

# 900.000 Sprachkombinationen

Seit längerer Zeit schon tüfteln EURAC-Forscher an der Entwicklung des Sprachlernbrowsers Gymn@zilla: Per Mausanzeige übersetzt er unbekannte Wörter herkömmlicher Internetseiten, erstellt Wortlisten und automatische Übungen.

Zuzana H. ist Deutschlehrerin in Brno, Tschechien. Englisch hat sie erst nach ihrem Studium 1993 gelernt, um auch an internationalen Projekten teilnehmen zu können. Da Zuzanas' Zeit eingeschränkt ist, vertieft sie ihre Englischkenntnisse abends im Internet. Helmut M. ist ein Ingenieur aus Südtirol. Im Studium hat er zwei Semester Russisch belegt und Delegationen nach Russland begleitet. Seine sprachlichen und landeskundlichen Kenntnisse frischt er auch heute noch gerne im Internet auf. Chiara P. studiert Sprachen in Bologna und bereitet sich derzeit auf ihren Studienaufenthalt in Klagenfurt vor. Im Internet liest sie regelmäßig deutschsprachige Zeitungen, um ihren Wortschatz zu erweitern.

verbindet Webseiten in verschiedenen Sprachen - zum Beispiel englische, russische und deutsche Nachrichten - mit elektronischen Wörterbüchern und automatisch generierten Sprachübungen. Eine Benutzerin wie Zuzana H. kann mit Gymn@zilla die englischsprachigen Nachrichten der BBC lesen. Die Übersetzung der Wörter ins Deutsche erscheint beim Lesen, sobald sie die Computermaus über das Wort bewegt (siehe Abb.). „Die automatische Lesehilfe ist viel schneller und effizienter als gedruckte Wörterbücher“, erklärt Zuzana H. Helmut M. liest mit Gymn@zilla auch russische Texte, an die er sich sonst nie heranwagen würde. Chiara P. wiederum schätzt die Qualität der deutsch-ita-

erklärt Judith Knapp, E-learning Spezialistin an der EURAC. Der Lernende kann im Anschluss an die Lektüre die Liste der unbekanntenen Worte abrufen und lernen. Außerdem erstellt das Programm automatische Übungen zu den Wortlisten als zusätzliche Lernhilfe. Aus pädagogischer Sicht erweist sich diese Art des Sprachlernens als besonders effizient: „Jeder Nutzer wählt sich seine Texte selber aus. Er ist also motiviert und am Thema interessiert“, erklärt Judith Knapp. Erst durch die eigenständige Erstellung von Wortlisten und das Üben in Form von Ratespielen werde der Lernprozess rationalisiert. Darin liege die Stärke von Gymn@zilla.

Das, was am Bildschirm in Hundertstelsekunden abläuft, bedurfte hinter den Kulissen einer langen Tüftelarbeit. Das System wird laufend verbessert und für weitere Sprachen aufgebaut. Als erster Schritt wurde Gymn@zilla wie ein normaler Browser programmiert. Das Programm klopft beim Server der BBC an und bittet über das Hypertext Transfer Protocol (HTTP) um die Herausgabe der aktuellen Nachrichten. Die erhaltene Seite wird von Gymn@zilla überarbeitet: Der Buchstabensalat des Internets wird in das einheitliche Unicode-Format überführt. Dann werden neue Verweise mit Wortklärungen auf die Seiten gesetzt und bestehende Verweise über Gymn@zilla umgelenkt.

Die Überarbeitung der Seiten beinhaltet auch eine sprachspezifische Programmierung, das so genannte Stemming. Flektierte Wörter wie „Kindern“, „ragazzi“ müssen vom Programm auf ihre Grundformen zurückgeführt werden können, also auf „Kind“ oder „ragazzo“, damit sie im elektronischen Wörterbuch gefunden werden. Hierfür arbeitet das Entwicklerteam im Moment noch mit einem kleinen, selbst gestrickten Programm.

lienischen Wörterbücher, auch wenn manchmal ein Wort falsch erkannt oder übersetzt wird. Gymn@zilla ist aber weit mehr als eine reine Lesehilfe. „Wir bieten den Nutzern die Möglichkeit, sich aus ihrer Lektüre individuelle Wortlisten zu erstellen“,



Die Plattform verbindet Webnachrichten mit elektronischen Wörterbücher

Seit kurzem nutzen alle drei die neu entwickelte Internet-Plattform der EURAC, Gymn@zilla. Dass Gymn@zilla als Internet Browser arbeitet, deutet der Namensteil „zilla“ an. Die Silbe „gymn“ steht für *üben*. Das Zeichen @ symbolisiert das Internet. Die einzigartige Plattform

## Glossar

<sup>1</sup> Ein **Tagger**, oder besser Part of Speech Tagger ist ein Programm, das eine eindeutige Entscheidung über die Zugehörigkeit eines Wortes zu einer Wortklasse trifft. Das italienische Wort „colpevole“, zum Beispiel kann ohne Kontext betrachtet sowohl Nomen als auch Adjektiv sein. Ein Tagger analysiert den nächsten Kontext (purtroppo/adverb è/verb colpevole/noun/verb), indem er diesen Kontext mit bereits gelernten Kontexten vergleicht und die wahrscheinlichste Möglichkeit auswählt.

<sup>2</sup> Ein **Parser** errechnet die hierarchische Struktur eines Satzes. Auch hierbei müssen Ambiguitäten (Mehrdeutigkeiten) aufgelöst werden:  
Franz beobachtete **den Vogel mit dem Fernrohr**.  
**Franz beobachtete** den Vogel **mit dem Fernrohr**.  
Beobachtet nun Franz mit dem Fernrohr oder besitzt der

Vogel eines? Der Parser erkennt, welche der beiden möglichen Analysen richtig ist.

<sup>3</sup> Auch nachdem Wortklasse und Satzstruktur erkannt sind, kann ein Wort noch mehrere Bedeutungen haben. Ein Programm zur **Bedeutungsdesambiguierung** schaut sich zum Beispiel die Wörter folgenden Kontexts an: „Rubinstein ist berühmt wegen seines zarten *Anschlags*“ und vergleicht diesen Kontext mit weiteren Kontexten, in denen das Wort Anschlag verschiedene Bedeutungen hat. So findet er etwa Kontexte wie „Irak, Terror, Bombe,...“, „Pianist, Klavier, Piano, zart“, „Artikel, Beitrag, Zeitung,...“, „Brett, Verwaltung, schwarz,...“, „Montage, Werkzeug, anbringen,...“. Die Bedeutung der ähnlichsten Kontextvektoren wird auf das Wort übertragen.

In Zukunft sollen aber noch stärker die reichhaltigen Open Source Möglichkeiten genutzt werden, also freie Programme, die kostenlos zu Verfügung stehen, und die große Sprachvielfalt abdecken. Ein besonderes Problem stellen die asiatischen Sprachen dar, da die Wortgrenzen nicht durch Leerzeichen markiert werden. Das chinesische Schriftbild 你不是昨天來的嗎 etwa setzt sich aus acht Zeichen und sieben Wörtern zusammen. Und so hat das Team für das Chinesische noch ein zusätzliches Programm entwickeln müssen, das den Text segmentiert.

Gymn@zilla steht allen Interessenten kostenlos zur Verfügung. „Im Moment sind wir noch in der Probephase und freuen uns über jede Rückmeldung der Benutzer“, meint das Entwicklerteam. Weitere Sprachmodule in Gymn@zilla sollen mit Hilfe internationaler Kooperationen erarbeitet werden. „Wir brauchen nicht nur Wörterbücher und Stemmer für viele Sprachen, sondern das gesamte Arsenal der heutigen Computerlinguistik, also Tagger<sup>1</sup>, Parser<sup>2</sup> und Bedeutungsdesambiguierung<sup>3</sup>“, erklären die Wissenschaftler. Diese sind nötig, um die Wörterbücher treffsicher anzusteuern. Ein *Anschlag* (ital. attentato) im Irak hat nichts mit dem *Anschlag* (ital. affissione) an einer Tafel zu tun, und ein Artikel mit 6000 *Anschlägen* (ital. battute) ist wieder etwas ganz anderes. Solange der Computer diese Unterschiede nicht erkennt, kommt

es immer wieder zu haarsträubenden Übersetzungsfehlern. Statische Lernmaterialien in Büchern oder im Internet haben diese Probleme für einige Texte bzw. eine Sprache gelöst, sind aber schnell veraltet, relativ teuer und überhaupt nur für Welt-sprachen wie Englisch erhältlich. Auf dem Markt gibt es derzeit kein mit Gymn@zilla vergleichbares Produkt. Zwar haben große Verlagshäuser PC-Wörterbücher entwickelt, wie beispielsweise *ifinger* von PONS, die als Lesehilfe eingesetzt natürlich weitaus präzisere Informationen liefern können als Gymn@zilla, doch sind derartige Softwareprogramme nicht für Nischensprachen wie etwa dem Ladinischen vorhanden. Außerdem bieten sie nicht die Funktion einer personalisierten Wortliste mit entsprechenden Übungen. Die im Handel erhältlichen Softwareprodukte wie *ifinger* sind eher Übersetzungshilfen als tatsächliche Lernprogramme. Freie im Internet erhältliche Übersetzungshilfen, wie etwa *Babel Fish Translation* bei Altavista, sind zwar auch in der Lage ganze Internetseiten zu übersetzen, allerdings nicht zum Zwecke des Sprachlernens.

Zurzeit deckt Gymn@zilla 17 Sprachpaare ab. „Gehen wir aber von den mindestens 3000 existierenden Sprachen aus, eröffnet sich uns ein Arbeitsfeld von 900.000 Sprachpaaren“, erläutert Teammitglied Oliver Streiter. Eine schier unermessliche Arbeit, die die EURAC-Forscher niemals alleine schaffen können. Deshalb arbeiten sie mit der Russischen

Akademie der Wissenschaften zusammen. Weitere Kooperationen sind im Gespräch. Die Kontakte ergeben sich meist bei internationalen Expertentreffen, auf denen auch dank Gymn@zilla die ein oder andere Sprachbarriere abgebaut wurde.

[www.eurac.edu/gymnazilla](http://www.eurac.edu/gymnazilla)

Oliver Streiter/EURAC  
Sprache und Recht  
oliver.streiter@eurac.edu

Leonhard Voltmer/EURAC  
Minderheiten und Autonomien  
leonhard.voltmer@eurac.edu



Das Gymn@zilla-Team setzt sich aus drei EURAC-Forschern mit unterschiedlicher Fachrichtung zusammen: Oliver Streiter (Mitte) ist Computerlinguist im Projekt BISTRO und Experte für Technologien, die es ermöglichen, Sprache elektronisch zu verarbeiten. Judith Knapp (links) ist Informatikerin im E-learning Projekt EL-DIT. Leonhard Voltmer (rechts) ist Jurist im Projekt MIRIS und unterrichtet juristische Fachsprache. Bei einem gemeinsamen Mittagessen entstand die Idee, die drei Projekte miteinander zu verknüpfen und einer breiten Öffentlichkeit zugänglich zu machen. Dies war die Geburtsstunde von Gymn@zilla.