L'ingénierie linguistique en Estonie

Qu'est-ce que l'ingénierie linguistique et la linguistique informatique ?
L'ingénierie linguistique est la branche de l'informatique qui traite des langues humaines. Le terme de **linguistique informatique** désigne un domaine hybride, dérivé de la linguistique et de l'informatique. Ingénierie linguistique et linguistique informatique concernent toutes deux le traitement automatique des langues naturelles, mais la linguistique informatique aborde les problèmes sous un angle plutôt théorique, l'ingénierie linguistique sous un angle plutôt pratique.

théorie	théorie	application
LINGUISTIQUE langue humaine		
	LINGUISTIQUE INFORMATIQUE traitement informatique d'une langue humaine →	INGÉNIERIE LINGUISTIQUE méthodes et procédés de traitement d'une langue humaine
INFORMATIQUE traitement de l'information par ordinateur		

(source : Ülle Viks, Les rapports entre linguistique et ingénierie linguistique, exposé au colloque « La linguistique appliquée en Estonie », 11-12 avril 2002)


Haldur Õim


Mare Koit


Ülle Viks

Quels sont les centres de linguistique informatique et d'ingénierie linguistique en Estonie ?

À l'université de Tartu fonctionnent une équipe de recherche en linguistique informatique et un groupe de travail d'ingénierie linguistique.
L'Institut de la langue estonienne travaille sur la synthèse de la parole et la lexicographie informatique.
L'Université technologique de Tallinn abrite un laboratoire de phonétique et de technologie de la parole, au sein de l'Institut de cybernétique.

La formation dans ce domaine est dispensée à l'université de Tartu, où la linguistique informatique est enseignée à la faculté des Lettres dans le cadre du programme de linguistique estonienne et finno-ougrienne, et l'ingénierie linguistique à la faculté de Mathématiques et d'Informatique dans le cadre du programme d'informatique. Les cours sur la technologie de la parole sont dispensés à l'Université technologique de Tallinn.

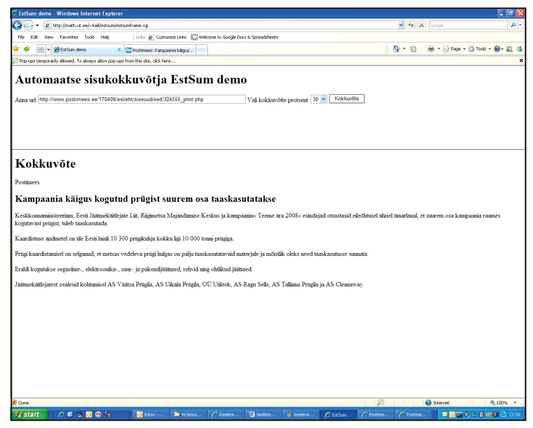
Quels problèmes traitent la linguistique informatique et l'ingénierie linguistique en Estonie ? Quelques exemples

Chaque personne saisissant un texte sur ordinateur a été en contact avec un correcteur orthographique automatique. À la base d'un tel correcteur se trouve un module d'analyse morphologique automatique – un programme identifiant, pour chaque mot du texte, le paradigme correspondant et la forme fléchie utilisée dans le texte. Les mots du texte qui ne sont pas reconnus par l'analyseur morphologique sont soulignés en rouge par le correcteur orthographique automatique. L'analyse morphologique automatique – ou la synthèse morphologique, qui en est un dérivé – sont utilisées également par les moteurs de recherche (par exemple pour déterminer les formes fléchies des mots composés, afin de rechercher aussi les textes contenant ces formes).
Outre le correcteur orthographique, beaucoup aimeraient pouvoir utiliser également un correcteur grammatical, qui saurait par exemple corriger les fautes de ponctuation dans un texte. Il serait nécessaire pour cela de disposer d'un programme d'analyse syntaxique capable de découper une phrase en propositions et d'analyser la structure syntaxique de ces propositions.
Les rédacteurs de dictionnaires utilisent avec profit un environnement de travail ouvert sur Internet et intégrant divers outils d'ingénierie linguistique (logiciels et ressources linguistiques), afin de rendre la rédaction et la révision des dictionnaires plus simples et plus efficaces.
Les systèmes dialoguants, c'est-à-dire les interfaces utilisateur permettant l'usage des langues naturelles, sont promis à un grand avenir. Par exemple, un ordinateur peut répondre à la place d'un être humain dans un système de renseignements téléphoniques. Pour créer un système dialoguant qui fonctionne via le téléphone, il faut tout d'abord un module de saisie de la parole qui transforme celle-ci en texte, puis un programme capable de « traduire » la question humaine en requête adressée à une base de données et, inversement, le résultat de cette requête en texte compréhensible par l'utilisateur. Pour cela, il faut savoir comment les gens s'expriment au téléphone : comment ils posent leurs questions, comment on leur répond. Enfin, il faut disposer d'un module de synthèse vocale, pour transmettre la réponse à l'utilisateur via le téléphone.

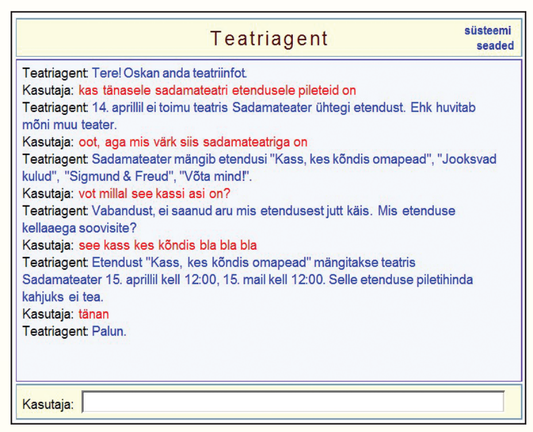
Adresses Internet :

Groupe de recherche en linguistique informatique : http//www.cl.ut.ee ; groupe de travail en ingénierie linguistique : http://www.cs.ut.ee/~koit/KT/ ; Institut de la langue estonienne : http://www.eki.ee ;
Laboratoire de phonétique et de technologie de la parole (Université technologique de Tallinn, Institut de Cybernétique) : http://www.ioc.ee/ ;
Systèmes dialoguants : http://www.dialoogid.ee
Synthèse texte-parole : http://www.eki.ee/keeletehnoloogia/projektid/syntees/ ; http://www.phon.ioc.ee/ -> projektid -> tekst-kõne süntees pimedatele
Portails regroupant des dictionnaires électroniques : http://www.keeleveeb.ee ; http://www.keelevara.ee
EELex, système de dictionnaires de l'Institut de la langue estonienne : http://exsa.eki.ee/ ; rédacteur automatique de résumés : http://math.ut.ee/~kaili/estsum/estsumframe.cgi

La version Internet du quotidien « Postimees », avec l'article « La majeure partie des déchets ramassés pendant la campagne de collecte seront réutilisés », que le rédacteur automatique de résumés va condenser. L'utilisateur peut choisir le taux de compression : 10, 20, 30, 40 ou 50 % de l'article original.

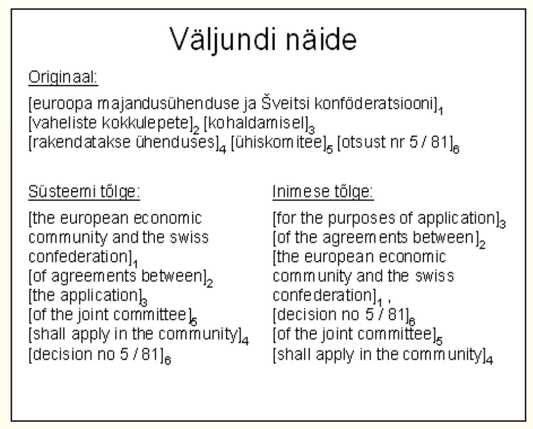


Résultat renvoyé par le rédacteur automatique de résumés : résumé à 30 % de l'article « La majeure partie des déchets ramassés pendant la campagne de collecte seront réutilisés ».

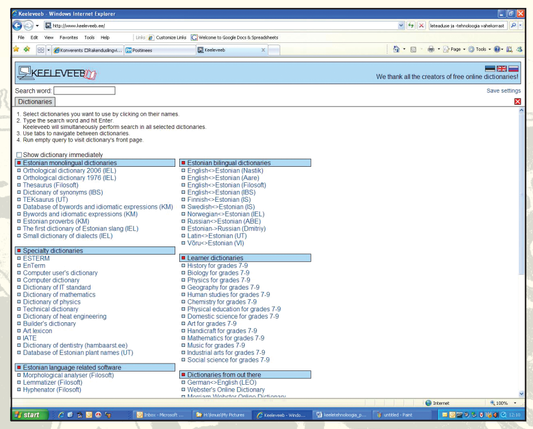


Le système de dialogue « Agence théâtrale » donne des informations sur la programmation des salles de théâtre en Estonie.

Le système de dialogue « Agence de voyage » délivre des informations sur les vols au départ de l'aéroport de Tallinn. Il est capable de donner ses réponses par synthèse vocale. Dans l'exemple illustré, la question était : « Est-il possible de faire des réservations ? » et le système a répondu : « Les réservations sont traitées au guichet. Je ne connais que les heures de départ des vols.»



On peut voir que le système de traduction statistique donne de bons résultats quand il traduit des propositions indépendantes, et aussi parfois dans le réagencement des propositions voisines. Toutefois, si l'ordre des propositions doit être différent dans la langue cible de ce qu'il est dans la langue source, le système n'est pas capable de les disposer dans le bon ordre.



Le portail lexicographique Keeleveeb a lui aussi une interface en langue anglaise.

Panorama préparé par le Musée littéraire estonien, le ministère estonien de l'Éducation et de la Recherche et la Société pour la langue estonienne. Texte : Kadri Muischnek. Mise en page : Margus Nõmm, Centre Multimédia de l'université de Tartu, 2008.