den Feldern mehrerer Kategorien befinden. Sie erlaubt es also, sowohl nach der (groben oder feinen) Valenzkonfiguration von ausgewählten Verben zu suchen als auch die Verben zu ermitteln, die eine ausgewählte Valenzkonfiguration haben, und die entsprechenden Frequenzberechnungen und *tokens* zu erhalten.

Einige Probleme, welche die Forschungsgruppe bei der Bearbeitung des Datenmaterials bewußt in Kauf genommen hat, seien hier noch kurz vermerkt. Das Corpus ist trotz aller Vielfalt noch immer etwas unausgewogen, insbesondere was die gesprochene Sprache (diamedial) und die Varietäten (diatopisch und diastratisch) angeht (Rojo 1993: 16). Bei aller Bemühung um vollständige Texte wurden die Corpora der gesprochenen Sprache aus Buenos Aires, Madrid und Sevilla nicht vollständig übernommen (Rojo 1993: 16). Zudem wurden die Theaterstücke vollständig, d. h. zusammen mit den Bühnenanweisungen übernommen und analysiert, was den entsprechenden Statistiken ihren (möglichen) Wert als

Nachahmung gesprochener Sprache nimmt, es sei denn man scheidet manuell Anweisungen und Text. Auch wurde im Laufe der Bearbeitung klar, daß die Texte von José Luis Sampedro (SONRISA) stark idiolektisch geprägt sind. Schließlich wurde nur eins von fünf *tokens* der fünf häufigsten Vollverben *ser*, *estar*, *haber*, *querer* und *decir* in die Datenbank aufgenommen, weil diese sehr häufig mit denselben wenigen Valenzkonfigurationen vorkamen (Rojo 2001).

Die Bearbeiter des Corpus waren entsprechend geschult und haben äußerst sorgfältig gearbeitet.[10] Trotzdem sind bei der Korrektur zahlreiche Fehler entdeckt worden, und es ist nicht auszuschließen, daß auch die aktuelle fertige Version von BDS, die die Grundlage dieser Arbeit bildet, weiterhin Fehler enthält. Dies liegt zum einen daran, daß gewisse Begrifflichkeiten oder Kategorien, die bisher in der Grammatik üblich waren, in der Praxis sehr schwer anzuwenden waren. So wurden z.B. möglichst viele Präpositionalphrasen als *suplemento* (valenznotwendige Präpositionalobjekte) aufgenommen. Viele davon hätten vielleicht bei näherer Betrachtung genau so gut als nicht valenznotwendige Umstandsangaben angesehen werden können.[11] Es liegt auch zum anderen daran, daß viele Sätze in gewissen Bereichen nicht eindeutig analysiert werden können (syntaktische Ambiguität). So war z.B. in Feld [21] CDIR der Schlüssel <8> für die Fälle vorgesehen, bei denen es nicht möglich war zu entscheiden, ob das entsprechende Argument als «direktes Objekt» oder als «indirektes Objekt» aufzufassen war. Dies offenbart nicht zuletzt, daß die Datenbank nur qualitative Einzelanalysen geordnet wiedergibt und letztlich von der Qualität der theoretischen Voraussetzungen und deren Anwendung abhängig ist. Es gehört aber zu der ursprünglichen Aufgabe der Sprachwissenschaft, im Spannungsfeld zwischen theoretischer Reflexion und praktischer Anwendung daran zu arbeiten, und dies darf ihr deshalb auch nicht in diesem Fall zum Vorwurf gemacht werden. Deshalb soll schon an dieser Stelle davor gewarnt werden, den hier vorgelegten Statistiken einen Wert beizumessen, den sie nicht besitzen und auch aus den genannten Gründen nicht beanspruchen können.

3. Zur Auswertung der aufbereiteten Datenbasis

Die dritte Frage, die noch gestellt werden muß, ist folgende: Welche Informationen zur Satzgliedabfolge können und sollten der Datenbank entnommen werden? Hier ist das Feld «Orden» [15] die wichtigste Quelle der Information. Darin sind die Satzglieder in der Anordnung ihres Vorkommens nach folgendem Schlüssel eingegeben:

(10) Schlüssel für Feld «Orden» [15] in BDS (Grupo de Sintaxe do Español o. J.)
S = Subjekt D = 2. Präpositionales Objekt

[10] Bei meinem zweiten Forschungsaufenthalt an der Universidade de Santiago de Compostela im September 2000 hatte ich die Möglichkeit, einen kursorischen Blick in die Protokollhefte einzelner Mitarbeiter des Projektes zu werfen. Dort sind alle Fragen und Ungereimtheiten, die aufgefallen sind, sowie ihre Besprechung im Team und die darauffolgenden Entscheidungen verzeichnet.

[11] Diese Informationen wurden mir teilweise von verschiedenen Mitgliedern der Forschungsgruppe während meiner beiden Forschungsaufenthalte mitgeteilt.

V =	Prädikat (d. h. Verb)	A =	Umstandsangaben
C =	Direktes Objekt	P =	Prädikative
I =	Indirektes Objekt	G =	Agens (im Passivsatz)
U =	1. Präpositionales Objekt	E =	«*Extraposición*»

Folgende Richtlinien galten bei der Bearbeitung dieses Feldes:

(11) Richtlinien für die Bearbeitung des Feldes «Orden» [15] in BDS (Grupo de Sintaxe do Español 1996: 37-39)
  a. Einzugeben sind:
    i. Prädikat (V, Verb bzw. Hilfsverb) und Argumente (S, C, I, U, D, G, valenziell)
    ii. Umstandsangaben, in jeder Stellung, auch mehmals (A, nicht valenziell)
    iii. Prädikative (P, nicht valenziell)
  b. Nicht einzugeben sind:
    i. Implizite Subjekte bzw. weggelassene Objekte
    ii. Relativ-, Interrogativ- und Klitikpronomina,
    iii. Satzeinleitungen in fixierter Stellung (z.B. *con respecto a...*)
    iv. Diskurskommentare (z.B. *ciertamente*, *análogamente*, *indudablemente*)
    v. Sätze in direkter Rede, als direktes Objekt eines *verbum dicendi* aufgefaßt
    vi. Angaben zu unterbrochenen Sätze

Für unsere Fragestellung genügt also die Richtlinie (11ai). Daß nach (11aii) Umstandsangaben undifferenziert aufgenommen werden, mag vielleicht bedauerlich sein, für die Fragestellung spielt dies keine Rolle. Auch die Beschränkungen in (11b) sind für die Untersuchung sehr sinnvoll. Das Fehlen von (11bi) hätte die Datenbank für unsere Zwecke völlig unbrauchbar gemacht: wir erhalten damit also nur Informationen über explizit realisierte Partizipanten. Die Richtlinie (9bii) läßt pronominal realisierte Partizipanten in festgelegten Positionen weg, welche sonst die Statistiken verfälscht hätten. Allerdings wurde schon festgestellt, daß diese Richtlinie nicht durchgehend beachtet worden ist. Außerdem erfaßt sie nicht die Fälle, in denen Interrogativ- und Relativpronomina als Determinanten eines Nomens auftreten und das ganze Satzglied an eine feste Stelle am Anfang des Satzes binden, was bei der Untersuchung der Voranstellungen beachtet werden muß. Die Richtlinie (11biii) vermeidet, daß Satzeinleitungen zu den Argumenten oder zu den Umstandsangaben gerechnet werden, erlaubt aber im Gegenzug nicht, die von ihnen eingeleiteten Topiks genauer zu untersuchen (vgl. Kapitel 4 und 7). Richtlinie (11bv) verhindert, daß die Angaben zum «direkten Objekt» zu sehr aufgebauscht werden, denn Sätze in direkter Rede mögen semantisch dem *verbum dicendi* zugeordnet werden, doch sind sie syntaktisch von ihm weitgehend unabhängig. (11bvi) mag zunächst einmal für die Untersuchung gesprochener Sprache hinderlich sein; trotzdem ist sie sinnvoll, denn unterbrochene Sätze erlauben keine Rückschlüsse auf eine nur intendierte, aber nicht realisierte Struktur.

In (10) wurde auch Schlüssel <E> für «*Extraposición*» eingeführt. Diese Kategorie erhält folgende Definition: «Por *extraposición* entendemos la bipartición de un elemento de la secuencia por interposición de otro(s) elemento(s).» (Grupo de Sintaxe do Español 1996: 39) Folgende Beispiele werden u. a. gebracht:

(12) Kodierung «geteilter Elemente» <E> in BDS (nach Grupo de Sintaxe do Español 1996: 39)

a.	«Subjekthebung»:	La operación *parece* que fue un éxito.	EVS
b.	«Hyperbaton»:	El tabaco *fumaste* de tu amigo.	EVC

In (12a) besteht die Subjekthebung gerade darin, daß das «Subjekt» des Nebensatzes vor das Verb des Hauptsatzes gestellt wird. Trotzdem wird es weiterhin als Teil des Nebensatzes gesehen, der das «Subjekt» des Verbs *parecer* ist. In (12b) wird das Komplement der «Objekt»-NP *in situ* gelassen, während Determinant und Kopf vorangestellt werden. Hier kann zudem gefragt werden, ob es nicht besser gewesen wäre, es als CVE zu notieren. Diese Art des Hyperbatons ist jedoch ein künstliches Mittel, das in Anlehnung an die klassische Literatur besonders in der Lyrik immer wieder gerne benutzt wird. Sie ist aber nicht Gegenstand dieser Arbeit.

Bei den meisten Untersuchungen wurde die Suchfunktion *SCAN FOR* genutzt, die es ermöglicht, die Datenbank nach bestimmten Informationen gezielt abzufragen. Für die Gewinnung der untersuchten Satzanordnungen von Verb <V>, «direktes Objekt» <C> und «indirektes Objekt» <I> wurden folgende Befehle eingegeben:

(13) Abfrage von Satzanordnungen in BDS: SCAN FOR...

a. Voranstellungen:

i. CSV/SCV: at("C",campo15)<at("V",campo15).and.at("S",campo15)<at("V",campo15).and.at("C",campo15)≠0.and.at("S",campo15)≠0.and. campo8="1"

ii. ISV/SIV: at("S",campo15)<at("V",campo15).and.at("I",campo15)<at("V",campo15).and.at("S",campo15)≠0.and.at("I",campo15)≠0.and. campo8="1"

iii. CIV/ICV: at("C",campo15)<at("V",campo15).and.at("I",campo15)<at("V",campo15).and.at("C",campo15)≠0.and.at("I",campo15)≠0.and. campo8="1"

iv. SCIV/SICV/CSIV/CISV/ISCV/ICSV: at("S",campo15)<("V",campo15).and.at("C",campo15)<at("V",campo15).and.at("I",campo15)<at("V",campo15).and.at("S",campo15)≠0.and.at("C",campo15)≠0.and.at("I",campo15)≠0.and. campo8="1"

b. Nachstellungen:

i. VCI: at("V",campo15)<at("C",campo15).and.at("C",campo15)<at("I",campo15).and.at("V",campo15)≠0.and.at("C",campo15)≠0

ii. VIC: at("V",campo15)<at("I",campo15).and.at("I",campo15)<at("C",campo15).and.at("V",campo15)≠0.and.at("I",campo15)≠0

c. Voran- und Nachstellungen:

i. CVI: at("C",campo15)<at("V",campo15).and.at("V",campo15)<at("I",campo15).and.at("C",campo15)≠0.and.at("V",campo15)≠0.and. campo8="1"

ii. IVC: at("I",campo15)<at("V",campo15).and.at("V",campo15)<at("C",campo15).and.at("I",campo15)≠0.and.at("V",campo15)≠0.and. campo8="1"

Es wurde die Möglichkeit genutzt, nach gewissen Zeichen ("S", "V", "C", "I") in einem bestimmten Feld (campo15) unter Berücksichtigung ihrer Stellung im Feld (Befehl *AT*) zu suchen. Bei (13a) wurde nach vorangestellten Argumenten gesucht, d. h. die Stellen von S, C und I sollten jeweils vor (<) der Stelle von V liegen. Bei (13b-c) wurde nach bestimmten Anordnungen gesucht, weswegen Ketten gebildet wurden (z.B. bei VCI sollte V vor C liegen und C vor I liegen). In allen Fällen mußte eine Bedingung zusätzlich eingefügt werden, nämlich daß das erste Glied der Ungleichheitsrelation nicht Null (nicht existent) sein

durfte. In (13a, c) wurde noch gefordert, daß es sich um deklarative Sätze handeln mußte (Feld (8) Modalidad, Schlüssel <1> deklarativ), um Voranstellungen von Argumenten auszuschließen, die von einem Interrogativpronomen als Determinant begleitet werden.

Die Funktion *SCAN FOR* ist zwar die flexibelste Suchmöglichkeit, die das Programm anbietet, doch nichtsdestotrotz beschränkt. Sie erlaubt nicht die automatische Neueingabe der zuletzt gemachten Suche und ist auf 264 Zeichen begrenzt. Da die Suchbefehle für gewisse Anordnungen schon ziemlich lang gewesen sind, konnten diese nur mit sehr kurzen Angaben weiter eingegrenzt werden, so daß komplexere Suchen mit mehreren Faktoren nicht möglich waren.

4. Zur Analyse der gewonnenen Information

Die letzte Frage, die es zu beantworten gilt, lautet: Wie kann und sollte die gewonnene Information analysiert werden? In 2.3 wurde festgestellt, daß die Datenbank zweierlei Typen von Informationen bereithält, die natürlich auch für das eben beschrieben Feld «Orden» gültig sind: quantitative Information (Frequenzberechnungen) zu jeder nominalen Kategorie sowie qualitative Information (Kategorialanalyse) zu jedem *tokens*, d. h. Satz, der im Corpus vorkommt. In den folgenden Abschnitten sollen beide Informationstypen zueinander in Beziehung gesetzt werden.

4.1 Quantitative Analyse der gewonnenen Information

Eine erste Annäherung an die Analyse der aus der Datenbank gewonnenen Informationen besteht darin, die quantitativen Informationen statistisch aufzubereiten und sie als Bestätigung oder Widerlegung von Hypothesen zu nutzen. Wir müssen aber bedenken, daß die Datenbank mit qualitativen Daten, nämlich Angaben zu sog. «nominalen Kategorien» oder Variablen, gespeist wurde. Solche Informationen sind naturgemäß an sich nicht quantitativ meßbar (z.B. «Wortstellungsmuster»), auch wenn die möglichen Varianten manchmal untereinander eine quantitative Beziehung unterhalten (z.B. Zahl der offenen Stellen eines Verbs) oder auf eine ordinalen Skala gebracht werden können, wenn sie komplexer Natur sind (z.B. «Prototypikalitäts»-Einteilungen, wie bei der «semantischen Transitivität»; vgl. Hopper/Thompson 1980; und allgemein zum Thema Moure 1996). Diese Informationen können nur auf der absoluten diskreten Skala der natürlichen Zahlen gezählt werden (Frequenzberechnung). Allerdings erlaubt die Frequenzberechnung verschiedener unabhängiger nominaler Kategorien bzw. Variablen auch Kreuzklassifikationen, bei denen komplexe Varianten entstehen (z.B. [[+belebt]&[+bestimmt]], die zu komplexen Variablen gehören (z.B. /Animation/&/Bestimmtheit/; vgl. z.B. Lowry 1998-2005a). Diese Möglichkeit kann ausgenutzt werden, um festzustellen, ob zwei oder mehrere Variablen assoziiert sind bzw. korrelieren. Eine sehr einfache und häufig genutzte Möglichkeit besteht darin, den sog. «Chi Quadrat Test» (χ^2) durchzuführen (Lowry 1998-2005a: 8.2; Butler 1985: 112-123; Woods/Fletcher/Hughes 1986: 137-151; Oakes 1998: 24-29).

Dieser Test überprüft die statistische Signifikanz der beobachteten Verteilungen, d. h. die Wahrscheinlichkeit, mit der es in einer beliebig zusammengesetzten Stichprobe dazu kommen kann, daß zwei oder mehrere Variablen durch Zufall zueinander im beobachteten Verhältnis stehen («Nullhypothese»). Dabei werden die beobachteten Verteilungen mit den Verteilungen verglichen, die man erwarten würde, wenn die Nullhypothese zutreffen würde. Die Formeln lauten:

(14) Chi Quadrat Assoziationstest (Lowry 1998-2005a: 8.2; Butler 1985: 112-123; Woods/Fletcher/ Hughes 1986: 137-151;.Oakes 1998: 24-29):

a.	Chi-Quadrat-Formel:	$\chi^2=\sum(O - E)^2/E$
b.	χ^2 mit Yates Korrektur:	$\chi^2=\sum(/O - E/-0,5)^2/E$
c.	Erwartete Frequenz:	Durch ein Modell oder eine Beobachtung vorgegeben. E=(Total Reihe x Total Kolumne)/Gesamt
d.	Freiheitsgrade	df=(Zahl der Reihen–1)x(Zahl der Kolumnen–1)

Yates Korrektur wird bei df=1 benutzt. Wenn die erwartete Frequenz einer Variante niedriger ist als 5, gilt der Test als ungültig, d. h. nicht aussagekräftig (genug). Dies wird in unseren Tabellen häufiger der Fall sein; diese Tatsache wird mit dem Wort «ungültig» gekennzeichnet. Den Ergebnissen des Tests werden Wahrscheinlichkeitswerte zugeordnet; eine Verteilung wird als signifikant angesehen, wenn die Wahrscheinlichkeit, daß sie durch Zufall zustande gekommen ist, weniger als 5% (<0,05) beträgt.[12]

4.2 Qualitative Analyse der gewonnenen Information

Da BDS im Grunde genommen eine Valenzdatenbank ist, kann gefragt werden, wie diese Eigenschaft für unsere Untersuchung nutzbar gemacht werden kann. Zwei Wege bieten sich an, um valenzrelevante Informationen zur Satzgliedstellung aus der Datenbank zu erhalten. Auf der einen Seite können untersuchungsrelevante Valenzkonfigurationen hinsichtlich ihrer Satzgliedstellung unter die Lupe genommen werden. Auf der anderen Seite können Verben, die wegen ihrer Valenzeigenschaften besondere Beachtung verdienen, hinsichtlich ihrer Satzgliedstellung untersucht werden. Jeder dieser Wege führt zu andersartigen Erkenntnissen; im Folgenden sollen also beide ansatzweise und exemplarisch beschritten werden.

Der erste Weg beinhaltet folgende vier Schritte:

a. Auflistung der Valenzkonfigurationen, die Träger der untersuchten syntaktischen Relationen enthalten
b. Unterklassifizierung der Valenzkonfigurationen nach den Satzgliedanordnungen, in denen die syntaktischen Relationen vorkommen
c. Feststellung von Anordnungstendenzen in jeder Konfigurationsklasse
d. Untersuchung der Rolle der Verbsemantik in jeder Konfigurationsklasse hinsichtlich der ausgemachten Anordnungstendenzen

---

[12] Die Ergebnisse des χ^2 wurden durch das frei zugängliche Internet-Programm VassarStats von Richard Lowry (1998-2005b, Vassar College, Poughkeepsie) <http://faculty.vassar.edu/lowry/VassarStats.html> errechnet.

Der dritte Schritt eröffnet uns die Möglichkeit, die These der «Affinität» gewisser syntaktischer Relationen zu gewissen pragmatischen Beziehungen oder Satzgliedstellungen zu überprüfen, so wie es in letzter Zeit postuliert wird (vgl. z.B. Gil 1999 in Rückgriff auf Oesterreicher 1991). Der vierte Schritt ergibt sich aus folgendem Gedankengang: Wenn die Valenzkonfiguration gleichbleibt, so ist ihr möglicher Einfluß auf die Satzgliedstellung ausgeschaltet, so daß es möglich wird, die Rolle der Verbsemantik genauer zu betrachten.

Die Liste der 15 häufigsten Valenzkonfigurationen (mehr als 1% im Corpus) nach Rojo (2001) fördert die ersten Ergebnisse zu Tage:

(15) Liste der fünfzehn häufigsten Valenzkonfigurationen in ARTHUS (Rojo 2001):

Diathese	Valenzkonfiguration	*Tokens*	% Sätze	Verben	% Verben
Aktiv	*S-V-D*	*64.638*	*40,20*	*2.434*	*70,50*
Aktiv	S-V	16.819	10,50	1.168	33,80
Aktiv	S-V-PS	10.190	6,34	106	3,07
Pronominal	S-V	9.588	5,96	1.352	39,20
Aktiv	*S-V-D-I*	*8.987*	*5,59*	*593*	*17,20*
Aktiv	S-V-AD	7.057	4,39	305	8,83
Aktiv	S-V-SP	5.121	3,19	435	12,60
Pronominal	S-V-SP	4.906	3,05	599	17,30
Aktiv	*S-V-I*	*4.467*	*2,78*	*267*	*7,73*
Aktiv	*S-V-D-PD*	*3.700*	*2,30*	*114*	*3,30*
Pronominal	S-V-AD	3.128	1,95	313	9,06
Aktiv	*S-V-D-AD*	*3.032*	*1,89*	*273*	*7,91*
Pronominal	*S-V-D*	*3.013*	*1,87*	*426*	*12,30*
Pronominal	S-V-PS	2.777	1,73	136	3,94
Aktiv	*S-V-D-SP*	*1.995*	*1,24*	*367*	*10,60*
		149.418	93,94		

Anmerkung: AD = Adverbiale Ergänzung; D = «Direktes Objekt»; I = «Indirektes Objekt»; V = Verb; PS = Subjektprädikativ; S = «Subjekt»; SP = «Suplemento» (Präpositionalobjekt); *Kursiv*: Untersuchte Valenzkonfigurationen

(16) Häufigste M-Konfigurationen mit «direktem» und/oder «indirektem Objekt» (Rojo 2001):

Konfiguration	Diathese	Valenzkonfiguration	*Tokens*	% Sätze	Verben	% Verben
Mindestens D	Aktiv	S-V-D	64.638	40,20	2.434	70,50
	Aktiv	*S-V-D-I*	8.987	5,59	593	17,20
	Aktiv	S-V-D-SP	1.995	1,24	367	10,60
	Aktiv	S-V-D-PD	3.700	2,30	114	3,30
	Aktiv	S-V-D-AD	3.032	1,89	273	7,91
	Pronominal	S-V-D	3.013	1,87	426	12,30
			85.365	*53,67*	*3781*	
Mindestens I	Aktiv	S-V-I	4.467	2,78	267	7,73
	Aktiv	*S-V-D-I*	8.987	5,59	593	17,20
			13.454	*8,46*	*860*	

Einerseits machen die ersten fünf Valenzkonfigurationen 69,30%, die ersten fünfzehn 93,94% der Sätze aus (vgl. Rojo 2001). Von denen kommen sieben in Frage; das ist noch 56,48% des Corpus. Hieraus kann man ersehen, daß zwischen möglichen und häufig realisierten Valenzkonfigurationen ein Mißverhältnis besteht. Der Tatsache, daß eine kleine Anzahl von Valenzkonfigurationen (15 von ca. 150) über 90% des Corpus ausmachen, wurde bisher wenig Aufmerksamkeit geschenkt (vgl. aber Ashby/Bentivoglio 1993 und Bentivoglio 1994 für das Spanische).

Nachdem also die zu untersuchenden Valenzkonfigurationen isoliert worden sind, geht es darum, Statistiken zu den verschiedenen Stellungsmustern zu erstellen, welche die Konfigurationen benutzen können. In diesem Punkt wirken sich zwei Entscheidungen beim Aufbau der Datenbank als besonders nachteilig aus. Zum einen ist es sehr schwierig, Valenzkonfigurationen und Stellungsmuster in Deckung zu bringen; denn die Valenzkonfiguration ergibt sich aus der Information von mehreren Feldern, während das Stellungsmuster in einem Feld angegeben ist. Es muß also ein Programm geschrieben werden, das beide Informationen in Beziehung setzt und die entsprechenden Statistiken kalkuliert. Zum zweiten sind die Argumenten in beiden Fällen anders kodiert; unter der Angabe von «*Extraposición*» werden außerdem undifferenziert verschiedene Träger syntaktischer Beziehungen aufgenommen, wie dies oben gezeigt wurde. Es ist uns also zum gegenwärtigen Zeitpunkt nicht möglich, verläßliche Statistiken zu den Stellungsmustern der verschiedenen Valenzkonfigurationen vorzulegen.[13] Es dürfte aber klar sein, daß Aussagen über die Stellung von «direktem» und «indirektem Objekt» eine ganz andere Tragweite haben als die über die Stellung des «Subjekts». Ein «Subjekt» haben alle Verben, auch wenn dieses nicht immer explizit gemacht wird. Ein «Objekt» ist weit weniger häufig, weil die nötigen Valenzkonfigurationen viel seltener anzutreffen sind, wie oben gezeigt, und, selbst wenn dieses explizit gemacht würde, könnte es nicht immer «frei» plaziert werden, denn sehr häufig handelt es ich um Klitika, Interrogativ-/Exklamativ- oder Relativpronomina, die festgelegte Positionen einnehmen.

Obwohl die Einteilung der Valenzkonfigurationen nach ihren Stellungsmustern fehlt, kann exemplarisch weiter verfahren werden. Ein Teil der Information zu den verschiedenen Stellungsmöglichkeiten der expliziten, nicht pronominalen Argumente von mindestens zwei- und dreiwertigen Verben werden durch die folgenden Tabellen veranschaulicht. Sie zeigen die Stellungsmöglichkeiten von prä- und postverbalen «direktem» und «indirektem Objekt» bezüglich eines weiteren Partizipanten unabhängig von der übrigen Valenzkonfiguration und von weiteren dazwischen liegenden nicht valenziellen Elementen:

---

[13] Während meines Forschungsaufenthaltes an der Universidade de Santiago de Compostela in September 2000 war Herr Prof. Dr. Guillermo Rojo so freundlich, ein Programm für die Untersuchung dieser Frage zu schreiben, das allerdings bisher noch nicht richtig funktionierte. Ihm gilt trotzdem mein Dank für die Mühe, die er sich gemacht hat.

(17) Beschreibung der möglichen Grundstellungen bei explizitem nicht pronominalem «direkten Objekt» und «indirekten Objekt» in Mindestkonfigurationen:

Grundstellung	Partizipant	Gesamt	Deklarativsätze	Verhältnis
Voranstellung	*C*	*1.625*	*1.285*	*2,71%/2,10%*
	C, S	123		
	I	*1.004*	*951*	*27,00%/25,94%*
	I, S	84		
	C, I (s. o.)	7		
	C, I, S	0		
Nachstellung	*C*	*59.809*		*97,35%/97,90%*
	C, S	1.257		
	C, U	3.552		
	C, D	14		
	C, A	10.221		
	C, P	2.416		
	I	*2.714*		*73,00%/74,05%*
	I, S	197		
	I, U	76		
	I, D	0		
	I, A	370		
	I, P	43		
	C, I	1.557		
	C, I, S	12		

(18) Stellung von vorgestelltem «direkten Objekt» und «indirekten Objekt» im Verhältnis zum «Subjekt»

«Direktes Objekt»			«Indirektes Objekt»		
Anordnung	Anzahl	Frequenz	Anordnung	Anzahl	Frequenz
S<C<V	69	56,09%	S<I<V	16	19,04%
C<S<V	54	43,90%	I<S<V	68	80,95%

(19) Stellung von nachgestelltem «direkten Objekt» und «indirekten Objekt» im Verhältnis zu anderen Partizipanten

«Direktes Objekt»			«Indirektes Objekt»		
Anordnung	Anzahl	Frequenz	Anordnung	Anzahl	Frequenz
V<S<C	*658*	*52,34%*	*V<S<I*	*111*	*56,34%*
V<C<S	*599*	*47,65%*	*V<I<S*	*86*	*43,65%*
V<U<C	861	24,23%	V<U<I	38	50,00%
V<C<U	2691	75,76%	V<I<U	38	50,00%
V<A<C	*4684*	*45,82%*	*V<A<I*	*177*	*47,83%*
V<C<A	*5537*	*54,17%*	*V<I<A*	*193*	*52,16%*
V<P<C	1279	52,93%	V<P<I	29	67,44%
V<C<P	1137	47,06%	V<I<P	14	32,55%
V<I<C	*566*	*36,35%*	*V<C<I*	*991*	*63,64%*

Die letzte Tabelle zeigt eine leichte Tendenz des «direkten Objekts» <C> zur Verbnähe bei Kombination mit Angaben <A>, eine stärkere bei Valenzkonfigurationen mit «indirektem Objekt» <I> und eine ganz eindeutige bei «Präpositionalobjekten» <U>. Das «indirekte Objekt» tendiert nur in Kombination mit Angaben leicht zum Verb hin. «Subjekte» <S> neigen leicht bei Inversion zur Verbnähe, was auch von «Präpositionalobjekten» gesagt werden kann. Ob dies Folge der unterschiedlichen «Rhematizität» dieser Partizipanten ist, kann aufgrund dieser Daten nicht entschieden werden (vgl. Kapitel 6).

Innerhalb einer Valenzkonfiguration sollte diese als Faktor der Satzgliedstellung weitgehend ausgeschaltet sein, so daß die Rolle der Verbsemantik herausgearbeitet werden kann. Die zehn Verben der folgenden Tabelle mögen als Beispiel dienen:

(20) Stellung von nachgestelltem «direkten» und «indirekten Objekt» zueinander bei ausgewählten Verben (>250 *tokens*) mit der minimalen Valenzkonfiguration S-D-I

Verbtyp	Verb	*Tokens*	S-D-I	%	Alle	%	V<C<I	%	V<I<C	%
Dicendi	contar	682	312	45,74	22	7,25	8	36,36	14	63,63
(A>R)	*decir* (!)	2686	610	22,71	43	7,05	9	20,93	34	79.06
	explicar	443	113	25,50	15	13,27	2	13,33	13	86,66
Dicendi	pedir	538	272	50,55	65	23,89	19	29,23	46	70,76
(R>A)	preguntar	901	222	24,63	42	18,91	2	4,76	40	95,23
Dandi	*dar*	3169	1315	41,49	320	24,33	281	87,81	39	12,18
	poner	1389	187	13,46	60	32,08	58	96,66	2	3,33
Recipiendi	quitar	302	122	40,39	24	19,67	17	70,83	7	29,16
Kausativa	*hacer*	5446	516	9,47	85	16,47	81	95,29	4	4,70
	permitir	285	128	44,91	15	11,71	0	0,00	15	100,00

Die Verben wurden in mehreren Kategorien eingeteilt (vgl. Kapitel 5). Von allen gibt es im Corpus mehr als 250 *tokens*, allerdings weisen im Schnitt nur 31,88% davon die erforderliche Valenzkonfiguration auf. Von diesen wiederum haben im Schnitt nur 16,11% davon explizite, nicht pronominale «Objekte». Eindeutig lassen sich *verba dicendi* von *verba*

dandi/recipiendi aufgrund ihrer Bevorzugung der Anordnung V<I<C trennen (Schnitt von 79,06% gegenüber einem Schnitt von 14,89%). Eine genauere Besprechung dieser Tabelle erfolgt in Kapitel 5.

Der zweite Weg beinhaltet folgendes Verfahren:

a. Erstellung der Liste der häufigsten und repräsentativsten Verben.
b. Auswahl der Verben mit der erforderlichen Valenzkonfiguration.
c. Auswahl der Verben, die weitere interessante semantische Merkmale aufweisen.
d. Untersuchung der Stellung der Argumente.

Auf den ersten Blick ist nicht leicht erkennbar, in welcher Reihenfolge diese Stationen durchlaufen werden müssen. Vorteil dieses Verfahrens könnte die Möglichkeit sein, die Verbsemantik bei der Untersuchung der Satzgliedanordnung weitgehend auszuschalten. Wenn das Verhalten semantisch ähnlicher Verben unter die Lupe genommen wird, wäre es denkbar, daß die Rolle der Valenzkonfiguration besser zur Geltung kommen kann.

Die Liste der 20 häufigsten Vollverben (Rojo 2001) bringt schon erste Erkenntnisse:

(21) Liste der zwanzig häufigsten Vollverben in ARTHUS (Rojo 2001)

Stelle	Verb	%
1	*ser*	12,50
2	decir	4,40
3	*estar*	3,70
4	tener	3,00
5	hacer	2,80
6	querer	2,70
7	ver	2,00
8	dar	1,60
9	saber	1,60
10	haber	1,50
11	*ir*	1,10
12	creer	0,97
13	pasar	0,96
14	parecer	0,91
15	hablar	0,81
16	dejar	0,80
17	pensar	0,74
18	poner	0,73
19	llevar	0,73
20	quedar	0,71

Einerseits machen die ersten zehn Verben 39,90%, die ersten zwanzig 47,76% der Sätze aus (Rojo 2001). Von ihnen kommen nur drei offensichtlich nicht in Frage (*kursiv* gesetzt: 17,30%); mit den restlichen siebzehn kann 30,46% des Corpus abgedeckt werden. Andererseits gibt es 748 Verben (21,04% der Verben), die nur einmal, und 1.679 Verben (47,24%), die weniger als fünfmal vorkommen (Rojo 2001). Damit wird klar, daß die häufigsten Verben nicht die ganze Bandbreite möglicher Valenzkonfigurationen repräsentieren und daß das Kriterium der Häufigkeit nur eingeschränkt Verwendung finden kann. Die Tatsache, daß eine kleine Anzahl von Verben fast die Hälfte der Sätze des Corpus ausmacht, wird meines Wissens in Untersuchungen häufiger mit Stillschweigen übergangen.

Die Auswahl der Verben mit den erforderlichen Valenzkonfigurationen, die oben angedeutet wurde, muß mit der Wahl der Verben mit interessanten Valenzkonfigurationen verbunden werden, denn es besteht kein 1:1 Verhältnis zwischen Verben und Valenzkonfigurationen, geschweige denn zwischen Verben und Bedeutungen. Im folgenden möchte ich die Ergebnisse einer solchen Auswahl am Beispiel zweier Verben unterschiedlicher Häufigkeit (841 vs. 3097 Vorkommen) darstellen, die die untersuchten Valenzkonfigurationen (s. u.) enthalten und unter dem etwas unscharfen Oberbegriff der «Erkenntnisverben» zusammengefaßt werden können: «*conocer*» und «*saber*» (s. (22); vgl. Blumenthal (1998: 14-24; 1999) für Italienisch und Französisch).

(22) Hauptbedeutungen der spanischen Erkenntnisverben «aprender», «conocer» und «saber» in der traditionellen Lexikographie (nach Moliner 1990: s. v.)
  a. «*conocer*»:
    i. «Tener alguien en la mente o tener la mente misma, la representación de las cosas o de cierta cosa.»
    ii. «Enterarse o estar enterado de cierto suceso o noticia.»
  b. «*saber*»:
    i. «Tener en la mente ideas verdaderas acerca de determinada cosa.»
    ii. «Tener los conocimientos o la habilidad necesarios para hacer [algo].»

Die logische Struktur dieser Verben kann folgendermaßen dargestellt werden (vgl. Kapitel 5):

(23) Prototypische logische Strukturen der spanischen Verben «conocer» und «saber» (in Anlehnung an Van Valin/LaPolla 1997: 102-113)
  a. «*conocer*»:
    i. BECOME **know´** (x, y)
    ii. **know´** (x, y)
  b. «*saber*»:
    i. **know´** (x, y)
    ii. **be´** (x, [**known´**(x, y)]

Die Valenzkonfigurationen werden anschließend dargestellt:

(24) Valenzkonfigurationen des spanischen Verbs «*conocer*» in ARTHUS (Belén López Meirama, Rechercheergebnisse, BDS, 22.07.1999/unkorrigiert)

Valenzkonfigurationen	*«Tokens»*	VK/Verb
A. Aktivisch	*756*	*89,89%*
S	5	0,59%
SD	732	87,04%
SD PD	7	0,83%
SD PR	10	1,19%
SDI	2	0,24%
B. Pronominal	*79*	*9,40%*
S	61	7,25%
S PR	1	0,12%
S I	4	0,48%
SD	12	1,43%
SD PR	1	0,12%
C. Passivisch	*6*	*0,72%*
S	2	0,24%
S PS	4	0,48%

«*conocer*»: 841 *tokens*; 12 Valenzkonfigurationen

(25) Valenzkonfigurationen des spanischen Verbes «*saber*» in ARTHUS (Belén López Meirama, Rechercheergebnisse, BDS, 22.07.1999/unkorrigiert)

Valenzkonfigurationen	*«Tokens»*	VK/Verb
A. Aktivisch	*2961*	*95,40%*
S	395	12,75%
S MD	3	0,10%
S PR	2	0,06%
S SP	35	1,13%
S I PS	1	0,03%
S IMD	1	0,03%
S ISP	4	0,13%
SD	*2437	78,69%
SD PD	2	0,06%
SD AD	3	0,10%
SD PR	13	0,42%
SD SP	61	1,97%
SDI	3	0,10%
SDIPR	1	0,03%
B. Pronominal	*132*	*4,26%*
S	87	2,81%
S PS	6	0,19%
S SP	8	0,26%
SD	26	0,84%
SD PD	5	0,16%

C. Passivisch	*4*		*0,13%*
S	3	0,10%	
S A	1	0,03%	

«*saber*»: 3097 *tokens*; 21 Valenzkonfigurationen

Die semantische Ähnlichkeit kann auch daran erkannt werden, daß sie aktivisch, pronominal und passivisch in ähnlicher Häufigkeit vorkommen.

Nach der groben Vorstellung der Verben werden in zwei Schritten die Valenzkonfigurationen herausgefiltert, die Gegenstand der Untersuchung sind. Zunächst zeige ich die Valenzkonfigurationen, in denen es ein «direktes Objekt», ein «indirektes Objekt» oder beides gibt:

(26) Direktes und indirektes Objekt in den untersuchten Valenzkonfigurationen des spanischen Verbes «*conocer*» in ARTHUS (Belén López Meirama, Rechercheergebnisse, BDS, 22.07.1999/unkorrigiert)

Valenzkonfigurationen (VK)	«*Tokens*»	VK/Verb	
D	*762*		*90,61%*
Aktivisch			
SD	732	87,04%	
SD PD	7	0,83%	
SD PR	10	1,19%	
Pronominal			
SD	12	1,43%	
SD PR	1	0,12%	
I	*4*		*0,48%*
Pronominal: S I	4	0,48%	
D+I	*2*		*0,24%*
Aktivisch: SDI	2	0,24%	

«*conocer*»: 768 *tokens* (91,33%); 12 Valenzkonfigurationen

(27) Direktes und indirektes Objekt bei den untersuchten Valenzkonfigurationen des spanischen Verbes «*saber*» in ARTHUS (Belén López Meirama, Rechercheergebnisse, BDS, 22.07.1999/unkorrigiert)

Valenzkonfigurationen	*«Tokens»*	VK/Verb	
I	*6*		*0,19%*
Aktivisch			
S I PS	1	0,03%	
S IMD	1	0,03%	
S ISP	4	0,13%	
D	*2547*		*82,24%*
Aktivisch			
SD	*2437	78,69%	
SD PD	2	0,06%	
SD AD	3	0,10%	
SD PR	13	0,42%	

SD SP	61		1,97%	
Pronominal				
SD	26		0,84%	
SD PD	5		0,16%	
D+I		*4*		*0,13%*
Aktivisch				
SDI	3		0,10%	
SDIPR	1		0,03%	

«*saber*»: 2557 *tokens* (82,56%); 12 Valenzkonfigurationen

In einem weiteren Schritt filtere ich die Verben heraus, für dessen «direktes Objekt» die Merkmalsopposition [+/–belebt] relevant ist. Damit scheiden alle «direkten Objekte» aus, die als Komplementsatz in der direkten oder der indirekten Rede realisiert werden und es bleiben nur die Fälle von nominalen und pronominalen «direkten Objekten». Für «*conocer*» und «*saber*» werden gleich die Statistiken angegeben.

(28) Direkte Objekte und Belebtheit beim spanischen Verb «conocer» in ARTHUS (Belén López Meirama, Rechercheergebnisse an der BDS, 22.07.1999/unkorrigiert)

Valenzkonfiguration (fein)	«*Tokens*»		VK/Verb	
Dan		*377*		*44,82%*
Aktivisch				
San Dan	360			
Sin Dan	1			
San Dan PDfadj( )	1			
San Dan PDfn (como)	1			
San Dan PDfn (por)	3			
San Dan PDfprp( )	2			
San Dan PRin (de)	4			
Pronominal				
San Dan	4			
San Dan PRin (con)	1			
Din		*364*		*43,28%*
Aktivisch				
San Din	345			
Sin Din	7			
Sinc Din	1			
San Din PRan (de)	2			
San Din PRin (de)	4			
San Din Ian	2			
Pronominal				
San Din	3			

«*conocer*»: 741 *tokens* mit direktem Objekt [+/–belebt] (88,10%)

(29) Direktes Objekt und Belebtheit beim spanischen Verb «saber» in ARTHUS (Belén López Meirama, Rechercheergebnisse, BDS, 22.07.1999/unkorrigiert)

Valenzkonfiguration (fein)	*«Tokens»*	VK/Verb
Dan	*23*	*0,74%*
Aktivisch		
San Dan	10	
Sin Dan	1	
San Dan PDfadj( )	2	
San Dan ADin (en)	3	
Scf Dan IxxxPRxxx( )	1	
Pronominal		
San Dan	1	
San Dan PDfadj( )	4	
San Dan PDfn ( )	1	
Din	*367*	*11,85%*
Aktivisch		
San Din	292	
Sin Din	1	
San Din PRan (de)	3	
San Din PRcf (de)	2	
San Din PRin (de)	5	
Sin Din PRin (de)	1	
San Din SPan (acerca)	1	
San Din SPan (de)	12	
San Din SPan (sobre)	2	
San Din SPcf (de)	1	
San Din SPin (acerca)	2	
San Din SPin (de)	25	
San Din SPin (sobre)	1	
Sin Din SPin (acerca)	1	
San Din Iin	1	
Pronominal		
San Din	17	

«*saber*»: 390 *tokens* mit direktem Objekt [+/- belebt] (12,59%).

An dieser Stelle sieht man, daß «*saber*», das neunthäufigste Verb im Corpus, nur in 12,59% der Fälle die Valenzkonfiguration und -realisierung aufweist, die für die Untersuchung benötigt wird. Bei *«conocer»* kommt ca. 90% in Frage. Dieses selegiert «direkte Objekte» mit den Merkmalen [+/–belebt] fast gleich häufig, während «*saber*» hauptsächlich solche mit dem Merkmal [–belebt] nimmt. Wir stellen fest, daß trotz einer großen Bandbreite an möglichen Valenzkonfigurationen eindeutige Verhältnisse hinsichtlich der dominanten Konfigurationen ausgemacht werden können. Aussagen über die Rolle verschiedener Valenzkonfigurationen bei der Satzgliedstellung erscheinen also wenig aussichtsreich.

In der Tat bieten beide Verben am Ende nur sehr wenig Material zur Voranstellung des «direkten Objekts»: Das Verb «*conocer*» bietet 25 Beispiele, von denen 16 gültig sind (1,90%) und «*saber*» 22, von denen 18 gültig sind (0,58%).

(30) Beispiele für verschiedenartige Voranstellungen mit den spanischen Verben «conocer» und «saber» nach ARTHUS/BDS.

a. MORO: 43, 15
(Alberto se pone tenso ante la alusión de Chusa a sus / relaciones.) / ALBERTO.- ¿A qué viene eso ahora? Es otra cosa, / ¿no? ***A ella*** no *la* conozco de nada. Tú a veces dirás también / que no, digo yo. ¿O es que te metes en la cama con / todo el que te lo pide? / CHUSA.- ¿Y a ti qué te importa con quién me meto?

b. SEVILLA: 94, 2
¿Qué te parece la televisión, qué programas te / gustan más de ella?. / Pues mira, yo ***la televisión***, en realidad / comparándola con otros países conozco muy poco, / porque el único país que conozco es Italia, y / francamente... . Así que no puedo decir qué me / parece [...]

c. OCHENTA: 66, 18
(Se interrumpe, quebrándosele la voz.) ...decía. / JUAN.- Sigue. / MARI ANGELES.- ...No puedo acabar de creer que ya no / le veré más. ***Que se ha muerto*** *lo* sé, lo acepto, pero que / nunca más..., nunca, nunca más le voy a volver a ver... Es / absurdo. Como si de pronto desaparecieran todos los árboles / o algo así.

d. COARTADA: 50, 6
BLANCA.- ¿Pensáis eso, Eminencia? Yo estoy de / acuerdo con vos. Los médicos... / CARDENAL.- Los médicos, mi señora Blanca, ***poco*** saben / de medicinas, menos de enfermedades y nada del / cuerpo humano. Confiemos en que pronto se recupere la / sabiduría de los antiguos.

Deshalb habe ich diesen Weg für weniger praktikabel erachtet, um allgemeingültige Aussagen zur Satzgliedstellung zu machen.

Als Fazit dieses Abschnitts kann gesagt werden, daß die Verbsemantik über die Valenzkonfiguration zum Ausdruck gebracht wird. Darum können nur gewisse Verbklassen innerhalb einer Valenzkonfiguration gefunden werden (erster Weg) und umgekehrt: gewisse Verbklassen, die unabhängig bestimmt werden, werden regelmäßig durch dieselben Valenzkonfigurationen vertreten (zweiter Weg). Dies aber ist eine Frage, die den Rahmen dieser Arbeit sprengt, und trotzdem als Hintergrundsinformation nicht unwichtig erscheint. Auch die Frage der positionellen Variabilität innerhalb einer Valenzkonfiguration ist nur durch einen Paradoxon zu beantworten. Je kleiner eine Valenzkonfiguration ist, desto kleiner sind die Variationsmöglichkeiten der Elemente; möglicherweise werden sie häufiger benutzt. Je größer die Valenzkonfiguration ist, desto rechnerisch größer sind Variationsmöglichkeiten der Elemente; möglicherweise werden sie aber nicht so häufig benutzt. Diese Betrachtungsweise führt uns allerdings nur in die Richtung der Beschreibung von «Wortfolgen»; Aspekte von Syntax, Semantik und Pragmatik werden dabei nicht berücksichtigt. Unser Interesse gilt vielmehr der «Wortstellung», der Stellung eines Elementes bezüglich eines Referenzelementes (das Prädikat). Diese wird selbstverständlich zunächst einmal auch von der Zahl der Elemente einer Valenzkonfiguration bestimmt. Das Zusammenspiel von Syntax, Semantik und Pragmatik wird darauf aufgebaut und soll der eigentliche Gegenstand dieser Arbeit werden, auch wenn die Methode darin besteht, Anordnungen aus dem Corpus herauszufiltern.