Friðrik Skúlason ehf.

Endurbætt tillögugerðar- og orðskiptiforrit Púka

Friðrik Skúlason ehf.

Endurbætt tillögugerðar- og orðskiptiforrit Púka

Með endurbótum á tillögugerðar- og orðskiptiforriti Púka var leitast við að bæta Púka og festa hann þannig í sessi sem öflugt íslenskt tungutæknitól fyrir almenna notendur. Endurbæturnar skiluðu sér inn í nýjustu útgáfu forritsins, Púka 2003. Verkefnin, sem tóku um fimm mannmánuði, voru unnin af Friðriki Skúlasyni.

Sérstaða Púka sem leiðréttingarforrits fyrir íslensku felst í sveigjanleika og nákvæmni við villuleit. Ólíkt öðrum stafsetningarforritum notar Púkinn mjög öfluga aðferð þar sem hann leiðir út orðmyndir frá grunnmynd orða, með öðrum orðum hann kann íslenskar beygingar- og orðmyndunarreglur. Auk þess geymir Púki ýmsar nánari upplýsingar um orðin sem varða m.a. merkingarfræði og nýtast við villuleit. Hefðbundin stafsetningarforrit sem byggjast á stórum uppflettiorðasöfnum eða tölfræðilegum aðferðum greina ekki á milli orðanna á þennan hátt sem eykur hættuna á að þeim yfirsjáist villur og þekki ekki rétt mynduð orð.

Sú vitneskja sem Púki býr yfir gerir það að verkum að hann yfirfer textann með mun betri árangri en almennt tíðkast, tillögugerð hans er nákvæmari og hann býður auk þess upp á þann möguleika að beygja orð, skipta orðum á milli lína og finna samheiti orða.

Púki hefur styrkt stöðu sína enn frekar með þeim endurbótum sem voru gerðar fyrir tilstilli tungutæknisjóðs.

Endurbætt tillögugerðarforrit Púkans

Púki er leiðréttingarforrit Friðriks Skúlasonar ehf. Púki skoðar eitt og eitt orð í senn og athugar hvort það sé rétt stafsett. Þegar hann finnur rangt orð inni í texta eða orð sem hann þekkir ekki staðnæmist hann við það og kemur með tillögur að réttu orði. Mörg orð eru tví- eða margræð og því getur reynst erfitt í sumum tilfellum að láta Púka stinga upp á því orði sem notandi ætlaði að skrifa. Ástæðan er sú að Púki leggur til orð út frá útliti orðanna en getur ekki metið efni eða innihald textans.

Tillögugerð Púka leyfði öll rétt mynduð samsett orð í íslensku óháð því hvort þau voru í raun notuð eða ekki. Þetta gerði það

að verkum að tillögugerðin stakk oft upp á orðum sem þóttu langsótt. Endurbætur á tillögugerð Púka fólust annarsvegar í að fækka bulltillögum sem Púki kom með þegar hann fann rangt stafsett orð og hinsvegar að bæta orðum við tillögugerðina þar sem ástæða þótti til.

Við endurbætur á tillögugerð var m.a. notast við textabanka sem byggist á orðasafni Morgunblaðsins frá síðustu þremur árum (u.þ.b. fjórar milljónir setninga – 450 MB). Endurbætur á Púka, sem varða bæði tillögugerð og orðskiptiforrit, hafa tvöfaldað orðasafn hans.

Helstu verkþættir:

1. Orðasafn bætt: Púki þekkti ekki viðkomandi orð og kom því með bullorð sem tillögur. Þetta var lagfært með því að bæta orðaforða Púka.

2. Bullorðum fækkað: Púki samþykkti ýmis orð sem eru í raun leyfileg út frá íslenskum orðmyndunarreglum en eru samt sem áður bullorð eða jafnvel erlend. Þetta var lagfært með því að taka um 50 eins til tveggja stafa orð t.d. 'il', 'te' 'á' og 'ari' og leyfa tillögugerðinni ekki að nota þau í seinni hluta samsettra orða þar sem þau víkka tillögugerðina of mikið.

3. Notkun viðskeyta þrengd: Viðskeyti sem finnast aðeins í ákveðnum orðum voru yfirfarin til að koma í veg fyrir að þau yrðu leyfð í öllum samsetningum.

4. Algeng orð sett inn sem ein heild: Um eitt þúsund algengustu samsettu orðin úr textabankanum voru sett inn sem ein heild. Ástæðan er sú að tillögugerðin notar fyrst orð sem eru í einu lagi. Ef hún finnur ekkert slíkt notar hún þau orð sem eru möguleg úr tveimur hlutum o.s.frv. Með því að setja algeng orð inn í einu lagi tekur tillögugerðin þessi orð fram yfir samsett orð sem minnkar líkurnar á bulltillögum.

5. Fleiri möguleikar prófaðir: Tillögugerðin prófar fleiri möguleika en áður. Ástæðan er sú að Púki var uppbyggður á þann hátt að hann gerði ráð fyrir aðeins einni villu í hverju orði. Nú prófar hann mun fleiri möguleika sem víkkar tillögugerðina til muna.

Endurbætt orðskiptiforrit Púkans

Orðskiptiforrit Púka var bætt með það að markmiði að Púki gæti skipt öllum orðum rétt og veldi alltaf líklegustu skiptinguna hverju sinni.

Helsti galli orðskiptiforritsins var að það skipti ekki orðum sem voru geymd heil í orðasafninu. Einnig skipti hann í einstaka tilfellum orðum vitlaust miðað við hefðbundna skiptingu, þá aðallega samsettum orðum. Sum orð eru tví- eða margræð og hægt að skipta þeim á fleiri en einn hátt. Sumir möguleikar eru þó ansi langsóttir þótt skiptingin sé möguleg, þessum tilfellum var fækkað til muna.

Helstu verkþættir:

1. Skipting sett í heil samsett orð: Orðasafnið var yfirfarið m.t.t. samsettra orða sem eru geymd heil og settar inn viðeigandi skiptingar. Áður skipti Púki einungis orðum sem voru samsett. Ef samsett orð var geymt heilt í orðasafni skipti hann því ekki.

2. Skiptimerki sett inn: Viðeigandi skiptimerki voru sett inn í ákveðin samsett orð. Orð með viðskeyttum greini voru t.d. geymd heil áður og því skipti Púki þeim ekki en gerir nú. Skiptingar á samsetningum hafa forgang í Púka. Skiptingar á atkvæðum eru ekki leyfilegar. Púki er byggður inn í enska Púkann sem leyfir ekki val þarna á milli eða hvorttveggja. Skiptingar á samsetningum eru teknar fram yfir atkvæðaskiptingu þar sem langflestir notendur Púka vilja skipta á þann hátt í ritvinnslu.

3. Líklegasta skiptingin valin: Yfirferð á þeim orðum sem eru tví- eða margræð og gefa möguleika á tveimur eða fleiri mismunandi skiptingum er lokið þar sem líklegasta skiptingin fyrir hvert orð var valin þar sem möguleiki var á.

Dæmi: ‘sjóðs-láni’ – ‘sjóð-sláni’
‘fisk-afli’ –‘fis-kafli’

Púki getur ekki skoðað setningar í heild heldur aðeins eitt orð í einu. Þó er unnt er að ná góðum árangri í skiptingum sem varða tví- eða margræðni. Í einhverjum tilvikum er engan veginn hægt að gera upp á milli möguleika. Sumt af því mun Púki ráða við í framtíðinni, þ.e. þau atriði sem setningafræðin ræður við, dæmi: ‘heims-enda’ eða ‘heim-senda’. Önnur vandamál í orðskiptingum sem varða merkingarfræði og samhengi munu seint verða leyst.