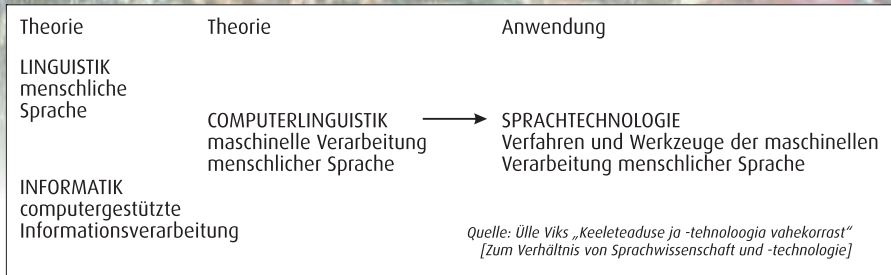


Estnische Sprachtechnologie

Was sind Sprachtechnologie und Computerlinguistik?

Die Sprachtechnologie ist ein Zweig der Informatik, der die menschliche Sprache bearbeitet. Die Computerlinguistik stellt einen Hybridbereich der Sprach- und Computerwissenschaft dar. Sowohl die Sprachtechnologie als auch die Computerlinguistik befassen sich mit der automatischen Verarbeitung natürlicher Sprache, wobei die Computerlinguistik theoretische Probleme zu ihrem Gegenstand macht, die Sprachtechnologie hingegen mehr anwendungsorientiert ist.



Forschungszentren der Computerlinguistik und Sprachtechnologie in Estland

An der Universität Tartu gibt es eine Forschungsgruppe Computerlinguistik und eine Arbeitsgruppe Sprachtechnologie.

Am Institut für Estnische Sprache wird an den Fragen der Sprachsynthese und der elektronischen Lexikografie gearbeitet.

Am Institut für Kybernetik der Technischen Universität Tallinn ist ein Labor für Phonetik und Redetechnologie angesiedelt.

Computerlinguisten werden an der Universität Tartu ausgebildet. Als Studienfach kann man Computerlinguistik im Rahmen der Finnougristik an der Philosophischen Fakultät studieren und Sprachtechnologie im Rahmen der Informationstechnologie an der Fakultät für Mathematik und Informatik. An der Technischen Universität Tallinn werden Kurse zur Sprachtechnologie (Redetechnologie) angeboten.

Womit befassen sich estnische Computerlinguisten und Sprachtechnologe? Einige Beispiele

Erfahrungen mit einem automatischen Rechtschreib-Korrekturprogramm hat jeder, der am Computer Texte verfasst hat. Als Grundlage für Rechtschreibprüfprogramme dient der morphologische Analysator und Disambiguierer – ein Programm, das die Grundform sowie die jeweilige Flexionsform eines Textwortes bestimmt. Die Textwörter, die das Korrekturprogramm nicht erkennen kann, werden auf dem Bildschirm rot unterstrichen. Auf der automatischen morphologischen Analyse und deren Umkehrprozess – der morphologischen Synthese – beruht auch die elektronische Informationssuche. Zum Beispiel wird dadurch ermöglicht, als Antwort auf das Suchwort *suveaeg* [Sommerzeit] u. a. die Phrase *suveajale üleminek* [Übergang zur Sommerzeit] zu erhalten.

Außer Rechtschreibprüfer brauchen viele Computerbenutzer auch einen automatischen Grammatikprüfer, um beim Schreiben eines Textes auch eventuelle Kommafehler o. Ä. zu entdecken. Für ein Grammatikprüfprogramm wird ein Programm der automatischen syntaktischen Analyse (Parsing) benötigt, das einen Satz in Teilsätze zerlegen und die syntaktische Struktur jedes einzelnen Teilsatzes ermitteln könnte.

Als gutes Hilfsmittel für Wörterbuchverfasser dient ein elektronisches Arbeitsmedium, in dem verschiedene sprachtechnologische Werkzeuge (Sprachsoftware und lexikalische Datenbanken) integriert sind, durch die das Verfassen sowie die Redaktion von Wörterbüchern erheblich erleichtert und effektiviert werden kann.

Eine große Zukunft haben Dialogsysteme, d. h. Benutzeroberflächen zur Ermöglichung von Datenbankabfragen in natürlicher Sprache. Beispielsweise könnten bei Telefonauskunft Computer anstatt Menschen eingesetzt werden. Bei einem Telefon-Dialogsystem braucht man dazu ein Programm der Spracherkennung, das die Sprache in Text umwandelt. Ein weiteres Programm soll die natürliche Frage in eine Datenbank-Abfrage „übersetzen“ und die Antwort in einem natürlichen Wortlaut zum Ausdruck bringen. Zu diesem Zweck muss man wissen, wie ein Telefongespräch abläuft – wie die Fragen und Antworten gewöhnlich formuliert werden. Schließlich braucht man noch einen Sprachsynthesizer, der dem Kunden die Antwort telefonisch vermittelt.



Abbildung 1.
Lexikografisches
Wörterbuchsystem des Instituts
für Estnische Sprache EELex:
Arbeitsfenster des
Textbearbeitungssystems.



Abbildung 2.
Lexikografisches
Wörterbuchsystem des
Instituts für Estnische
Sprache EELex: Fenster für
die Erstellung des
Benutzerwörterbuchs.



Haldur Öim



Mare Koit



Ülle Viks

Hier ist ein Online-Artikel der Tageszeitung Postimees mit dem Titel: „Kampaania käigus korjatud prügist suurem osa taaskasutatakse“ [„Ein Großteil des während der Müllaktion gesammelten Mülls wird recycelt“], von dem eine automatische Zusammenfassung erstellt wird. Der Benutzer kann den Umfang der Zusammenfassung frei wählen: entweder 10, 20, 30, 40 oder 50% des Artikels.



Das Ergebnis der automatischen Textzusammenfassung: eine Zusammenfassung von 30% des Artikels „Kampaania käigus korjatud prügist suurem osa taaskasutatakse“ [„Ein Großteil des während der Müllaktion gesammelten Mülls wird recycelt“].



Das Dialogsystem
Teatriagent [Theateragent]
informiert über den Spielplan
estnischer Theater.

Internetadressen

- Forschungsgruppe Computerlinguistik: <http://www.cl.ut.ee/>;
- Arbeitsgruppe Sprachtechnologie: <http://www.cs.ut.ee/~koit/KT/>;
- Institut für Estnische Sprache: <http://www.eki.ee/>;
- Labor für Phonetik und Speech Technology des Instituts für Kybernetik an der Technischen Universität Tallinn: <http://www.ioc.ee/>;
- Dialogsysteme: <http://www.dialogid.ee>
- Text-to-Speech-System: <http://www.eki.ee/keeletehnoloogia/projektid/syntees/>; <http://www.phon.ioc.ee/> -> Projekte -> Text-to-Speech Software für Blinde
- Sammelportale elektronischer Wörterbücher <http://www.keeleveeb.ee/>; <http://www.keelevara.ee/>
- Lexikografisches System EELex (Institut für Estnische Sprache) <http://exsa.eki.ee/>
- Automatische Textzusammenfassung <http://math.ut.ee/~kaili/estsum/estsumframe.cgi>



Das Poster wurde angefertigt vom Estnischen Literaturmuseum, vom Estnischen Ministerium für Bildung und Wissenschaft und von der Estnischen Gesellschaft für Muttersprache. Text: Jüri Viikberg. Layout: Margus Nõmm, Zentrum für Multimedia der Universität Tartu 2008.