

# Nutzung eines konventionellen Valenzwörterbuches in einem MÜ-System: Insegnare l'italiano a CAT2

Leo Wanner  
ISI, University of Southern California

## 1 Einleitung

Für jedes System im Bereich der maschinellen Sprachverarbeitung (im Englischen als *Natural Language Processing*, NLP bekannt), das das Experimentierstadium verlassen hat, stellt sich das Problem der Akquisition umfangreicher linguistischer Ressourcen—insbesondere der Lexika. Da eine manuelle Erstellung zeit- und kostenaufwendig ist, geht der Trend dahin, konventionelle maschinenlesbare Wörterbücher soweit wie möglich automatisch oder semi-automatisch in ein vom System nutzbares Format zu überführen; siehe z.B. entsprechende Studien zur Verwendung des *Oxford Advanced Learner's Dictionary* in [Heyn 92], des *Longman Dictionary of Contemporary English* in [Wilks *et al.* 89, Sanfilippo and Poznański 92], etc. Es werden aber auch Anstrengungen unternommen, in verschiedenen NLP-Anwendungen bereits genutzten lexikalischen Ressourcen in einer lexikalischen 'Wissensbasis' zu vereinen [Hovy and Knight 92, Knight 93, Knight and Luk 94, Viegas *et al.* 96].

Das CAT2-System, ein transfer-basiertes MÜ-System [Sharp91, SharpStreiter92, Haller93], hat das Experimentierstadium bereits vor längerem verlassen und nähert sich in großen Schritten dem industriellen Einsatz. Somit stellt sich auch für CAT2 das Problem der Akquisition lexikalischer Ressourcen für die behandelten Sprachen—u.a. Deutsch, Englisch, Französisch, Italienisch und Russisch.

Am *Institut für Linguistik/Romanistik* (ILR) der Universität Stuttgart wird ein detailliertes Verbvalenzwörterbuch für das Italienische erstellt [Blumenthal und Rovere in Druck]. Das Ziel des in diesem Bericht beschriebenen Experiments war es auszuloten, in welchem Maße das ILR-Valenzlexikon für CAT2 nutzbar gemacht werden kann (und welche Informationen gegebenenfalls noch hinzuzufügen wären). Das Wörterbuch wird in Kürze über das Internet frei verfügbar sein und kann dann ohne Einschränkungen genutzt werden.

Der Bericht gliedert sich thematisch in drei Teile. Im ersten Teil präsentieren wir CAT2 in der Welt der maschinellen Übersetzung. Dazu gehört (i) ein kurzer Überblick über die Ansätze in der MÜ und die Zuordnung von CAT2 zu einem dieser Ansätze (Abschnitt 2); (ii) Vorstellung der Architektur von CAT2 und seiner Arbeitsweise

(Abschnitt 3); (iii) Diskussion der Struktur der lexikalischen Einträge für Verben in CAT2, die uns in Verbindung mit dem Verbvalenzlexikon besonders interessieren.

Im zweiten Teil stellen wir das Valenzwörterbuch vor (Abschnitt 4).

Im dritten Teil diskutieren wir dann die Überführung des Valenzwörterbuches in das in CAT2 verwendete Format (Abschnitt 6, sowie die Möglichkeit, die im Valenzwörterbuch nicht vorhandene, von CAT2 jedoch benötigte Information zu akquirieren (Abschnitt 7).

Ein Anhang enthält ein einfaches Beispiel der deutsch-italienischen Übersetzung.

## 2 Ansätze in der Maschinellen Übersetzung

Im allgemeinen lassen sich drei verschiedene Strategien der Maschinellen Übersetzung unterscheiden: (i) direkte Übersetzung, (ii) transfer-basierte Übersetzung und (iii) interlingua-basierte Übersetzung. Eine detaillierte Einführung in die maschinelle Übersetzung findet man z.B. in [Hutchins 86, Hutchins and Somers 92] und mit Einschränkung auch in [Schwanke 91] und [Apresjan *et al.* 89]. Sehr aufschlußreich ist auch der von Nirenburg herausgegebene Sammelband [Nirenburg 87].

Was uns in dem gegebenen Zusammenhang besonders interessiert, ist welche Rolle spielt die Valenzinformation in den Lexika bei den verschiedenen Ansätzen.

### 2.1 Direkte Übersetzung

Die direkte Übersetzung ist die älteste Art der maschinellen Übersetzung. Sie ist dadurch gekennzeichnet, daß der Übersetzungsprozess monolytisch ist und im wesentlichen eine ‘Wort-für-Wort-Übersetzung’ darstellt. Das Herzstück eines solchen Prozesses sind bilinguale Wörterbücher, die Quell-Zielsprachewortpaare enthalten.

Bei den ältesten direkten Systemen bestand der Übersetzungsprozess aus (i) eventueller Textvorbereitung (einschließlich einer Vorformatierung, Vereinfachung, etc.), (ii) morphologischer Analyse der Quellsprache, (iii) Zugriffen auf bilinguale Wörterbücher, (iv) morphologischer Anpassung/evtl. lokalen Umordnungen lexikalischer Einheiten der Zielsprache (der Prozess ist wesentlich komplexer bei späteren direkten Systemen wie SYSTRAN). Die Ergebnisse dieser Systeme war—wie man sich denken kann—von sehr schlechter Qualität. Anbei zwei Beispiele krasser Fehlübersetzungen eines frühen direkten russisch-englischen Systems<sup>1</sup>:

(1) Rus. *My trebuem mira*

(2) a. Engl. ‘We require the world’ (‘Wir fordern die Welt’.)  
anstatt:

b. Engl. ‘We require peace’ (‘Wir fordern Frieden’)

---

<sup>1</sup>Die Beispiele sind [Hutchins and Somers 92, 72f] entnommen.

Die Fehlübersetzung resultiert hier aus der Ambiguität von *mir*, das ‘die Welt’ oder ‘der Frieden’ bedeuten kann.

- (3) Rus. *Nam nužno mnogo uglja, železa, elektronergii*
- (4) a. Engl. ‘To us much coal is necessary, gland, electric power’  
(‘Für uns viel Kohle ist notwendig, Drüse, elektrische Energie’)  
anstatt:  
b. Engl. ‘We need a lof of coal, iron and electricity’  
(‘Wir brauchen viel Kohle, Eisen und Elektrizität’)

Hier resultiert die falsche Übersetzung aus der Ambiguität von *nužno* und *železa*. *Nužno* kann ‘brauchen’ (wie in *Nam aetogo ne nužno* ‘Wir brauchen das nicht’) oder ‘notwendig sein’ (wie in *Nužno, čtoby vse javilis’* ‘Es ist notwendig, daß alle erscheinen’ heißen. *Železa* kann entweder für ‘Drüse’ im Nominativ oder ‘Eisen’ im Akkusativ (von *železo*) stehen.

Vetreter der direkten Übersetzung mit solcher Ausgabe benutzen keine Valenzinformation—obwohl sie zweifelsohne zur Verbesserung der Übersetzungsqualität beitragen würde (so haben z.B. die zwei Lesarten von *nužno* verschiedene Valenzrahmen). Dies muss aber nicht immer sein—wie z.B. das System SYSTRAN [van Slype and Pigott 79] zeigt. SYSTRAN verfügt über zwei Arten von Wörterbüchern: ein Stammwörterbuch und eine Reihe von Kontextwörterbüchern. Das Stammwörterbuch enthält für jedes Quellsprachenwort den Stamm (außer für English, wo volle Formen angegeben sind), die Wortart, Flexionstabellennummer, Genus, Kasus, Numerus, etc., Valenzrahmen, verschiedene semantische Codes, usw. Für jedes Zielsprachenwort wird einige morphologische und syntaktische Information angegeben. Unter den Kontextwörterbüchern befinden sich: 1. das idiomatische Wörterbuch, das Idiome entweder wiederum als Idiome oder als wörtliche Ausdrücke übersetzt; 2. das semantische (“limited semantics”) Wörterbuch, das Komposita spezifiziert; 3. das Homographenwörterbuch, das Information zur Auflösung von Homographen enthält; etc. Der im Stammllexikon spezifizierte Valenzrahmen wird im Rahmen einer Analysephase benutzt.

## 2.2 Interlingua

Das Prinzip der Interlingua setzt eine abstrakte Zwischenrepräsentationsebene voraus, die von den Merkmalen der Ausgangs- und Zielsprache weitgehend unabhängig ist. Das Ergebnis der Analyse des Ausgangstextes wird in Interlingua dargestellt. Ausgehend von dieser Interlinguarepräsentation wird der Zieltext generiert. Dabei wird angenommen, daß die Interlinguarepräsentation alle Information enthält, die von dem Zielgenerator eventuell benötigt werden könnte; sie stellt somit gleichzeitig die Repräsentation des Ausgangs- und Zieltextes dar.

In den frühen Jahren der Interlinguaforschung wurde daran gearbeitet, eine wirklich universale Interlingua zu entwickeln, die als Zwischenrepräsentation für praktisch

alle Sprachen dienen könnte. Heutzutage sind die Vorstellungen realistischer, und es wird die Entwicklung einer Interlingua für eine beschränkte Anzahl von Sprachen angestrebt. Kurz erwähnen wollen wir zwei Entwicklungen. In Rahmen des Systems DLT [Schubert 88] (entwickelt von der Firma BSO) wurde *Esperanto* als Interlingua verwendet. Esperanto ist eine “künstliche natürliche” Sprache, die Elemente vorhandener Sprachen enthält, sie jedoch in regulären und konsistenten Strukturen anordnet (was für die natürlich gewachsenen Sprachen nicht immer zutrifft). Dementsprechend wurden als Hauptgründe für die Verwendung von Esperanto als Interlingua angegeben:

- (i) das Ausdruckspotential (das einer sich natürlich gewachsenen Sprache entspricht),
- (ii) die Regularität und Konsistenz, die bei einer natürlich gewachsenen Sprache nicht gewährleistet ist).

Bemerkenswert ist auch der Trend, eine wissensbasierte Interlingua zu entwickeln. Die ersten Denkanstöße in dieser Richtung kamen von IBM Tokyo. Sie wurden am *Center for Machine Translation* an der *Carnegie Mellon University* aufgegriffen und in dem von IBM finanzierten Projekt KBMT weiter geführt [Nirenburg and Goodman 91, Nirenburg *et al.* 92]. Der Formalismus für die Beschreibung der Interlingua (genannt *InterLingua Text*, ILT) ist eine Art semantischen Netzwerks mit durch ‘frames’ spezifizierten Knoten und Kanten. Es werden drei verschiedene Arten von Knoten unterschieden: *Clauses*, *Propositions* und *Roles*. Die *Clauses* enthalten die oberflächennahe Information hinsichtlich einer Proposition (im folgenden Beispiel ist propositionid der Zeiger auf die entsprechende Proposition);:

```
(make-frame clause1
  (ilt-type (value clause))
  (clauseid (value clause1))
  (propositionid (value proposition1))
  (discourse-cohesion-marker (value (conditional clause2)))
  (speechactid (value speech-act1))
```

Die Propositions enthalten die Bedeutung einer Aktion, eines Zustandes, etc.:

```
(make-frame proposition1
  (ilt-type (value proposition))
  (propositionid (value proposition1))
  (clauseid (value clause1))
  (aspect (value aspect1))
  (complete (value yes))
  (is-token-of (value *connect))
  (agent (value unknown))
  (theme (value role2))
  (time (value time1))
```

Die Roles enthalten Objekte, die bei einer oder mehreren Propositionen eine Ak-  
tantenrolle (wie durch die ‘Case’-Theorien bekannt) ausfüllen:

```
(make-frame role2
  (ilt-type (value role))
  (clauseid (value clause1))
  (is-token-of (value *device))
  (r-quantifier (value universal))
  (reference (value definite)))
```

Da eine Interlinguarepräsentation notgedrungen die Semantik einer Äußerung  
widerspiegelt (sonst kann sie nicht sprachunabhängig sein), ist in interlinguabasierten  
Systemen neben der syntaktischen auch die semantische Valenz von Relevanz.

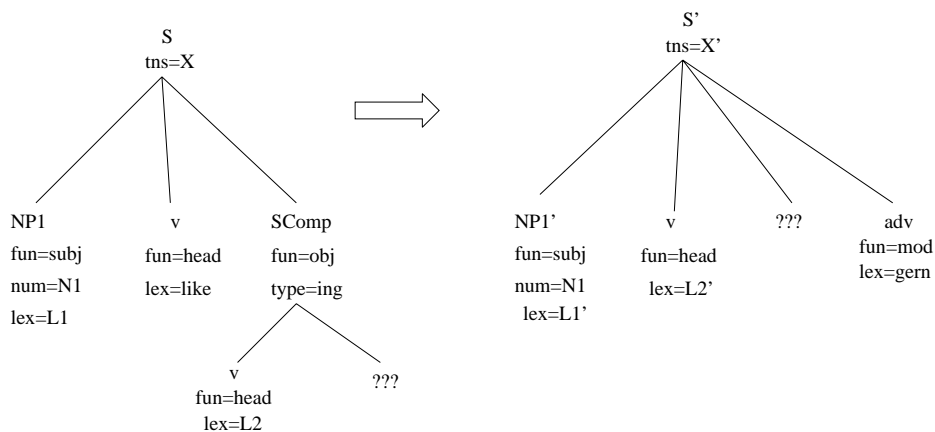
## 2.3 Transfer

Transfer ist auf mehreren Ebenen der Übersetzung möglich. In Anlehnung an E.  
Steiner [Steiner 1993], wollen wir im folgenden drei Transferebenen unterscheiden:

- (i) Situationstransfer,
- (ii) struktureller Transfer,
- (iii) lexikalischer Transfer.

Bei dem Situationstransfer werden Korrespondenzen zwischen Diskurseinheiten  
aufgebaut. D.h. wichtig ist hier nicht eine möglichst wörtliche, sondern eine stylistisch  
und kontextuell adäquate Übersetzung der Äußerung in der Quellsprache.

Bei dem strukturellen Transfer, werden Korrespondenzen zwischen syntaktischen  
Strukturen definiert. Das folgende Beispiel zeigt eine Strukturtransferregel für die  
Übersetzung des Verbs [to] *like* (wie in *Maria likes swimming*) als *gern* (wie in *Maria  
schwimmt gern*:<sup>2</sup>



<sup>2</sup>Das Beispiel ist [Hutchins and Somers 92, 114] entnommen.

Bei dem lexikalischen Transfer werden Korrespondenzen zwischen den lexikalischen Einheiten der Quell- und Zielsprachen definiert. Cf. eine lexikalische Transferregel zwischen dem französischen Verb *aborder* und dem deutschen *nähern*.<sup>3</sup>

$\text{aborder}=\{\text{lex}=\text{aborder}\}.[]\Rightarrow\{\text{lex}=\text{naehern},\text{head}=\{\text{pref}=\text{nil}\}\}.[]$ .

Eng verknüpft mit der Frage des Transfers ist das Problem der Metaphorik und das Problem der Kohärenz und Kohäsion im Diskurs.<sup>4</sup>

CAT2 ist ein bidirektionales, multilinguales transferbasiertes Übersetzungssystem, bei dem der Transfer auf der strukturellen und lexikalischen Ebenen durchgeführt wird. Im folgenden Abschnitt geben wir einen kurzen Überblick über CAT2. Ein Beispiel der Übersetzung eines Satzes aus dem Deutschen ins Italienische durch CAT2 findet sich im Anhang.

### 3 CAT2: Ein kurzer Überblick

Eine detaillierte Darstellung des CAT2-Systems ist in [Haller93] gegeben; eine Einführung in den CAT2-Formalismus in [Sharp91, SharpStreiter92, Streiter 96]. Eine umfangreiche Dokumentation ist ebenfalls erhältlich. Aus diesem Grund präsentieren wir hier lediglich einen kurzen Überblick, der für das Verstehen unseres Experiments notwendig ist.

#### 3.1 CAT2-Entwicklung: Etwas Hintergrundinformation

Die Entwicklung von CAT2 begann 1987 als ein Alternativvorschlag zu dem ursprünglichen Eurotra formalismus  $\langle C,A \rangle, T$  [Arnold *et al.* 86] als sich abzeichnete, daß  $\langle C,A \rangle, T$  die an ihn gestellten Anforderungen nicht erfüllen konnte. Ein kontrastiver Vergleich von CAT2 mit  $\langle C,A \rangle, T$  und mit dem später offiziell für Eurotra adaptierten Formalismus, dem *Eurotra Engineering Framework*, ist in [Sharp91] gegeben, so daß wir an dieser Stelle nicht darauf einzugehen brauchen.

CAT2 profitiert von Erkenntnissen, die im Rahmen verschiedener linguistischer Theorien und Formalismen gewonnen wurden. So spiegelt es einige Merkmale der *Head Phrase Structure Grammar* (HPSG) [Pollard und Sag 87, Pollard und Sag 94], der *Government and Binding Theory* (GB) [Chomsky 81], und sogar solcher dependenzorientierter Theorien wie die *Meaning-Text* Theorie [Mel'čuk 81] wider.

#### 3.2 Der CAT2-Formalismus

Wir stellen zuerst kurz die in CAT2 benutzten Strukturen und dann die Repräsentationsebenen vor.

---

<sup>3</sup>Das Beispiel ist dem CAT2-Transferlexikon entnommen.

<sup>4</sup>Eine zusätzliche Ebene stellt für Steiner der Funktionstransfer (i.e., Transfer der kommunikativen Ziele des Sprechers). Diese Ebene liegt orthogonal zu allen anderen Ebenen.

### 3.2.1 Strukturen in CAT2

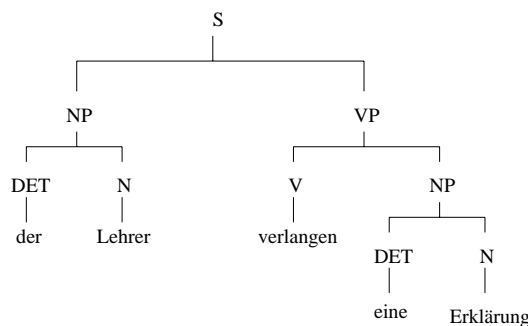
Die Grundstrukturen in dem CAT2-Formalismus sind Attribut-Wertpaare und Baumstrukturen (die wiederum als Attribut-Wertpaare dargestellt werden können). Ein Attribut-Wertpaar wird in Form einer Gleichung (das Attribut links vom Gleichheitszeichen, der Wert rechts davon) in geschweiften Klammern dargestellt:

$$\{\text{lex}=\text{verlangen}, \text{head}=\{\text{pref}=\text{nil}\}, \dots\}$$

Die Knoten einer Baumstruktur sind Attribut-Wertpaare. In einer linearen Darstellung sind die Töchterknoten eines Knoten links von diesem durch eckige Klammern gekennzeichnet; cf. die Struktur für den Satz *Der Lehrer verlangt eine Erklärung*:

$$\{\text{cat}=\text{s}\}[\{\text{cat}=\text{np}\}[\{\text{cat}=\text{d}, \text{lex}=\text{der}\}, \{\text{cat}=\text{n}, \text{lex}=\text{Lehrer}\}], \\ \{\text{cat}=\text{vp}\}[\{\text{cat}=\text{v}, \text{lex}=\text{verlangen}\}, \\ \{\text{cat}=\text{np}\}[\{\text{cat}=\text{d}, \text{lex}=\text{eine}\}, \{\text{cat}=\text{n}, \text{lex}=\text{Erklärung}\}]]]$$

Oft wird eine graphische Darstellung der folgenden Art benutzt:



Auf den Strukturen des Formalismus sind Regeln definiert. Insgesamt werden fünf verschiedene Arten von Regeln unterschieden:

1. **b-Regeln** ('b' steht für 'building'); b-Regeln definieren Baumstrukturen; cf. 8

@rule(b).  
 $\{\text{cat}=\text{s}\}[\{\text{cat}=\text{np}\}, \{\text{cat}=\text{vp}\}]$

die einen Knoten 's' mit den Töchtern 'np' und 'vp' definiert:

GRAPHISCHE DARSTELLUNG

2. **f-Regeln** ('f' steht für 'feature'); f-Regeln werden auf Strukturen angewendet, die durch b-Regeln spezifiziert sind; cf.

@rule(f).  
 $\text{acc} = \{\} . [\{\text{cat}=\text{v}\}, *, \{\text{cat}=\text{np}\}] >> \{\text{CASE}=\text{ACC}\}, *].$

Sie vergleichen die in der Regel spezifizierten Merkmale mit den Merkmalen einer Baumstruktur (und fügen evtl. Merkmale hinzu); so ist die obige Regel anwendbar auf Strukturen, die irgendwo einen 'v'-Knoten besitzen, dessen zweiter Tochterknoten eine SF NP ist, der zusätzlich die Einschränkung 'CASE=ACC' aufweist. f-Regeln können jedoch die Baumstrukturen nicht verändern.

3. **l-Regeln** ('l' steht für 'lexical'); l-Regeln sind im Grunde f-Regeln, die während der Lexikonkompilation angewendet werden (im Unterschied zu f-Regeln, die während des Übersetzungsprozesses zur Anwendung kommen);
4. **t-Regeln** ('t' steht für 'transfer'); t-Regeln werden benutzt, um entweder die Repräsentation auf einer Ebene in die Repräsentation auf der nächst höheren oder niedrigeren Ebene zu überführen oder die Repräsentation auf einer Ebene der Sprache X in die Repräsentation der gleichen Ebene abzubilden; cf.

@rule(t).  
 aborder={lex=aborder}.[]=>{lex=naehern,head={pref=nil}}.[].

wo das französische Lexem *aborder* in das deutsche Lexem *nähern* abgebildet wird.

5. **tf-Regeln** ('tf' steht für transfer feature'); tf-Regeln werden auf zwei Strukturen angewendet, die durch eine t-Regel verbunden sind, oder mit anderen Worten: auf eine Ausgangs- und Ergebnisstruktur einer t-Regel; cf.

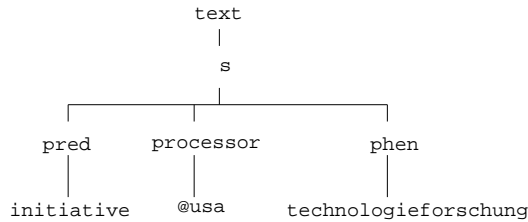
@rule(f).  
 {} . [{cat=np}, {cat=vp}]>>{tense≐nil}  
 =>  
 {} . [{cat=vp}, {}]>>{CASE=NOM}].

die anwendbar ist, wenn (i) die Ausgangs- und die Ergebnisstrukturen jeweils zwei Tochterknoten besitzen (die Ausgangsstruktur np und vp), die Ergebnisstruktur vp ist und einen weiteren Knoten); (ii) die Ausgangsstruktur die Einschränkung 'tense≐nil', und die Ergebnisstruktur 'CASE=NOM') aufweist.

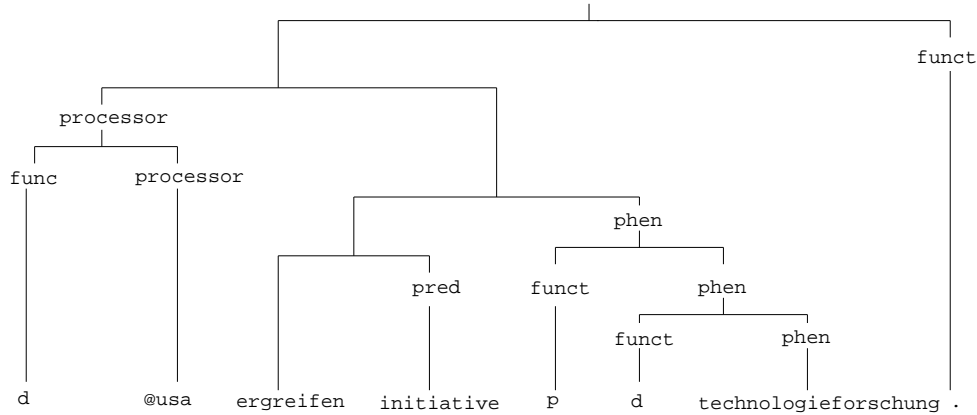
### 3.2.2 Repräsentationsebenen in CAT2

Im Rahmen von CAT2 werden drei sprachspezifische Repräsentationsebenen unterschieden: 1. die *Interface Structure* (IS-) Ebene, 2. die *Constituent Structure* (CS-) Ebene und 3. die *Morphology Structure* (MS-) Ebene.

Die IS-Ebene enthält die semantische Struktur eines Satzes; sie bildet die Schnittstellenebene zwischen der Quell- und Zielsprache. Als Beispiel einer IS zitieren wir aus [Streiter 96] die Struktur des Satzes *Die USA ergreifen die Initiative in der Technologieforschung*.



Die CS-Ebene enthält die syntaktische Struktur eines Satzes. Für den obigen Satz sieht sie wie folgt aus:



Und die MS-Ebene enthält die morphologische Struktur eines Satzes. Cf.

d usa ergreifen initiative p d technologieforschung

Der Ablauf der Übersetzung eines Satzes in CAT2 sieht dann wie im Schaubild 1 illustriert aus: der Satz wird analysiert und in seine MS-Struktur überführt. Mit Hilfe von Regeln der oben eingeführten Arten wird die MS-Struktur in die CS- und diese anschließend in die IS-Struktur überführt. Die IS-Struktur des Quellsatzes wird auf die IS-Struktur des Zielsatzes abgebildet.<sup>5</sup> Die Oberflächenstruktur des Zielsatzes wird durch die Schritte 'IS-Struktur → CS-Struktur → MS-Struktur → Oberflächenstruktur' gewonnen.

## 4 Struktur der lexikalischen Einträge für Verben in CAT2

Da CAT2 ein transfer-basiertes System ist, verfügt es über zwei verschiedene Arten von Lexika: Transferlexika (in unserem Falle *deutsch-italienisch* und *italienisch-deutsch*) und einsprachige Lexika (in unserem Falle für Italienisch). Es gibt zwei einsprachige Lexika: das "open-class lexicon" und das "closed-class lexicon". Wie der Name bereits andeutet, enthält das open-class Lexikon die Einträge für Nomina, Verben, Adjektive und Adverbien. Das closed-class Lexikon enthält Einträge für die Zeichen der Punctuation und Funktionswörter, d.h. Artikel, Präpositionen, Partikel, usw. Auch wenn

<sup>5</sup>Falls Information zur Erstellung der IS-Struktur fehlt, kann auch die CS-Struktur des Quellsatzes in die CS-Struktur des Zielsatzes überführt werden.

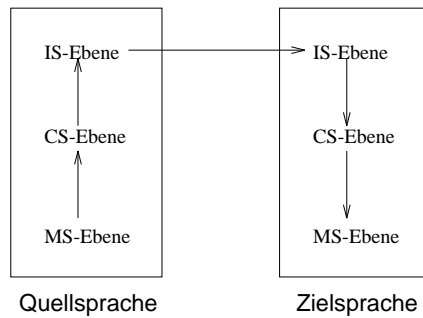


Figure 1: Der Übersetzungsablauf in CAT2.

für unser Unterfangen lediglich das open-class Lexikon von Interesse ist, wollen wir zwei Beispieleinträge für das closed-class Lexikon angeben—nämlich den von ‘,’ und ‘aber’:

Komma (‘,’):

```
punct = {role=funct,lex=',',head={restr=RESTR,cat=punct}}>>
  ({pos=pre,komma_l=yes,
  frame={arg1={frame={pred={xpred=nil,xpred=nil}},komma_l=no,cat~=v,
  head={restr=RESTR,ehead={type~=main,cat=v}}}}})
;{pos=post,komma_r=yes,
  frame={arg1={frame={pred={xpred=nil,xpred=nil}},cat~=v,
  head={restr=RESTR,ehead={type~=main,cat=v}}}}}). [] .
```

Aber:

```
coord_aber =
{role=funct,lex=coord,string=aber,pos=pre,emax=no,max=no,
head={cat=coord,ehead={index=coord,wh=WH,coord=adve,ref={'T'=T},
pvalue=PV,cat=CAT,num=NUM},restr=RESTR},
frame={arg1={max=yes,pos=pre,role=R,head={cat~=coord,restr=RESTR,
ehead={num=NUM,wh=WH,ref={'T'=T},pvalue=PV,cat=CAT}}},
arg2={max=yes,pos=post,role=R,head={cat~=coord,restr=RESTR,
ehead={wh=WH,coord=adve,ref={'T'=T},pvalue=PV,cat=CAT}}}}}. [] .
```

## 4.1 Open-Class Lexikon

Beim open-class Lexikon wollen wir uns im Folgenden auf die Verbeinträge konzentrieren. Die Verbeinträge in CAT2 bestehen aus der *lex*-, *head*- und *frame*-information; cf. den Eintrag für *verlangen*:

```
{lex=verlangen,head={pref=nil},
flex={paradigma=schwach,part={ge=no,end=t}},
head={cat=v,perf=haben}
```

```

;{cat=n,deriv={type=action,action=ing},
  ehead={type~=abs,gen=neuter,num=sing,sem={abstract={abstr=_},
    temp={ext=_},gran=nil,anim=nil}}}),
frame={arg1={role=processor,head={ehead={cat=n,sem={temp=nil,anim={'T'=hum}}}},
  arg2={role=phen,head=({ehead={cat=n}};{ehead={cat=v,tense~=nil},c=yes};
    {ehead={cat=v,tense=nil}}),head=({ehead={cat=n,pform=nil}}
      ;{ehead={cat=v,tense~=nil}})},
arg3={role=processor2,head={ehead={cat=n,sem={temp=nil,anim={'T'=hum}}}},
  head={ehead={pform=von,case=dat}},oblig=no}}}. [] .

```

Wir wollen nun auf diese Arten von Information kurz eingehen.

**Lex-Information.** Die *lex*-information besteht lediglich aus dem Lexem; zum Beispiel: *lex=stampare*.

**Head-Information.** Der Begriff des „head“ ist aus der *Head Phrase Structure Driven Grammar* (HPSG) [Pollard und Sag 87, Pollard und Sag 94] entliehen.<sup>6</sup>

Die *head*-Information enthält i.a.: (i) die Kategorie des Lexems, (ii) die Aspekt-Angabe ‘statisch/ nicht statisch’ (im Sinne von Vendler [Vendler 67]), (iii) Reflexivitätsangaben, (iv) “extended-head“-Information. Funktionsverben sind zusätzlich markiert. Eine typische *head*-Spezifikation sieht dann so aus:<sup>7</sup>

```
head={cat=v, stative=no, refl=yes}, ehead={...} (z.B. für rompere)
```

```
head={cat=v, stative=no, refl=no, vsup=yes}, ehead={...} (z.B. für
badare03)
```

Unter Umständen kann die *head*-Information zusätzlich noch Angaben zur *vform* enthalten (*vform=fin*, *vform=inf*, *vform=ptc*).<sup>8</sup>

**Extended Head-Information.** Der Begriff des “extended head” (*ehhead*) ist eine Adaption von *Extended Projection* in GB [Grimshaw 91]. Wir können (etwas vereinfachend) davon ausgehen, daß *ehhead* per default semantische Angaben folgender Art enthält:

```
sem={temp={ext=_}, abstract={abstr=_}, gran=nil, anim=nil}
```

<sup>6</sup>Und hier wiederum aus der *Generalized Phrase Structure Grammar* (GPSG) [Gazdar *et al.* 85].

<sup>7</sup>Die Numerierung der Lesarten der italienischen Verben bezieht sich immer auf das Stuttgarter Valenzwörterbuch.

<sup>8</sup>Im gegenwärtigen französischen Lexikon nicht enthalten.

‘temp=\_’ drückt durch “\_” aus, daß ein Verb immer auf der zeitlichen Achse definiert ist (eine zeitliche Ausdehnung ausdrückt); ‘abstr=\_’ drückt aus, daß das Verb immer ein Abstraktum bezeichnet; ‘gran=nil’ setzt das Merkmal *gran*,<sup>9</sup> daß nur für Nomina relevant ist, auf null; ‘anim=nil’ ist ebenfalls nur für Nomina (siehe auch unten die Diskussion der “frame”-Information).

Anstatt des Defaulteintrages für temporale Ausdehnung (‘\_’) kann auch eine Aspektspezifikation (*achievement/accomplishment*) angegeben sein (z.B. `temp={ext={'T'=bound,bounded=achiev}}`).<sup>10</sup>

**Frame-Information.** Die Frame-Information enthält die Spezifikation der Argumente des Verbs. Sie beinhaltet einerseits die *head*-Information; andererseits die semantische Rolle, die das Argument gemäß dem semantischen Typ des Verbs innehat. Beispielsweise sieht die Frameinformation für *dire* ‘sagen’ wie folgt aus:

```
frame={arg1={role=processor,
             head={ehead={cat=n,sem={temp=nil,anim={'T'=hum}}}},
arg2=({role=phen,head={cat=c,string=que}};
      {role=phen,
       head={ehead={cat=n,pform=nil,sem={abstract={abstr=_}}}})})}
```

#### 4.1.1 Transferlexikon in CAT2

Das Transferlexikon enthält direkte Übersetzungen; d.h. Lexempaare. Vgl. einige Beispiele, die dem französisch-deutschen Transferlexikon entnommen sind:

```
aborder={lex=aborder}. []=>{lex=naehern,head={pref=nil}}. [].
aborder={lex=aborder}. []=>{lex=erreichen,head={pref=nil}}. [].
aborder={lex=aborder}. []=>{lex=sprechen,head={pref=an}}. [].
aborder={lex=aborder}. []=>{lex=legen,head={pref=an}}. [].
aborder={lex=aborder}. []=>{lex=gehen,head={pref=an},
                             frame={arg2={pform=nil}}}. [].
accentuer={lex=accentuer}. []=>{lex=betonen,head={pref=nil}}. [].
accomplir={lex=accomplir}. []=>{lex=erledigen,head={pref=nil}}. [].
accord={lex=accord}. []=>{lex=vereinbaren,head={pref=nil}}. [].
accroitre={lex=accroi3tre}. []=>{lex=vergroessern,head={pref=nil}}. [].
acheter={lex=acheter}. []=>{lex=kaufen,head={pref=nil}}. [].
acquisition={lex=acquisition}. []=>{lex=erwerben}. [].
```

## 5 Das Valenzwörterbuch

Das italienische Verbvalenzwörterbuch, wie es gemeinsam von dem *Institut für Linguistik/Romanistik*, Universität Stuttgart und dem *Institut für Übersetzen*

<sup>9</sup>Gran steht für ‘granularity’; es kann die Werte ‘nil’, ‘unique’ oder ‘part’ annehmen.

<sup>10</sup>Wenn `ext` mit `{ext={'T' ...}}` angegeben ist, muß ‘T’ entweder ‘bound’ oder ‘unbound’ sein; wenn es ‘bound’ ist, ist eine weitere Spezifikation von ‘bounded’ erforderlich (z.B. ‘bounded=achiev’).

und *Dolmetschen*, Universität Heidelberg im Rahmen eines DFG-Projektes erstellt wurde, ist ein zweisprachiges Wörterbuch, das in erster Linie für den deutschen Fachübersetzer aus dem Deutschen ins Italienische gedacht ist. In seiner Makrostruktur lehnt es sich stark an das französische Valenzwörterbuch von Busse und Dubost [Busse and Dubost 83] an. Dieser Makrostruktur liegt die Annahme zugrunde, daß der ideale Benutzer Italienisch gut genug beherrscht, um die italienische Übersetzung der Verben im deutschen Original zu kennen, jedoch über ihre Valenzschemata im Zweifel ist. Daher dominiert die Richtung ‘Italienisch $\implies$ Deutsch’. Die Richtung ‘Deutsch $\implies$ Italienisch’ ist über einen Index am Ende des Wörterbuches zugänglich.

Jedem italienischen Lexem wird ein Lemma zugeordnet; die italienischen Lemmata sind alphabetisch geordnet. Ein Lemma ist i.a. in mehrere Sublemmata unterteilt. Die Kriterien für die Einführung eines Sublemmas sind sowohl semantischer als auch (valenz-)syntaktischer Natur, nämlich (i) die eindeutige Unterscheidung einer neuen Lesart, und (ii) Unterscheidung hinsichtlich des Valenzschemas. Das erste Kriterium für die Einführung eines Sublemmas, nämlich die eindeutige Unterscheidung einer neuen Lesart, wird jedoch nicht immer durchgehalten—wie man anhand des dritten Sublemmas für *combinare* sieht:<sup>11</sup>

1. *combinare* 03
2. N-V-N1
3. \*zusammenmischen, \*mischen, \*kombinieren, \*zusammenfügen, \*organisieren, \*abschließen, \*vereinbaren, \*sich einigen auf, \*machen, \*schaffen, \*anrichten, \*zustande bringen, \*wählen, \*verbinden (chem.)

(so signalisieren, z.B. ‘organisieren’, ‘vereinbaren’ und ‘zustande bringen’ deutlich unterscheidbare Lesarten).

Um den Aufbau des Wörterbuches vorstellen zu können, diskutieren wir im folgenden Abschnitt zuerst kurz die Präsentation der Valenzrahmen.

## 5.1 Valenzrahmen und ihre Notationen

In den Valenzrahmen (oder *Satzbauplänen*,<sup>12</sup> wie sie projektintern genannt werden) werden folgende Notationen verwendet:

1. ‘V’ steht für das Verb.

---

<sup>11</sup>Das Zeichen ‘\*’ kennzeichnet die deutschen Lexeme, über die im deutschen Index das betreffende italienische Lexem zu finden ist.

<sup>12</sup>Die Benutzung des Begriffs ‘Satzbauplan anstatt ‘Valenzrahmen’ halten wir jedoch für irreführend. Es ist wichtig, zwischen Valenzrahmen und Satzbauplänen zu unterscheiden: in der Literatur wird der Begriff ‘Satzbauplan’ oft zur Bezeichnung morphosyntaktischer Satzrahmen, die **duch** die Valenz des Verbs festgelegt werden, herangezogen. Cf. “Die Satz-Baupläne regeln das Vorkommen der Ergänzungen. Die Fähigkeit des Verbs, Ergänzungen zu fordern und damit den Satzbauplan festzulegen, nennt man seine Valenz [Engel 88, 185].

2. Für nominale Aktantenrollen wird der Buchstabe ‘N’ verwendet (‘N’ für die erste Aktantenrolle, i.e., das Subjekt), ‘N1’ für die zweite Aktantenrolle, i.e., das direkte Objekt, ‘N2’ für die dritte Aktantenrolle, i.e., das indirekte Objekt, usw. Desweiteren wird Npred benutzt um ein prädikatives Nomen, das als eine zusätzliche Aktantenrolle auftritt, zu markieren. Vrgl. einige Beispielvalenzrahmen (die italienischen Beispiele stammen ebenfalls aus dem Wörterbuch):

N-V

für *combinare***02** ‘handeln’: *Io non so >>combinare<< e ho pagato ciò che mi ha chiesto.*

N-V-N1

für *combinare***03** ‘kombinieren, verbinden, etc.’: *... una terza che >>combina<< gli aspetti diplomatici, econmici e militari.*

N-V-N1-(*da* Npred)-*a* Npred

für *abbassare***12** ‘degradieren zu’ (die Klammern signalisieren, daß das Argument optional ist): *L’indicazione ...importante per evitare che si sviluppino nuovi disguidi sempre dipendenti dall’opzione utilitaristica, tra cui quello che >>abbassa<< l’etica a ingrediente necessario dell’efficienza; una certa etica degli affari non ...esente da tale equivoco.*

3. Für die Umstandsrollen wird die Bezeichnung ‘Avv’ verwendet. Diese wird erweitert, um die genaue Rolle zu erfassen: ‘Avvmisura’ für Grad (cf. *>>abbassare<< al 5 per cento*), ‘Abbloc’ für Ort (*incrociandosi sopra il Polo Nord e >>abbassandosi<< successivamente verso il Polo Sud*), usw.

N-V-N1-(Avvmisura)

4. ‘Inf’ steht für das Verb im Infinitiv (wie in *>>abbassarsi<< a chiedere l’elemosina*)

*si* V-*a* Inf

Wie im letzten Beispiel bereits angedeutet, enthalten Valenzrahmen Präpositionen falls diese obligatorisch sind oder häufig auftreten; cf.:

N-*si* V-(*verso*Avvloc)

Das *si* vor dem Verb (wie in diesem Beispiel) signalisiert die Reflexivität des Verbs.

Die Namen der einzelnen Argumente können u.U. auch morphologische Restriktionen (wie ‘plural’ oder ‘singular’) widerspiegeln; cf.

Nplur-V

für *combinare***01** ‘zusammenpassen’: *Questi colori >>combinano<< bene.*

## 5.2 Die Mikrostruktur des Wörterbuches

Die wichtigsten Felder innerhalb eines Sublemmas enthalten folgende Information:

1. das italienische Lexem (mit der Nummer des Sublemmas),
2. das Valenzschema des italienischen Lexems,
3. die deutsche Übersetzung,
4. semantische Restriktionen für die Argumente,
5. Stil- und Feldrestriktionen des betreffenden Lexems,
6. Beispiele
7. Idiome und Kollokationen, die das Lexem enthalten.

Wie bereits oben in einer Fußnote erwähnt, ist in der Übersetzungzone das Stichwort im deutschen Index über das der betreffende Eintrag zu finden ist durch das Zeichen ‘\*’ markiert:

abbassare 09  
o N-si V-a N3  
o sich nicht zu \*gut sein für  
...

Das erste Sublemma für *abbassare* sieht wie folgt aus (in Klammern ist die jeweilige Nummer des Feldes angegeben).<sup>13</sup>

- o (1) *abbassare* 01
- o (2) N-V-N1
- o (3) \*herunterlassen
- o (4) N1: konkret
- o (5) —
- o (6) **abbassare** la tenda, il sipario (D)  
**Abbassò** il ponte levatoio. (DIR)  
Ancora il Pertini [...] fa fermare la vettura **abbassa** il finestrino e in dialetto savonese spiega [...] che il Parlamento ha fatto il suo dovere [...].  
\*öffnen  
**abbassare** le vele, le bandiere (D) die Segel \*streichen, die Fahnen \*einholen  
**abbassare** il capo, lo sguardo, gli occhi (D) \*senken  
**abbassare** un quadro (D) \*tiefer hängen  
**abbassare** un muro, una siepe (D) eine Mauer niedriger machen, eine Hecke \*zurückschneiden \*niedriger machen
- o (7) [*abbassare* le armi ‘die Waffen \*strecken’] [*abbassare* la cresta/la coda/le orecchie/le ali/le piume/le corna ‘\*demütig werden’] [*abbassare* i fari ‘\*abblenden’]

---

<sup>13</sup>Die Abkürzung in Klammern hier einem Beispiel gibt die Quelle des Beispiels im italienischen Korpus an.

Wie man anhand von *abbassare la cresta* lit. ‘das Kreuz senken’ und *abbassare i fari* lit. ‘die Lichter senken’ sieht, werden Kollokationen und Idiome nicht voneinander unterschieden—was eindeutig ein Problem hinsichtlich einer Automatisierung der Extraktion von Kollokationen aufwirft.

Der Eintrag für *abbassare* enthält folgende zwölf Sublemmata (wir geben hier nur die prominentesten Felder wider, d.h. das italienschie Original, sein Valenzschema, die deutsche Übersetzung und semantische Restriktionen für die Argumentstellen):

1. abbassare 01
2. N-V-N1
3. \*herunterlassen
4. N1: konkret
1. abbassare 02
2. N-V-N1
3. \*herunterholen
6. math., geom.
1. abbassare 03
2. N-V-N1-(Avvmisura)
3. \*senken
4. N1: abstr.
1. abbassare 04
2. N-V-N1-a N3
3. \*herabsetzen
6. fig.
1. abbassare 05
2. N-si V-(Avvmisura)-(sino a,a N3)
3. \*sinken
1. abbassare 06
2. N-si V
3. sich \*bücken
4. N: belebt
1. abbassare 07
2. N-si V-(versoAvvloc)
3. \*sinken
6. Bewegung im Raum
1. abbassare 08
2. N-si V-(N2)
3. \*sinken
1. abbassare 09
2. N-si V-a N3
3. sich nicht zu \*gut sein für

4. N: menschl.

1. abbassare 10
2. N-si V-a Inf
3. sich nicht zu \*gut sein für
5. N: menschl.

1. abbassare 11
2. N-V
3. \*fallen

1. abbassare 12
2. N-V-N1-(da Npred)-a Npred
3. \*degradieren zu

Das Lemma für *mettere* besitzt 59 Sublemmata, das von *andare* 77. Diese Zahlen spiegeln die Feinheit der im Wörterbuch angebotenen Valenzinformation wider.

### 5.2.1 Unregelmäßige Information

Einige Information wird im Valenzwörterbuch nicht regelmäßig erfaßt, taucht jedoch sporadisch immer wieder auf. So erscheinen als Sublemmata neben einzelnen Lesarten eines Verbs auch einzelne phraseologische Ausdrücke, wie im folgenden Beispiel *andarne* im Sinne von *auf dem Spiel stehen*:

1. andare:andarne
2. V-N1/di N3
3. auf dem \*Spiel stehen
7. **Ne va** la vita. (VLI) Das Leben steht auf dem Spiel.
- Ne va** della nostra vita. (Z) Unser Leben steht auf dem Spiel.
- Ne va** del nostro buon nome. (DD) Unser Ruf steht auf dem Spiel.

## 6 Überführung des ILR Wörterbuchs in das CAT2-Format

Im Rahmen des von uns durchgeführten Experiments stand die Erstellung von Transferlexika und des italienischen open-class Lexikons im Fordergrund. Die verwendete Programmiersprache war *awk*.

### 6.1 Erstellen der Transferlexika

Zu erstellen waren zwei Transferlexika: deutsch-italienisch und italienisch-deutsch. Bei der Erstellung des deutsch-italienischen Lexikons wurde wie folgt vorgegangen:

für jede Zeile im deutsch-italienischen Index:

falls nur ein italienisches Äquivalent vorhanden ist:

1. erstelle den Grundeintrag im Transferlexikon:  
<d.verb>={lex=<d.verb>,head={pref=nil}}. []=>{lex=<i.verb>}. [].
2. gleiche die 'pref'-Information des deutschen Äquivalents mit dem vorhandenen deutschen open-class Lexikon ab.
3. falls das deutsche Äquivalent im dt. Lexikon einen Reflexivmarker hat:  
kopiere diese Information: <d.verb>={lex=<d.verb>,head={pref=nil,refl={pform=??}}}. []
4. falls das italienische Äquivalent einen Reflexivmarker aufweist:  
erweitere die rechte Seite wie folgt:  
{lex=<i.verb>,head={refl=yes}}. [].

falls mehr als ein italienisches Äquivalent vorhanden ist:

führe (1) bis (4) für jedes Äquivalent aus.

Beispiele resultierender Einträge sind:

```
abbiegen={lex=biegen,head={pref=ab}. []=>{lex=declinare}. []  
abfallen={lex=fallen,head={pref=ab}. []=>{lex=declinare}. []  
aufpassen={lex=aufpassen,head={pref=nil}. []=>{lex=badare}. []
```

Für das italienisch-deutsche Lexikon wurde wie folgt vorgegangen:

für jedes Lemma im Valenzwörterbuch:

für jedes Sublemma:

für jedes deutsche Äquivalent des Sublemmas:

1. erstelle den Grundeintrag im Transferlexikon:  
<i.verb>={lex=<i.verb>}. []=>{lex=<i.verb>,head={pref=nil}}. [].
2. gleiche die 'pref'-Information des deutschen Äquivalents mit dem vorhandenen deutschen open-class Lexikon ab.
3. falls das deutsche Äquivalent im dt. Lexikon einen Reflexivmarker aufweist: kopiere diese Information:  
{lex=<d.verb>,head={pref=nil,refl={pform=??}}. []
4. falls das italienische Äquivalent einen Reflexivmarker aufweist:  
erweitere die linke Seite wie folgt:  
{lex=<i.verb>,head={refl=yes}}. [].

Beispiele resultierender Einträge sind:

```
combinare={lex=combinare}. []=>{lex=passen,head={pref=zusammen}}. []  
combinare={lex=combinare}. []=>{lex=passen,head={pref=zueinander}}. []  
combinare={lex=combinare}. []=>{lex=mischen,head={pref=nil}}. []  
combinare={lex=combinare}. []=>{lex=kombinieren,head={pref=nil}}. []  
combinare={lex=combinare}. []=>{lex=f{\u}gen,head={pref=zusammen}}. []  
...
```

## 6.2 Erstellen des italienischen Open-Class Lexikon

Während die italienischen Transferlexika im Vergleich zu den anderen vorhandenen Transferlexika in CAT2 noch relativ vollständig erstellt werden können, enthält das einsprachige italienische open-class Lexikon nur einen Teil der benötigten Information. Um es zu erstellen wurde wie folgt verfahren:

für jedes Lemma im Valenzwörterbuch  
für jedes Sublemma:

**Lex-Information.** Die Erstellung des 'lex-'-Teiles ist einfach:

1. generiere: 'lex=<verb>'

**Head-Information.** Die Erstellung des 'head'-Teiles sieht so aus:

2. generiere: 'head={cat=v}'
3. falls das Verb einen Reflexivmarker trägt:  
erweitere den Eintrag wie folgt: '{cat=v,refl=yes}'

**Extended Head-Information.** Im allgemeinen können aus dem Valenzwörterbuch keine Angaben hinsichtlich des 'extended head'-Teiles extrahiert werden. Deswegen wird dieser Teil nicht generiert.

**Frame-Information.** Die Erstellung des 'frame'-Teiles sieht so aus:

4. generiere 'frame={}'
5. für  $i := 1$  bis  $n$  ( $n$  = Anzahl der Argumente im Valenzrahmen)  
innerhalb von '{}':
  - 5.1. generiere: 'arg $i$ ={head={ehead={cat=nil}}}'
  - 5.2 ersetze das 'nil' bei 'cat' durch:  
falls das Argument ein Nomen ist: 'n'  
ein Verb ist: 'v'  
ein Adverb ist: 'ad'  
ein Adjektiv ist: 'a'
  - 5.3. falls die Argumentstelle eine Präposition besitzt:  
füge an: ',pform=?'
  - 5.4. falls die Argumentstelle Numerusrestriktionen besitzt:  
falls das Argument nur in Singular auftritt: füge an ',num=sing'  
anderenfalls: füge an ',num=plu'
  - 5.5. falls für das Argument eine semantische Beschränkung vorhanden ist:  
füge ein: ',sem={}'  
wenn die Beschränkung lautet:  
'menschl': füge innerhalb von 'sem' ein: 'anim={'T'=hum}'  
'abstr': füge innerhalb von 'sem' ein: 'abstract={abstr=\_}'  
...

Nach den Schritten (1) bis (5) erhalten wir somit Einträge wie im folgenden für *abbassare*<sup>9</sup> gezeigt:

```
{lex=combinare},
head={cat=v,refl=yes}
frame={arg1={head={ehead={cat=n,sem={anim={'T'=hum}}}}},
      arg2={head={ehead={cat=n,pform=a}}}}
```

### 6.3 Zusammenfassung: Ist-Stand

Das ideale Resultat unseres Experiments wäre es, die gesamte in CAT2 benötigte Information in den aus dem Valenzwörterbuch gewonnenen Lexika wiederzufinden. Die vorhergehenden Abschnitte haben jedoch deutlich gemacht, daß dies jedoch nicht möglich ist. Auch wenn die beiden erstellten Transferlexika in ihrer Vollständigkeit in etwa den in CAT2 genutzten französischen Lexika entsprechen, fehlt, wie bereits erwähnt, in dem open-class Lexikon einige wesentliche Information. Dies betrifft insbesondere:

- (i) die Funktionsverbgefüge: im Valenzwörterbuch sind sie nicht systematisch erfasst (auch wenn einige FVG im Kollokationsfeld aufgeführt werden);
- (ii) die semantische Klassifikation von Argumentstellen in dem 'frame'-Teil der Einträge: sie ist im Valenzwörterbuch nicht vorhanden.

Die Frage, die man sich nun stellen muß, ist auf welche Weise die fehlende Information gewonnen werden könnte. Die Information bzgl. der FVGs läßt sich nur bedingt aus externen Quellen (wie anderen on-line vorhanden italienischen Lexika) extrahieren. Somit bleibt nur die Option der manuellen Erweiterung. Die Information bzgl. der Semantik von Argumenten läßt sich beschränkt aus vorhandenen Quellen gewinnen. Wie, skizzieren wir kurz im folgenden Abschnitt.

Hinsichtlich des erstellten open-class Lexikons wäre nochmals zu erwähnen, daß zum Abgleich der Information über deutsche Verben das in CAT2 bereits vorhandene deutsche Lexikon benutzt wurde. Dies ist im Rahmen eines Experiments akzeptabel, führt jedoch zu folgenden zwei Einschränkungen:

- (i) eventuelle Modifikationen der deutschen Einträge (wie Behebung von entdeckten Fehlern) müssen in den italienischen Transferlexika getrennt nachvollzogen werden;
- (ii) die Einträge der im deutschen Lexikon nicht vorhandenen Verben haben in den italienischen Transferlexika keine zusätzliche Information ((z.B. über die Abtrennbarkeit des Präfixes) erhalten

## 7 Anreicherung des Open-Clas Lexikons

Wie bereits erwähnt, ist es ein großer Nachteil der aus dem Valenzwörterbuch gewonnenen Lexika, daß sie keine semantische Information hinsichtlich der Argumente enthalten. Am ILR wird zwar an der Verbsemantik des Italienischen gearbeitet, es handelt sich jedoch um Grundlagenforschung, die notwendig ist, um das Niveau der im Valenzwörterbuch erreichten Qualität beizubehalten. Was aber gegenwärtig gebraucht wird, ist eine vereinfachte Semantik, die die tatsächliche Nutzung des Valenzwörterbuches im Rahmen von CAT2 in vollem Umfang möglich macht. Eine mögliche Quelle der fehlenden Information ist das *Generalized Upper Model* (GUM) [Bateman *et al.* 94] (eine Präsentation des GUM aus der Sicht des italienischen Materials findet man in [Magnini 94]).

Im folgenden stellen wir zuerst kurz das GUM vor und diskutieren dann eine mögliche Intergration der darin enthaltenen semantischen Information in das italienische open-class Lexikon.

### 7.1 Generalized Upper Model

Das GUM  
ist eine multilinguale Weiterentwicklung des *Upper Model* [Bateman *et al.* 90], wie es am ISI in Los Angeles entwickelt wurde. Das UM enthält eine semantisch motivierte Klassifikation syntaktischer Strukturen der englischen Sprache. Im Rahmen dieser Klassifikation erhalten die Argumente einer verbalen Prädikation semantische Namen—entsprechend dem Typ der Prädikation. Die ursprüngliche semantische Klassifikation der Argumentstellen im ‘frame-Teil’ geht auf das UM zurück.

Das GUM reflektiert die Erkenntnis, daß einige Teile der semantischen Klassifikation für mehrere Sprachen gleich, während andere sprachspezifisch sind. Dementsprechend enthält das GUM sprachspezifische und allgemeingültige Fragmente. Die Sprachen, die bei der Entwicklung des GUM in Betracht gezogen wurden, sind Englisch, Deutsch, Niederländisch, Französisch und Italienisch. Das Schaubild 2 zeigt ein Fragment, das für alle betrachteten Sprachen verwendet wird sowie ein Fragment, das spezifisch für Italienisch ist.

Es ist offensichtlich, daß diese semantische Information gerade die in dem Valenzwörterbuch fehlende Information ist—auch wenn (i) es einige Abweichungen in der Benennung der Rollen gibt, und (ii) der beschränkte Umfang des bestehenden GUM dem Ausbau des open-class Lexikons Grenzen setzt. Im nächsten Abschnitt beschreiben wir, wie die Extraktion der Information vonstatten gehen könnte.

### 7.2 Extraktion semantischer Information für das Valenzwörterbuch

Der Algorithmus zur Extraktion der semantischen Information aus dem GUM lautet wie folgt:

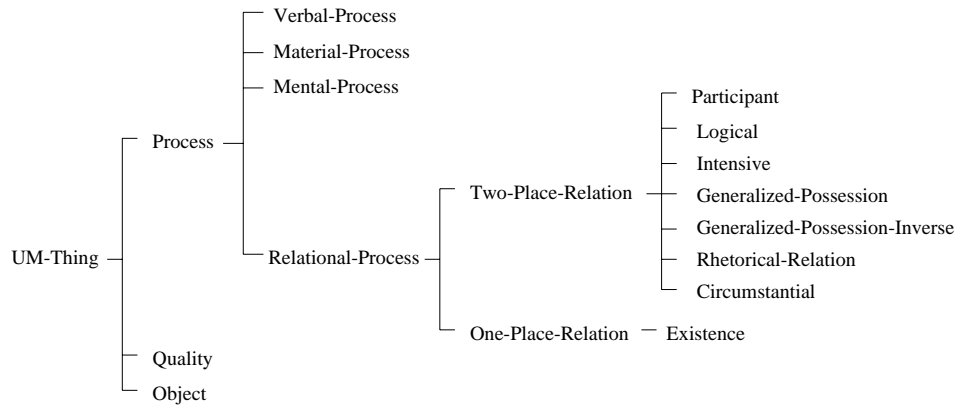


Figure 2: Ein Fragment des Generalized Upper Model.

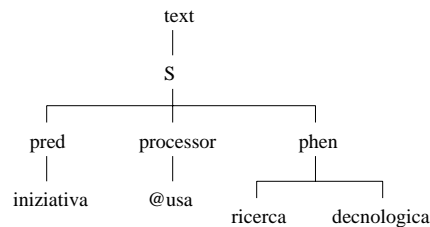
1. für alle Einträge im open-class Lexikon:
2. hole aus dem GUM den semantischen Typ des Verbs
3. für jedes Argument des Verbs:
  - hole aus dem GUM den Typ des Arguments
  - falls es das erste Argument ist:
    - füge in den frame-Teil ein: ‘,role=<typ>’
  - sonst
    - füge ein: ‘role=<typ>’

Wie bereits erwähnt, gibt es jedoch Abweichungen zwischen den Namen der Rollen im GUM und im CAT2-System. Diese Abweichungen müssen manuell korrigiert werden.

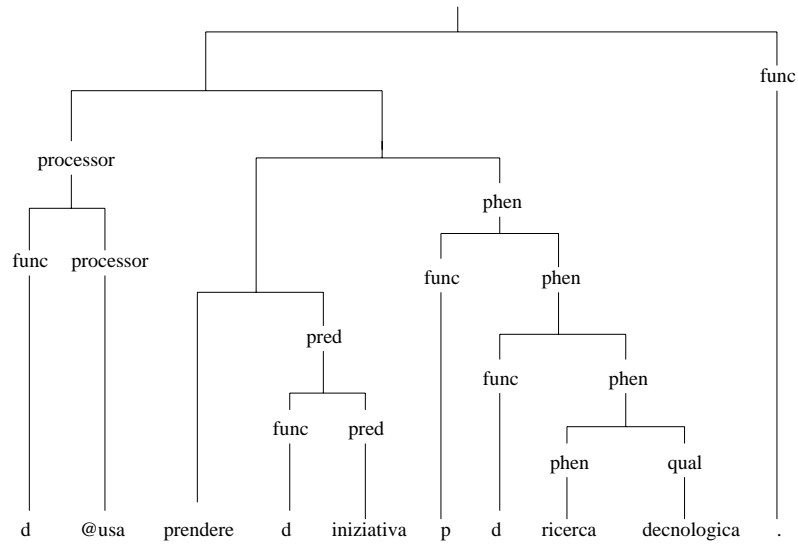
## Anhang: Beispiele italienischer Strukturen in CAT2

Um die Einsatzfähigkeit der erstellten Lexika zu demonstrieren, haben wir zum Abschluß unseres Experiments aus der vorhandenen französischen CAT2-Grammatik eine kleine Grammatik für das Italienische abgeleitet und damit ein paar Sätze aus dem Deutschen ins Italienische übersetzt. Einer davon war *Die USA ergreifen Initiative bei der Technologieforschung*—im Italienischen: *Gli USA prendono l’iniziativa nella ricerca tecnologica*. Anbei die IS- und die CS-Strukturen dieses Satzes.

IS-Struktur:



CS-Struktur:



## References

- [Apresjan *et al.* 89] Apresjan, Yu. *et al.* 1989. *Lingvističeskoje obespečenie sistemy ETAP-2*. Nauka: Moskau.
- [Arnold *et al.* 86] Arnold, D. *et al.*. The <C,A>,T Framework in EUROTRA: A Theoretically Committed Notation for MT. In Proceedings of COLING '86. 297–303.
- [Bateman *et al.* 90] Bateman, J.A. *et al.*. 1990. A General Organization of Knowledge for Natural Language Processing: The PENMAN Upper Model. Technical Report. USC/ISI.
- [Bateman *et al.* 94] Bateman, J.A. *et al.* 1995. The Generalized Upper Modl Knowledge Base: Organization and Use. In *Proceedings of the ECAI '94 Workshop on Implemented Ontologies*. Amsterdam.
- [Blumenthal und Rovere in Druck] P. Blumenthal und G. Rovere. In Druck. Semantisch-syntaktische Beschreibung italienischer Verben unter übersetzungswissenschaftlichen Gesichtspunkten (italienisch/deutsch).
- [Busse and Dubost 83] Busse, W. and J.-P. Dubost. 1983. *Französischers Verblexikon*. Klett: Stuttgart.
- [Chomsky 81] Chomsky, N. 1981. *Lectures on Government and Binding*. Foris Publications: Dordrecht.
- [Engel 88] Engel, U. 1988. *Deutsche Grammatik*. Julius Groos Verlag: Heidelberg.
- [Gazdar *et al.* 85] Gazdar, G. *et al.* 1985. *Generalized Phrase Structure Grammar*. Basil Blackwell: Oxford.
- [Grimshaw 91] Grimshaw, J. 1991. *Extended Projection*. Memo. Brandheis University.
- [Haller93] J. Haller. 1993. CAT2, Vom Forschungssystem zum präindustriellen Prototyp. In H.P. Pütz und J. Haller. (eds). *Sprachtechnologie: Methoden, Werkzeuge, Perspektiven*. Hildesheim. GLDV. 282–303.
- [Heyn 92] Heyn, M. 1992. *Zur Wiederverwendung maschinenlesbarer Wörterbücher*. Tübingen: Niemeyer.
- [Hovy and Knight 92] Hovy, E.H. and K. Knight. 1992.
- [Hutchins 86] Hutchins, W.J. 1986. *Machine translation: past, present, future*. Chichester: Ellis Horwood.
- [Hutchins and Somers 92] Hutchins, W.J. and H.L. Somers. 1992. *An Introduction to Machine Translation*. London, etc.: Academic Press.

- [Knight 93] Knight, K. 1993. Building a Large Ontology for Machine Translation. In *Proceedings of the ARPA Human Language Conference*.
- [Knight and Luk 94] Knight, K. and S. Luk. 1994. Building a Large Knowledge Base for Machine Translation. In *Proceedings of the Conference of AAAI*.
- [Magnini 94] Magnini, B. 1994. Specification of the Upper Model. GIST Project. IRST.
- [Mel'čuk 81] Mel'čuk, I. 1981. *Meaning-text Models: a Recent Trend in Soviet Linguistics*. In *Annual Review of Anthropology*. 10:27–62.
- [Nirenburg 87] Nirenburg, S. (ed.) 1987. Machine Translation. Cambridge: Cambridge University Press.
- [Nirenburg *et al.* 92] Nirenburg, S., J. Carbonell, M. Tomita, and K. Goodman. 1992. Machine Translation: A Knowledge-Based Approach. San Mateo: Morgan Kaufmann.
- [Nirenburg and Goodman 91] Nirenburg, S. and K. Goodman (eds.). 1991. *The KBMT Project: A case study in Knowledge-Based Machine Translation*. San Mateo: Morgan Kaufmann.
- [Pollard und Sag 87] Pollard, C. und I. Sag. 1987. *Information-based Syntax and Semantics: Volume 1*. Chicago University Press: Chicago.
- [Pollard und Sag 94] Pollard, C. und I. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago University Press: Chicago.
- [Sanfilippo and Poznański 92] Sanfilippo, A. and V. Poznański. 1992. The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Sources. In *Third Conference on Applied Natural Language Processing*. Trento. 80–87.
- [Schubert 88] Schubert, K. 1988. The Architecture of DLT—interlingual or double direct? In Maxwell *et al.* (eds.) *New Directions in Machine Translation*. Dordrecht: Foris. 131–144.
- [Schwanke 91] Schwanke, M. 1991. *Maschinelle Übersetzung*. Berlin, etc.: Springer Verlag.
- [Sharp91] R. Sharp. 1991. CAT2: An Experimental Eurotra Alternative. In *Machine Translation*, 6(3):215–228.
- [SharpStreiter92] R. Sharp und O. Streiter. 1992. Simplifying the Complexity of Machine Translation. *Meta*, 37(4):681–692.
- [Steiner 1993] Steiner, E.H. Skriptum zur Vorlesung “Einführung in die Sprachwissenschaft”. Universität des Saarlandes.

- [Streiter 96] Streiter, O. 1996. *Linguistic Modeling for Multilingual Machine Translation*. PhD Dissertation. Universität des Saarlandes: Saarbrücken.
- [van Slype and Pigott 79] van Slype, G. and I. Pigott. 1979. Description du système de traduction automatique. Szstran del Commission des Communautés Européennes. *Documentaliste*. 16:150–159.
- [Vendler 67] . Vendler, Z. 1967. *Linguistics in Philosophy*. Ithaca: Cornell University Press.
- [Viegas et al. 96] Viegas, E. et al. 1996. From *Submit* to *Submitted* via *Submission*: On Lexical Rules in Large Scale Lexicon Acquisition. In *Proceedings of the 34th Annual Meeting of the Association of Computational Linguistics*. 32–39.
- [Wilks et al. 89] Wilks, Y. et al.. 1989. A tractable machine dictionary as a resource for computational semantics. In B. Boguraev and T. Briscoe (eds.) *Computational Lexicography for Natural Language Processing*. Longman: London. 193–231.