

# LILA

## Linguistisch intelligente Lehrwerksanalyse

Prof. Johann Haller  
Fachrichtung 4.6. Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen und  
Institut für Angewandte Informationsforschung (IAI)  
Universität des Saarlandes  
Martin-Luther-Str. 14  
66111 Saarbrücken  
email: hans@iai.uni-sb.de

Johann Haller

geb. 1948 in München

Studium der Romanistik und Linguistik in Deutschland und Spanien

Computerlinguistische Forschung und Lehre

In der Industrie bei SIEMENS München, Systran

(Kalifornien/Luxemburg)

sowie an Universitäten in der BRD, Frankreich, Spanien und Lateinamerika;

Wissenschaftliche Leitung des Eurotra-D-Projekts

(Forschungsprojekt der EU-Kommission zur Maschinellen Übersetzung 1985-1992)

Seit 1990 Professor für Maschinelle Übersetzung an der Universität des Saarlandes und Direktor des Instituts für Angewandte Informationsforschung (IAI); zahlreiche Projekte für BMBF, BMWi, EU-Kommission und Industrie.

Forschungsschwerpunkte:

Multilinguale Sprachverarbeitung, Maschinelle Übersetzung, Technische Dokumentation, Kontrollierte Sprachen, Information Retrieval

### Abstract

In the era of information technology and new publication methods, it seems realistic to support the production of language learning material with software packages. This holds not only for the printing and publishing process but also for the ‚content‘ and learner interface part.

LiLa (Linguistic intelligent Learner's Book Analysis) is an attempt to use NLP (Natural Language Processing) Methods for progression control in vocabulary and grammar, as well as for a first feedback to learners when they perform exercises with relatively free (or unforeseeable) answers. A special version of spelling and grammar checking is envisaged which detects typical errors of foreign language learners, using a general dictionary and general grammar as a basis.

The first instance of these instruments will be realized within the **uni-deutsch.de** project which aims at a multimedia package for advanced German language learners at universities.

## **Zusammenfassung**

Im Zeitalter der Informationstechnologie und neuer Publikationsmethoden von Lehrbüchern für Fremdsprachen liegt es nahe, nicht nur die 'physikalische' Herstellung dieser Bücher mit dem Computer zu unterstützen, wie dies mit Publikationssoftware (QuarkExpress, PowerPublisher etc.) schon seit einigen Jahren geschieht. Auch die Erzeugung des Inhalts (Texte, Übungen, Vokabelverzeichnisse, zusätzliche Lernsoftware) kann sprachlich überprüft und unterstützt werden, wenn automatische Verfahren zur Textanalyse eingesetzt werden. Der folgende Beitrag stellt einige dieser Möglichkeiten vor, wie sie am Institut für Angewandte Informationswissenschaft an der Universität Saarbrücken prototypisch entwickelt wurden.

Bei einer für die Zukunft zu erwartenden weiteren Verlagerung von Lehr- und Lernmaterialien auf Datenträger und Internet werden solche Methoden entscheidende Beiträge zur Qualität und zum Bedienkomfort der Materialien leisten.

## **Einleitung: Elektronische Sprachverarbeitung in Saarbrücken**

Die automatische Analyse von Texten in natürlicher Sprache hat an der Universität des Saarlandes eine lange Tradition. Bereits 1967 bestand eine Arbeitsgruppe Elektronische Sprachforschung, als deren Gründer und Mentor besonders H. Eggers bekannt wurde. Er war auch der erste Sprecher des SFB100 'Elektronische Sprachforschung', an dem sich eine Reihe von philologischen Fächern und die Informatiker beteiligten. Die Arbeit dieser Gruppe begann mit dem Programmpaket SATAN (Saarbrücker Textanalyse, für Deutsch), dem dazu notwendigen SADAW (Deutsches Analyse-Wörterbuch) und dem ersten in Deutschland entwickelten Maschinellen Übersetzungssystem SUSY, das Material für eine Reihe von Sprachen (Deutsch, Englisch, Französisch, Russisch, Esperanto) erstellte. Das SADAW enthielt morphologische und syntaktische Informationen zu Wortstämmen, während für die Ermittlung übersetzungsrelevanter Lesarten eines Wortes ein eigenes Disambiguierungswörterbuch mit semantischen Angaben notwendig war (SEDAM). Der Sonderforschungsbereich wurde nach 1986 nicht mehr weitergeführt, aber eine Reihe seiner Mitglieder hatte sich sehr aktiv an der Vorbereitung und der Forschungsphase des europäischen Forschungsprojekts EUROTRA beteiligt und setzte die Arbeit dann in diesem Rahmen fort.

Im Jahre 1984/85 wurde das Institut für Angewandte Informationsforschung an der Universität des Saarlandes (IAI) gegründet, das den deutschen Anteil an EUROTRA als erstes Großprojekt übernahm. Am IAI werden auf der Basis der im SFB 100 und in EUROTRA entwickelten Grundlagen anwendungsbezogene Projekte durchgeführt, die zu einsatzfähigen Prototypen führen. Dazu gehören Übersetzungshilfen, Korrektur- und Prüfprogramme für Rechtschreibung, Grammatik und Terminologie, mehrsprachige Informationssysteme, z.B. für die Mediengesetze in Europa oder die Bewertung kommerzieller maschineller Übersetzungssysteme.

Theoretische Vorüberlegungen zu einem Einsatz automatischer Textanalysen in verschiedenen Phasen des Fremdsprachenunterrichts sowie bei der Erstellung und Kontrolle von Lehrwerken wurden in den letzten Jahren begonnen. Ausgangspunkt waren eventuelle Einsatzmöglichkeiten eines Grammatik-Prüfprogramms in Sprachlernprogrammen sowie Unterstützungsmöglichkeiten für Autoren von fremdsprachlichen Lehrwerken. Diese Überlegungen haben zu einigen prototypischen Softwaresystemen geführt, die heute praktisch demonstriert werden können.

## **Linguistische Analyse von Texten in Lehrwerken für Fremdsprachen**

Grundvoraussetzung für Anwendungswerkzeuge, wie sie oben angeführt sind, ist eine automatische Analyse der textlichen Elemente eines Lehrwerks (Texte, Übungen, Vokabelverzeichnisse etc.). Hierzu muss das Werk in einer maschinenlesbaren Form vorliegen, was bei modernen Publikationsmethoden eine Selbstverständlichkeit ist. Die Extraktion des Textmaterials, d.h. seine Trennung von Formatanweisungen, Bildern, Grafiken etc., ist zwar nicht immer trivial, kann aber je nach verwendetem Publikationsprogramm meist ohne größere Probleme durchgeführt werden.

In sehr einfacher Form (bestimmt für Texte, die ein Englisch-Lehrer zusätzlich zu einem Lehrwerk benutzen will) existiert ein solches Werkzeug in Form der 'Toolbox English G' (Cornelsen Verlag). Diese Toolbox enthält das Lexikon mit dem gesamten Wortschatz eines bestimmten Lehrwerks, wobei neben

den Grundformen von Verben auch (regelmäßige und unregelmäßige) Partizipien aufgeführt sind. Beim Vorliegen einer flektierten Form (-s als Endung der 3. Person Singular oder als Plural eines Substantivs) reicht es im Englischen ja aus, wenn man im Wörterbuch an die alphabetisch entsprechende Stelle geführt wird. Komplexe Verben (*turn on*) oder andere zusammengesetzte Ausdrücke sind nicht im Demo-Wörterbuch enthalten (das nur das Vokabular der ersten beiden Jahrgänge enthält). So kann der Lehrer beispielsweise rasch einen neuen Text darauf überprüfen, welche (Einzel-)Wörter in diesem Text mit welcher Bedeutung in welcher Lektion schon gelernt wurden. Wörter, die nicht zum Wortschatz English G gehören, werden als unbekannt markiert.

Eine Voraussetzung für komplexere Analysen bzw. für die Anwendung solcher Werkzeuge auf stärker flektierende Sprachen ist das Vorliegen relativ gut entwickelter und getesteter Analyseprogramme für Morphologie und Syntax sowie ausreichender Wörterbücher der entsprechenden Fremdsprache. Satzerkennungsprogramme zerlegen zunächst den Text in einzelne Sätze und Wörter; Analyse- und Lemmatisierungsprogramme erkennen flektierte Formen und Mehrwortlexeme und führen sie auf standardisierte Darstellungen zurück, wie sie in Lexika und Vokabelverzeichnissen zu finden sind (frz. *répétons* -> *répéter*, *sac* + *à* + *dos* -> *sac\_à\_dos*, etc.). Die Programme erkennen auch, welche Konjugationsmuster hier angewandt werden. Zur Reduzierung von Mehrdeutigkeiten durch den Kontext (z.B. *fait* Partizip vs. *fait* Nomen vs. *fait* 3. Pers. Sing.) ist auch eine (zumindest flache) syntaktische Analyse notwendig, die die Wörter des Satzes in sinnvolle Gruppen (Syntagmen) zusammenfasst. Das Beispiel im Anhang zeigt einen Ausschnitt aus der Bearbeitung eines französischen Lehrbuchtextes; die Prüf- und Suchprogramme sind jedoch sprachunabhängig.

Linguistische Analysen stehen für eine Reihe von Fremdsprachen zur Verfügung; mit abnehmender Abdeckung (d.h. Größe des Wörterbuchs) sind dies derzeit:

Deutsch, Englisch, Französisch, Spanisch, Schwedisch, Russisch, Niederländisch, Bulgarisch, Italienisch, Portugiesisch, Slowenisch, Esperanto, Latein.

Vorhandene Wörterbücher können jedoch leicht integriert werden; auch weitere, zumindest westeuropäische, Sprachen bedeuten keinen größeren Aufwand.

Die aktuellen Umfänge der elektronischen Wörterbücher sowie Beschreibungen anderer Anwendungen der linguistischen Analyse finden sich im Internet unter [www.iai.uni-sb.de](http://www.iai.uni-sb.de)

## **Werkzeuge für den Lehrbuchautor bzw. Redakteur**

Eine solche Computeranalyse des Textes bietet eine Reihe weiterer Möglichkeiten bei Erstellung und Einsatz fremdsprachlicher Lehrwerke. Als erstes sollen kurz die Möglichkeiten der Unterstützung für den Autor oder Redakteur eines Verlags dargestellt werden, die sich durch den Einsatz solcher Algorithmen ergeben.

### **Prüfprogramm für einzelne Lektionen**

Sind für die einzelnen Lektionen eines Lehrbuchs solche Analysen erstellt worden, so kann man eine Lektion gegen die "bekannten" Vokabeln und grammatischen Phänomene der vorausgehenden Lektionen (und auch vorausgehender Lehrwerksbände) prüfen lassen. Zu jedem Wort bzw. grammatischen Phänomen im Text kann das Programm bestimmen, ob dies schon gelernt wurde. Farbliche Markierungen können anzeigen, ob das Wort überhaupt nicht zur entsprechenden Sprache gehört (z.B. Schreibfehler, neue Personennamen, falsch gesetzte Übungsanweisungen - durch Vergleich mit einem allgemeinen Lexikon, s.u.), ob das Wort als solches neu ist, ob seine Form (z.B. unregelmäßiger Plural oder Futurbildung) neu ist, ob das Wort eventuell mehrdeutig interpretiert werden kann (Gefahr von Missverständnissen beim Lerner!) etc. Dies ist im Anhang an einem (französischen) Beispiel veranschaulicht.

### **Automatische Erstellung und Kontrolle von Vokabelverzeichnissen**

Das LiLa-System verfügt über relativ umfangreiche Wörterbücher der behandelten Sprache und vermag deshalb auch Wörter und Ausdrücke zu klassifizieren, die nicht im Lehrbuch enthalten sind. Für den Autor kann LiLa dadurch einen ersten Entwurf eines Vokabelverzeichnisses in der gewünschten Form erstellen; natürlich muss eine intellektuelle Nachkontrolle erfolgen, die aber mit sehr viel geringerem

Aufwand verbunden ist. Bei fertigen Lehrwerken kann eine systematische Kontrolle durchgeführt werden. Wenn ein (z.B. im selben Verlag vorhandenes) Wörterbuch als Basis oder Bezugspunkt genommen werden soll, kann auch hiermit ein automatischer Abgleich vorgenommen werden.

## Einsatz bei Wörterbüchern auf CD-ROM

Bei der Verwendung eines Wörterbuchs auf dem PC ist es oftmals nur möglich, Grundformen nachzuschlagen. Zwar bringen Suchsysteme (wie in der oben geschilderten Toolbox oder in der Duden CD) den Benutzer eines solchen Wörterbuches wenigstens an einen dem gesuchten Wort ähnlichen Eintrag (bei der Suche nach dem Verb *ahmen* immerhin zu *ahnen* - im Text steht *ahmen .... nach*), aber bei der Suche nach *gegangen* auf den Eintrag *gegen*.

Beim Einsatz einer linguistischen Analyse kann der Benutzer im Falle von *gegangen* auf *gehen* geführt werden, im Falle von *gehört* mindestens auf die beiden Einträge *hören* und *gehören*, und im Falle von *ahmen ... nach* auf den richtigen Eintrag *nachahmen*.

## Einsatz in Sprachlernsoftware-Paketen

### Nachschlagen von unbekanntem Wörtern in Texten und Übungen

Die im vorigen Absatz geschilderte Methode kann auch bei Erklärungen von Wörtern in Sprachlernsoftware-Paketen verwendet werden, bei denen dann nicht alle Flexionsformen von Hand mit der Grundform verbunden werden müssen; wenn die Zuordnung eindeutig erkennbar ist, kann auch in Zweifelsfällen die richtige Erklärung automatisch präsentiert werden.

Beispiel: *Das Auto gehört mir : gehören*

### Komplexe Analyse von Lernerantworten

In den meisten der heute verfügbaren 'Grammatiktrainer' und verwandter Software werden die Antworten der Lerner mit vorgegebenen Möglichkeiten verglichen oder sogar nur mit richtig/falsch gekennzeichnet. Aus diesem Grund können bei Lückentexten meist auch nur ein bis zwei Wörter eingesetzt werden; Rechtschreibfehler werden hier von vornherein als falsche Antworten interpretiert.

Die Integration eines Rechtschreibprüfprogramms, wie sie in einzelnen Trainingspaketen vorgenommen ist, bringt hier bereits wesentliche Verbesserungen.

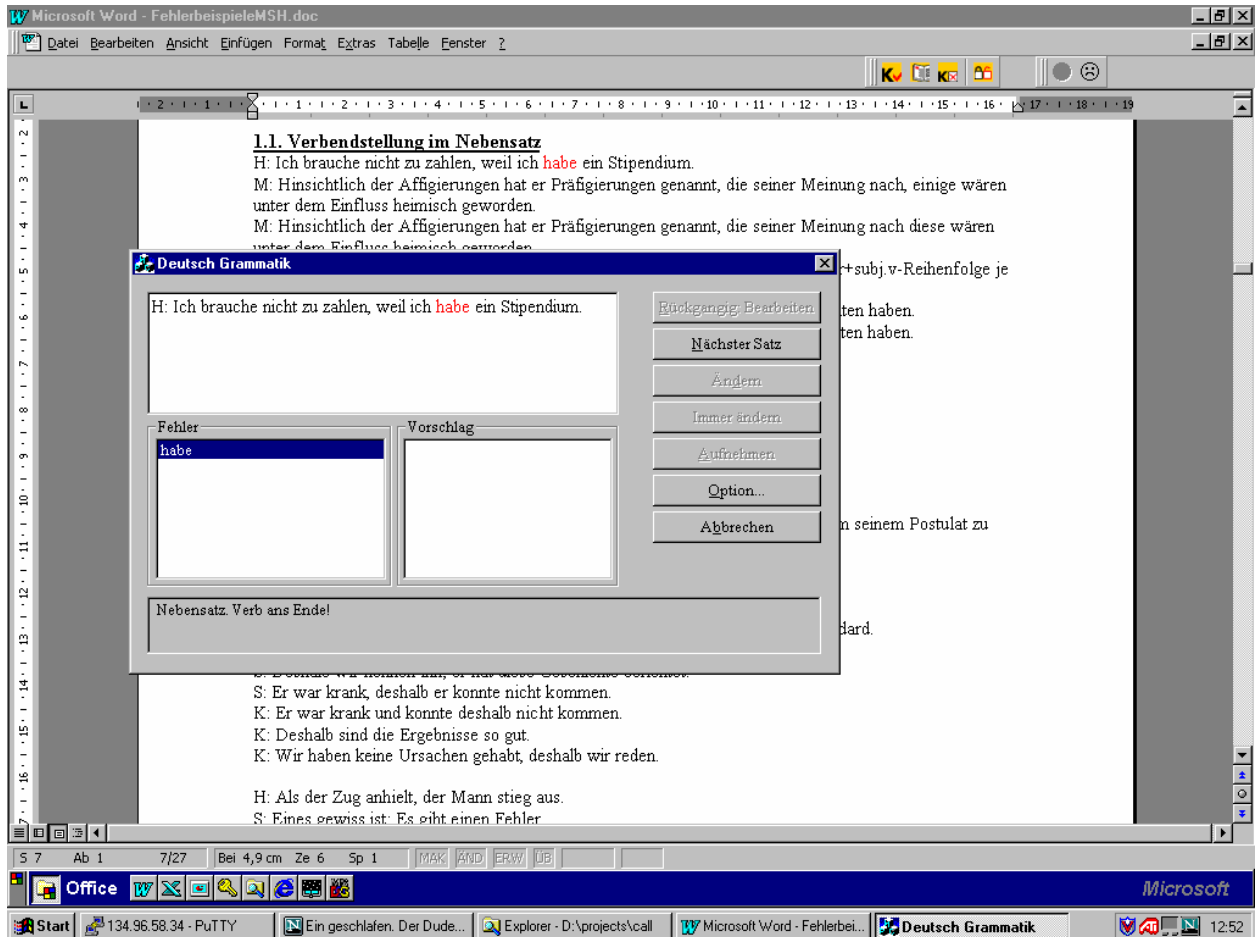
Das IAI hat einen Formalismus entwickelt, mit dem auf der Basis der oben geschilderten linguistischen Analyse individuelle syntaktische Patterns geprüft und mit individuellen Fehlermeldungen und Hinweisen auf Grammatikparagrafen und Übungen verbunden werden können. Das Prüfprogramm in LiLa erfüllt einerseits die Funktion eines Rechtschreibprüfprogramms auf Wortebene und überprüft andererseits grammatische Korrektheit auf Satzebene. Als fehlerhaft erkannte grammatische Strukturen werden mit einer Fehlerdiagnose versehen. Behandelte Fehlerklassen umfassen Kongruenz, Wortstellung, Kasusreaktion und Zeichensetzung, z.B.:

- *Das Programm suchen den Fehler.*  
"Das Programm": Subjekt und Verb müssen im Numerus übereinstimmen.
- *Sie findet den Fehler, nachdem ihr hilft das System.*  
"nachdem": Nach dieser Konjunktion muss das finite Verb am Satzende stehen.
- *Rettet dem Dativ!*  
"dem Dativ": Diese Nominalgruppe ist vom Verb nicht erlaubt.

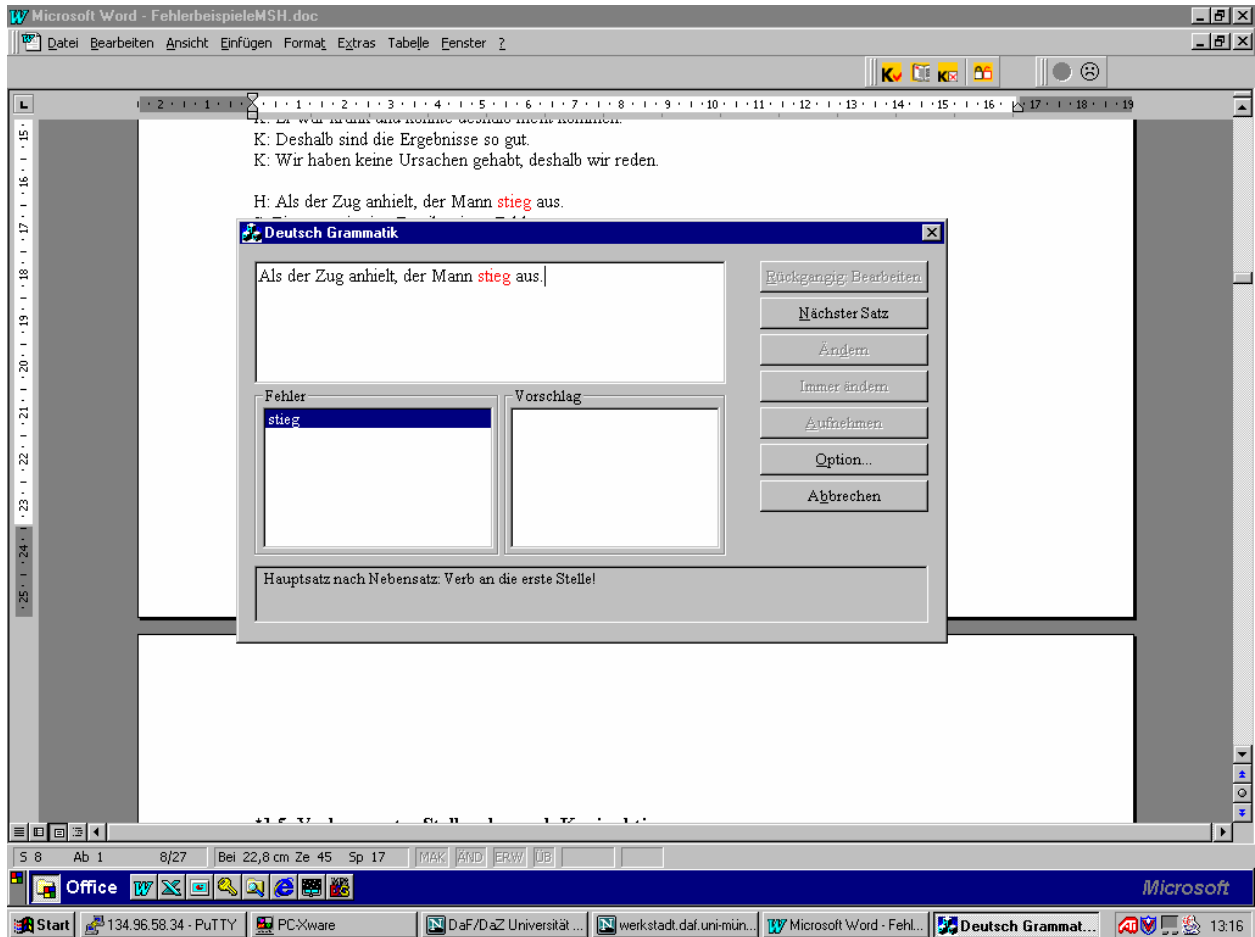
Integriert in ein CALL System kann LiLa zur automatischen Überprüfung von schriftlichen Übungen am Computer eingesetzt werden. Es ermöglicht eine wesentlich flexiblere und genauere Fehlerdiagnose als gängige Pattern-Matching-Techniken. Auch von Lernern relativ frei formulierte Texte können auf grammatische Korrektheit geprüft werden. Dies eröffnet die Möglichkeit, auch anspruchsvollere Übungsformen anzubieten, wie z.B. Fragen zum Text, Zusammenfassungen, Bildergeschichten oder Übersetzungen, da eine erste Fehlerdiagnose automatisch durchgeführt werden kann.

Einige einfache, aber syntaktisch interessante Beispiele aus der unten zitierten Fehlersammlung sollen dies veranschaulichen.

Der jeweils rot gekennzeichnete Fehler wird unten im Kommentarfeld erläutert; so wird im ersten Beispiel gegen die „Verbendstellung im Nebensatz“ verstoßen:



Auch Verstöße gegen die Verbstellung in einem Hauptsatz können erkannt werden, dem ein Nebensatz vorausgeht:



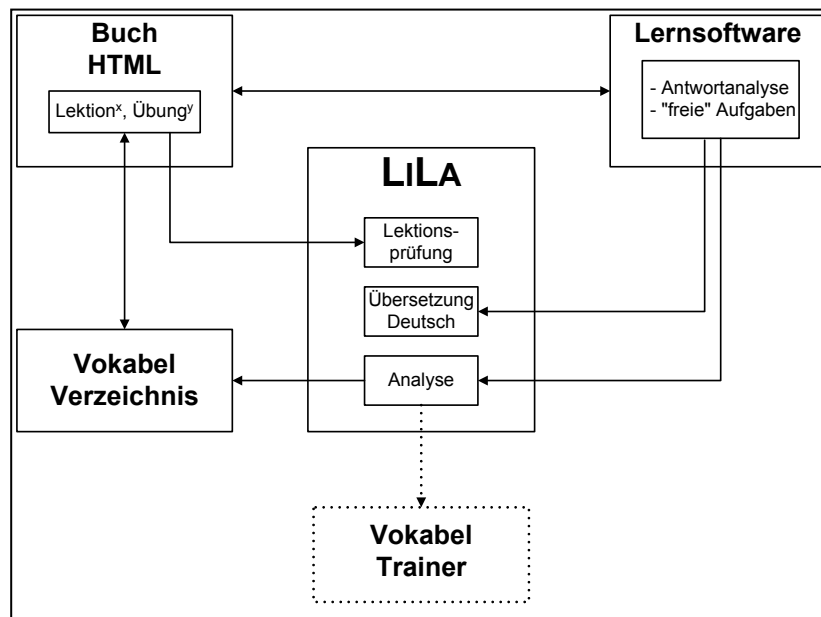
Natürlich lassen sich auch Links zu einer Hypertext-Grammatik oder Verweise auf entsprechende Kapitel in einem konventionellen Lehrwerk bzw. auf Übungssequenzen. Die von LiLa erstellten Fehlerdiagnosen können an die Bedürfnisse (z.B. Muttersprache, Kenntnisstand) der Lerner angepasst werden. Eine optimale Anpassung ist zu erreichen, indem über die technische Integration hinaus Grammatik und Lexikon von LiLa gezielt auf ein bestimmtes computergestütztes Lehrwerk zugeschnitten werden. Bei einer Verbindung mit den in den ersten Kapiteln beschriebenen Methoden kann der Prozess der Anpassung automatisch unterstützt werden.

Eine Weiterentwicklung und erste Erprobung dieses Konzepts finden innerhalb des Projekts **uni-deutsch.de** (<http://www.iai.uni-sb.de/pilot/pilot.html>) statt, das unter der Federführung der „Werkstadt“ Deutsch als Fremdsprache (<http://werkstadt.daf.uni-muenchen.de/home2.htm>) der Universität München ein Lehrwerk zum Online-Unterricht in der Wissenschaftssprache Deutsch zur Studienvorbereitung und Forschungsbetreuung ausländischer Studierender und Wissenschaftler an deutschen Hochschulen entwickelt.

Als erste Schritte wurden neben dem Abgleich der Vokabeln mit der Liste des Zertifikats „Deutsch als Fremdsprache“ die Erzeugung von Lexika in bestimmten Formaten sowie die systematische Analyse der bei Heringer gesammelten und klassifizierten typischen Fehler solcher Deutschlerner (<http://www.philhist.uni-augsburg.de/Faecher/GERMANIS/daf/forschung/fehler/index.html>) in Angriff genommen.

## Zukunftsvision: Das integrierte Lehrwerk auf CD-ROM und/oder Internet

Wie in diesem Projekt schon versuchsweise angedacht, könnte sich in der Zukunft ein Szenario ergeben, wo das gesamte Lehr- und Lernmaterial nicht mehr auf Papier existiert, sondern wo alle Komponenten nur in elektronischer Form erstellt und über CD und/oder Internet vertrieben werden: Lehrbuch, Arbeitsbuch, Grammatik, Lernsoftware, Wörterbuch. Eine solche Konstellation wird dann optimal genutzt werden können, wenn sie maximal miteinander vernetzt ist (wie in der folgenden Abbildung dargestellt), und wenn eine intelligente Verbindung von einer Lernerproduktion (Einzelwort, Wortliste beim Vokabeltest, Satz bei der grammatischen Übung, Übersetzung, freie Nacherzählung, Rollenspiel, Aufsatz) zu dem vom Lehrbuch- und Grammatikautor geschaffenen Material besteht. Eine solche intelligente Verbindung ist die linguistische Analyse. Sie wird auch für die Schaffung und Erhaltung der Vokabel- und Grammatikkonsistenz innerhalb eines solchen 'Hyperlehrwerks' unabdingbar sein.



## Anhang: Lektion 5 eines Französisch-Lehrwerks.

### Ausschnitt aus Lektionstext

**Blau:** In Lektion 1-4 noch nicht eingeführte Einzelwörter

**Rosa:** In Lektion 4 noch nicht eingeführte feste Mehrwortbegriffe

**Grün:** In Lektion 1-4 noch nicht eingeführte variable Mehrwortbegriffe

**Rot:** Von der linguistischen Analyse nicht erkannte Wörter (Schreibfehler oder durch die Verletzung der SGML in den Textbereich geratenen Übungsanweisungen, nicht im Lexikon enthaltene Eigennamen, Phantasiebegriffe)

Die unterstrichenen Einheiten sind mit Links zu ihren linguistischen Analyseergebnissen versehen (s.u.).

Luc et Nathalie

Il est sept heures.

Nathalie est dans la salle de séjour.

M. et Mme Noblet sont dans la cuisine.

Elle regarde la télévision.

Ils font le repas.

Sept heures et quart: Nathalie entre dans la cuisine. Elle a soif.

Mme Noblet donne un verre à Nathalie.

Mme Noblet : Nathalie, tu regardes trop la télé.

Nathalie : Moi, je ne trouve pas.

Mme Noblet : Tu aimes la télé, d'accord. Mais tu as des devoirs, non?

Nathalie . Oui, j'ai des devoirs, comme toujours!

Mme Noblet : Et tu ne travailles pas?

Nathalie . Si, mais pas maintenant. Il y a les "Bidibules" à la télé.

Elle crie : "Et puis, Luc n'est pas là. Il est avec des copains. Monsieur ne travaille pas. Alors, moi, je regarde la télé."

**Lesen Sie bitte den Text laut!**

### **Ausschnitt aus der linguistischen Analyse eines französischen Textes**

Die Merkmale der ersten Zeile enthalten neben der Original-Textform (ori) Verwaltungsinformationen, die auf die Stellung des Wortes im Satz sowie seine Gross- oder Kleinschreibung hinweisen; die Merkmale der zweiten Zeile enthalten die eigentlich 'linguistischen' Informationen sowie das Merkmal 'new=yes', wenn sie nicht in der Bezugslektion (und den vorhergehenden Lektionen) enthalten waren.

Für die Mehrwortausdrücke wird eine deutsche Übersetzung angegeben; Mehrwortausdrücke sind immer sprachpaarspezifisch zu analysieren, da nur im Kontrast zu entscheiden ist, welche Wortkombinationen als Ausdruck zu lernen sind. Wenn die Zielsprache über dieselbe Kombinatorik verfügt, muss das Wortpaar nicht im Vokabular aufgenommen und auch nicht eigens gelernt werden.

```
{ori=Luc,wnra=2,snr=2,gra=cap,pctr=yes,pctl=no,last=no,lu=Luc,s=fname,c=noun,ehead={nb=sg,g=m},wnrr=1,new=yes}
```

Textform: Luc

Status: Einzelwort

Grundform: Luc

Wortklasse: Substantiv

Semantik: Vornamen

---

```
{ori=Nathalie,wnra=4,snr=2,gra=cap,pctr=no,pctl=yes,last=yes,lu=Nathalie,s=fname,c=noun,ehead={nb=sg,g=f},wnrr=3,new=yes}
```

Textform: Nathalie

Status: Einzelwort

Grundform: Nathalie

Wortklasse: Substantiv

Semantik: Vornamen

---

```
{ori=M.et_Mme,wnra=21,snr=8,gra=cap,pctr=no,pctl=no,last=no,lu=M._et_Mme,c=noun,s=title,ehead={nb=plu,g=m},wnrr=1,new=yes}
```

Textform: M.et Mme

Status: Fester Mehrwortausdruck

Grundform: M. et Mme

Wortklasse: Substantiv

Semantik: Titel

---

```
{wnra=30,snr=10,gra=small,pctr=no,pctl=no,last=no,tns=pres,mode=ind;subj,per=3;1,nb=sg,c=verb,vtyp=fiv,ori=regarde_la_télévision,lu=regarder_la_télévision,de=fern_$sehen,source=mwu,wnrr=2,new=yes}
```

Textform: regarde la télévision

Status: Variabler Mehrwortausdruck

Grundform: regarder la télévision

Übersetzung: fern\_sehen  
Wortklasse: Verb

---

```
{wnra=53,snr=16,gra=small,pctr=no,pctl=no,last=no,  
c=w,sc=verb,vtyp=fiv,tns=pres,mode=ind,per=3,nb=sg,ori=a_soif,  
lu=avoir_soif,de=Durst_haben,source=mwu,wnrr=2,new=yes}
```

Textform: a soif

Status: Variabler Mehrwortausdruck

Grundform: avoir soif

Übersetzung: Durst\_haben

Wortklasse: Verb

---

```
{ori=crie,wnra=147,snr=40,gra=small,pctr=yes,pctl=no,last=no,tns=pres,  
mode=ind;subj,per=3;1,nb=sg,lu=crier,c=verb,vtyp=fiv,wnrr=2,new=yes}
```

Textform: crie

Status: Einzelwort

Grundform: crier

Wortklasse: Verb

---

```
{ori=puis,wnra=151,snr=41,gra=small,pctr=yes,pctl=no,last=no,  
lu=puis,c=adv,s=temp,wnrr=3,new=yes}
```

Textform: puis

Status: Einzelwort

Grundform: puis

Wortklasse: Adverb

Semantik: temp