Ðàçìå÷åííûé êîðïóñ äèàëîãîâ êàê ðåñóðñ ìîäåëèðîâàíèÿ äèàëîãà: îðãàíèçàöèÿ è ðàçìåòêà Ý

Àâòîð Îëüãà Àëåêñàíäðîâíà Ãåðàñèìåíêî
18.07.2008 ã.
Ïîñëåäíåå îáíîâëåíèå 20.07.2008 ã.

Òåêñòû ïå÷àòíîãî èçäàíèÿ â ôîðìàòå PDF

Ýñòîíñêèé êîðïóñ äèàëîãîâ ñîçäàí ëèíãâèñòàìè è êîìïüþòåðíûì òåõíîëîãàì Òàðòóñêîãî óíèâåðñèòåòà ñ öåëüþ èññëåäîâàòü ðå÷åâîå ïîâåäåíèå êîìóíèêàöèè.

1. Ñîñòàâ êîðïóñà

Îñíîâíàÿ ÷àñòü êîðïóñà ñîñòîèò èç åñòåñòâåííûõ óñòíûõ äèàëîãîâ (758 ñïîíòàííî çàïèñàííûõ äèàëîãîâ è 106 íåïîñðåäñòâåííûõ äè

Óñòíûå äèàëîãè ñîõðàíÿþòñÿ â öèôðîâîì èëè îöèôðîâàííîì âèäå .wav, à â òåêñòîâîì âèäå ïðåäñòàâëåíû â òðàíñêðèïöèè Äæåôô Òðàíñêðèïöèÿ ôèêñèðóåò ÿâëåíèÿ, ïîçâîëÿþùèå ïðîñëåäèòü äèíàìè÷åñêîå ïîñòðîåíèå äèàëîãà èç ðåïëèê è èíòîíàöèîííûõ åäèíèö: ä

Ìàëàÿ, íî âàæíàÿ ÷àñòü êîðïóñà — 22 ïèñüìåííûõ äèàëîãà îáùåìî â 2500 ñëîâ — ïðåäñòàâëÿåò ñîáîé êîìïüþòåðíûå ñ åñòåñòâåííî ÿçûêå, íî â äåéñòâèòåëüíîñòè ðîëü ïðîãðàììû èñïîëíÿë ÷åëîâåê. Äèàëîãè ïðåäñòàâëÿþ ñîáîé ëîãè áåñåä ïîëüçîâàòå

2. Óðîâíè àííîòàöèè

2.1. Ìîðôîëîãèÿ

×àñòü óñòíûõ äèàëîãîâ êîðïóñà (20000 ñëîâ) ðàçìå÷åíà ìîðôîëîãè÷åñêè. Àâòîìàòè÷åñêàÿ ðàçìåòêà îñóùåñòâëåíà ñ ïîìîùüþ ìîðôîë îáñóæäåíèå, êîòîðîìó ñîïóòñòâîâàëî ïðîáëåìàìè ñíÿòèÿ ìîðôîëîãè÷åñêîé îìîíèìèè â óñòíîé ýñòîíñêîé ðå÷è â ñðàâíåíèè ñ

2.2. Ðå÷åâûå àêòû

Íè îäíà èç ðàññìîòðåííûõ ñõåì ðàçìåùåíèå íå îòâå÷àëà öåëÿì ìîäåëèðîâàíèÿ è èñïëåäîâàíèÿ åñòåñòâåííûõ äèàëîãîâ ñ òî÷êè çðåíèÿ
òèïîëîãèé [Gerassimenko et al. 2004].

Òèïîëîãèÿ ñîñòîèò èç 126 äèàëîãîâûõ àêòîâ è îñíîâûâàåòñÿ íà ðàñïîçíàâàíèè äèàëîãà êàê ñîöèàëüíîãî âçàèìîäåéñòâèÿ. Ðå÷åâûå
îäèíî÷íûå àêòû (íàïðèìåð, ïîáóäèòåëüíûå ðåïëèêè îáðàòíîé ñâÿçè). Ñ äðóãîé ñòîðîíû, àêòû ïîäðàçäåëÿþòñÿ íà òðåõñùèåñÿ ê ïðåäì

Âîñåìüñîò äâàäöàòü óñòíûõ äèàëîãîâ è äâàäöàòü îäèí äèàëîã ÂÎÇ ðàçìå÷åíû äâóìÿ àíîòàòîðàìè, ðàçíîãëàñèÿ êîòîðûõ óñòðàíåíû
âñïîìîãàòåëüíîé ïðîãðàììû.

Ïðèìåð 1 (åñòåñòâåííûé äèàëîã â ïåðåâîäå íà ðóññêèé).

((çâîíîê)) | RIE: ÇÂÎÍÎÊ |

V: `ñïðàâî÷íàÿ | RIJ: ÎÒÂÅÒ ÍÀ ÇÂÎÍÎÊ | | RY: ÈÄÅÍÒÈÔÈÊÀÖÈß |

`Kðèñòà | RY: ÈÄÅÍÒÈÔÈÊÀÖÈß |

çäðàâñòâóéòå? | RIE: ÏÐÈÂÅÒÑÒÂÈÅ |

(.)

H: çäðàâñòâóéòå, | RIJ: ÏÐÈÂÅÒÑÒÂÈÅ |

=ÿ `õîòåë áû ((íàçâàíèå ôèðìû)) (.) < äèñ`ïåò÷åðà > èëè `êàêîé-íèáóäü `íî- íîìåð. | DIE: ÄÈÐÅÊÒÈÂ |

(0.8)

V: ì ìèíóòî÷êó. | DIJ: ÎÒÑÐÎ×ÊÀ |

(5.5)

V: òåëåôîí îá`ñëóæèâàíèÿ [òî åñòü,] äà? | KYE: ÎÁÙÈÉ ÂÎÏÐÎÑ | | VTE: ÓÒÎ×ÍÅÍÈÅ ÓÑËÎÂÈÉ ÎÒÂÅÒÀ |

H: [äà.] | KYJ: ÄÀ | | VTJ: ÓÒÎ×ÍÅÍÈÅ ÓÑËÎÂÈÉ ÎÒÂÅÒÀ |

V: ïî íåïîäòâåðæäåííûì äàííûì íîìåð < ÷åòûðå ÷åòûðå ñåìü? (0.5) äåâÿòü âîñåìü? (.) [îäèí] íîëü. > | DIJ: ÈÍÔÎÐÌÀÖÈß |

H: [.hh] | YA: ÈÍÎÅ |

(.)

H: ñïàñèáî. | RIE: ÑÏÀÑÈÁÎ |

V: ïîæàëóéñòà? | RIJ: ÏÎÆÀËÓÉÑÒÀ |

2.3. Êîììóíèêàòèâíûå ñòðàòåãèè

Ìû èñõîäèì èç ïîíÿòèÿ êîììóíèêàòèâíîé ñòðàòåãèè è êîíñòðóêòèâíîé ìîäåëè äèàëîãà (Constructive Dialogue Model, CDM) [Jokinen ó÷àñòíèêîì äèàëîãà äëÿ ïîñòðîåíèÿ ñëåäóþùåé ðåïëèêè êàê ðåàêöèè íà ïðåäûäóùóþ ðåïëèêó ïàðòíåðà. Êîììóíèêàòèâíûå ñòðàòåã îáîáùåííî óðîâå.

Äëÿ ïîðàæäåíèÿ êîììóíèêàòèâíûõ ñòðàòåãèé â CDM èñïîëüçóþòñÿ ÷åòûðå êîíòåêñòóàëüíûõ ôàêòîðà:

– îæèäàåìîñòü ðåïëèêè;

– ñâÿçü ðåïëèêè ñ òåìîé;

– äîñòèãíóòîñòü öåëåé ãîâîðÿùåãî;

– èíèöèàòèâà ãîâîðÿùåãî èëè ïàðòíåðà.

Âñå êîíòåêñòóàëüíûå ôàêòîðû áèíàðíû, ñîîòâåòñòâåííî, êîììóíèêàòèâíûõ ñòðàòåãèé 2*4=16.

Íàèáîëåå óäà÷íîé ñòðóêòóðîé ïðåäñòàâëåíèÿ êîììóíèêàòèâíûõ öåëåé ÿâëÿåòñÿ ìàãàçèí öåëåé (stock) è ïðèíöèï «ïîñëåäíèì âîøåë, ïîñëåäíèì âûøåë». Ïî ìåðå äîñòèæåíèÿ âåðõíèõ öåëåé îíè óäàëÿþòñÿ; âñå öåëè äèàëîãà ñ÷èòàþòñÿ äîñòèãíóòûìè,

Êîììóíèêàòèâíûå ñòðàòåãèè ðàçìå÷åíû â 20 åñòåñòâåííûõ è 20 ñèíóëÿòèîííûõ äèàëîãàõ.

Ïðèìåð 2 (ñèíóëÿöèîííûé äèàëîã â ïåðåâîäå íà ðóññêèé ÿçûê).

Ðåïëèêà

Ñòðàòåãèÿ

Ìàãàçèí

Êëèåíò:

Êàê äîåõàòü äî Ïýðíó èç Òàðòó äî ïîëóäíÿ?

îæèäàåì.-ñâÿçí.-äîñòèãí.-ãîâîðÿù.-íà÷àëî/êîíåö

–

Êîìïüþòåð:

Àâòîáóñ âûõîäèò â 5 óòðà.

Àâòîáóñ âûõîäèò â 8 óòðà.

îæèäàåì.-ñâÿçí.-íåäîñòàòî÷íî.-ãîâîðÿò.-ïðîäîëæåíèå òåìû

Òàðòó-Ïÿðíó

Èíòåðåñóåò ëè âàñ âðåìÿ ïðèáûòèÿ? ...

íåîæèäàåì.-ñâÿçí.-äîñòàòî÷íî.-íîâûé äèàëîã

Òàðòó-Ïÿðíó

Ðàçìå÷åííûé êîðïóñ èñïîëüçóåòñÿ â òåîðåòè÷åñêèõ è ïðèêëàäíûõ èññëåäîâàíèÿõ (íàïð., [Hennoste et al. 2005]) è â ðàáîòå íàä àâ
[Fishel 2004]), à òàêæå â ïîñòîÿííûé ïåðâûõ äèàëîãîâûõ ñèñòåì íà ýñòîíñêîì ÿçûêå. Ýòè ñèñòåìû, Òðàíñëÿòîðíûé è Òåàòðàëüíûé À
ïðèìåíåíèì çíàíèé î ñòðóêòóðå èíôîðìàöèîííûõ äèàëîãîâ è ïîääèàëîãîâ.

Ñïèñîê ëèòåðàòóðû

Mark Fishel 2005. Dialogue Act Recognition in Estonian Dialogues Using Artificial Neural Networks. In: Proceedings of the Second Baltic Conference on Human Language Technologies, Tallinn, 4–5 April 2005, 249–254.

Olga Gerassimenko, Tiit Hennoste, Mare Koit, Andriela Rääbis, Krista Strandson, Maret Valdisoo, Evely Vutt. Annotated Dialogue Corpus as a Language Resource: An Experience of Building the Estonian Dialogue Corpus. The First Baltic Conference “Human Language Technologies. The Baltic Perspective”. Commission of the Official Language at the Chancellery of the President of Latvia, Riga, 2004, 150–155.

Tiit Hennoste, Olga Gerassimenko, Riina Kasterpalu, Mare Koit, Andriela Rääbis, Krista Strandson, Maret Valdisoo. Questions in Estonian Information Dialogues: Form and Functions. Text, Speech and Dialogue. 6th International Conference TSD 2005. Springer, 2005, 420–427.

Gail Jefferson 2004. Glossary of transcript symbols with an introduction. In Lerner, G.H. (Ed). Conversation Analysis: Studies from the first generation. Amsterdam/Philadelphia: John Benjamins, 13–31.

Kristiina Jokinen. 1995. Rationality in Constructive Dialogue Management; URL: http://cl.aist-nara.ac.jp/lab/papers/kris/aaai.ps (used 20.05.2006).

Áëàãîäàðíîñòè

Ðàáîòó ïîääåðæèâàåò Ýñòîíñêèé íàó÷íûé ôîíä (ãðàíò 5685).

Summary

The Estonian Dialogue Corpus is collected with the aim of developing the dialogue system using the natural language. The spoken dialogues (884 dialogues, 155000 running words) are used to study the rules and norms of the human-human communication; the corpus also includes human-computer dialogues (21, 2500 running words) collected by the Wizard of Oz method used to study the role behaviour of the users and information provider. The presentation considers the means and levels of transcription and annotation dialogues and also the application of the corpus.

Ïðèëîæåíèå

1. Òðàíñêðèïöèîííûå çíàêè

ñïàä èíòîíàöèè .

ïîëóñïàä èíòîíàöèè ,

ïîäúåì èíòîíàöèè ?

êîðîòêàÿ ïàóçà (ìàêñ. 0.2 ñ.) (.)

äëèíà ïàóçû â ñåêóíäàõ (2.0)

íàëîæåíèå [text]

ñëèÿíèå íåçàâèñèìûõ åäèíèö =

ðàñòÿíóòûé çâóê ::

óäàðíîå ñëîâî `

ïðåðâàííîå ñëîâî do-

âäîõ .hhh

óáûñòðåíèå òåìïà > text <

çàìåäëåíèå òåìïà < text >

ñîìíèòåëüíûé îòðåçîê {text}

íåðàçáîð÷èâûé îòðåçîê {---}

2. Òèïîëîãèÿ ðå÷åâûõ àêòîâ

I Àêòû, ñîñòàâëÿþùèå ñìåæíûå ïàðû

1.1 Àêòû óïðàâëåíèÿ äèàëîãîì

1.1.1 Kîììóíèêàöèÿ

1.1.1.1 Ðèòóàëû (RIE RIJ)

1.1.1.2 Ñìåíà òåìû

1.1.2 Ðàçðåøåíèå ïðîáëåì

1.1.2.1 Èñïðàâëåíèå, èíèöèèðîâàííûå ïàðòíåðîì

1.1.2.2 Ïðîâåðêà êîíòàêòà

1.1.2.3 Óòî÷íåíèå óñëîâèé îòâåòà (VTE VTJ)

1.2 Èíôîðìàöèîííûå àêòû

1.2.1 Äèðåêòèâû (DIE DIJ)

1.2.2 Âîïðîñû (KYE KYJ)

KYE: îáùèé âîïðîñ

KYE: îáùèé âîïðîñ, îæèäàþùèé ðàçâåðíóòîãî îòâåòà

KYE: àëüòåðíàòèâíûé âîïðîñ

KYE: ñïåöèàëüíûé âîïðîñ

KYE: èíîå

KYJ: ää

KYJ: íåò

KYJ: ñîãëàñíîå íåò

KYJ: èíîé îòâåò íà îáùèé âîïðîñ

KYJ: àëüòåðíàòèâà: îäíà

KYJ: àëüòåðíàòèâà: îáå

KYJ: àëüòåðíàòèâà: òðåòèé âûáîð

KYJ: àëüòåðíàòèâà: îòðèöàíèå

KYJ: àëüòåðíàòèâà: èíîå

KYJ: ðàçâåðíóòûé îòâåò

KYJ: îòñóòñòâèå èíôîðìàöèè

KYJ: îòêàç

KYJ: èíîå

1.2.3 Ìíåíèå

II Îäèíî÷íûå àêòû

2.1 Àêòû óïðàâëåíèÿ äèàëîãîì

2.1.1 Êîìóíèêàöèÿ

2.1.1.1 Ðèòóàëû (RY)

2.1.1.2 Îáðàòíàÿ ñâÿçü

2.1.2 Ðàçðåøåíèå ïðîáëåì

2.1.2.1 Èñïðàâëåíèå

2.2 Èíôîðìàöèîííûå àêòû

2.2.1 Îñíîâíûå àêòû (YA)

2.2.2 Äîïîëíåíèÿ îñíîâíûõ àêòîâ (ïîÿñíåíèå, óòî÷íåíèå)

Ïðèìå÷àíèå: Ïîäðîáíî ðàñïèñàíà ãðóïïà âîïðîñîâ è îòâåòîâ — íàèáîëåå ÷àñòîòíûõ ðå÷åâûõ àêòîâ; ýñòîíñêàÿ àááðåâèàòóðà,

Summary

The Estonian Dialogue Corpus is collected with the aim of developing the dialogue system using the natural language. The spoken dialogues (884 dialogues, 155000 running words) are used to study the rules and norms of the human-human communication; the corpus also includes human-computer dialogues (21, 2500 running words) collected by the Wizard of Oz method used to study the role behaviour of the users and information provider. The presentation considers the means and levels of transcription and annotation dialogues and also the application of the corpus.