Learner Corpora: Design, Development and Applications

Development of NLP tools for CALL based on learner corpora (German as a foreign language)

Sandrine Garnier, Youhanizou Tall, Sisay Fissaha, Johann Haller

Institut für angewandte Informationsforschung
Institute of Applied Information Sciences (IAI)

Martin-Luther-Str. 14, D-66111 Saarbrücken
E-Mail: info@iai.uni-sb.de, http://www.iai.uni-sb.de
Telefon: +49 (0)681-38951-0 Fax: +49 (0)681-38951-40

Introduction

The project unideutsch.de is carried out by the DAF Institute in Munich (German as foreign language) in collaboration with IAI which is responsible for the parts described in this paper.

IAI was founded as a non-profit-making Research & Development (R&D) organisation in 1985 in Saarbrücken, Germany, to carry out the European EUROTRA machine translation project for the German language. Meanwhile IAI is an internationally acknowledged R&D institute in the field of multilingual information processing covering all aspects of natural language processing, computer-aided translation, and information management and knowledge management in advanced information technology environments. This is mainly done through strategic alliances with industry and national and European authorities.

At present, the institute's major R&D activity is in the area of multilingual language technology, in particular machine translation technology with special emphasis on domain-specific terminology and concept-based methodologies, multilingual information retrieval and information filtering, as well as general knowledge management.

1 Project uni-deutsch.de: description

Currently IAI is working on learning projects such as uni-deutsch.de, which combines our experience of NLP and the results of the research in foreign languages learning. The use of the technical advanced NLP tools provides another means of learning.

The main objective of the project is to devise an autonomous, long-distance, language learning system for advanced learners. The project is aimed further at:

- improving the language performance of advanced learners.
- boosting the learning process by giving vocabulary in certain specified domains such as science or education.
- checking automatically a text produced by a foreign language learner allowing them to learn by an interactive process which combines a human corrector and a machine corrector.
  The machine corrector has the purpose of correcting spelling and grammar errors whereas the human corrector has to answer learner´s questions concerning errors found by the machine and to comment on other semantic errors.
  The combination of these both methods allows for accelerated language acquisition through error based learning.

2 Computer based learning of vocabulary

In order to achieve the above mentioned aims, IAI has developed a programme, LiLa, - *Linguistisch intelligente Lehrwerksanalyse* – the task of which is to analyse textbooks with linguistic intelligence (texts, exercices and vocabulary listing).[1]

The programme is a combination of morphosyntactic taggers and parsers based on partial parsing techniques (Constraint Grammar) with developed linguistic resources in German. The different programmes break down the text into sentences and words.

This programme allows students to have a list of new vocabulary which is the result of comparing the existing vocabulary of a lexicon containing a list of basic words (level B1: *Zertifikat Deutsch*) with the vocabulary in a text.

The new entries are produced with grammatical information. LiLa indicates new nouns together with their gender information and plural form, and for verbs infinitive, present, imperfect and past participle. Other words such as conjunctions are associated with their grammatical categories. All this information is essential for learners in order to improve their use of vocabulary. Consequently the student will then see from the list what they have still got to learn and what they should already know. LiLa provides a second lexicon which is the addition of the first lexicon to the other new words. This second lexicon can be then compared with another text.

Example of an automatically produced vocabulary listing:

abkühlen, kühlt ab, kühlte ab, abgekühlt
Abkühlung, die, Abkühlungen
absinken, sinkt ab, sank ab, abgesunken
adiabatisch (Adjektiv)
aerob (Adjektiv)
aggressiv (Adjektiv)
Alge, die, Algen
altern, altert, alterte, gealtert
Ammoniak, das
andauernd (Adjektiv)
anhand (Präposition)
anpassen, passt an, passte an, angepasst
Anreicherung, die, Anreicherungen
Argon, das

The listing can also be shown as one text where new words appear in color. Other words can be tagged, e.g. a spelling error, a proper noun etc.

3 Terminology

Another programme which isn´t part of the project uni-deutsch.de, but which could also be useful for learning vocabulary is the programme AUTOTERM. This programme, based on statistics, automatically gives the terminology from specialised texts.

This could be useful when students are working in specialized fields in English, German and French.

[1] **http://www.spz.tu-darmstadt.de/projekt_ejournal/jg_07_1/beitrag/haller1.htm**

Example of a text from the online newspaper Wissenschaft-aktuell.de (http://www.wissenschaft-aktuell.de/):

Eintrittspforten in der Hülle von Nervenzellen nachgewiesen [Neurowissenschaften]

Durham (USA) - Bisher glaubte man, dass die äußere Zellmembran gleichförmig aufgebaut ist, so dass an jeder Stelle der Zelloberfläche Moleküle in eine Zelle eindringen können. Jetzt haben Wissenschaftler der Duke University bei Nervenzellen ganz definierte Eintrittsstellen nachgewiesen. Die Regulation des Stofftransports in diesen Membranzonen ist für die Funktion der Nervenzelle wichtig. Weitere Untersuchungen der neu entdeckten Strukturen könnten daher helfen, Wirkstoffe gegen verschiedene Nervenerkrankungen zu entwickeln, schreiben die Forscher im Fachblatt "Neuron".

Nervenzellen geben über so genannte Synapsen Signale in Form von Botenstoffen an Nachbarzellen weiter. Diese Neurotransmitter binden an Rezeptorproteine an der Außenseite der Empfängerzelle. Die Rezeptoren werden ständig durch neue ersetzt, also in beiden Richtungen durch die Membran transportiert. Indem die Zelle diesen Austausch reguliert, kontrolliert sie die Zahl der Rezeptoren und damit die Empfindlichkeit, mit der sie auf die Neurotransmitter reagiert. Wirkstoffe, die diesen Transport hemmen oder beschleunigen, könnten als Medikamente zur Behandlung von Depressionen, Epilepsie oder Suchterkrankungen eingesetzt werden.

Die Arbeitsgruppe von Michael Ehlers wollte verfolgen, wie die außen sitzenden Rezeptormoleküle wieder in die Zelle gelangen. Dazu markierten die Wissenschaftler das Protein Clathrin mit einem Fluoreszenzfarbstoff und gaben es zu Kulturen von Nervenzellen. Clathrin lagert sich an die Stellen der Membran, an denen sie sich einstülpt, um Proteine nach innen zu transportieren. Die mikroskopische Auswertung ergab, dass sich das Clathrin nicht gleichmäßig verteilt sondern nur in bestimmten Zonen auf der Membran ablagerte und damit definierte Eintrittsstellen markierte. "Wir haben herausgefunden, dass die Nervenzelle mit einem Zimmer vergleichbar ist, in das nur ganz bestimmte Türen führen", sagt Ehlers. Bisher habe man die Zellmembran eher für einen an vielen Stellen durchlässigen Vorhang gehalten.

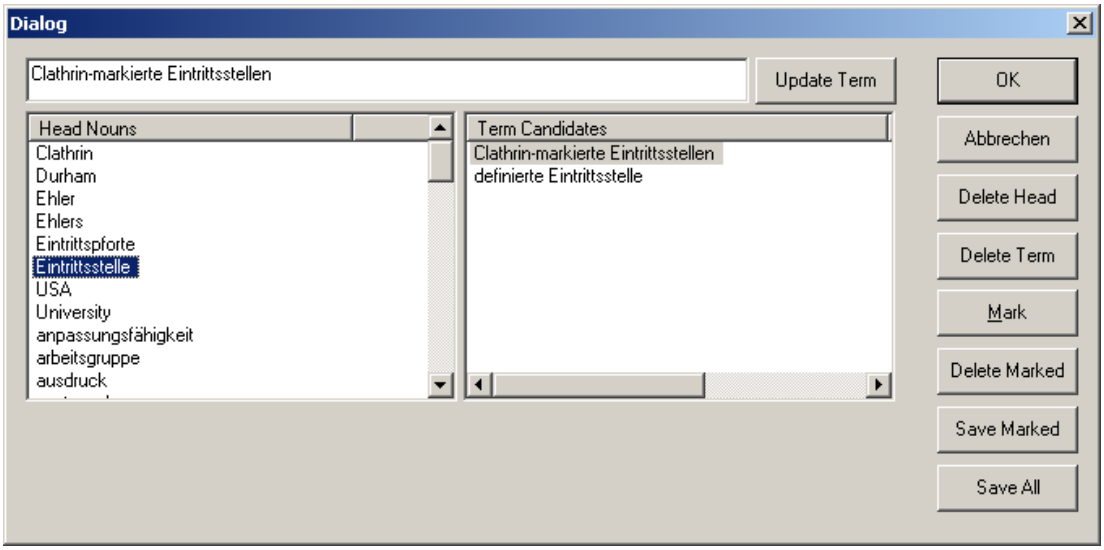
Während sich die Verteilung der Clathrin-markierten Eintrittsstellen bei jungen Nervenzellen häufig veränderte, blieb sie in reifen Neuronen weitgehend konstant. Möglicherweise sei das ein Ausdruck für die im Alter nachlassende Anpassungsfähigkeit des Gehirns, vermutet Ehlers. Die jetzt nachgewiesene Membranzone könnte nur die erste von weiteren noch unbekannten speziellen Membranstrukturen sein, glauben die Forscher.

Author: Joachim Czichos
Source: EurekAlert

Extract of an automatically produced terminology listing:

Anpassungsfähigkeit
Anpassungsfähigkeit des Gehirns
Ausdruck
Austausch
Außenseite
Außenseite der Empfängerzelle
Behandlung
Behandlung von Depressionen
Behandlung von Epilepsie
Botenstoff
Clathrin
Clathrin-markierte Eintrittsstellen
bestimmte Tür
bestimmte Zone
definierte Eintrittsstelle

Example of the dialoge interface showing the nouns and noun groups found in the text:



4 Error-based self learning

According to some authors (Prof. H. J. Heringer) students make the same or similar errors which are repeated. If the student knows what type of errors he constantly makes, he could then accelerate the process of language acquisition. This idea is the basis of the development of specialized programmes for foreign language learners.

Several programmes are connected together in order to give the structure of a German sentence. The different information which is given by the programmes is finally processed by an additional programme, KURD, which has the task of finding syntactically or morphologically incorrect structures in the German sentence. KURD is "a formalism for shallow post morphological processing" and uses "the input from the morphological analyser MPRO".[2] It works on grammar checking and style control (German & English).

This programme uses several operators such as unification and deletion to add and delete information in the linguistic representation of the sentence in order to mark one or several errors in the sentence. Formalism rules have been developed to add information in the analysis of the most common structural mistakes made by learners of German. The new information is then processed by another programme which tags the errors found in the text.

[2] **http://www.iai.uni-sb.de/iaide/en/kurd.htm**

Example: The position of the verb in a subordinate clause in German, where the verb must be at the end of the sentence. The error message is "Subordinate clause: verb at the end".

4.1 Error messages

This objective aims at developing the necessary tools for the system to evaluate the task done by the learners and give them feedback on their performance. Qualitative methods (morphosyntactic tagger and syntactic parser) allows a linguistically based analysis of the texts produced by learners.

Error messages can be more precise due to linguistic analysis. This method replaces the straightforward right/wrong answer or the comparison of pattern structures and instead provides a detailed message to help the students correct themselves. The type of error message can be adapted to the linguistic knowledge of the user.

Example 1: the genitive form in German.

> Sentence: “Der Stil des schweizerischen Autors des langen Textes ist langweilig.“
> The error message is the following: “too many genitive forms.”

Example 2: wrong participle.

> Sentence: “Ich habe das gebrach.”
> The error message is the following: “Wrong participle”.

4.2 Work basis: learner corpora

In order to develop rules adapted to foreign learners we had to collect errors made by speakers of other languages. The sentences collected were written by Socrates DAF students and students from Burkina Faso, Vietnam and Russia.

We have also used material from Professor Heringer. He has listed different incorrect types of sentences in the book *Fehlerlexikon* and on his web-site.[3] From these sources we have classified different error types in order to help us formalize them.

A lot of errors made can be automatically found:

- spelling errors
- grammatical errors
- some valence errors.

What we can´t find at the moment in most cases are the valence. Work on this is currently in progress and we hope to resolve it during the coming year.

Classification: structure and some examples
H stands for examples from Prof. Heringer, I for IAI, S for Socrates students and K for correct structures. The sign + means that IAI is able to detect the wrong structure and gives a corresponding error message, whereas the sign – means that we are not able at the moment to detect the error.

2.4. Wortbildung

2.4.1. Partizip von starken Verben

+H: Ich habe Münzen zu den Leuten *geworft*.
+H: Danach hat er ein Rezept *geschreiben*.
+I: Ich habe das *gebrach*.

2.4.2. Partizip von schwachen Verben

+H: Am Sonntag bin ich um 10 Uhr *aufgewach*.
+H: Sofort habe ich eine Wohnung *gemiet*.
-V: Die Stadt wurde *überschwommen*.

2.4.3. Kein ge- bei nicht trennbaren Präfixverben und Verben auf -ieren

+H: Er hat ein Zimmer für seinen Vater *gebestellt*.
+H: Er hat dort Medizin *gestudiert*.

2.4.4. Falsche Konjugation

+S: Man verglicht.
+U: Mit Sicherheit *geb* es auch künftig große Kraftfahrzeuge, aber die Besitzer von kleinen Autos würden lächeln, weil Sie *durchkommen* viel schneller durch *den* Straßenchaos.

2.4.5. Komposition

+K: Beschwerdeführer

+I: Beschwerdenführer
-I: Beispielssätze
-U: Statt eines 20-Minuten-Taktes sei *in Basel* ein 6-Minute-Takt eingeführt werden.

The result of the analysis can be seen in unix as well as in winword. The userfriendly interface shows the tagged wrong structure and gives a message which helps learners to correct themselves. The programme can also give a correct example or several options of correction.

You can find the site at: http://www.uni-deutsch.de.
The internet address of the DAF Institute is the following: http://werkstadt.daf.uni-muenchen.de/home2.htm.

[3] **http://www.philhist.uni-augsburg.de/Faecher/GERMANIS/daf/Forschung/fehler/analyse.html**

Conclusion

The NLP tools will be a powerful additional means for the acquisition of a foreign language. Computer-assisted-language-learning seems to be a good solution today to the individual e-learning of languages and helps the human corrector, who can then be more aware of semantic and style problems. The CALL must be seen as another option and be added to the other traditional means of acquiring a foreign language.
What we have described in this paper, is just a beginning. Special grammar rules for specific grammar or vocabulary exercises, integration and comparison with predefined answers into grammar chapters and dictionary could be the next task in order to customize the learning process.

Books

Garnier S, Tall Youhanizou, 2003 *Uni-deutsch.de Report*, forthcoming, IAI

Heringer H. J. 2000 *Fehlerlexikon Deutsch als Fremdsprache*, Berlin, Cornelsen Verlag.

Internet pages

Haller J 2002 *Linguistisch intelligente Lehrwerksanalyse*. In *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, 7(1), 6 pp.
http://www.spz.tu-darmstadt.de/projekt_ejournal/jg_07_1/beitrag/haller1.htm (stand 02.13.2003)

Michael C 2001 *KURD*, 10 pp .
http://www.iai.uni-sb.de/iaide/en/kurd.htm (stand 02.13.2003)

Heringer Hans Jürgen, 1999, *Aus Fehlern lernen* (Multimediales Computerprogramm zur Fehleranalyse, mit Grammatikteil), Augsburg
http://www.philhist.uni-augsburg.de/Faecher/GERMANIS/daf/Forschung/fehler/analyse.html
(stand 02.13.2003)