KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
KOMPIUTERINIŲ TINKLŲ KATEDRA

Jurgita Lasytė

Skaitmenizuoto teksto automatinio rengimo santraukų sistemos: technologijų tyrimas ir diegimas lietuvių kalbai

Magistro darbas

Darbo vadovas
Doc. B.Tamulynas

Kaunas, 2011

KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
KOMPIUTERINIŲ TINKLŲ KATEDRA

Jurgita Lasytė

Skaitmenizuoto teksto automatinio rengimo santraukų sistemos: technologijų tyrimas ir diegimas lietuvių kalbai

Magistro darbas

Recenzentas
Doc.dr. R.Butkienė

Vadovas:
doc. B.Tamulynas
2011-05-30

Atliko:
IFN 9/1, Jurgita Lasytė
2011-05-30

Kaunas, 2011

Turinys

Summary

Summarizer technology research and adaptation in Lithuanian language

Summarizer is useful system, which allows compressing text and represents it in a shorter way. In Lithuania summarizers are new technology and information about them is poor, although these systems are created and developed by computational linguistics specialists in other countries. In summarizers' research English and Lithuanian texts were used, which allowed seeing differences between received summaries. In the English texts, summaries using all summarizers were quite informative, useful and did not loose main information from original texts, except cases when a high compression rate was used. While using for summarization Lithuanian texts, some summarizers confront with language recognition problem – summaries in Lithuanian are with unrecognizable symbols, or in individual cases these symbols are missing. Summaries in Lithuanian language are not of the same quality as summaries in English. It was also noticed, that Lithuanian summaries with higher compression rate are loosing original text meaning, information and their usefulness. Analyzed summarizers are best fitted for English language, therefore using summarizer for other language than English stipulates various problems – from language recognition to rendering of text meaning. It is very important to adapt different languages to summarizers, as this would significantly improve summarizers. Thats why were tries to make a summarizer to Lithuanian language texts. Results of summarizer creation lets to make conlusion, that summarizers designed to Lithuania language should use more than one method of summarization, and sentence length method not allways can be effective. Findings of this paper can be used in creating Lithuanian summarizer or improving other summarizers.

1. Įvadas

Šiuolaikinis pasaulis pilnas informacijos, todėl vis dažniau ir dažniau susiduriame su informacijos perkrova. Pastaruoju metu mokslininkai vis daug dėmesio skiria kompiuterizuotoms priemonėms, kurios padėtų sumažinti informacijos perkrovą arba leistų sutrumpinti gaunamą informacijos kiekį iki minimalaus lygio. Tam tikslui pasiekti buvo pradėtos kurti vadinamosios santraukų sistemos – sistemos, kurios pagal vartojo reikalavimus iš tam tikro teksto parengia santraukas. Santraukų sistemos nuo savo atsiradimo pradžios iki šių dienų pakito – pirmosios santraukų sistemos galėjo parengti santraukas tik iš tekstinių skaitmeninių dokumentų, šiuolaikinės specializuotos sistemos santraukas gali parengti ne tik iš skaitmeninio teksto, bet ir iš skaitmeninių garso bei vaizdo šaltinių.

Santraukų sistemos (*summarizers*) pradininku galima laikyti mokslininką Luhn, kuris 1958 m. viename iš savo darbų paminėjo sąvoką *summarizer*, ir aprašė pirmosios santraukų sistemos veikimą. Lunh pirmąją santraukų sistemą sukūrė dirbdamas IBM kompanijoje 1950 m, tačiau kaip minėta platesnei auditorijai informacija apie tokias sistemas pateikta po aštuonerių metų. Luhn laikomas pirmuoju mokslininku, kuris sukūrė daugybę šiuo metu informacijos moksle naudojamų technologijų. Luhn sukurta sistema tuo metu nesudomino kitų mokslininkų ir buvo primiršta. Šios sistemos prisimintos vėl, kai vartotojai pradėjo susidurti su informacijos pertekliumi ir atsirado poreikis įrankiui, kuris padėtų vartotojui padėtų sutrumpinti laiką, skiriamą informacijos peržiūrai. Santraukų sistemų populiarumas paskutiniuoju metu augęs, dabar yra kiek nustumtas į šalį, kadangi atsirado kitų NLP reikalaujančių sričių. Mokslininkai yra sukūrę ar bent jau yra bandę sukurti santraukų sistemas, kurios padėtų atsirinkti informaciją, sutrumpinti ilgus tekstus, atrinkti pačią reikalingiausią informaciją. Kadangi šios santraukų sistemos yra pakankamai naudingos, galima bandyti jas pritaikyti lietuvių kalbai arba kurti tokią sistemą, kuri būtų tam tikrais aspektais pritaikyta lietuvių kalbos tekstams.

Automatinės teksto santraukų rengimo sistemos yra viena iš kompiuterinės lingvistikos sričių, todėl mokslininkai ir toliau plėtoja bei tobulina santraukų sistemas bei algoritmus. Daugiausia apie santraukų sistemas darbų yra parašę mokslininkai Mani I., Hovy E., Mitkov R. ir Moens M.. Tačiau santraukų rengimo sistemoms populiarėjant, įvairių šalių universitetų mokslininkai paskelbė savo idėjas, sukurtas sistemas bei gautus rezultatus, kurie neretai pristatomi tarptautinėse konferencijose. Pastaruoju metu domėjimasis automatinėmis santraukų sistemomis kiek priblėso, kadangi šios sistemos neretai yra priskiriamos NLP, kuri yra labai plati bei įvairių tyrimų reikalaujanti sritis. Internetinėje erdvėje galima rasti ne tik straipsnių apie

santraukų sistemas, ištraukų iš Hovy E. ar Mani I. Knygų atskirų skyrių, bet ir paskaitų, kuriose universitetų mokslininkai pasakoja apie savo kuriamus arba sukurtus santraukų rengimo įrankius, jų privalumus, trūkumus bei gautus rezultatus.

Oficialiai yra skelbiama jog komercinių automatinių teksto santraukų rengimo sistemų šiuo metu yra 16. Nekomercinių tokių sistemų skaičius yra panašus, bet jis nuolat auga, kadangi vis daugiau mokslininkų ar kompiuterinės lingvistikos entuziastų įsijungia į tokių sistemų kūrimą. Nekomercinės santraukų rengimo sistemos tam tikrais atvejais ne tik, kad nenusileidžia komercinėms sistemoms, bet ir funkcionalumo atžvilgiu lenkia komercines – vartotojui patogiau internetinės nemokamos, o ne diegiamos ir mokamos sistemos, turinčios ir tam tikrus apribojimus, bei keliančios sudėtingesnius reikalavimus, nei analogiškos nemokamos sistemos.

Uždavinio aktualumas

Internetinėje erdvėje atsirandant vis didesniam informacijos kiekiui iškyla poreikis tą informaciją sumažinti, atrinkti ar dar kaip nors nufiltruoti. Tam į pagalbą pasitelkiamos santraukų rengimo sistemos, kurios padeda iš ilgų tekstų parengti trumpą santrauką, kuri atspindi pagrindines originalaus teksto mintis, todėl šiam darbui keliamas uždavinys – sukurti santraukų sistemą, kuri parengtų lietuviško teksto santrauką. Santraukų sistemų analizės rezultatai pristatyti Kauno technologijos universiteto Humanitarinių mokslų fakulteto surengtoje konferencijoje „Kalbos teorija ir praktika“, šia tema parašytas straipsnis „Skaitmenizuotų tekstų santraukų rengimo sistemos: panašumai, skirtumai ir taikymas lietuvių kalbai“, kuris išspausdintas žurnale „Kalbų studijos“. Bandomosios santraukų sistemos kūrimas turėtų būti aktualus kompiuterinės lingvistikos specialistams, entuziastams, kadangi tai viena iš sričių, kurios domina NLP tyrėjus. Santraukų sistemos sukūrimas yra aktualus ir todėl, kad oficialiai nėra paskelbtas joks darbas apie Lietuvoje sukurtą tokią sistemą, kuri gebėtų atpažinti lietuviškus simbolius bei nepradėti gilesni tyrimai šioje srityje.

Darbo tikslas

Lyginant įvairias skaitmenizuoto teksto automatinio rengimo sistemas išskirti bei apibūdinti teksto santraukų rengimo metodus ir lyginant sistemas aptarti jų pliusus bei minusus. Pagal pasirinktus kriterijus sukurti santraukų rengimo sistemą programavimo kalboje JAVA bei pademonstruoti jos veikimą.

Darbo uždaviniai

- Surinkti informaciją apie šiuo metu populiariausias ir lengviausiai pasiekiamas automatines santraukų rengimo sistemas;
- Surinkti informaciją bei ištirti automatinių teksto santraukų sistemų naudojamus metodus;
- Susisteminti žinias apie automatines santraukų rengimo sistemas, suskirstyti jas į grupes, pagal naudojamus metodus;
- Nustatyti, kokiais metodais kurtos automatinių teksto santraukų sistemos yra efektyviausios;
- Aptarti santraukų sistemų metodus ir juos susieti su santraukos kokybe;
- Pasirinkti metodus bei programos prototipus, kuriais remiantis bus projektuojama ir kuriama sistema;
- Sukurti sistemą, kuri pateiktų santrauką;
- Išbandyti sukurtą santraukų sistemą bei pateikti gautus rezultatus;
- Pateikti darbo rezultatų išvadas.

Darbo tyrimo objektas – automatinės skaitmeninio teksto santraukų sistemos (angl. *summarizer*).

Darbo metodai – aprašomasis, lyginamasis.

2. Skaitmenizuoto teksto automatinio rengimo santraukos

Šiame baigiamojo darbo skyriuje aptariamos santraukų sistemos bei santraukų tipai. Aprašomi dažniausiai naudojami santraukų rengimo metodai, kurie priskiriami atitinkamoms santraukų rengimo sistemoms. Pateikiama 9 santraukų sistemų bei jų sugeneruotų 54 santraukų analizė bei įvertinamas skaitmenizuoto teksto santraukų sistemų efektyvumas, naudingumas ir kokybė.

2.1. Santraukų automatinio rengimo problematika

Esant dideliam informacijos kiekiui interneto erdvėje iškilo poreikis sukurti sistemą, kuri padėtų atrinkti pagrindinę, svarbiausią informaciją. Tokios sistemos, kurios padeda atrinkti informaciją vadinamos santraukų sistemomis (angl. *summarizer*). Santraukų sistemos kuriamos taip, kad padėtų atrinkti, nufiltruoti pagrindinę informaciją. Šios sistemos informacijai atrinkti naudoja įvairius metodus – sakinio ilgio, dažniausiai pasitaikančių žodžių, reikšminių žodžių. Santraukų sistemos yra ne tik skirtos tekstui, bet ir kitai informacijai, tokiai kaip vaizdo ar garso informacijai, sutrumpinti. Šiuo metu labiausiai išvystytos yra teksto santraukų sistemos, kurios kuriamos padėti žmonėms atsirinkti informaciją, nufiltruoti nereikalingus duomenis, kuriuos skaitant tik veltui eikvojamas laikas.

Kita santraukų sistemų problema yra ta, jog sukurtos sistemos dažniausiai yra orientuotos į tam tikrą kalbą, t.y. arba į kūrėjų gimtąją kalbą, arba populiariausias, labiausiai paplitusias kalbas – anglų, prancūzų, ispanų. Kitomis kalbomis, tokiomis, kaip suomių ar estų, sukurtos sistemos taip pat orientuotos būtent į tas kalbas, todėl iškyla kalbos suderinamumo kalba. Konkretinant problemą – priderinti santraukų sistemas ar jų įrankius, ar sukurti santraukų sistemą, skirtą arba priderintą lietuvių kalbai.

Kuriant santraukų sistemas dažnai susiduriama su kalbos atpažinimo problemomis. Automatinė skaitmenizuoto teksto santraukų rengimo sistema turėtų atpažinti lietuviškus simbolius. Kita problema, su kuria susiduriama - santraukų rengimo kriterijai, įvairios nuostatos, kurios reikalingos kurti santraukai. Nuostatų, kriterijų apsibrėžimas yra vienas pirmųjų etapų siekiant suformuoti sistemą, gebančią sugeneruoti naudingą santraukų sistemą. Tam, kad būtų išspręsta kriterijų ir nuostatų problema reikia tiksliai apibrėžti kokiems tekstams yra skirta santraukų sistema: ar visų stilių tekstams ar tik vieno konkretaus stiliaus. Jei nutariama sistemą

skirti visų tipų tekstams, reikia atrinkti metodus, kurie tinkamai atrinktų sakinius atspindinčius teksto prasmę į santrauką. Bet kokiu atveju, santraukų rengimo sistema turi skaičiuoti sakinius, pagal galimybes skaičiuoti sakinyje esančius žodžius ir pagal tam tikrus kriterijus, pvz. sakinio ilgį, atrinkti sakinius. Kiekvienas teksto žanras pasižymi skirtingu vidutiniu žodžių sakinyje kiekiu, todėl reikia apibrėžti konkrečius skaičius, kuriais remiantis, veiks programa. Žinoma, jei sistema remiasi žodžių skaičiumi sakinyje. Taip pat galima taikyti antrą sprendimo variantą, kai visiems tekstams, nesvarbu kokio žanro jie bebūtų, nustatyti bendrą sakinio ilgio skaičių, atrinkti raktinius žodžius, kurie išrinkti remiantis statistika bei kitus kriterijus, kurie reikalingi, jog santraukų sistema funkcionuotų normaliai, be problemų. Neretai tokios sistemos naudoja ir papildomą informaciją apie žodžius, kurie sakiniuose neįgauna jokios prasmės. Bet kokiai kuriamai sistemai nėra privalomi visi kriterijai, todėl neretai naudojama sakinio ilgio ir raktinio žodžio kriterijai.

Svarbia problema būtų galima įvardinti vartotojo sąsają. Dažniausiai santraukų sistemos būna dviejų tipų: diegiamos ir nediegiamos. Nediegiamos sistemos yra įkomponuojamos į tinklalapius, kuriuos interneto vartotojas gali lengvai pasiekti. Jei programa kuriama taip, jog ją reikėtų įdiegti automatiškai atsiranda grafinės sąsajos bei suderinamumo problemos. Reikia atsižvelgti ir į tai, jog programą esant reikalui būtų galima pataisyti ir vartotojui tai nesukeltų nepatogumų. Jei santraukų sistema įkomponuojama į internetinį tinklalapį reikia nepamiršti, jog ir šioje erdvėje sistema turi turėti tam tikrą apipavidalinimą, atlikti jai priskirtas funkcijas bei nuolat stebėti ar internetinis tinklalapis veikia taip, kaip ir turėtų veikti.

Santraukų sistema pasinaudodama įvairiais metodais bei funkcijomis pateikia originalaus teksto santrauką, kuri padeda vartotojui taupyti laiką. Tikslui pasiekti panaudojamos funkcijos, metodai, kurie skaičiuoja sakinių skaičių, sakinio ilgį, lygina su programoje gautais sakinio ilgio skaičiais, atrenka sakinius. Rezultatas, kuris turėtų tenkinti vartotoją – sugeneruota santrauka, pateikta programos arba interneto naršyklės lange. Jei reikalingas produkto įvertinimas, sukuriama anketa, ir atliekamas tyrimas. Tyrimo metu įvertinimas santraukų sistemos efektyvumas ir naudingumas.

2.2. Santraukų tipai

Santraukos, kaip ir daugelis tekstų gali būti skirstomi į tam tikrus tipus. Dažniausiai yra išskiriamos ***tiesioginės*** arba ***nurodomosios*** santraukos (kurios atskleidžia pagrindinę teksto mintį nepateikdamos turinio) ir ***informatyviosios*** santraukos (pateikiančios trumpesnę turinio versiją).

Santraukoms kurti yra naudojamos **ištraukos** (žodžiai, sakiniai ir kiti teksto elementai) bei **konspektai**. Vadinamosios **ištraukos** daugeliu atvejų pakartotinai panaudojamos paimant jas iš įvesto teksto, o **konspektai** yra sukuriami iš naujo generuojant išrinktą turinį. Ištraukų tipo santrauką galima bandyti apibrėžti taip: „tai tokia santrauka, kai teksto dalys yra nukopijuotos nuo originalaus teksto“. Konspekto tipo santrauka – kitaip nei ištraukos tipo santrauka, turi papildomos informacijos, nesančios originaliame tekste, todėl ją apibrėžti yra pakankamai sunku. (Mani 2001:6)

Antras santraukų grupavimo būdas yra įprastinių skirtumų tarp tiesioginių ir informatyvių santraukų radimas. Tiesioginės santraukos atlieka informacijos funkciją pasirenkant dokumentus gilesniam skaitymui. Šio tipo santraukos yra skirtos padėti vartotojui nuspręsti ar skaityti duotąjį informacijos šaltinį ar ne. Priešingai nei tiesioginė santrauka, informatyvioji santrauka aprėpia visą tyliąją informaciją šaltinyje tame pačiame detalumo lygyje. Skirtumai tarp tiesioginių ir informuojančiųjų santraukų dažnai yra išplečiami į skirtumus tarp informuojančiųjų ir kritiškai įvertinamųjų santraukų. Kritiškos santraukos vertina temos aktualumą šaltinyje, darbo kokybę. Šios santraukos įtraukia atsiliepimus, taip pat įtraukiamos ir nuomonės, atsiliepimai, silpnybių nustatymas, rekomendacijos ir t.t., viskas, kas yra randama už šaltinio (Mani 2001:6-7).

Pačios SARS gali būti skirstomos į dvi kategorijas:

- internetines, kai tiesiog pakanka nukopijuoti norimą tekstą, ar puslapio adresą, kuriame yra dominantis fragmentas, pasirinkti norimą sakinių kiekį, ar santraukos dydį procentais ir paspausti vieną mygtuką;
- įdiegiamas kompiuteryje, kurios turi tokias pat galimybes kaip ir internetinės. Žinoma, dalis tokio tipo santraukų kūrimo programų yra mokamos, kai kada pateikiama santrauka gali būti tik kita spalva išskirtas tekstas.

SARS paruoštos santraukos pagal paruošimo būdą gali būti skirstomos į dvi grupes:

- bendrąsias;
- reikšminiais žodžiais pagrįstomis.

Bendrosios kompiuterizuotos teksto santraukos technologijos ir sistemos santrauką kuria naudodamosi programiniais kodais, parašytais taip, kad būtų atrenkami svarbiausi sakiniai, kuriose atsispindi pagrindinės teksto mintys. Reikšminiais žodžiais pagrįstos teksto santraukos technologijos ir sistemos veikia kiek kitaip nei bendruoju principu veikiančios sistemos. Reikšminiais žodžiais pagrįstos sistemos vadovaujasi tuo, kokius žodžius santraukoje nori matyti vartotojas. Šioje sistemoje vartotojas įveda reikšminius žodžius, kurie jį domina, todėl sakiniai su

raktiniais žodžiais automatiškai pakliūna į rengiamos santraukos tekstą, kurį pateikia vartotojui. (Mani 2001: 26).

2.3. Automatinių santraukų sistemų prototipai

Internete šiuo metu yra daugiau nei 10 automatinių santraukų sistemų programų, kurios gali būti kaip būsimos sistemos prototipas.

Automatic text summarizer. Apdoroja įvairių sričių tekstus. Santrauka rengiama pagal sakinių skaičių (šiuo metu sistema neveikia). Kokybiniu atžvilgiu gaunamas rezultatas ne visada gali tenkinti vartotoją. Parengtoje santraukoje paprastai vyrauja ilgiausi originalaus teksto sakiniai.

GreatSummary. Apdoroja įvairių sričių tekstus. Kaip ir **Automatic text summarizer** santrauką rengia pagal sakinių skaičių. Į metodų tyrimą ši sistema neįtraukiama ir plačiau neanalizuojama.

Copernic summarizer. Santraukoms generuoti naudoja statistinius (pagal žodžio ar žodžių dažnumą) ir lingvistinius (pagal ieškomą žodį) algoritmus. Galima taikyti įvairiems angliškiems tekstams, tačiau kitų kalbų tekstams nėra pilnai pritaikyta.

Intellexer summarizer. Santraukoms generuoti kaip ir **Copernic summarizer** naudoja statistinius ir lingvistinius metodus. Rekomenduojama taikyti ekonomikos, politikos, bendrojo tipo, mokslinio tipo tekstams. Kai kurių kalbų tekstams ši sistema nėra pritaikyta.

Pertinence summarizer. Gali apdoroti tekstus, kurių žanras nėra aiškus, taip pat chemijos, finansų, teisės, spaudos, telekomunikacijų, medicinos tekstus. (Šiuo metu sistemos internetinis puslapis neveikia). Kalbos atpažinimo problemų nepasitaiko.

Tool4Noobs. Apdoroja įvairių sričių tekstus. Sistema daug kuo panaši į **Automatic Text Summarizer, GreatSummary** sistemas. Ši sistema išsiskiria iš kitų sistemų tuo, jog galima rinktis ar santrauka bus gaunama pagal glaudinimą, ar pagal eilučių skaičių. Taip pat šioje sistemoje galima pasirinkti ir minimalius sakinio bei žodžio ilgius, kurie gerokai pagerina santraukos kokybę. Dar vienas paminėtinas dalykas, jog ši sistema pateikia ir dažniausiai pasitaikančių žodžių sąrašą, kuris pakankamai gerai iliustruoja santrauką.

QuickJist summarizer. Gali apdoroti bet kokio žanro ar stiliaus tekstus. Santrauka rengiama pagal procentinį sakinių skaičių.

Shvoong summarizer. Gali apdoroti įvairaus žanro tekstus. Santrauka rengiama pagal procentinį sakinių skaičių, kuris taip pat remiasi sakinio ilgiu.

Subject search summarizer. Gali apdoroti įvairaus žanro ir stiliaus tekstus. Santrauka rengiama pagal žodžių kiekį.

Microsoft Office MS Word 2003 AutoSummarize. Microsoft kompanijos **MS Office Word** santraukos rengimo funkcija. Teksto žanras ar stilius neturi reikšmės. Pagrindinis reikalavimas, kuris sukelia ir nepatogumus – tekstas turi būti anglų kalba. Naudoja statistinius algoritmus.

2.4. SARS kategorijos ir jų palyginimas

Iš tyrimui pasirinktų sistemų – penkios atvirosios ir nemokamos sistemos (*Automatic Text Summarizer, Pertinence Summarizer Online, Shvoong Summarizer, Subject Search Summarizer, Tools 4 Noobs*), trys demonstracinės mokamų sistemų versijos (*Copernic Summarizer, Intellexer Summarizer, QuickJist summarizer*) bei *MS Office Word 2003 AutoSummarize*, kuri įdiegiama įdiegus *MS Office Word 2003*. Pirmosios sistemos yra diegiamos į kompiuterį, tai – *Copernic Summarizer, Intellexer Summarizer, QuickJist summarizer, Subject Search Summarizer*, prie šių sistemų taip pat galima priskirti ir *MS Office Word 2003 AutoSummarize AutoSummarize* (žr. 2.1 lentelė). Likusios trys sistemos – veikia tik internete[1]. Visas šias santraukų kūrimo sistemas bendrai galima grupuoti į dvi grupes:

1. Pagal kainą:
    a. Nemokamos;
    b. Mokamos;
2. Pagal įdiegimą:
    a. Įdiegiamos;
    b. Neįdiegiamos.

2.1 lentelė. Santraukų sistemų palyginimas

Sistema	Kaina	Įdiegimas	Dydis
Automatic Text Summarizer	Nemokama	Neįdiegiama	-
Copernic Summarizer	59.95$	Įdiegiama	~5 MB
Intellexer Summarizer	59.95$	Įdiegiama	~ 10 MB
MS Office Word 2003 AutoSummarize	Mokama	Įdiegta	-

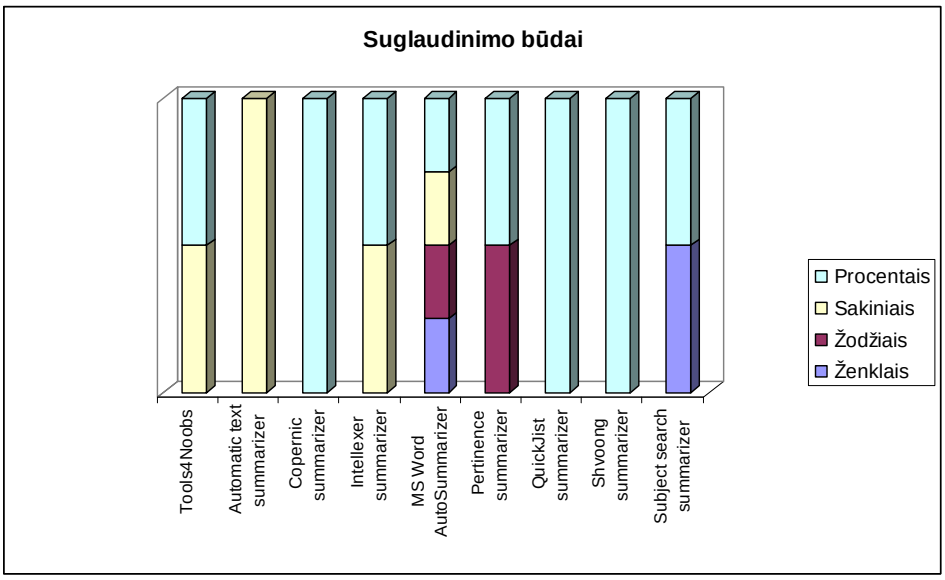
[1] 2011-05-17 tikrinimu internetinis Pertinence Summarizer Online tinklalapis neveikia.

Tools 4 Noobs	Nemokama	Neįdiegiama	-
Pertinence Summarizer	Nemokama[2] / 59.68 €	Neįdiegiama/įdiegiama	- /neskelbiama
QuickJist Summarizer	9.95$	Įdiegiama	578 KB
Shvoong Summarizer	Nemokama	Neįdiegiama	-
Subject Search Summarizer	Nemokama	Įdiegiama	1,2 MB

Teksto santraukų glaudinimas (santraukos ilgis)

Kiekviena automatinės santraukos sistema turi suspaudimo (kompresijos) lygį. Dauguma šių sistemų turi vienodus suspaudimo (kompresijos) lygius. Kai kurios iš jų glaustumo lygį vertina procentais, kitos – žodžių ar sakinių skaičiumi (žr. į 2.1 pav.). Todėl santraukų kūrimo sistemos pagal glaudinimo būdą skirstomos pagal:

1. Glaudinimą procentais;
2. Glaudinimą sakiniais;
3. Glaudinimą žodžiais;
4. Glaudinimą ženklų skaičiumi;
5. Mišrios.



2. 1 pav. Santraukų sistemų teksto suglaudinimo metodai

Dažniausias ir populiariausias teksto santraukos glaudinimo matas yra procentinis. Procentiniu glaudinimu galima vadinti tokį glaudinimą, kai originaliam pradiniam tekstui yra parengiama norimo procentinio dydžio santrauka. Kai kurios santraukų sistemos, kurios

[2] Naudota nemokama Pertinence summarizer internetinė versija. Versijos, kuri diegiama, kaina yra parašyta už pasvyrojo brūkšnio.

naudojasi tokiu glaudinimo būdu leidžia pasirinkti ir įvairesnius suglaudinimo lygius, tokius kaip 6%, 12% ir kitokius, kokių pageidauja vartotojas. Dažniausiai tokius papildomus pasirinkimus turi mokamos ir pakankamai aukštos kokybės santraukų sistemos.

Procentiniu glaudinimu paremtos yra:

- Copernic Summarizer;
- Intellexer Summarizer;
- Pertinence Summarizer;
- QuikJist Summarizer;
- Shvoong Summarizer;
- Tools 4 Noobs
- MS Office Word AutoSummarize.

Dar vienas populiarius santraukos glaudinimo būdas yra santraukos ilgio nustatymas sakiniais. Dalis santraukų sistemų leidžia pasirinkti sakinių skaičių, iš kurių bus sudaryta santrauka. Glaudinimą sakinių skaičiumi naudoja:

- Automatic Text Summarizer;
- Intellexer Summarizer;
- Subject Search Summarizer;
- Tools 4 Noobs (atskirais atvejais);
- MS Office Word AutoSummarize.

Vartotojas pats pasirenka kiek sakinių nori matyti iš tam tikro teksto – tai gali būti nuo 3 sakinių iki 100 ir daugiau, priklausomai nuo teksto apimties. Toks santraukos rengimo būdas yra patogus siekiant gauti kuo trumpesnę santrauką, tačiau dėl to neretai nukenčia santraukos informatyvumas ir kokybė.

Kitas populiarus būdas, kai vartotojas santraukos ilgį pats pasirenka pagal žodžių skaičių. Paprastai tokioms santraukoms parengti siūlomi 100, 250, 1000 bei kitokie žodžių skaičiaus variantai. Yra santraukų sistemų, kurios leidžia pačiam vartotojui nustatyti norimą žodžių skaičių būsimoje santraukoje. Santraukos rengimas pagal žodžių skaičių yra panašus į santraukos parengimą paremtą sakinių skaičiumi. Santraukų pagal žodžių skaičių glaudinimo būdą naudoja dvi santraukų rengimo sistemos - *Pertinence Summarizer* ir *MS Office Word AutoSummarize*. *Pertinence Summarizer* šį būdą naudoja, nes suglaudžiant tekstą procentiniu teksto glaudinimo metodu automatiškai pateikiamas ir žodžių skaičius, kuris atitinka procentinę žodžių išraišką. *MS Office Word AutoSummarize* tiesiog leidžia pasirinkti iš pateikiamų sakinių skaičiaus mums

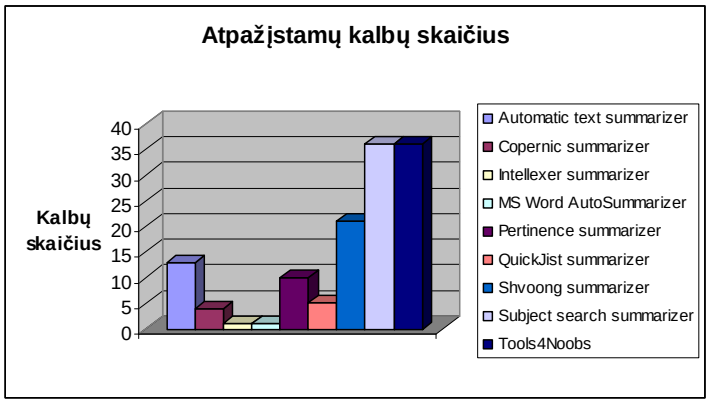
tinkamą skaičių sakinių. Tačiau jei santraukos rengimas sakiniais mūsų netenkina, galima rinktis procentinį ar žodinį būdus.

Ne toks paplitęs būdas yra santraukos rengimas pagal ženklų skaičių. Jį naudoja tik kelios sistemos. Šiuo būdu rengiamos santraukos paprastai yra nuo 1000 ženklų iki 90000 ženklų. Taip pat rengiant santrauką šiuo būdu galima pasirinkti neribotą ženklų skaičių, tačiau tada jau nebebus kuriama santrauka, o tik bus atkartojamas tekstas. Tam, kad būtų sukurta santrauka atitinkanti tam tikrą procentinę išraišką patartina žinoti kiek originaliame tekste yra spaudos ženklų. Santraukas glaudina pagal spaudos ženklų skaičių - *Subject Search Summarizer* ir *MS Office Word AutoSummarize*. Didžiausias skirtumas tarp šių sistemų yra tai, kad *MS Office Word AutoSummarize* leidžia pasirinkti *100 arba mažiau žodžių* ir *500 arba mažiau žodžių*.

Kaip parodė tyrimai, dalis santraukų rengimo sistemų turi du glaudinimo būdus iš kurių vartotojas gali pasirinkti jam tinkamiausią. Dažniausia jos naudoja procentinį glaudinimo būdą, bet šalia pateikiama glaudinimo sakiniais ar žodžiais galimybė. Reikia paminėti ir tai, kad *Pertinence Summarizer* pateikdama santrauką procentais automatiškai nurodo ir kiek žodžių parengtoje santraukoje. *Tools 4 Noobs* sistema pateikdama santrauką pateikia dažniausiai pasikartojusius žodžius, ir kiek kartų jie pasikartoja santraukoje. *MS Office Word AutoSummarizer* nurodo ne tik parengtos santraukos ilgį sakiniais ir žodžiais, bet ir pradinio dokumento ilgį taip pat išreikštą sakiniais ir žodžiais, bei suteikia daugiau santraukos glaudinimo būdų, nei kitos santraukų sistemos.

Kalbos atpažinimas

Santraukų rengimo sistemos taip pat atsižvelgia į originalaus teksto kalbą. Beveik visos santraukų rengimo sistemos atpažįsta daugiau nei vieną kalbą (žr. į 2.2 pav.). Daugiausiai kalbų atpažįsta ir palaiko *Subject Search Summmarizer* – 36 kalbas, tarp jų ir lietuvių kalbą. Pagrindinis jos, kaip ir kitų sistemų, reikalavimas, kad hieroglifai (kinų, japonų kalbų rašmenys) būtų užrašomi lotyniškų alfabetu, kitaip santrauka nebus padaroma, kadangi sistema nesupras parašyto teksto.



2.2 pav. Kalbų, kurias palaiko santraukų sistemos, palyginamoji diagrama

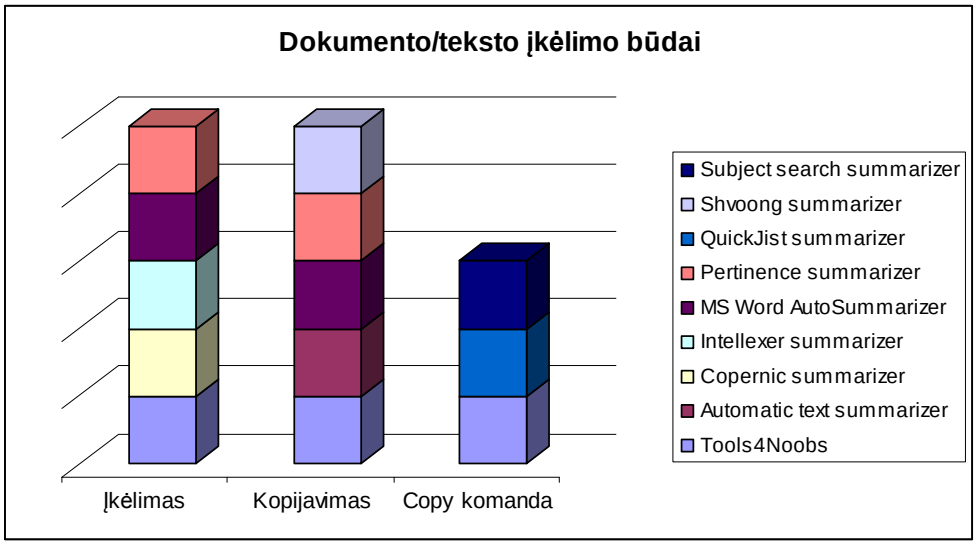
Mažiau kalbų atpažįsta *Shvoong Summarizer* – jis atpažįsta 21 kalbos tekstus. Šis sistema taip pat atpažįsta ir lietuvių kalba rašytus tekstus. Trečia pagal atpažįstamų kalbų skaičių yra *Automatic Text Summarizer* atpažįsta ne tik 13 kalbų tekstus, bet taip pat ir įvairių kitų kalbų tekstus parašytus lotynišku alfabetu. *Tools 4 Noobs* sistema nenurodo konkretaus atpažįstamų kalbų skaičiaus, tačiau iš atlikto tyrimo galima teigti, jog santraukos pateikiamos iš visų tekstų, kurie yra parašyti lotynų abėcėle, bei lietuvių kalbos abėcėle. *Pertinence Summarizer* atpažįsta 10 kalbų tekstus, tarp jų ir lietuvių kalbos tekstus. *Copernic Summarizer* atpažįsta tik 4 kalbų tekstus, o neatpažintus ženklus ši sistema paverčia neatpažįstamais ženklais. *QuickJist Summarizer* atpažįsta viena kalba daugiau nei *Copernic Summarizer* – 5 kalbas. Radęs neatpažįstamų ženklų tekste *Copernic Summarizer* tuos ženklus paverčia neįskaitomais ženklais. Taip pat ši sistema gali sugeneruoti santraukas ir kitiems tekstams, parašytiems lotynišku alfabetu. Vienas iš mažiausiai kalbų atpažįstančių *Intellexer Summarizer*, kurio pagrindinė ir vienintelė kalba yra anglų kalba. Bet tai tik tuo atveju, jei naudojama *English* versija. Be angliškosios *Intellexer Summarizer* versijos taip pat yra prancūzų ir vokiečių, kurios santraukas atlieka taip pat tikr viena kalba. Radęs neatpažįstamų ženklų tekste *Intellexer Summarizer*, kaip ir *Copernic Summarizer* ženklus gali paversti neatpažįstamais ženklais. Prieduose pateikta lietuviško teksto santraukos ir jų prisitaikymas prie lietuviškų rašmenų ir gautų santraukų fragmentai. *MS Office Word AutoSummarize* taip pat tekstą atpažįsta tik tada, kai tekstas yra angliškas arba teksto kalba pakeičiama į anglų kalbą, paliekant originalo rašmenis. Kitu atveju *MS Office Word AutoSummarize* nepateikia teksto santraukos.

Teksto įkėlimas

Kiekviena iš nagrinėtų santraukų parengimo sistemų tekstą įkelia skirtingais būdais. Jų patogumą gana sunku nustatyti, nes tai priklauso nuo vartotojo pasirinkimo ir poreikio.

Santraukų rengimo sistemos tekstus skaito keliais būdais:

- Įkeliamas dokumentas, kurio santrauką norima gauti;
- Nukopijuojamas dokumentas, kurio santrauką norima gauti;
- Pažymimas teksto turinys ir paspaudžiama komanda COPY arba greitųjų klavišų kombinacija Ctrl+C;
- Atidaromas Word dokumentas, kurio santrauką norime gauti.



2.3 pav. Teksto, skirto generuoti santraukai, įkėlimo būdai į santraukos sistemas

Automatic Text Summarizer ir *Shvoong Summarizer* (žr. į 2.3 pav.) tekstą atpažins ir panaudos tik tada, jei nukopijuojamas į naršyklės langą. *Copernic Summarizer* ir *Intellexer Summarizer* tekstą nuskaitys iš Word‘o ar kito tekstinio dokumento, taip pat ir internetinio puslapio. *Pertinence Summarizer* santrauką pateiks keliais atvejais: jei Word‘o ar tekstinis failas bus įkeltas į internetinį sistemos langą ar nukopijuotas. *Pertinence Summarizer* taip pat suteikia galimybę parengti ir internetinio puslapio santrauką, tereikia kaip ir į *Intellexer Summarizer* ar *Copernic Summarizer* langus įrašyti ar nukopijuoti internetinio puslapio adresą. O *Subject Search Summarizer* ir *QuickJist Summarizer* santraukas pateiks tik tuo atveju, jei bus pažymėtas visas tekstas ir pasirinkta COPY komanda arba greitųjų klavišų Ctrl+C kombinacija. *MS Office Word AutoSummarize* santrauką pateiks meniu eilutėje pasirinkus *Tools – AutoSummarize* (liet. –

Įrankiai – Automatiškai apibendrinti). *Tools 4 Noobs* leidžia tekstą įkelti dviem būdais: įkopijuoti į naršyklės langą arba nurodyti tinklalapio adresą.

2.5. Santraukų sistemų teksto analizės ir apdorojimo metodai

Modeliuojant bei kuriant santraukų sistemą turi būti atsižvelgiama į tai, kokiais būdais vadovaujasi žmogus rengdamas apibendrinimus. Žmogiškojo faktoriaus atveju remiamasi tiek lingvistikos, tiek filosofijos, tiek kitomis sritimis, kurios gali būti susijusios su tekstu. Tačiau kompiuterizuotai santraukų rengimo sistemai remtis šiais dalykais yra gana sunku – tai reikalauja daug papildomos informacijos bei išteklių, atitinkamai susiduriama su tokiomis problemomis kaip tinkamos realizacijos nebuvimas. Kartais tokios universalios sistemos tampa per ne lyg sudėtingomis ir paprastam vartotojui nesuprantamos. 2.4 pav. pateikiama bendroji santraukų rengimo schema. Žmogus rengdamas santrauką remiasi turimu tekstu ir turimomis žiniomis, atsirinkdamas jam tinkamą informaciją, tą patį stengiasi padaryti ir automatinės teksto santraukų rengimo sistemos, bandydamos imituoti arba atspėti tinkamą santraukos esmę.

Tekstui apdoroti neretai taikomas statistika paremtas modelis – renkami statistiniai duomenys apie žodžių vartojimą, ir sukauptų duomenų pritaikymas teksto apdorojimui. Santraukų sistemai dar reikia ir iš anksto sukurto semantinio tinklo – terminų sistemos su aprašymais ir tarpusavio ryšiais. Kuo „protingesnė“ sistema, tuo santrauka geresnė. Pereinant į kitą teksto analizės lygmenį siekiama atrinkti dažniausiai pasikartojančius terminus, jų grupes, remiantis įvairiais aspektais – ar terminai yra vienas šalia kito, ar ne, remiasi žodžių ar žodžių junginių žodynais bei lingvistiniais įrankiais (tekstynais ir pan.).

Žodžių statistika yra svarbus etapas, be kurio, tam tikrais atvejais, sunku gauti gerą santrauką, nes būtent žodžiai ir padeda perteikti pagrindinę originalaus teksto prasmę. Pagrindiniai santraukos parengimo ir rengimo būdai bei metodai atliekami lingvistiniame lygyje. Paprastai galima išskirti ir atskirti šiuos būdus:

1. **Paviršutinis** santraukos rengimo būdas neišeina už sintaksinio atvaizdavimo ribų, tačiau skirtingi elementai gali būti atvaizduojami skirtingais lygiais.

2. **Giluminis** santraukos rengimo būdas yra sakinio semantinio lygmens reprezentacija.

Santraukai generuoti ir formuoti paprastai naudojami šie metodai:

1. **Pozicinis arba vietos metodas.** Daugumos žanrų tekstai paklūsta būtent tam žanrui priskiriamoms taisyklėms (antraštės, pavadinimai, pastraipos ir pan.). Naudodama šį metodą sistema vadovaujasi principu – kas yra pavadinime, antraštėje ir/ar pirmoje pastraipoje –

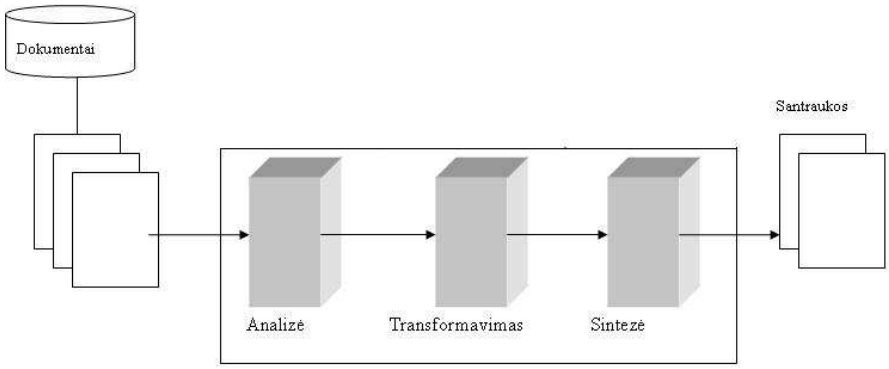
didžiausią svarumą turintys dalykai/žodžiai. Kitose pastraipose esanti informacija savo svarbumą praranda, nes paprastai esminė ir reikalingiausia informacija pateikiama teksto pradžioje, o teksto pabaigoje neretais atvejais gali būti pateikiama informacija, kuri visiškai nesusijusi su siekiamu/pageidaujamu gauti rezultatu. Pozicinį metodą galima derinti su raktinio žodžio metodu.

2. **Raktinio/reikšminio žodžio metodas**. Įvairiuose žanruose tam tikri žodžiai ir frazės aiškiai parodo svarbumą. Tokie žodžiai vadinami raktiniais arba reikšminiais žodžiais. Raktiniais žodžiais galime vadinti kuriam nors žanrui būdingus žodžius ar žodžių junginius (pvz.: „reikšmingas“, „šiame darbe“, „svarbiausias“, „pavyzdžiui“ ir pan.). Remiantis įvairiomis sistemomis galima teigti, jog sakiniai su tokiais žodžiais turėtų būti paimti į santrauką.

3. **Žodžio ir frazės dažnumo metodas**. Šis metodas pagrįstas besikartojančių žodžių dažnumu. Kuo dažniau žodis kartojasi, tuo svarbesni sakiniai su tais žodžiais. Šiam metodui pritaikyti, kaip ir kitiems metodams, reikalingas kiekvieno teksto elemento (šiuo atveju žodžio(-ių)) lyginimas. Atvejais, kai tekste daug vienodų jungtukų, jaustukų, toks metodas neefektyvus, todėl patartina rašant kodą, užfiksuoti, kokie teksto elementai nebus skaičiuojami.

4. **Reikšminių žodžių (užklausos) ir dalinio pavadinimo sutapimo metodas**. Paprastas metodas, kuris pažymi sakinį pagal reikalingų žodžių kiekį sakinyje. Reikalingi žodžiai yra tokie, kurie yra teksto pavadinimuose, antraštėse, ar sutampa su vartotojo užklausoje įvestais žodžiais (reikšminiais žodžiais). Jei tekste yra tokių atitikmenų – santraukos sistemos pateikia juos kaip rezultatus.

5. **Rišlumo arba leksinio sujungimo metodas.** Žodžiai gali būti sujungiami įvairiais būdais, įskaitant kartojimą, sinonimiką, semantinius ryšius, kurie išreiškiami žodynuose. Naudojantis šiuo metu pasitelkiamas papildomas informacijos šaltinis (žodynas), ir atrenkami atitinkami sakiniai. Neretai šiuo metodu parengtų santraukų sakiniai būna ilgiausi originalo sakiniai. Galima teigti jog į santrauką sakiniai priimant, jog ilgiausi santraukos sakiniai yra svarbiausi bei informatyviausi tekste.



2.4 pav. Santraukų automatinio rengimo struktūrinė seka

Santraukų sistemos, kurios naudoja:

- Reikšminio žodžio metodus - *Microsoft Office Word 2003 AutoSummarize, SubjectSearchSummarizer, Copernic Summarizer, Pertinence summarizer* ir *Intellexer Summarizer*.
- Rišlumo arba leksinio sujungimo (šiuo atveju – atrenkami ilgiausi sakiniai) metodą - *Automatic text summarizer, Shvoong summarizer, Microsoft Office Word 2003 AutoSummarize, QuickJist Summarizer, Tools 4 Noobs (*atskirais atvejais).
- Reikšminių žodžių (užklausos) ir dalinio pavadinimo sutapimo metodą naudoja *Pertinence Summarizer, Intellexer summarizer, Copernic Summarizer, SubjjectSearch Summarizer, Tools 4 Noobs.*

Akivaizdu, jog tos sistemos, kurios naudoja daugiau nei metodą santraukoms parengti yra geresnės nei sistemos, kurios naudoja tik vieną metodą. Naudodamos du ar keletą metodų, sistemos gali parengti geresnę santrauką, kuri geriau atspindi teksto prasmę ir vartotojui yra geresnė kokybės atžvilgiu, tačiau tokias sistemas yra sunkiau realizuoti. Sistemos, kurios naudoja vieną metodą nusileidžia kokybės atžvilgiu, kadangi atitinkamai pagal naudojamą metodą gauta santrauka gali būti prastesnės kokybės, gali nutikti taip, kad nebus perteikta pagrindinė teksto mintis, kas nulems, jog santrauka bus netinkama.

Mokslininkai, kurie tiria bei kuria santraukų sistemas paprastai išskiria 3 žingsnius, kurie atliekami reziumuojant tekstą. Pirmasis žingsnis vadinamas *analizės žingsniu/lygiu*. Šiame lygyje analizuojamas tekstas ir nustatoma pagrindinė teksto mintis arba turinys. Neretai šiame lygmenyje naudojami vadinamieji analizatoriai (angl. *parsers*). Įvairūs analizatoriai taikomi,

siekiant gauti kuo geresnės kokybės, aiškesnį ir nuo temos nenukrypusį tekstą. Yra atvejų, kai santraukų sistemos šiame žingsnyje naudoja tekstynus. Antrasis žingsnis vadinamas *išrinkimo žingsniu/lygiu*, kadangi šiame lygyje pasitelkiamos priemonės, padedančios nustatyti, kas yra svarbiausia analizuojamame tekste. Šiame lygyje naudojami įvairūs metodai, kurie padeda atrinkti santraukai reikalingus sakinius pagal žodžių dažnumą, sakinio ilgį, raktinius žodžius ar kitus kriterijus, kuriais remiantis kuriama sistema. Šiame žingsnyje vienas iš elementų yra santraukos ilgis, kuris neretai ir nulemia galutinį rezultatą (kuo trumpesnė santrauka, tuo didesnė tikimybė, jog ji nebus informatyvi). Paskutinis, trečiasis žingsnis/lygis vadinamas *sinteze*, kadangi šiame etape visa tai, kas buvo atliekama pirmame ir antrame lygyje yra sujungiama ir pateikiama vartotojui jam suprantama ir tinkama forma – programos arba naršyklės lange. Šiame žingsnyje svarbiausia tinkamai sujungtus metodus apjungti ir pateikti į ekraną. O gautojo rezultato kokybę sprendžia vartotojas.

2.6. Santraukų analizė

Tiriant gautas tekstų santraukas pagal kompresijos lygį gautų santraukų ilgis sakinių skaičiumi yra skirtingas angliškiems ir lietuviškiems tekstams. Naudojant 50% kompresijos lygį angliško teksto santraukos ilgis skiriasi nuo 2 sakinių iki 36 sakinių. Ilgiausias angliško mokslinio teksto santraukas parengia *Copernic Summarizer* ir *Intellexer Summarizer*, o trumpiausią – *Tools4Noobs summarizer* bei *QuickJist Summarizer* (žr. į 2.2 lentelę). Ilgiausias santraukas *Copernic Summarizer* ir *Intellexer Summarizer* parengia lietuviškiems ir angliškiems tekstams. Iš tyrime naudotų sistemų *Shvoong Summarizer* vienintelė santraukų sistema, kuri nepateikė mokslinio teksto santraukos naudojant 50% bei 30% kompresiją, nors kitų lietuviškų tekstų santraukos buvo pateikiamos ir jokių kliūčių neiškildavo. Ši sistema taip pat nepateikė ir kitokios kompresijos lietuviško mokslinio teksto santraukų. Kadangi *Automatic Text Summarizer* ir *Subject Search Summarizer* negali atlikti procentinės kompresijos, sakinių skaičius, kuris sudaro santrauką buvo gautas paėmus 50% originalaus teksto sakinių, kurie kai kuriais atvejais skiriasi nuo kitų SARS pateikiamų sakinių santraukoje skaičiaus. *Subject Search Summarizer* pateikia papildomus sakinius tinkamus santraukai, reikšminius žodžius ar kitą papildomą teksto informaciją, todėl santraukos ilgis, gali būti ilgesnis 2-3 sakiniais nei pasirinktas pradinis santraukos ilgis. Iš 2.2 lentelėje pateikiamų duomenų galima teigti, jog angliškų tekstų santraukų ilgiai sakiniais santraukų sistemose yra vienodi, arba skiriasi nežymiai. Kitaip nei anglų kalbos santraukose, lietuvių kalbos santraukos ilgis sutampa tik publicistinio teksto santraukose, o mokslinio ir grožinio tektų santraukos ilgiai skiriasi, nors naudota tokia pati kompresija ir tas

pats tekstas. Taip yra todėl, kad lengviau pateikti trumpesnio ir paprastesnio teksto santraukas, nei ilgesnio ir sudėtingesnio. Žinoma, santraukų ilgiai skiriasi nežymiai, bet nebuvo sistemų pateikiančių vienodo ilgio santraukas, kas pasitaikė atliekant angliškų tekstų suglaudinimą.

Vidutinė angliško grožinio teksto santrauka tyrimo metu, naudojant 50% kompresijos lygį, apytiksliai lygi 38,75 sakinio, angliško mokslinio teksto vidutinis santraukos ilgis apytiksliai lygus 29,88 sakinio, angliško publicistinio teksto santrauka apytiksliai lygi 14,44 sakinio. Vidutinis lietuviško grožinio teksto santraukos ilgis naudojant 50% kompresijos lygį yra 53,33 sakinio, mokslinio – apytiksliai 30,75 sakinio, o publicistinio – 6,25 sakinio. Nors originalus angliškas grožinis tekstas buvo ilgesnis nei lietuviškas grožinis tekstas, tačiau pagal gaunamus skaičiavimus matosi, jog lietuviška grožinio teksto santrauka yra ilgesnė nei atitinkamo angliško teksto santrauka, nors naudojami tokie patys metodai informacijos generavimui ir pateikimui.

2.2 lentelė. Teksto santraukos ilgis remiantis sakinių skaičiumi, naudojant 50% kompresiją

Programa	**Teksto santraukos ilgis (sakinių skaičius) (50%)**					
	Grožinio teksto (angl.)	**Mokslinio teksto (angl.)**	**Publicistinio teksto (angl.)**	**Grožinio teksto (liet.)**	**Mokslinio teksto (liet.)**	**Publicistinio teksto (liet.)**
Tools4Noobs	6	7	14	22	19	7
MS Word AutoSummarize	38	31	15	56	31	5
Automatic text summarizer	40	35	15	65	35	6
Copernic summarizer	42	35	16	67	40	9
Intellexer summarizer	42	34	16	66	34	6
Pertinence summarizer	40	23	11	39	26	6
QuickJist summarizer	30	24	14	53	21	5
Shvoong summarizer	38	23	14	44	-	7
Subject search summarizer	40	35	15	68	40	5

Iš 2.2 lentelėje pateiktų santraukos sakinių skaičiaus pastebima, jog ne visose santraukose naudojančiose tą patį suglaudinimo lygį sakinių skaičius yra vienodas. Veikimo ir santraukų pateikimo principais panašios *Copernic Summarizer* ir *Intellexer Summarizer* ne visada pateikia tokio paties ilgio santraukas. Grožinių tekstų anglų ir lietuvių kalbomis santraukos, kurias pateikia minėtosios sistemos yra sudarytos iš vienodo skaičiaus sakinių, tačiau kitų žanrų santraukų ilgiai skiriasi – *Copernic Summarizer* pateikia trumpesnes santraukas, nei *Intellexer*

Summarizer. Iš 2.2 lentelėje pateikiamų duomenų išsiskiria *Tools4Noobs summarizer* sistema, kurios parengtos santraukos pagal sakinių skaičių yra trumpiausios, tačiau vidutinis žodžių skaičius sakiniui yra panašus į kitų sistemų gautus rezultatus.

Pagal procentinę 25% kompresiją (žr. į 2.3 lentelę**)** grožinio angliško teksto santraukos sakinių skaičiumi daugeliu atveju yra vienodos ar panašios. Sakinio ilgiu iš kitų santraukų sistemų labai išsiskiria *Tools4Noobs.* Šios sistemos angliškos grožinio ir mokslinio, lietuviškos grožinio, mokslinio ir publicistinio tekstų santraukos išsiskiria sakinio ilgiu žodžiais. Pagal pateiktus duomenis pastebima, jog sistema naudojasi sakinio ilgio metodu. Reikia paminėti, jog *Tools4Noobs* santraukų sistema šalia santraukos pateikia ir kelis žodžius, kurie dažniausiai sutinkami tekste. Tačiau šie žodžiai skaičiuojant santraukos ilgį nebuvo įtraukiami, o laikomi antraeile informacija. Iš kitų analizuojamų duomenų pastebima, kad *Pertinence Summarizer* parengta santrauka yra tik vienu sakiniu ilgesnė, nors buvo naudojama 30% kompresija. Tačiau *Shvoong Summarizer* parengta santrauka yra trumpesnė už 25% atitinkamo teksto santraukos sakinių skaičių, nors naudojama 30% kompresija. Trumpiausią teksto santrauką sakinių skaičiumi pateikia *QuickJist Summarizer*, nors jam šio bandymo metu buvo priskirta 25 % kompresija. *Subject Search Summarizer* santrauka yra ilgesnė dviem sakiniais, dėl to, kad pridedama pora papildomų sakinių, apibūdinančių kurią nors teksto dalį ar veiksmą, ar pateikiantys papildomą informaciją apie tekstą ir reikšminius jo žodžius. Į *Copernic Summarizer* ir *Intellexer Summarizer* santraukos ilgį neįskaičiuota reikšminių žodžių eilutės, kurios nurodo pagrindinius teksto žodžius.

2.3 lentelė Teksto santraukos ilgis, remiantis sakinių skaičiumi, naudojant 25-30%

Programa	**Santraukos sakinių ilgis (žodžių skaičius) (vid) (25-30%)**					
	Grožinio teksto (angl.)	**Mokslinio teksto (angl.)**	**Publicistinio teksto (angl.)**	**Grožinio teksto (liet.)**	**Mokslinio teksto (liet.)**	**Publicisti-nio teksto (liet.)**
Tools4Noobs	103	46	17,3	19	40,71	19,67
MS Word AutoSummarizer	31,81	17,6	16,11	11,26	25	13,66
Automatic text summarizer	35,33	23,86	18,56	12,12	24,25	16,33
Copernic summarizer	32,05	25,65	21,88	13,97	26,15	14
Intellexer summarizer	34,85	20,53	19,22	11,67	19,25	10,5
Pertinence summarizer	35,59	30,38	18,43	21,90	30,6	16
QuickJist summarizer	28,73	22,58	15,17	14,21	15,91	9,5
Shvoong summarizer	55,27	33,08	20,89	26,78	-	12,5
Subject search	44,36	22,42	21,22	18,85	15,27	17

summarizer						

Pastebėta, kad *QuickJist Summarizer* pateikiamos 25% kompresijos santraukos lietuvių ir anglų kalbų tekstuose yra trumpiausios, o ilgiausiomis galima laikyti *Subject Search Summarizer* pateikiamas santraukas. Tačiau *Subject Search Summarizer* santraukos ilgis buvo pasirinktas pagal *Word AutoSummarize* pateiktų sakinių 25% santraukos suglaudinimo kiekį, kadangi pati sistema nepateikia jokių procentinių glaudinimo būdų.

Kiekvienos SAR sistemos pateikiamos santraukos nors glaudinamos tokiu pačiu – 50 % suglaudinimo lygiu, bet kiekvienoje gaunamoje santraukoje žodžių skaičius (žr. į lentelė 2.4) yra skirtingas. Kaip minėta anksčiau *Shvoong Summarizer* nepateikia lietuviško mokslinio teksto santraukos, todėl šios sistemos santrauka nėra nagrinėjama. Iš turimų duomenų matyti, jog daugiausiai santraukos ilgis žodžiais skiriasi 653 žodžiais, o mažiausiai - 6 žodžiais. Vidutiniškai visų tekstų santraukų ilgis yra 654 žodžiai santraukai, tačiau santraukos ilgis priklauso nuo teksto ilgio bei kompresijos lygio, naudoto tekstams. Dalyje gaunamų santraukų (ir lietuviškų, ir angliškų) žodžių skaičius tam tikro stiliaus tekste yra panašus, skiriasi vidutiniškai 30 žodžių. Tačiau santraukos, kurios yra ilgesnės, dažniausiai yra suprantamesnės ir prasmingesnės

.

2.4 lentelė. Angliškų ir lietuviškų santraukų ilgis žodžiais, naudojat 50% suglaudinimą

Programa	**Santraukos sakinių ilgis žodžiais (žodžių skaičius) (50%)**					
	Grožinio teksto (angl.)	**Mokslinio teksto (angl.)**	**Publicistinio teksto (angl.)**	**Grožinio teksto (liet.)**	**Mokslinio teksto (liet.)**	**Publicisti-nio teksto (liet.)**
Tools4Noobs	423	187	298	348	604	160
MS Word AutoSummarize	1353	687	288	741	569	107
Automatic text summarizer	1508	971	307	869	818	107
Copernic summarizer	1382	863	342	911	869	144
Intellexer summarizer	1502	737	309	908	670	63
Pertinence summarizer	1321	683	269	*732*	*567*	*85*
QuickJist summarizer	855	495	245	716	355	80
Shvoong summarizer	1348	702	305	752	-	116
Subject search summarizer	1586	954	363	994	650	92

Paėmus mažesnį 30% suglaudinimo lygį naudojančios sistemos *Shvoong Summarizer* ir *Pertinence Summarizer* santraukų ilgis skiriasi 46 žodžiais. *Pertinence Summmarizer* pateikiamos santraukos ilgis žodžiais ir kitos sistemos, tokios kaip *Intellexer Summarizer* žodžių skirtumas yra palyginti labai mažas (51 žodis, o palyginus su *Automatic Text Summarizer* dar mažesnis – 41 žodis).

2.5 lentelė. Angliškų ir lietuviškų santraukų ilgis žodžiais, naudojat 25-30% suglaudinimą

Programa	**Santraukos sakinių ilgis žodžiais (žodžių skaičius)**					
	Grožinio teksto (angl.)	**Mokslinio teksto (angl.)**	**Publicistinio teksto (angl.)**	**Grožinio teksto (liet.)**	**Mokslinio teksto (liet.)**	**Publicisti-nio teksto (liet.)**
Tools4Noobs	103	92	52	133	285	59
Automatic text summarizer	742	525	167	400	485	49
Copernic summarizer	673	436	175	461	523	56
Intellexer summarizer	732	390	173	427	385	63
MS Word AutoSummarize	668	352	145	383	275	41
Pertinence summarizer	783	395	129	438	306	16
QuickJist summarizer	431	271	91	398	175	19
Shvoong summarizer	829	430	188	482		75
Subject search summarizer	976	583	191	622	336	34

Skirtumai tarp tokio paties suglaudinimo lygio teksto santraukų svyruoja priklausomai nuo to, kokius santraukų rengimo būdus naudoja sistemos. Sistemos, kurios santraukos rengimui naudoja sakinių skaičių automatiškai pateikia didesnį žodžių kiekį, nes dažniausiai į santrauką įtraukiami ilgiausi teksto sakiniai. Nors *Automatic Text Summarizer* ir *Subject Search Summarizer* suglaudinimui pasirinktas vienodas sakinių skaičius, atitinkantis 25% viso teksto, tačiau šių sistemų pateiktos santraukos skiriasi 234 žodžiais. Toks skirtumas sistemoms naudojančioms tą patį suglaudinimo būdą yra pakankamai didelis. Tačiau ir tarp kitokį suglaudinimo lygį naudojančių sistemų pastebimi gana dideli žodžių kiekio santraukoje skirtumai. 2.5 lentelė rodo, kad trumpiausias santraukas žodžiais pateikia *QuickJist Summarizer* bei *Tools4Noobs* sistemos, kurios, kaip ir likusios sistemos santrauką rengia naudodamas 25%

suglaudinimo lygį. Vidutiniškai nuo kitų santraukų šių sistemų parengtos santraukos savo ilgiu skiriasi nuo 260 iki 400 žodžių. Tai palyginus su kitomis sistemomis ir kitokiais suglaudinimo lygiais yra didelis skirtumas, nes vidutinis skirtumas yra ~27,33 žodžiai.

Kaip ir žodžių kiekis santraukoje, taip ir vidutinis santraukos ilgis kiekvienoje santraukoje yra skirtingas. Skirtumai, kaip jau minėta, dažniausiai priklauso nuo suglaudinimo lygio, santraukos ilgio ir sistemos parametrų.

2.6 lentelė. Vidutinis angliškos ir lietuviškos santraukos sakinių ilgis žodžiais, naudojant 50% kompresiją

Programa	Vidutinis santraukos sakinių ilgis žodžiais (žodžių skaičius) 50%					
	Grožinio teksto (angl.)	Mokslinio teksto (angl.)	Publicistinio teksto (angl.)	Grožinio teksto (liet.)	Mokslinio teksto (liet.)	Publicistinio teksto (liet.)
Tools4Noobs	70,5	26,71	21,29	15,81	31,78	22,86
MS Word AutoSummarize	35,61	22,16	19,2	13,23	18,36	21,4
Automatic text summarizer	37,7	20,47	20,47	13,37	23,37	17,83
Copernic summarizer	32,90	24,66	21,375	13,60	21,73	16
Intellexer summarizer	35,76	21,06	19,31	13,76	19,71	10,5
Pertinence summarizer	33,025	29,70	24,46	18,77	21,81	14,17
QuickJist summarizer	28,5	20,625	17,5	13,51	16,91	16
Shvoong summarizer	35,47	30,52	21,79	17,09	-	16,57
Subject search summarizer	44,36	22,42	21,22	14,62	16,25	18,4

2.6 lentelėje pateikiami duomenys apie vidutinį santraukos ilgį žodžiais taikant 50% kompresijos lygį. Kaip matome iš pateiktos lentelės skirtumai tarp vidutinio santraukos sakinių ilgio žodžiais nėra dideli – kai kuriais atvejais mažesni nei naudojant 25% ar 30% suglaudinimo lygį (žr. į 2.7 lentelė). Naudojant 50% suglaudinimo lygį vidutinis sakinio ilgis skiriasi nuo 2 iki 10 dešimties žodžių sakiniui, kas palyginus nėra daug, nors skiriasi bendras sakinių skaičius santraukose. Iš pateiktų duomenų galima teigti, jog santraukose ilgiausius sakinius pateikia skirtingos santraukų sistemos, priklausomai nuo teksto stiliaus, ilgio bei kalbos. Tendencija, jog trumpiausias 50% suspaudimo lygio santraukas pateikia *QuickJist Summarizer* išlieka beveik

visuose lietuvių ir anglų kalbų santraukose. *QuickJist Summarizer* pateikiamos santraukos yra trumpesnės nei kitos, tačiau dėl to santrauka nepraranda prasmingumo ir suprantamumo.

2.7 lentelė. Vidutinis angliškos ir lietuviškos santraukos sakinių ilgis žodžiais, naudojant 25-30% kompresiją

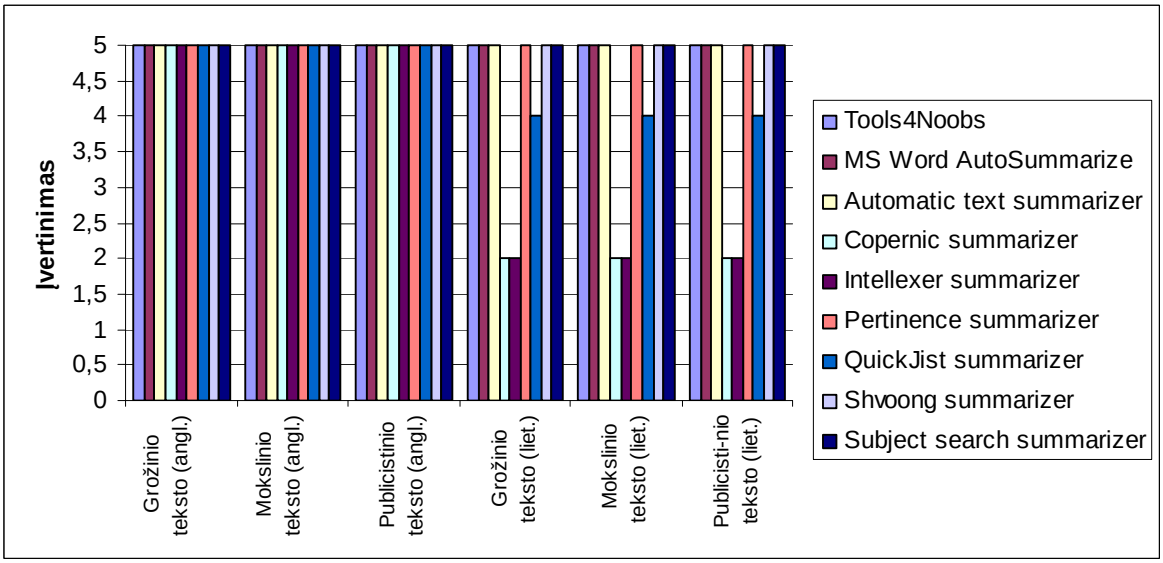
Programa	**Vidutinis santraukos sakinių ilgis žodžiais (žodžių skaičius)**					
	Grožinio teksto (angl.)	**Mokslinio teksto (angl.)**	**Publicistinio teksto (angl.)**	**Grožinio teksto (liet.)**	**Mokslinio teksto (liet.)**	**Publicisti-nio teksto (liet.)**
Tools4Noobs	103	46	17,3	19	40,71	19,67
Automatic Text Summarizer	37,7	20,47	20,47	13,37	23,37	17,83
Copernic Summarizer	32,9	24,66	21,375	13,6	21,73	16
Intellexer Summarizer	35,76	21,06	19,31	13,76	19,71	10,5
MS Word AutoSummarize	35,61	22,16	19,2	13,23	18,36	21.4
Pertinence Summarizer	33,025	29,7	24,46	18,77	21,81	14,17
QuickJist Summarizer	28,5	20,625	17,5	13,51	16,91	16
Shvoong Summarizer	35,47	30,52	21,79	17,09	-	16,57
Subject Search Summarizer	38,68	27,26	22,69	14,62	16,25	18,4

Iš pateiktų duomenų **(**žr. į 2.7 lentelė) matome, jog vidutiniškai ilgiausius sakinius santraukoje pateikia *Tools4Noobs* bei *Shvoong Summarizer*. Kaip jau minėta anksčiau, taip yra todėl, kad šiuo atveju *Shvoong Summarizer* naudoja 30% suglaudinimą, ir ima didesnę teksto dalį, nei kitos santraukų sistemos. Nors *Pertinence Summarizer* taip pat naudoja 30% suglaudinimą, tačiau šios sistemos pateikiama santrauka nuo 25% teksto santraukos skiriasi nežymiai. *Automatic Text Summarizer* ir *Subject Search Summarizer* vidutinis sakinio ilgis skiriasi 8,67 žodžio, tai yra palyginus su *Pertinence summarizer* ir *Shoong Summarizer*, kur skirtumas lygus 24,84 žodžiams.

2.7. Anketinių duomenų analizė

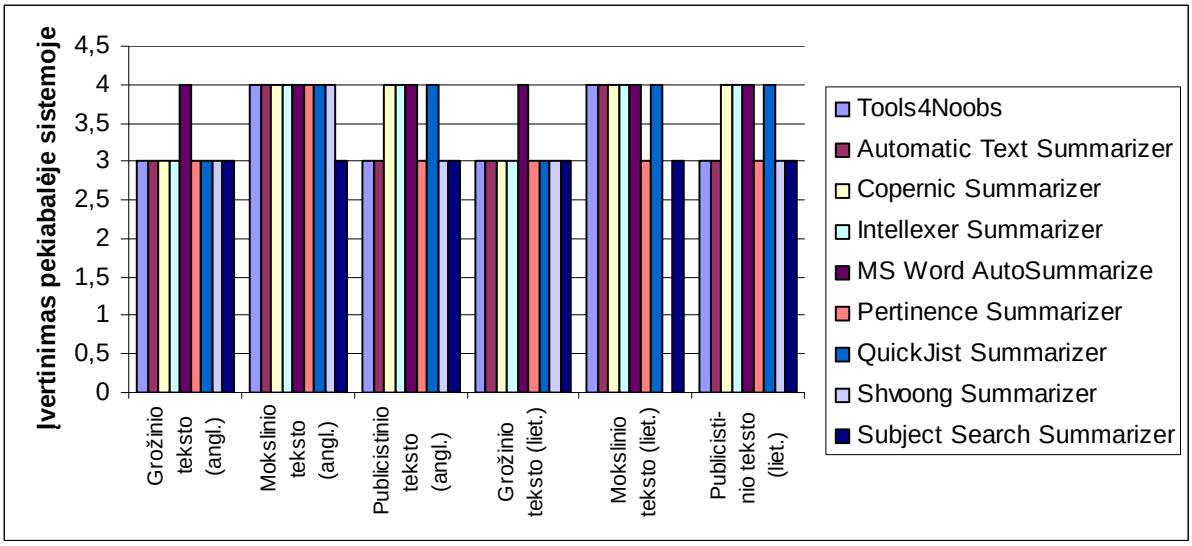
Santraukų sistemų kokybei ir kitiems kriterijams įvertinti buvo pateikti tekstai, kurių suglaudinimo lygis 50% ir 25-30%. Perskaitę tekstus respondentai užpildė jiems pateiktas anketas. Tekstus skaitė 20 respondentų, kurių amžius nuo 18 m. iki 50 m. 50% respondentų sudarė moterys, 50% vyrai. Anketą sudarė klausimai apie santraukos įskaitomumą, prasmės

perteikimą bei santraukos kokybės įvertinimą. Respondentas įvertina pateikto teksto santraukos kokybę remdamasis tuo, ar jis suprato santraukos pagrindinę mintį (t.y. prasmės perteikimą), ar santrauka buvo įskaitoma bei ar buvo informatyvi. Tuo atveju, kai nagrinėjamų anketų vertinimai nesutampa vedamas vidurkis. Jei vidurkis gaunamas ne sveikasis skaičius, tada rašomas dažniausiai pasikartojantis vertinimo balas. Anketoje kiekvienas punktas buvo vertinamas penkiabalėje sistemoje, kur 5 – puikiai, 4 – gerai, 3 – vidutiniškai, 2 – blogai, 1 – labai blogai.



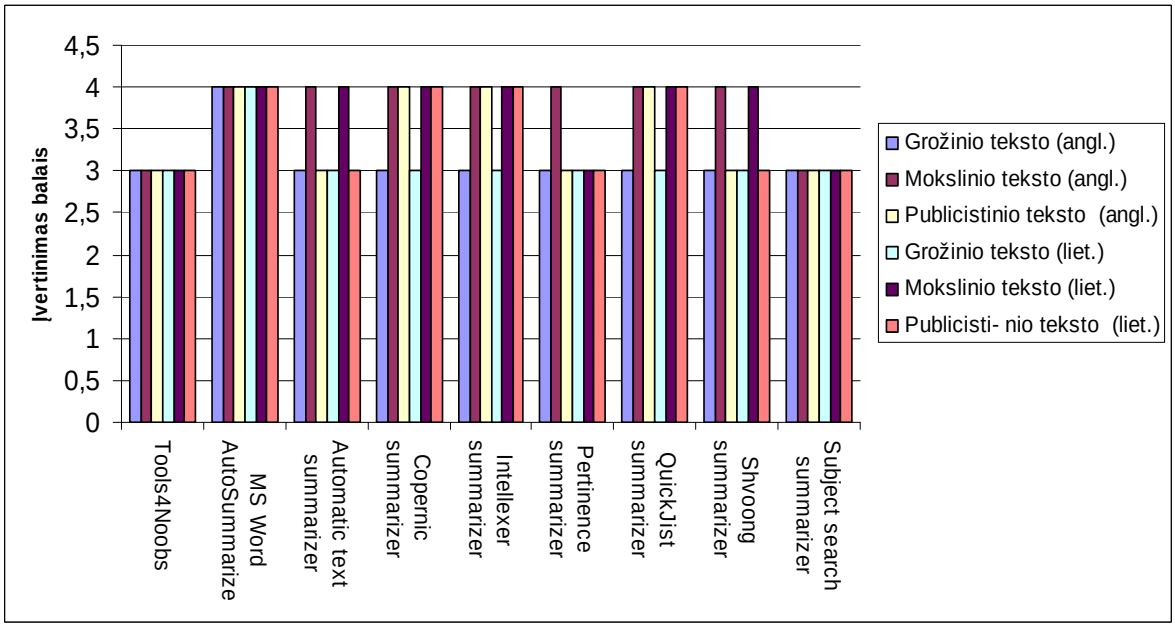
2.5 pav. Santraukų įskaitomumo diagrama

Iš 2.5 pav. diagramos matome, jog visi respondentai vieningai atsakė, jog visos sistemos pateikė įskaitomas anglų kalbos tekstų santraukas. Lietuvių kalbos tekstų įskaitomumas atskirose sistemose vertinamas aukščiausiais balais – *MS Word Summarizer, Automatic text Summarizer, Pertinence summarizer, Shvoong summarizer, Subject Search Summarizer*. *QuickJist summarizer* sistema parengtus lietuvių kalbos tekstus respondentai vertina prasčiau, nes sistema neatpažįsta kai kurių rašmenų. Blogiausiai respondentai vertina *Copernic Summarizer* ir *Intellexer summarizer* sistemomis parengtų sistemų įskaitomumą, kadangi sistemos visiškai neatpažįsta lietuviškų simbolių. Tokie pat įskaitomumo įvertinimai išlieka ir santraukose, kurioms naudojama mažesnė kompresija – lietuviški simboliai neatpažįstami.

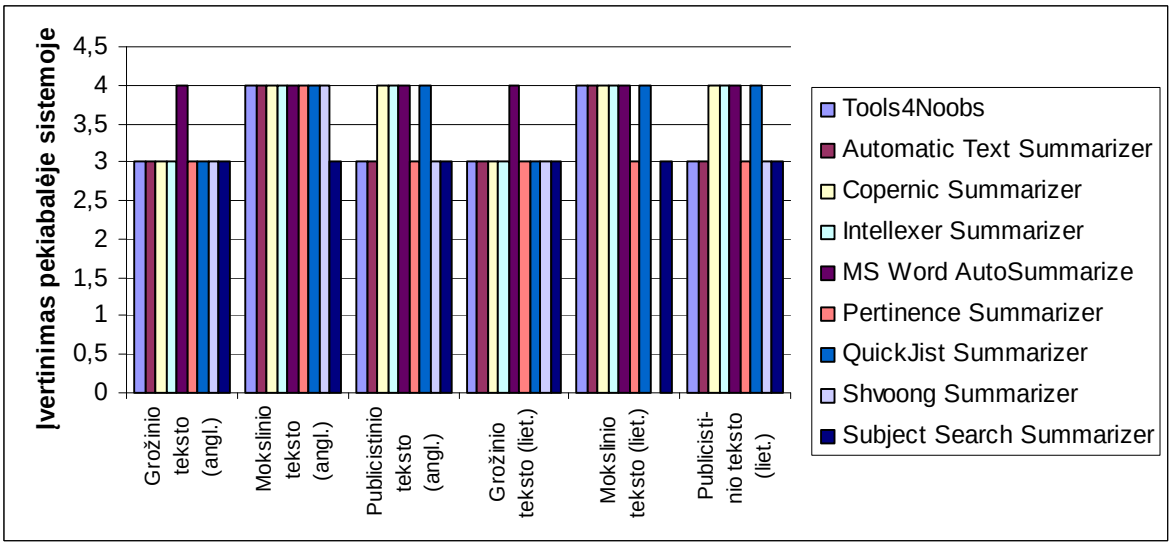


2.6 pav. Teksto prasmės perteikimas naudojant 50% kompresiją.

Pateiktame 2.6 pav. matome, kad respondentų nuomone jiems pateikta santrauka nėra labai gera, trūksta informacijos, nes bendras vidurkis yra 4 (vertinama gerai), tačiau nė vienas tekstas nėra įvertintas 5 (t.y. puikiai). Kaip matome iš pateiktos diagramos respondentai grožinio teksto santraukas vertino 3 (vidutiniškai), o publicistinio ir mokslinio tekstų santraukų tekstus įvertino 4 (t.y. gerai). Tačiau vertinimai pasikeičia, kai pateikiamas 25-30% santraukos tekstas (žr. į 2.7 pav.). Respondentų nuomone beveik visų sistemų santraukų tekstai vidutiniškai perteikia teksto prasmę – jie vertinami vidutiniškai. Respondentų nuomone prasčiausiai prasmė perteikta *Subject Search Summarizer* bei *Tools4Noobs* gautose grožinio ir mokslinio angliškuose ir lietuviškuose tekstuose. Publicistinių tekstų santraukų tekstus respondentai vertina, kaip vidutiniškai suprantamus.

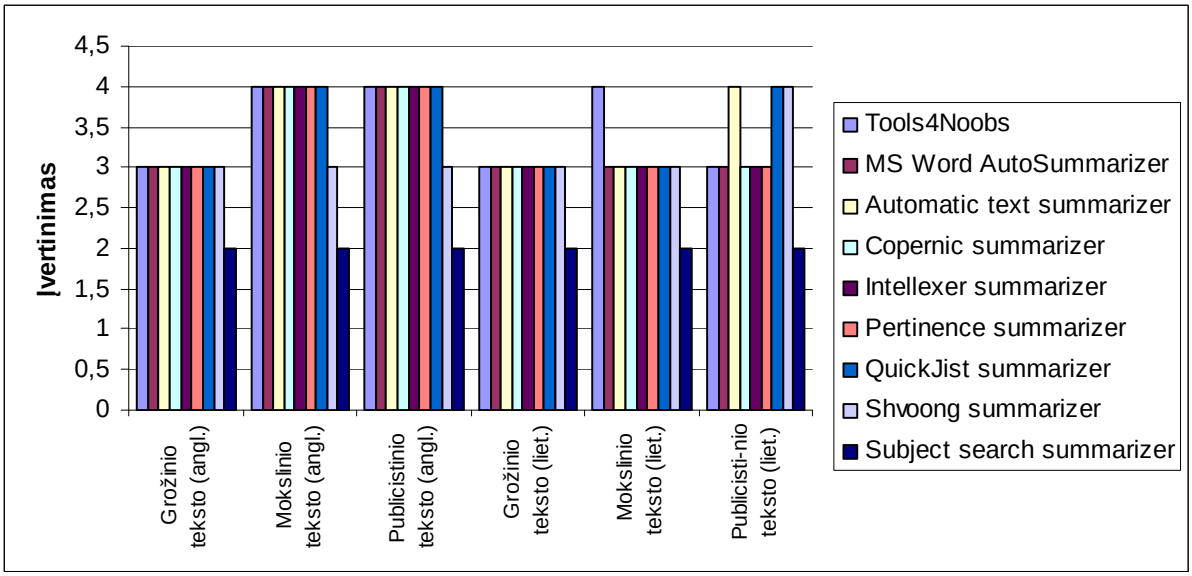


2.7 pav. Teksto prasmės perteikimas naudojant 25-30 % kompresiją.



2.8 pav. Santraukų kokybės diagrama, naudojant 50% kompresiją.

Respondentų nuomone, teksto santrauka, kuri yra 50 % suspausta(žr. į 2.8 pav.), daugiausiai vertinama kaip gera. Vidutiniškai respondentai įvertino lietuviškas santraukas, parengtas *Copernic Summarizer, Intellexer Summarizer, Tools4Noobs* sistemomis. *Shvoong summarizer* sistema parengtas angliškų mokslinių bei publicistinių, o taip pat ir lietuviškų grožinių bei mokslinių tekstų santraukas, respondentai taip pat įvertino kaip vidutiniškos kokybės. Blogiausiai kokybės atžvilgiu išskirti kurią nors sistemą sunku, kadangi susumavus respondentų atsakymus, gaunama, jog dauguma santraukų yra vidutiniškos ir iš dalies tenkina vartotoją.



2.10 pav. Santraukų kokybės įvertinimas, naudojant 25-30% kompresiją.

Santraukų kokybę 25-30% teksto santraukai, respondentai vertino jau kiek kitaip (žr. į 2.10 pav.). Palyginus su 50% suglaudinimu krito *MS Word Summarizer, Automatic text summarizer, Copernic summarizer, Intellexer summarizer, Pertinence Summarizer, QuickJist summarizer, Tools4Noobs* ir *Shvoong summarizer* tekstų santraukos kokybės įvertinimas krenta iki 3 balų (t.y. vertinamas kaip vidutiniškos kokybės). Taip pat prasčiau vertinamos *QuickJist summarizer* parengtos lietuvių kalbos grožinio ir mokslinio stiliaus santraukos, kokybės atžvilgiu. *Subject Search Summarizer* sistema respondentų nuomone pateikia blogiausias santraukas palyginimus su kitomis sistemomis.

Rengiant santraukas svarbu atsižvelgti į numatomos santraukos ilgį. Jei reikalinga kuo trumpesnė santrauka, ji gali būti gaunama prastos kokybės bei neinformatyvi, kadangi ne visada sistemose taikomi metodai ir technologijos pasiteisina. Kokiu metodu remiantis generuotos santraukos gaunamos prastenės kokybės išskirti sunku, kadangi tai priklauso nuo teksto žanro bei ilgio, sakinių ilgio bei žodžių skaičiaus. Vienais atvejais metodai pasiteisina, kitais atvejais – ne.

3. Skaitmenizuoto teksto automatinio rengimo santraukų projektavimas

Santraukų sistema (modelis) taikomas apdoroti tekstui ir pateikti to teksto santrauką. Modelis remiasi prototipinių sistemų pavyzdžiu (konkrečiai *Automatic Text Summarizer, Open Text Summarizer*). Sistemoje bus taikomi tokie paties arba panašūs metodai, kuriuos naudoja prototipas.

Sistemai keliami funkciniai reikalavimai:

- Sistema turi pateikti parengtą santrauką;
- Santrauka turi būti pateikiama be programos klaidų.

Nefunkciniai reikalavimai:

- Programa turi veikti Windows OS aplinkoje;
- Programa turi veikti minimalius išteklius turinčiame kompiuteryje;
- Programos metu įvykius klaidai darbas nutraukiamas ir programa išjungiama;
- Patogi ir aiški vartotojo sąsaja;
- Jei programa yra įdiegiama: turi būti lengvas bei suprantamas įdiegimo procesas.

3.1. Reikalavimų specifikavimas

Vartotojas – asmuo ar asmenys, besinaudojantys santraukų rengimo sistema per interneto prieigą ar kompiuteryje. Vartotoju gali būti bet kuris asmuo, kuriam reikalinga specifinio teksto (patenkančio į santraukų rengimo sistemos ribas) santrauka. Vartotojas turi turėti elementarius naudojimosi kompiuteriu gebėjimus.

Kompiuterizuota santraukų rengimo sistema – santraukų rengimo sistema, kuri iš vartotojo pateikto originalaus teksto parengia atitinkamo ilgio santrauką, perteikdama pagrindines originalaus teksto mintis. Santraukų rengimo sistema – programinis produktas, veikiantis Windows operacinėje aplinkoje.

SARS programinis modulis turi gebėti: nuskaityti skaitmenizuotą tekstą, jį analizuoti, skaičiuoti sakinius, atrinkti sakinius tinkamus santraukai, sudėti sakinius kaip vientisą tekstą, pateikti gautą rezultatą programos lange.

Norėdami nustatyti, kokius reikalavimus turi atitikti projektas buvo atlikta apklausta 20 interneto vartotojų. Apklausos metu siekiama sužinoti, kokius reikalavimus kelia vartotojas, kuris naudojasi programine įranga, ir gautus apklausos rezultatus panaudoti apibendrinant reikalavimus, kuriuos turėtų atitikti projektas. Atlikus apklausą, nustatyti ir suformuluoti tokie reikalavimai:

3.1.1. Funkciniai reikalavimai

1. **Programinė įranga**

1.1. Programinė įranga turi veikti vartotojo kompiuteriuose. PĮ neturi lėtinti kompiuterio darbo ir apkrauti kompiuterio bei programinės įrangos nereikalingais darbais. Respondentų pageidavimų sistema turėtų būti nediegiama, bet veikti be interneto ryšio. Programinė įranga turėtų būti lengvai valdoma ir suprantama (neturėtų būti terminų, funkcijų, kurie sunkiai suvokiami eiliniam vartotojui). Respondentų pageidavimu – kuo mažiau spalvų, nereikalingų funkcijų. Programinė įranga turi leisti įkelti tekstą iš duomenų failo ir atlikti teksto santraukos parengimą, bei, jei reikia suteikti vartotojui pagalbą. Vartotojui pasirodantis programos langas turi būti su lengva navigacija ir nesukelti problemų atliekant navigacinius veiksmus. Programinė įranga turi būti pritaikyta minimaliam kompiuterio galingumui, kuris nurodomas žemiau esančiuose punktuose. Vartotojui atsivėrus programos langą langas išjungiamas tada, kai vartotojas nusprendžia, kad jis nebenori toliau dirbti su programa. Pageidaujama, jog prieš išeinant iš programos lango, nebūtų pateikiama užklausa ar tikrai norima išeiti iš programos lango (60% respondentų teigia, jog tokios užklausos vargina vartotoją). Programai baigus darbą (uždarius programos langą) – programos veikimas nutraukiamas ir kuriamos santraukos duomenys yra ištrinami.

1.2. Programinė įranga turi leisti įkelti tekstinius failus, atlikti veiksmus reikalingus sugeneruoti santraukai ir pateikti rezultatus programos lange.

1.3. Jei programinė įranga tinkamai neveiks, santraukos generavimas ir kūrimas nebus tinkamai įvykdytas, gautas rezultatas gali netenkinti vartotojo. Jei tolimesniuose žingsniuose atsiranda klaidų, su tam tikromis išlygomis sistema veikia, tačiau gaunamas rezultatas su klaidomis. Jei duomenų failas nebus tinkamai nuskaitytas – automatiškai nebus generuojama santrauka. Jei generuojant santrauką įvyks klaida, sistema turi dirbti toliau „peršokdama“ klaidą. Jei tekstas nebus įkeliamas naudojantis mygtuko pagalba, įkopijuotas į teksto lauką, jis nebus nuskaitomas ir naudojamas generuoti santraukai.

1.4. Programinė įranga turi veikti, kai kompiuterio operatyvinė atmintis ne mažesnė už 512 MB. Kietajame diske programa turi užimti ne daugiau kaip 200 MB. Grafika turi būti pritaikyta taip, kad būtų tinkama ir gerai matoma įvairaus senumo vaizduokliuose. Programinės įrangos spalvos ir formos neturi varginti vartotojo akių ir sukelti problemų dirbant minėta sistema. Meniu punktai turi turėti aiškius pavadinimus.

2. Duomenų įvedimas

2.1. Duomenų įvedimas - duomenų failas, kurio formatas yra *.txt nuskaitomas iš nurodytos vietos, esančios kompiuteryje.

2.2. Duomenys turi būti įvedami pagal keliamus kriterijus: duomenų failai turi būti tam tikro formato, kuris nurodomas programos apraše, pateikiamame iš karto atidarius programos langą. Duomenys negali būti parašyti kinų, japonų ar kitomis kalbomis, kurios nėra suderintos su PĮ. Taip pat negali būti apdorojami hieroglifai, įvairūs piešiniai, garsiniai failai. Duomenų failas turi būti *.txt formatu, bei su kuo mažiau formatavimo žymų (jei jos lieka).

2.3. Jei duomenų faile bus simbolių ar kitokių duomenų, kurių sistema neatpažins, programa pateiks neatpažintų simbolių seką arba visą duomenų failą.

3. Duomenų patikra

3.1. Duomenų tikrinimas – tikrinama, ar duomenų faile esantys duomenys yra žodžiai. Atliekamas žodžių dažnio lyginimas, kuris programoje nepateikiamas. Jei ieškomi žodžiai yra faile – atitinkamai jie atrenkami į santrauką. Duomenys tikrinami pagal atskirus punktus (sakinio ilgį ir kitus). Jei duomenys atitinka keliamus reikalavimus, jie atrenkami santraukai.

3.2. Jei duomenys bus netikrinami, netinkama informacija bus saugoma bei atrenkama su tinkamais duomenimis, su jais bus atliekami veiksmai - taip bus apkraunama programinė įranga, atliekamas nereikalingas darbas, kuris neturės jokios naudos, veltui bus naudojami kompiuterio resursai.

4. Duomenų atranka

4.1. Duomenų atrinkimas – duomenys atrenkami pagal nurodytus kriterijus - sakinio ilgį. Atrinkimas turi būti vykdomas nurodytu būdu, arba prijungiant antrąjį pasirinktą metodą.

4.2. Jei nebus įvykdomas šis punktas, duomenys nebus atrinkti ir tai reikės atlikti rankiniu būdu. Tokiu būdu santraukos kūrimas atims laiko, o programinė įranga tinkamai

nefunkcionuos. Atrinkime gali pasitaikyti klaidų, kurios buvo padarytos netyčia ar per neapsižiūrėjimą, tačiau santraukos rengimo procesui tai įtakos neturėtų turėti.

5. **Duomenų išsaugojimas**

Veiksmai su duomenų atrinkimu nėra saugomi. Vartotojas pats šalina nereikalingą duomenų failą (rankiniu būdu). Duomenų failai (t.y. santraukos) nėra saugomos kietajame diske. Jei vis dėl to norima saugoti santrauką, gautą jos tekstą, esantį antrame programos teksto lauke reikia nusikopijuoti į atitinkamą programą ir išsaugoti. Santraukos failą, jei jį sukuria pats vartotojas, galima ištrinti rankiniu būdu.

3.1.2. Technologiniai reikalavimai vartotojui

Tai nefunkciniai reikalavimai, kurie keliami projektuojamai programai:

1. **Suderinamumas**

Programa turi būti lengvai suderinama su turima kompiuterine įranga, nekelti papildomų reikalavimų tinkamam funkcionavimui, nelėtinti kompiuterių sistemos darbo, bei papildomai neperkrauti kompiuterio įvairiomis nereikalingomis programos funkcijomis.

2. **Aiškumas**

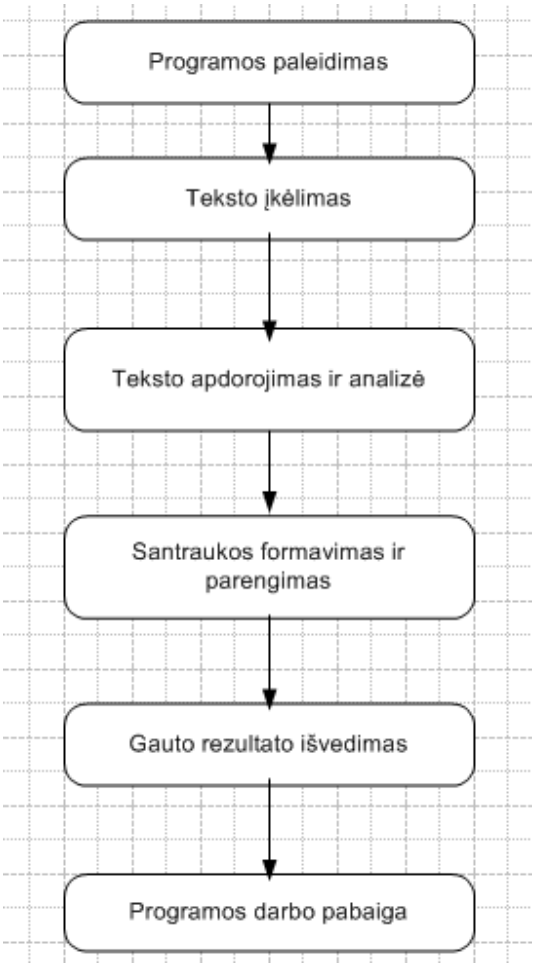
Aiškus ir suprantamas naudojimasis programine įranga. Raidžių šriftas turi būti *Times New Roman,* siūlomas raidžių dydis nuo 12 iki 16 punktų. Teksto spalva turėtų būti juoda. Programos langas turi būti nerėžiantis akių – šviesiai pilkas, šviesiai rudas, ar šviesus fonas. Programos lango dydis turėtų keistis pagal vartotojo pageidavimus – vartotojas gali padidinti, sumažinti programos langą.

3. **Valdymo paprastumas**

Programa turi nereikalauti sudėtingo valdymo Nereikia rašyti įvairių programos kodų, užtenka tik paspausti atitinkamus programos valdymo mygtukus. Programos valdymas turėtų būti sukurtas atsižvelgiant į vidutinio žmogaus poreikius ir gebėjimus.

4. **Sauga**

Programa nenaudos asmeninių duomenų, ar kitų duomenų apie asmenis, kuri galėtų patekti į pašalinių asmenų rankas. Programa yra atvirojo kodo, todėl leidžiamas šios sistemos kodo redavimas bei papildymas papildomais metodais, nurodant „Sanla“ sistemą kaip pradinį šaltinį.



3.1 pav. SARS programos paketo struktūrinė schema

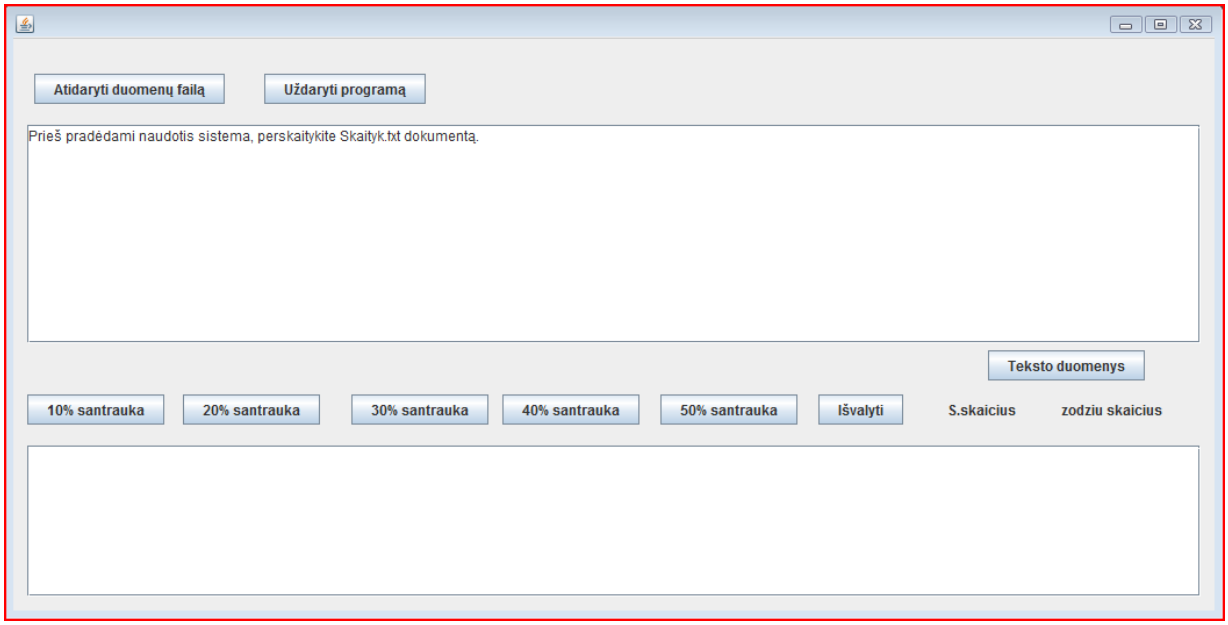
3.2. Programinių modulių ar objektų specifikacijos.

Santraukų sistemai kurti, buvo naudojami šie moduliai:

Klasė Sanla.**Main.** Pagrindinė sistemos klasė, kurioje vykdomi visi sistemos veiksmai.

Pavadinimas	Ką atlieka	Savybės tipas
skaitoTekstaIsFailo	Nuskaitomas tekstinis duomenų failas ir paverčiamas masyvu. Taip pat skaitant duomenų failą vykdomas TreeMap klasė, kuri nurodo, kiek kartų duomenų faile pasikartojo	Ženklų masyvas

	žodis.	
SkaiciuojaSakinius	Skaičiuoja kiek tekstiniame duomenų faile (masyve) yra sakinių. Nurodomi sakinių skyrikliai: taškas, šauktukas, klaustukas, daugtaškis. Rezultatą išveda į programos lango elementą.	Skaičius
zodziuSkaiciavimas	Skaičiuoja kiek tekstiniame duomenų faile (masyve) yra žodžių. Rezultatas išvedamas į programos lango elementą.	Skaičius
SakinioIlgioVidurkioNustatymas	Nustatoma viso teksto (masyvo) vidutinis sakinio ilgis. Rezultatas išvedamas į programos lango elementą.	Skaičius
Isrinkimas1	Išrenkami sakiniai, kurie yra ilgesni už vidutinį turimo teksto (masyvo) ilgį.	masyvas
sanlaf	Grafinės sąsajos aprašai bei parametrai. Mygtukų pavadinimai bei jiems priskirtos komandos.	Grafinė sąsaja



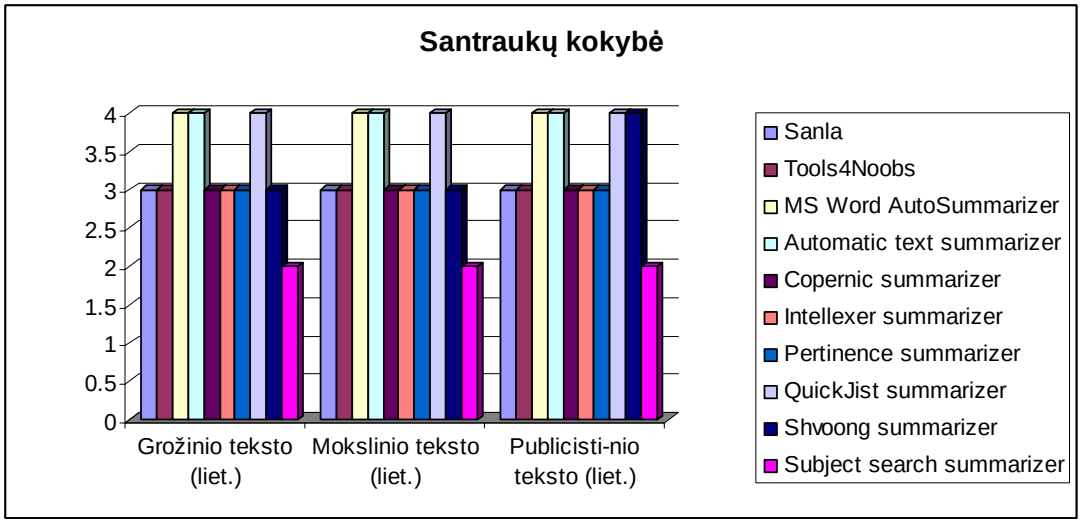
3.2 pav. Sistemos "Sanla" langas

3.3. Testavimo medžiaga

Santraukų sistemos testavimui buvo pasirinkti lietuvių kalbos tekstai, kurių santraukas yra parengusios kitos santraukų sistemos. Kitų santraukų sistemų gautos santraukos pateikiamos prieduose.

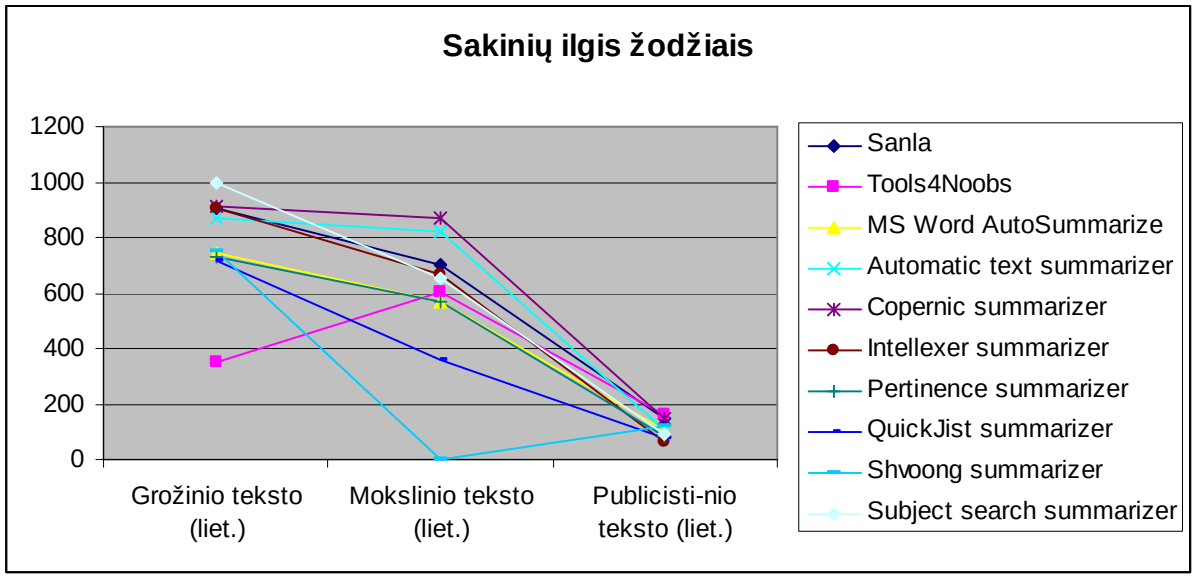
Santraukų vertinimas buvo vykdomas apklausiant respondentus, kurie santraukas vertino balais nuo 1 iki 5 (čia 1- žemiausias ir blogiausias įvertinimas, 5- aukščiausias ir geriausias įvertinimas).

Palyginus gautą santraukų kokybę, *Sanla* sistemos santraukos kokybė vidutiniškai vertinama 3 balais (vertinimo sistema: 5 balai – labai gerai, 4 – gerai, 3 – vidutiniškai, 2- blogai, 1 – labai blogai). Iš 3.3 paveikslo matome, jog savo santraukos kokybe sistema priklauso vidutinės kokybės santraukų daliai. Tačiau „Sanla" sistema savo kokybe neprilygsta *MS Word AutoSummarizer, Automatic text Summarizer* bei *QuickJist Summarizer,* todėl šiuo aspektu sistemą reikėtų tobulinti ir peržiūrėti parinktą metodą bei prijungti papildomą metodą, kuris leistų *Sanla* sistema parengtai santraukai įgauti geresnę kokybę kitų sistemų atžvilgiu.



3.3 pav. Santraukų kokybė

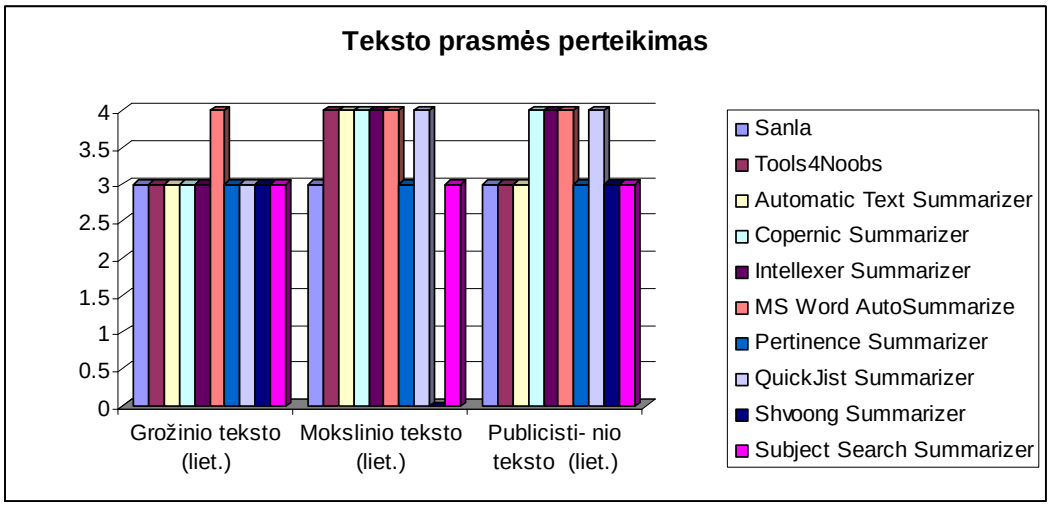
Sakinių ilgio atžvilgiu *Sanla* sistemos santraukos ilgį žodžiais, taikant 50% kompresiją galima pavadinanti vidutišku, kadangi yra santraukų sistemų, pateikiančių santraukas su didesniu žodžių skaičiumi (žr. į 3.4. pav.). Nors *Sanla* pateikiamų santraukos sakinių ilgis ir priskiriamas vidutiniškam, tačiau negalima teigti, jog santrauka išsaugo visą svarbiausią informaciją, kurią turi perduoti vartotojui.



3.4. pav. Sakinių ilgio žodžiais diagrama, naudojant 50% kompresiją.

Sanla sistema gautos santraukos palyginamos su kitų automatinių teksto santraukų rengimo sistemų parengtomis santraukomis prasmės perteikimo atžvilgiu (žr. 3.5. pav.). Iš pateitkos diagramos matome, jog respondentai gautas santraukas vertina vidutiniškai. Šiuo atžvilgiu *Sanla* vidutiniškai perteikia mokslinio žanro teksto prasmę. Kitaip nei *Shvoong*

Summarizer Sanla pateikia santrauką, todėl galima teigti jog šiuo atveju, *Sanla* sistema yra pranašesnė už *Shvoong Summarizer*. Tačiau palyginus *Sanla* sistemą, su kitų sistemų santraukų prasmės perteikimo įvertinimais, *Sanla* sistemą dar reikia tobulinti.



3.5. pav. Santraukų prasmės perteikimo įvertinimas

Iš pateiktų diagramų galima teigti, jog *Sanla* sistema nusileidžia komercinėms santraukų sistemoms, tačiau neretai santraukos kokybe ar prasmės perteikimu prilygsta nekomercinėms santraukų sistemoms. Iš atliktos analizės daroma išvada, jog vienu metodu paremta santraukų sistema efektyvi ne visada, todėl rekomenduotina šią sistemą tobulinti priungiant papildomą metodą.

4. Vartotojo dokumentacija

4.1. Bendrasis aprašas

Projekto užsakovas. Programa kuriama kaip KTU informacinių technologijų magistrantūros baigiamasis darbas.

Projekto vykdytojas. Kauno technologijos universiteto Informatikos fakulteto IFN 9/1 grupės magistrantė Jurgita Lasytė.

Produkto vartotojas. Bet kuris asmuo, besinaudojantis informacinėmis technologijomis, dirbantis su įvairiais informacijos šaltiniais, siekiantis sumažinti nereikšmingos informacijos kiekį informacijos šaltiniuose bei atrinkti tik jam aktualią informaciją.

Reikalavimai vartotojo sąsajai
Vartotojo sąsaja turi būti lengvai valdoma, aiški, suprantama.

Projekto finansavimas. Projektas ir jo įgyvendinimas nėra finansuojamas. Projektą įgyvendina ir finansuoja (jei to reikia) darbą atliekantis studentas.

Patikimumas ir kokybė. Kuriama programa yra bandymas (bandomoji), todėl šiam projektui nėra keliami aukšti kokybės reikalavimai, kaip Copernic Summarizer ar kitoms panašioms sistemoms. Projektas realizuojamas kaip bandymas ir tolimesnių tyrimų objektas.

Projekto nesėkmės tikimybė. Lyginant įvairių panašių projektų patirtį, dažnai prieinama išvada, jog tobulai veikiančią santraukos sistemą sukurti yra pakankamai sunku, kadangi atsiremiama į įvairius teksto prasmės perteikimo bei kitus niuansus. Iš turimos informacijos aišku, jog sukurtos panašios sistemos iš dalies gerai vykdo tam tikras funkcijas. Įgyvendinant projektą yra nemaža tikimybė, jog dalies užsibrėžtų uždavinių nepavyks realiai įgyventi, ar juos įgyvendinus – veiks neteisingai.

4.2. Sistemos funkcinis aprašymas, paskirtis

Sistema gali parengti nesudėtingą santrauką, naudodamasi suprogramuotais algoritmais. Santrauką galima parengti pagal užprogramuotus metodus, kurie taikomi sistemoje.

Sistema gebės atrinkti sakinius pagal jų ilgį, ir pateikti vartotojui santrauką. Sistema rinkdama sakinius, pagal sakinio ilgį, atrinks į santrauką atrenka sakinius, kurie yra lygūs ar ilgesni už nurodytą sakinio ilgį algoritme.

Programa apdoroja duomenis, kurie įrašyti *.txt tipo dokumentuose. Kito tipo dokumentai nėra palaikomi.

Sistema skirta parengti teksto santraukai, taip sutrumpinant laiką, skiriamą tekstams skaityti ir pagrindinei informacijai ieškoti.

Vartotojo sąsaja ir valdymas

1. Vartotojo sąsaja turi būti kuo paprastesnė ir kiek įmanoma lengviau valdoma.
2. Vartotojo sąsajos paprastumas, neturi būti kliūtis įrankio funkcionalumui.
3. Vartotojo sąsajos valdymui taip pat turi būti naudojamos ir trumpos klavišų kombinacijos.
4. Įvairūs vartotojo sąsajos elementai (meniu, dialogo langai ir kt.) turi būti paprasti, kad supažindinant vartotoją su sistema, apmokymo laikas būtų kuo trumpesnis.

Programa valdoma naudojant įrankių juostą bei klavišų kombinacijas. Tekstas įkeliamas ir parodomas teksto lauke. Tame pačiame lauke pateikiama santrauka (generavimo metu senas tekstas išvalomas ir pakeičiamas nauju).

4.3. Sistemos vadovas

1. Santraukų rengimo sistema "Sanla", toliau vadinama sistema, skirta įvairių žanrų tekstų santraukos rengti.
2. Ji yra sukurta mokslo bandomaisiais tikslais, todėl jos pateikiami rezultatai nebūtinai turi atitikti analogiškų profesionalių sistemų pateikiamus rezultatus.
3. SARS darbo eiga

Parsisiunčiame suarchyvuotą sistemos failą ir jį išarchyvuojame. JAR failas paprastai saugomas DIST kataloge, pavadinimu „sanla“. Jei tokio failo neradote, komandinėje eilutėje parašykite java -jar "sanla.jar". Sistemos archyviniame faile gali būti įdėtas ir Sanla.exe failas. Jį paleidus

automatiškai pasileidžia sistema. Jei neradote nė vieno iš minėtų failų – praneškite Skaityk.txt faile esančiu elektroniniu adresu autorei, ir jums bus atsiųstas kitas suarchyvuotas failas.

4. Sistemos valdymas

Sistema valdoma sukurtais mygtukais.

5. Tekstą į atsivėrusį sistemos langą galite įkelti meniu esančia komanda arba tiesiog nukopijuoti į pirmąjį teksto lauką.

6. Sistemos lange matomi 9 mygtukai, ant kurių užrašyta:

"10% santrauka" - pateikiama santrauka, kuri yra ~ 10% pradinio teksto;

"20% santrauka"- pateikiama santrauka, kuri yra ~ 20% pradinio teksto;

"30% santrauka"- pateikiama santrauka, kuri yra ~ 30% pradinio teksto;

"40% santrauka"- pateikiama santrauka, kuri yra ~ 40% pradinio teksto;

"50% santrauka" - pateikiama santrauka, kur yra ~ 50% pradinio teksto;

"Išvalyti" - išvalomi sistemos teksto langai.

"Atidaryti duomenų failą" – iškviečiamas failo atidarymo dialogas, kurio pagalba pasirenkamas duomenų šaltinis.

"Išeiti iš programos" – programa baigia darbą. Darbo rezultatai nėra saugomi.

"Teksto duomenys" – išvedami į atitinkamus programos laukus duomenys apie įkelto teksto sakinių skaičių, žodžių kiekį tekste bei vidutinį sakinio ilgį.

4.4. Sistemos instaliavimo dokumentas ir administratoriaus vadovas

Norint paleisti sistemą, reikia, jog jūsų kompiuteris atitiktų šiuos keliamus reikalavimus:

a) Minimalūs reikalavimai kompiuteriui: 512 RAM operatyviosios atminties, Windows XP ar naujesnė operacinė sistema, 1 GHz procesorius ir 50 MB laisvos vietos diske (jei vis dėl to sistemą diegiate).

b) Kompiuteryje turi veikti Java platforma, kitaip sistema neveiks. (Rekomenduojama įdiegti Java Runtime Environment Version 6).

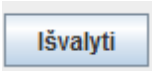
c) Sistema netestuota atvirojo kodo operacinėse sistemose, todėl sistemos kūrėjas už galimus nesklandumus neatsako.

3. Sistemos paleidimas

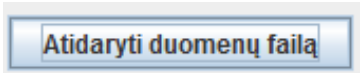
Parsisiunčiame suarchyvuotą sistemos failą ir jį išarchyvuojame. JAR failas paprastai saugomas DIST kataloge, pavadinimu „sanla“. Jei tokio failo neradote, komandinėje eilutėje

parašykite java -jar "sanla.jar". Sistemos archyviniame faile gali būti įdėtas ir Sanla.exe failas. Jį paleidus automatiškai pasileidžia sistema. Jei neradote nė vieno iš minėtų failų – praneškite Skaityk.txt faile esančiu elektroniniu adresu autorei, ir jums bus atsiųstas kitas suarchyvuotas failas.

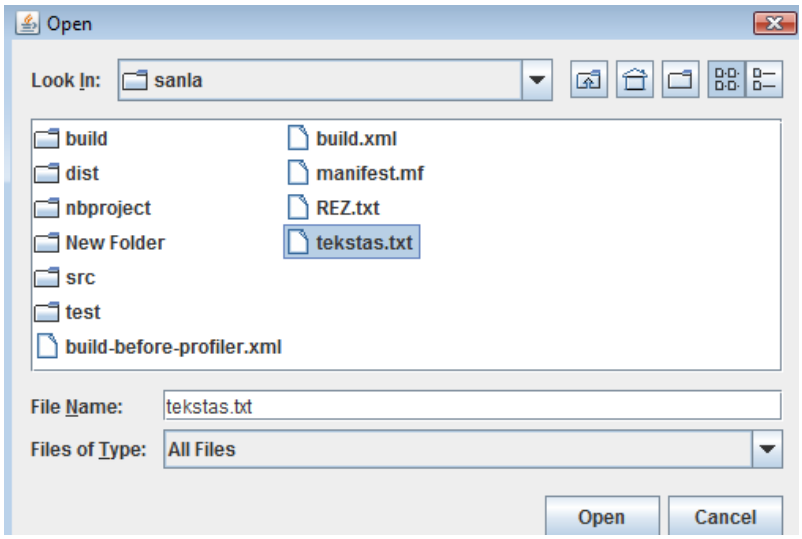
4. Sistemos valdymas

Paleidus sistemą tekstiniame lauke prašoma perskaityti Skaityk.txt failą. Pašalinti tekstiniame laukelyje esančius duomenis galima paspaudus mygtuką 

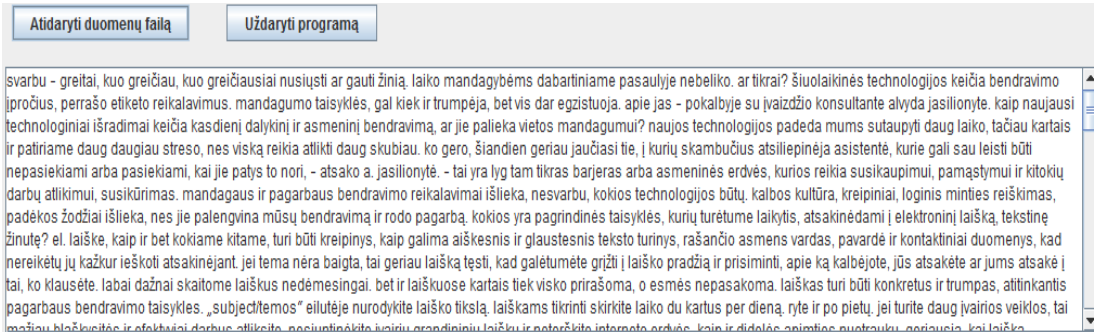
Duomenų failas privalo būti **txt** formato, kitaip įkeltas tekstas yra iškraipomas ir nesuprantamas. Jei įkėlus tekstą į programos langą vietoj lietuviškų simbolių matote neaiškius ženklus – patikrinkite, ar jūsų duomenų failas išsaugotas UTF-8 koduote. Jei vis dėl to sistema neatpažįsta lietuviškų simbolių

Tekstinis failas įkeliamas paspaudus mygtuką 

ir atsivėrusiame dialoge pasirinktus tekstinį duomenų failą:



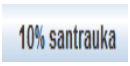
Tekstiniame lauke atsiranda tekstas iš tekstinio failo. Jei tekstas netelpa į numatytąjį teksto lauko plotį, pridedama vertikalioji juosta, kuri leidžia peržiūrėti ar įkeltas visas tekstas.



Jei norite sužinoti, kiek tekste yra sakinių, žodžių ir koks vidutinis originalaus teksto sakinio ilgis spauskite 

. Teksto ilgis sakiniais ir žodžiais pateikiamas iš karto po mygtuku, o vidutinio sakinio ilgio duomenys pateikiami antrame teksto lauke.

Norėdami gauti santrauką, spauskite vieną iš pateiktų mygtukų:

Paspaudę 

mygtuką, gausite 10% santrauką (atitinka ~10% originalaus teksto ilgio).

Paspaudę 

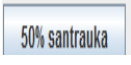
mygtuką, gausite 20% santrauką (atitinka ~20% originalaus teksto ilgio).

Paspaudę 

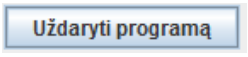
mygtuką, gausite 30% santrauką (atitinka ~30% originalaus teksto ilgio).

Paspaudę 

mygtuką, gausite 40% santrauką (atitinka ~40% originalaus teksto ilgio).

Paspaudę 

mygtuką, gausite 50% santrauką (atitinka ~50% originalaus teksto).

Norėdami išeiti iš sistemos, spauskite 

mygtuką arba 

5. Sistema yra atvirojo kodo, todėl esant būtinybei administratorius ar vartotojas gali keisti sistemos programinį kodą.

Išvados

1. Atliktas kokybinis esamų skaitmenizuoto teksto automatinio rengimo santraukų tyrimas, kuriame palyginti testuojamų tekstų rezultatai ir nustatyti metodai, kuriais remiantis, parengiama informatyvi santrauka. Efektyviomis santraukų sistemomis galima laikyti tas, kurios naudoja du, vienas kitą papildančius, metodus. Mažiausiu efektyvumu išsiskiria nepilnas sakinio ilgio metodas, kai jis nėra tinkamai suformuotas ir neprijungtas ar nepapildantis antrojo metodo.
2. Tyrimas vykdytas dviem etapais. Pirmajame išrinktos kelios automatinio rengimo santraukų sistemos ir jomis sugeneruoti įvairios kompresijos tekstai. Gauti duomenys apie santraukas palyginami pagal santraukos suprantamumą, kokybę, įskaitomumą, sakinių ilgį, skaičių, žodžių kiekį sakinyje. Nustatyta, jog naudojant skirtingus metodus gaunamas labai panašus rezultatas – sakinių skaičius, ilgis ar žodžių skaičius skiriasi nežymiai. Kuriamos santraukų sistemos prototipu buvo pasirinktos sistemos naudojančios sakinio ilgio metodą. Metodas buvo įgyvendintas JAVA aplinkoje. Sukurta aplinka vartotojui, kuri leidžia parengti santrauką. Sukurtoji sistema pagal atliktą analizę bei gautus rezultatus nežymiai nusileidžia kitiems analogams. Vienu metodu paremta santraukų sistema ne visada pateikia informatyvią santrauką.
3. Sukurtoji sistema naudinga tolimesniuose su santraukų rengimo sistemų bei metodų tyrimams bei taikymams. Ją galima naudoti rengiant publicistinių tekstų santraukas, bei kitų žanrų tekstų santraukų generavimui ir gautų rezultatų analizei.
4. Pagal gautus taikymo ir analizės rezultatus bandomoji sistema nusileidžia esamoms komercinėms sistemoms, todėl reikalingas tolimesnis tobulinimas, kuris gali būti sėkmingas tik atlikus papildomus sisteminius visų esamų metodų tyrimus ir santraukų technologijas paremtas šiais metodais. Toks būdas racionaliai suprojektuoti skaitmenizuotų santraukų automatinio rengimo prototipą ir atidžiau atsirinkti metodus ir algoritmus.
5. Santraukų sistemų analizės rezultatai pristatyti Kauno technologijos universiteto Humanitarių mokslų fakulteto rengiamoje konferencijoje „Kalbos teorija ir praktika“ bei ir straipsnyje „Skaitmenizuotų tekstų santraukų rengimo sistemos: panašumai, skirtumai ir taikymas lietuvių kalbai“ žurnale „Kalbų studijos. 2010. Nr.17, 46-52 psl.“

Literatūra

1. Bravan, K.R.S (2008), High compression rate text summarization. Massachusetts Institute of Technology.
2. Hassel, M., (2004) Evaluation of automatic text summarization, Stockholm, Universitetsservice US AB
3. Hassel, M.. Automatic text summarization evaluation. A survey of methods and and tools. Stockholm, Royal Institute of Technology
4. Yang, C.C., Wang, F.L.,(2008) Hierarchical summarization of large documents. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 59(6):887–902.
5. Mani, I. (2001), Automatic Summarization. Philadephia, John Benjamins Publishing Company. 300 p.
6. Mani, I. (1999), Maybury, M.T. Advances in Automatic Text Summarization. MIT Press. 434 psl.
7. Mani, I. Recent developments in text summarization. The MITRE Corporation
8. Mani, I. Summarization Evaluation: An Overview. The MITRE Corporation
9. Mitkov, R., (2005). The Oxford Handbook of Computational Linguistics. Oxford, Oxford University Press. 583-598 p.
10. Moens, M.F., (2000). Indexing and abstracting of document texts. Kluwer Academic Publishers. 290 p.
11. Moens, M.F., Szpakowicz, S., (2004) Text summarization branches out. Barselona, Forum Convention Centre

Šaltiniai

1. Automatic text summarizer. Search and Information Extraction Research Lab. Language technologies research center International institute of information technology. Puslapis neveikia. Žiūrėta [2011-05-01]. Prieiga per internetą: http://search.iiit.ac.in/~jags/summarizer/index.cgi
2. Copernic summarizer. Copernic Inc. Interaktyvus. Žiūrėta [2011-05-01]. Prieiga per internetą: http://www.copernic.com/en/products/summarizer/
3. Intellexer summarizer. Intellexer. Interaktyvus. Žiūrėta [2011-05-01]. Prieiga per internetą: http://summarizer.intellexer.com/
4. Pertinence summarizer. Pertinence Mining. Puslapis neveikia. Žiūrėta [2011-05-01]. Prieiga per internetą: http://www.pertinence.net/ps/main.jsp?ui.lang=
5. QuickJist summarizer. Vitaly Demin. Puslapis neveikia. Žiūrėta [2011-05-01]. Prieiga per internetą: http://www.quickjist.com/
6. Shvoong summarizer. Shvoong Ltd. USA Interaktyvus. Žiūrėta [2011-05-01]. Prieiga per internetą: http://www.shvoong.com/summarizer/
7. Subject Search Summarizer. Kryloff technologies, Inc. Interaktyvus. Žiūrėta [2011-05-01]. Prieiga per internetą: http://www.kryltech.com/summarizer.htm
8. MS Word AutoSummarize – Word Help. Microsoft Corpeoration. Interaktyvus. Žiūrėta [2011-05-01]. Prieiga per internetą: http://www.microsoft.com/education/autosummarize.mspx
9. Tools4Noobs. Tools4Noobs. Interaktyvus. Žiūrėta [2011-05-01]. http://www.tools4noobs.com/summarize/

Priedai

1 priedas. Straipsnis „Skaitmenizuotų tekstų santraukų rengimo sistemos: panašumai, skirtumai ir taikymas lietuvių kalbai“

Kalbų studijos. 2010. 17 NR., 46-52 psl.

ISSN 1648-2824

Skaitmenizuotų tekstų santraukų rengimo sistemos: panašumai, skirtumai ir taikymas lietuvių kalbai

Jurgita Lasytė, Bronius Tamulynas

Anotacija. Informacinei atskirčiai mažinti bei žmogaus laikui taupyti yra sukurti automatiniai santraukų rengimo instrumentai, kurių esmę sudaro specialiosios kompiuterinės santraukų komponavimo programos. Automatinis teksto santraukos rengimas yra vienas iš svarbių kompiuterinės lingvistikos tyrimo objektų, siekiant gauti prasmingesnį ir kokybiškesnį rezultatą. Apie šias technologijas Lietuvoje kol kas turime mažai informacijos, nors užsienyje kompiuterinės lingvistikos specialistai jas kuria ir nuolat tobulina. Šiuo metu didelė dalis santraukų automatinio rengimo sistemų (SARS) neapsiriboja vien skaitmeniniu tekstu, bet siekia apimti įvairialypę informaciją, cirkuliuojančią pasauliniame interneto tinkle – tekstą, vaizdinę ir garsinę medžiagą, paveikslėlius ir pan. Tyrimui buvo atrinktos automatinės **teksto** santraukų rengimo sistemos ir technologijos, kurios yra daugiau artimos ir reikšmingos kompiuterinės lingvistikos sričiai. Nustatyta jų praktinė nauda ir galimybės jas taikyti lietuvių kalbos tekstams. Atrinkę ir susipažinę su keliomis populiariausiomis SARS ir technologijomis (*summarizers*), atskleisime skaitmeninio teksto santraukų rengimo skirtumus, atsirandančius kuriant santraukas įvairių kalbų ar žanrų tekstams su skirtingo tipo sistemomis.

Reikšminiai žodžiai: Kompiuterinė lingvistika, automatinis santraukų rengimas, teksto glaudinimas, santraukos įskaitomumas, santraukos prasmingumas.

1. Įvadas

Pastaraisiais metais ypač padaugėjo informacijos, pasiekiančios žmogų per internetą. Informacinei atskirčiai mažinti ir žinių suvokimo efektyvumui padidinti buvo sukurti automatiniai santraukų rengimo instrumentai. Jų esmę sudaro specialiosios kompiuterinės santraukų komponavimo programos. Pastaruoju metu SARS tampa ypač svarbiu kompiuterinės lingvistikos tyrimo objektu. Apie panašias sistemas Lietuvoje kol kas turime mažai informacijos, nors užsienyje jos intensyviai kuriamos ir tobulinamos. Kadangi mūsų šalyje kol kas nėra sukurta panašaus tipo sistemų, todėl yra svarbu susipažinti su jų taikymo praktine nauda. Dabartinės automatinės santraukų rengimo technologijos neapsiriboja vien skaitmeniniais tekstais, bet siekia apimti įvairialypę informaciją, cirkuliuojančią pasauliniame interneto tinkle – tekstą, vaizdinę ir garsinę medžiagą, paveikslėlius ir pan. Šiame straipsnyje aprašomos ir tiriamos automatinės **teksto** santraukų rengimo technologijos ir sistemos, kurios labiausiai artimos ir reikšmingos kompiuterinės lingvistikos sričiai. Tyrimui buvo atrinktos 8 plačiau žinomos ir dažniau naudojamos SARS technologijos ir sistemos (*summarizers)* bei pabandyta jas taikyti lietuvių kalbos tekstams. Atskleisti skaitmenizuoto teksto santraukų sistemų skirtumai, kurie atsiranda kuriant santraukas skirtingų kalbų ar žanrų tekstams.

2. Skaitmenizuoto teksto analizės ir apdorojimo metodai santraukoms rengti

Pirmieji įrašai apie kompiuterizuotas tekstų santraukos technologijas ir sistemas randami XX amžiaus šeštojo ir septintojo dešimtmečių moksliniuose darbuose bei publikacijose. Jas aprašė Luhn ir Edmundson (Mitkov 2005). Po keleto dešimtmečių pertraukos, patobulėjus kalbos apdorojimo technologijoms, išsiplėtus kompiuterio atminčiai ir padidėjus kompiuterio spartai bei atsiradus poreikiui skaitmenizuoti didelius tekstų rinkinius (tekstynus, internetinės informacijos srautus ir pan.), susidomėjimas SARS technologijomis (Mitkov 2005) vėl atgijo. Vienas iš aktualiausių ir daugiausiai ginčų keliantis klausimas yra pačio žodžio *santrauka* samprata. Mokslininkai dirbantys, bei rašantys apie SARS pačią santrauką dažnai apibūdina skirtingai. Paprastai naudojama bendroji santraukos

samprata. *Santrauka - tekstas, kuris gaunamas iš vieno ar daugiau tekstų, glaustai pateikiantis svarbiausią originalo informaciją ir yra ne ilgesnis kaip pusė pradinio (originalaus) teksto* (Mitkov 2005). SARS kūrimo tikslai nuo XX amžiaus 6-ojo dešimtmečio mažai pasikeitė – tebesiekiama pateikti ilgų tekstų trumpas, bet informatyvias santraukas. Gal būt išskirtina šiuolaikinių SARS technologijų savybe reikėtų laikyti jų gebėjimą apimti ne tik tekstą, bet ir garsinę bei vaizdinę medžiagą, internetinius straipsnius ar portalus (Mitkov 2005).

Tekstui apdoroti dažniausiai taikomas statistika paremtas modelis – renkami statistiniai duomenys apie žodžių vartojimą, kurie naudojami teksto apdorojimo procese. Santraukoms generuoti ir jas formuoti įprasta išskirti šiuos metodus.

1. Pozicinis arba vietos metodas. Daugumos žanrų tekstai paklūsta būtent tam žanrui priskiriamoms taisyklėms (antraštės, pavadinimai, pastraipos ir pan.). Naudodama šį metodą sistema vadovaujasi principu – kas yra pavadinime, antraštėje ir pirmoje pastraipoje – svarbūs dalykai. Kitose pastraipose esanti informacija laikoma ne tiek svarbia, nes deklaruojama, kad esminė informacija pateikiama teksto pradžioje.

2. Raktinio/reikšminio žodžio metodas. Įvairiuose žanruose tam tikri žodžiai ir frazės aiškiai parodo svarbumą (pvz.: „reikšmingas“, „šiame darbe“, „svarbiausias“ ir pan.), todėl sakiniai su reikšminiais žodžiais įtraukiami į santrauką.

3. Žodžio ir frazės dažnio metodas pagrįstas besikartojančių žodžių dažniu. Kuo dažniau žodis kartojasi, tuo svarbesni sakiniai su tais žodžiais.

4. Reikšminių žodžių (užklausos) ir dalinio pavadinimo sutapimo metodas pažymi sakinį pagal reikalingų žodžių kiekį sakinyje. Reikalingi žodžiai yra tokie, kurie yra teksto pavadinimuose, antraštėse, ar sutampa su vartotojo užklausoje įvestais reikšminiais žodžiais.

5. **Rišlumo arba leksinio sujungimo metodas.** Žodžiai gali būti sujungiami įvairiais būdais, įskaitant kartojimą, sinonimiką, semantinius ryšius, kurie atsispindi žodynuose. Naudojant šį metodą pasitelkiamas papildomas informacijos šaltinis (žodynas), ir atrenkami atitinkami sakiniai (Mitkov 2005:585-587, Mani 2001:13-14, 18).

3. Praktinis santraukų automatinio rengimo sistemų taikymas

Straipsnyje pateikiami rezultatai, gauti atlikus tyrimus su lietuviškais ir angliškais mokslinio, publicistinio ir grožinio stiliaus tekstais. Kiekvienos santraukų sistemos reikalavimai yra skirtingi, todėl jos geriausiai veikia galingesniuose ir talpesniuose naujesnės kartos kompiuteriuose, kurie naudoja tiek senesnės kartos Windows XP ar naujausios kartos Windows 7 OS operacines sistemas su interneto ryšiu[3]. Tyrimai buvo atliekami su sistemomis, kurios gali apdoroti anglų ir lietuvių kalbos tekstus, išskyrus naujienų portalus bei garso ir vaizdo medžiagą: *Automatic Text Summarizer*; *Copernic Summarizer*; *Intellexer Summarizer, Pertinence Summarizer*; *QuickJist Summarizer*; *Shvoong Summarizer*; *Subject Search Summarizer*; *Microsoft Office MS Word 2003 AutoSummarize*. Iš pasirinktųjų – keturios laisvai prieinamos ir nemokamos sistemos. Kitos trys veikia tik internete. Jas galima suskirstyti į dvi grupes: pagal kainą – nemokamos ir mokamos; pagal įdiegimą – įdiegiamosios ir internetinės.

Didesnė dalis kompiuterizuoto teksto santraukų sistemų skiria tam tikrus tekstų žanrus ar stilius, kuriems jos geba rengti santraukas. Žanrų bei stilių neskiria *Automatic Text Summarizer, Copernic Summarizer, QuickJist Summarizer, Subject Seach Summarizer, Shvoong Summarizer ir MS Office Word 2003 AutoSummarize*. Kitos dvi sistemos (*Intellexer Summarizer, Pertinence Summarizer*) skiria tam tikrus teksto žanrus ar stilius: *Intellexer Summarizer* skirtas bendrojo, mokslinio, ekonomikos, politikos ir licencijoms; *Pertinence Summarizer* (internetinis) – be žanro (nežinomam žanrui), chemijos, finansų, teisės, spaudos, telekomunikacijoms, medicinos bei kitoms sritims.[4] Galima teigti, kad SARS pagal teksto žanrą yra skirstomos į priklausomas arba nepriklausomas nuo teksto žanro.

Kiekvienai SARS yra būdingas atitinkamas suspaudimo (glaudinimo) lygis. Dauguma jų turi vienodus teksto glaustumo lygius. Kai kuriose iš jų šis lygis vertinamas procentais, kitose – žodžių, sakinių ar tiesiog ženklų skaičiumi. Dažniausias ir populiariausias teksto santraukos glaudinimo matas yra procentinis. Procentiniu glaudinimu vadinamas atvejis, kai originaliam pradiniam tekstui yra parengiama norimos procentinės apimties santrauka. Kai kurios santraukų sistemos, leidžia pasirinkti ir įvairesnius glaudinimo lygius, tokius kaip 6%, 12%, t.y. tokius, kokių pageidauja vartotojas. Dažniausiai tokie papildomi pasirinkimai naudojami mokamose santraukų sistemose. Procentiniu glaudinimu yra paremtos: *Copernic Summarizer*; *Intellexer Summarizer*; *Pertinence Summarizer*; *QuikJist Summarizer*; *Shvoong Summarizer*; *MS Office Word AutoSummarize*.

Antras pagal populiarumą santraukos glaudinimo būdas yra santraukos ilgio nustatymas sakiniais. Dalis SARS leidžia pasirinkti sakinių skaičių, iš kurių bus sudaryta santrauka. Glaudinimą pagal sakinių skaičių naudoja: *Automatic Text Summarizer*; *Intellexer Summarizer*; *Subject Search Summarizer*; *MS Office Word AutoSummarize.* Vartotojas pats gali pasirinkti kiek sakinių nori matyti iš tam tikro teksto – tai gali būti nuo 3 sakinių iki 100 ir daugiau, priklausomai nuo teksto apimties. Toks santraukos rengimo būdas yra patogus siekiant gauti trumpesnę ir informatyvesnę santrauką.

[3] Interneto ryšys reikalingas toms SARS, kurios pateikia santraukas naršyklės lange.

[4] Dėl kitų sričių tekstų santraukų rengimo reikia konsultuotis su Pertinence Summarizer kūrėjais.

Trečiasis būdas, kai vartotojas santraukos ilgį pats pasirenka pagal žodžių skaičių. Paprastai tokioms santraukoms parengti siūlomi 100, 250, 1000 bei kitokie žodžių skaičiaus variantai. Santraukos rengimas pagal žodžių skaičių yra panašus į santraukos parengimą paremtą sakinių skaičiumi. Santraukų pagal žodžių skaičių glaudinimo būdą naudoja dvi santraukų rengimo sistemos - *Pertinence Summarizer* ir *MS Office Word AutoSummarize*. *Pertinence Summarizer* šį būdą naudoja, nes spaudžiant tekstą procentiniu teksto glaudinimo metodu automatiškai pateikiamas ir žodžių skaičius, kuris atitinka procentinį žodžių santykį. *MS Office Word AutoSummarize* tiesiog leidžia pasirinkti iš visų pateiktų sakinių mus tenkinantį skaičių. Tačiau, jei santraukos rengimas sakiniais mūsų netenkina, galima rinktis pagal procentinę teksto apimtį arba žodžių kiekį.

Kiek mažiau žinomas būdas yra santraukos rengimas pagal ženklų skaičių. Jį naudoja tik kelios sistemos. Taip rengiamos santraukos paprastai naudoja nuo 1000 ženklų iki 90000 ženklų. Galima pasirinkti neribotą ženklų skaičių, tačiau tada jau nebus santrauka, o tik atkartotas tekstas. Tam, kad būtų sukurta santrauka, atitinkanti tam tikrą procentinį santykį, patartina žinoti kiek originalo tekste yra spaudos ženklų. Santraukas glaudina pagal spaudos ženklų skaičių - *Subject Search Summarizer* ir *MS Office Word AutoSummarize*. Didžiausias skirtumas tarp šių sistemų yra tai, kad *MS Office Word AutoSummarize* leidžia atitinkamai pasirinkti *100 arba mažiau žodžių* ir *500 arba mažiau žodžių*.

Kaip parodė tyrimai, dalis SARS turi du glaudinimo būdus, iš kurių vartotojas gali pasirinkti jam tinkamesnį. Dažniausia jos naudoja procentinį glaudinimo būdą, bet šalia pateikiama glaudinimo sakiniais ar žodžiais galimybė. Reikia paminėti ir tai, kad *Pertinence Summarizer,* pateikdama santrauką procentais, automatiškai nurodo ir žodžių kiekį parengtoje santraukoje. *MS Office Word AutoSummarizer* nurodo ne tik parengtos santraukos apimtį sakiniais ir žodžiais, bet ir pradinio dokumento apimtį sakiniais ir žodžiais, bei suteikia daugiau santraukos glaudinimo būdų, nei kitos santraukų sistemos.

Kalbos atpažinimas. Santraukų rengimo sistemos taip pat atsižvelgia į originalo teksto kalbą. Beveik visos SARS atpažįsta daugiau nei vieną kalbą. Daugiausiai kalbų atpažįsta ir palaiko *Subject Search Summmarizer* – 36 kalbas, tarp jų ir lietuvių kalbą. Pagrindinis jos, kaip ir kitų sistemų, reikalavimas, kad simboliai-hieroglifai (kinų, japonų kalbų rašmenys) būtų užrašomi lotynišku alfabetu, kitaip santrauka nebus daroma, nes ji nesupras parašyto teksto.

Mažiau kalbų atpažįsta *Shvoong Summarizer* – 21 kalbos tekstus. Ji taip pat atpažįsta ir lietuvių kalba rašytus tekstus. Trečia pagal atpažįstamų kalbų skaičių yra *Automatic Text Summarizer* atpažįsta ne tik 13 kalbų tekstus, bet taip pat ir įvairių kitų kalbų tekstus parašytus lotynišku alfabetu. *Pertinence Summarizer* atpažįsta 10 kalbų tekstus, tarp jų ir lietuvių kalbos tekstus. *Copernic Summarizer* atpažįsta tik 4 kalbų tekstus, o neatpažintus ženklus ši sistema paverčia neatpažįstamais ženklais. *QuickJist Summarizer* atpažįsta viena kalba daugiau nei *Copernic Summarizer* – 5 kalbas. Suradusi neatpažįstamų ženklų tekste, *Copernic Summarizer* juos paverčia neįskaitomais simboliais. Taip pat ši sistema gali sugeneruoti santraukas ir kitiems tekstams, parašytiems lotynišku alfabetu. Mažiausiai kalbų atpažįsta *Intellexer Summarizer*, kurios pagrindinė ir vienintelė kalba yra anglų kalba. Bet tai tik tuo atveju, jei naudojama *English* versija. Be angliškosios *Intellexer Summarizer* versijos taip pat yra prancūzų ir vokiečių, kurios santraukas atlieka taip pat tik viena kalba. Neatpažintus simbolius tekste *Intellexer Summarizer* gali interpretuoti kitais atsitiktiniais simboliais. *MS Office Word AutoSummarize* taip pat tekstą atpažįsta tik tada, kai tekstas yra angliškas arba teksto kalba pakeičiama į anglų kalbą, paliekant originalo rašmenis. Kitu atveju *MS Office Word AutoSummarize* visiškai neteikia teksto santraukos.

Teksto įkėlimas. Kiekviena minėta SARS tekstą įkelia skirtingais būdais. Įkėlimo patogumą gana sunku nustatyti, nes tai priklauso nuo vartotojo pasirinkimo ir poreikio. Galima išskirti kelis įkėlimo būdais: įkeliamas dokumentas, kurio santrauką norima gauti; kopijuojamas dokumentas, kurio santrauką norima gauti; pažymimas teksto turinys ir su komanda COPY arba greitųjų klavišų kombinacija Ctrl+C išskiriamas MS Word dokumentas, kurio santrauką norime gauti.

Automatic Text Summarizer ir *Shvoong Summarizer* tekstą atpažins ir panaudos tik tada, jei jis kopijuojamas į naršyklės langą. *Copernic Summarizer* ir *Intellexer Summarizer* tekstą gali nuskaityti iš MS Word‘o ar kito tekstinio dokumento bei internetinio puslapio. *Pertinence Summarizer* santrauką pateikia keliais atvejais, jei MS Word‘o ar tekstinis failas bus įkeltas į internetinį sistemos langą ar nukopijuotas. *Pertinence Summarizer* taip pat gali parengti ir internetinio puslapio santrauką, tereikia kaip ir į *Intellexer Summarizer* ar *Copernic Summarizer* langus įrašyti ar kopijuoti internetinio puslapio adresą. Tuo tarpu *Subject Search Summarizer* ir *QuickJist Summarizer* santraukas pateikia tik tada, kai pažymėtas visas tekstas ir pasirinkta COPY komanda arba greitųjų klavišų Ctrl+C kombinacija. *MS Office Word AutoSummarize* santrauką pateikia meniu eilutėje pasirinkus *Tools – AutoSummarize* (liet. – *Įrankiai – Automatiškai apibendrinti*).

Santraukų saugojimas. SARS, sukūrusi santrauką gali ją ir išsaugoti. Šią savybę turi visos anksčiau minėtos sistemos, tačiau tai atlieka skirtingai. Santraukų saugojimo būdai: MS Word‘o dokumentai; HTML dokumentai; tekstiniai dokumentai; XML dokumentai; kopijos iš programos ar interneto lango į MS Word dokumentą; programos langų kopijos kaip ekrano vaizdas (angl. *printscreen*).

Kiekviena SARS suteikia skirtingas saugojimo galimybes. *Automatic Text Summarizer, QuickJist Summarizer, Pertinence Summarizer* leidžia tik pasižymėti parengtą santrauką ir ją nukopijuoti į MS Word‘o ar

tekstinį dokumentą. *Copernic Summarizer* leidžia santrauką išsaugoti keturiais formatais – MS Word'o, tekstinio dokumento, HTML dokumento, XML dokumento. *Intellexer Summarizer* leidžia saugoti tik dviem formatais – tekstiniu dokumentu ir HTML dokumentu. *Subject Search Summarizer* ir *Shvoong Summarizer* santrauką saugo tik kopijuotą aktyvaus programos lango vaizdą. Taip pat galima kopijuoti ir *Pertinence Summarizer* bei kitomis santraukų sistemomis parengtą tekstą. *MS Office Word AutoSummarize* leidžia tekstą išsaugoti įvairiais formatais – ne tik kopijuojant, XML, HTML, *.txt ar *.doc bet ir *.rtf formatu, MS Word 2007, Works 4.0–8.0 formatais, senesnės versijos MS Word formatais – MS Word 97 ir t.t.

SARS praktinio taikymo rezultatų analizė. Praktinio taikymo tyrimui buvo panaudoti 6 tekstai – 3 angliški tekstai, ir 3 lietuviški trijų žanrų tekstai – moksliniai, publicistiniai ir grožiniai. Angliškų ir lietuviškų tekstų apimtis sakiniais ir žodžiais pateikiamas 1 lentelėje. Santraukos kokybės vertinimą atliko 5 nepriklausomi ekspertai, kurie perskaitę pateiktas tekstų santraukas, jas vertino pagal įvairius kokybės rodiklius. Ekspertinę analizę atliko 18–22 metų amžiaus studentai, studijuojantys pagal verslo vadybos, turizmo ir viešbučių vadybos, maisto technologijos ir inžinerijos, automobilių techninio eksploatavimo ir automatikos studijų programas.

Santraukos kokybei ištirti buvo naudojami šie kriterijai: teksto suspaudimo laipsnis, gautos santraukos apimtis ir prasmės perteikimas, teksto suprantamumas ir įskaitomumas. Taip pat buvo įvertinta bendra santraukos kokybė, atsižvelgiant į visus minėtus kriterijus.

1 lentelė. Tyrime naudotų angliškų ir lietuviškų tekstų suvestinė

Kalba	Mokslinis tekstas		Grožinis tekstas		Publicistinis tekstas	
Angliškas tekstas	1417 ž.	70 sak.	2699 ž.	81 sak.	580 ž.	31 sak.
Lietuviškas tekstas	1132 ž.	71 sak.	1489 ž.	130 sak.	220 ž.	13 sak.

Praktiniam SARS tyrimui buvo paimtos aštuonios, anksčiau aprašytos ir nagrinėtos sistemos: *Automatic Text Summarizer, Shvoong Summarizer, Copernic Summarizer, Intellexer Summarizer, Pertinence Summarizer, QuickJist Summarizer, Subject Search Summarizer* ir *MS Word AutoSummarize.* Kiekvienai SARS buvo naudojami tie patys tekstai bei panašus glaustumo laipsnis. Tiriamiems ir analizuojamiems tekstams buvo taikomas 25%, 30% ir 50% glaustumas, kuris buvo parenkamas atsižvelgiant į sistemos galimybes ir leistinas parinktis. Tiriamiems tekstams buvo taikoma ir 5% bei 10% suspaudimas, siekiant surasti santraukų kokybės skirtumus tarp gautų 50%, 30%, 25%, 10% ir 5% santraukų.

Kadangi *Pertinence Summarizer, Shvoong Summarizer, Subject Search Summarizer* ir *Automatic Text Summarizer* negali glaudinti pagal procentinę išraišką, todėl joms buvo nurodomas sakinių skaičius, kuris atitinka 25 % ar 50 % viso sakinių skaičiaus arba naudojama 30 % bei 50 % suspaudimas. Sakinių glaudinimo būdas buvo tiriamas *Automatic Text Summarizer* ir *Subject Search Summarizer* programoms, o 30% glaustumas buvo taikomas *Pertinence Summarizer* ir *Shvoong Summarizer* sistemoms. Kitos sistemos gali parengti 10%, 25%, 30%, 50% ir kitokio laipsnio glaustumo santraukas. *MS Word AutoSummarizer* negali atlikti kitokių glaudinimo nei 10%, 25%, 50%. Likusios SARS leido pasirinkti suspaudimo laipsnį 5%.

Santraukų apimtys, kai glaudinama pagal sakinių skaičių, skiriasi angliškiems ir lietuviškiems tekstams. Naudojant 50% suspaudimą, angliškojo teksto santraukos apimtis skirėsi nuo 2 iki 12 sakinių. Ilgiausias angliško mokslinio teksto santraukas parengia *Copernic Summarizer* ir *Intellexer Summarizer*, o trumpiausią – *QuickJist Summarizer*. Ilgiausias santraukas *Copernic Summarizer* ir *Intellexer Summarizer* parengia lietuviškiems ir angliškiems tekstams. Iš tyrime naudotų sistemų *Shvoong Summarizer* vienintelė santraukų sistema, kuri nepateikė mokslinio teksto santraukos, kai taikėme 50% bei 30% glaustumo kriterijų, nors kitokio žanro lietuviškų tekstų santraukos buvo pateikiamos be jokių kliūčių. Ši sistema taip pat nepateikė ir įvairesnio glaudinimo atvejų lietuviškam moksliniam tekstui. Kadangi *Automatic Text Summarizer* ir *Subject Search Summarizer* negali atlikti procentinio glaudinimo, sakinių skaičius, kuris sudaro santrauką buvo gautas paėmus 50% originalaus teksto sakinių, kurie kai kuriais atvejais skyrėsi nuo kitų SARS pateikiamų sakinių santraukoje skaičiaus. *Subject Search Summarizer* pateikia papildomus sakinius tinkamus santraukai, reikšminius žodžius ar kitą papildomą teksto informaciją, todėl santraukos ilgis, gali būti ilgesnis 2-3 sakiniais nei pasirinktasis pradinis santraukos ilgis. Iš gautų duomenų galima teigti, jog angliškų tekstų santraukų apimtis sakiniais santraukų sistemose yra vienodi, arba skiriasi nežymiai. Kitaip nei anglų kalbos santraukose, lietuvių kalbos santraukos ilgis sutampa tik publicistinio teksto santraukose, o mokslinio ir grožinio tekstų santraukos savo apimtimi skiriasi, nors naudota toks pats glaustumo kriterijus ir tas pats tekstas. Taip yra todėl, kad trumpesnio ir paprastesnio teksto santraukas parengti lengviau, nei ilgesnio ir sudėtingesnio. Iš tikrųjų santraukų apimtys skyrėsi nežymiai ir nebuvo sistemų, pateikiančių vienodo ilgio santraukas, kas pasitaikė atliekant angliškų tekstų glaudinimą.

Vidutinė angliško grožinio teksto santrauka tyrimo metu, naudojant 50% glaustumo laipsnį, apytikriai gali būti prilyginta 38,75 sakiniams, angliško mokslinio teksto vidutinė santrauka – 29,88 sakiniai, angliško publicistinio teksto santrauka – 15,125 sakiniai. Vidutinis lietuviško grožinio teksto santraukos ilgis naudojant 50% glaustumą yra 57,25 sakiniai, mokslinio – 30,71, o publicistinio – 6,125 sakiniai. Nors originalus angliškas grožinis tekstas buvo

ilgesnis nei lietuviškas grožinis tekstas, tačiau gauta lietuviška grožinio teksto santrauka yra ilgesnė nei atitinkamo angliško teksto santrauka.

Pastebėta, jog ne visose santraukose su tuo pačiu glaudinimo laipsniu sakinių skaičius yra vienodas. Tarkime, veikimu ir santraukų rengimo principais panašios *Copernic Summarizer* ir *Intellexer Summarizer* ne visada pateikia vienodos apimties santraukas. Grožinių tekstų anglų ir lietuvių kalbomis santraukos, kurias pateikia minėtosios sistemos yra sudarytos iš vienodo skaičiaus sakinių, tačiau kitų žanrų santraukų apimtys skiriasi – *Copernic Summarizer* pateikia trumpesnes santraukas, nei *Intellexer Summarizer*.

Pagal procentinį 25% glaustumą grožinio angliško teksto santraukos sakinių skaičiumi daugeliu atveju yra vienodos ar panašios. Pastebima, kad *Pertinence Summarizer* parengta santrauka yra tik vienu sakiniu ilgesnė, nors buvo naudojama 30% glaudinimas. Tačiau *Shvoong Summarizer* parengta santrauka yra trumpesnė už 25% atitinkamo teksto santraukos sakinių skaičių, nors naudojama 30% glaudinimas. Trumpiausią teksto santrauką sakinių skaičiumi pateikia *QuickJist Summarizer*, nors jam šio bandymo metu buvo priskirta 25% glaustumas. *Subject Search Summarizer* santrauka yra ilgesnė dviem sakiniais, dėl to, kad pridedama pora papildomų sakinių, apibūdinančių kurią nors teksto dalį ar veiksmą, ar pateikiant papildomą informaciją apie tekstą ir reikšminius jo žodžius. Į *Copernic Summarizer* ir *Intellexer Summarizer* santraukos ilgį neįskaičiuojamos reikšminių žodžių eilutės, kurios nurodo pagrindinius teksto žodžius.

Pastebėta, kad *QuickJist Summarizer* pateikiamos 25% santraukos lietuvių ir anglų kalbų tekstuose yra trumpiausios, o ilgiausiomis galima laikyti *Subject Search Summarizer* pateikiamas santraukas. Tačiau *Subject Search Summarizer* santraukos ilgis buvo pasirinktas pagal *MS Word AutoSummarize* pateiktų sakinių 25% santraukos glaudinimą, kadangi pati sistema nenaudoja jokių procentinių glaudinimo būdų.

2 lentelė. Lietuviško teksto santraukų ilgis remiantis sakinių skaičiumi, naudojant 50% kompresiją.

Programa	Grožinis tekstas	Mokslinis tekstas	Publicistinis tekstas
Automatic Text Summarizer	65 sak.	35 sak.	6 sak.
Copernic Summarizer	67 sak.	40 sak.	9 sak.
Intellexer Summarizer	66 sak.	34 sak.	6 sak.
MS Word AutoSummarize	56 sak.	31 sak.	5 sak.
Pertinence Summarizer	39 sak.	26 sak.	6 sak.
QuickJist Summarizer	53 sak.	21 sak.	5 sak.
Shvoong Summarizer	44 sak.	-	7 sak.
Subject search Summarizer	68 sak.	28 sak.	5 sak.

Kiekvienos SARS pateikiamos santraukos nors glaudinamos tokiu pačiu – 50 % laipsniu, bet kiekvienoje santraukoje žodžių skaičius yra skirtingas. Kaip minėta anksčiau, *Shvoong Summarizer* nepateikia lietuviško mokslinio teksto santraukos, todėl šios sistemos santrauka nėra nagrinėjama. Iš turimų duomenų matyti, jog santraukos apimtis žodžiais daugiausia skiriasi 653 žodžiais, o mažiausiai - 6 žodžiais. Vidutiniškai visų tekstų santraukų ilgis yra 654 žodžiai santraukai, tačiau santraukos ilgis priklauso nuo teksto ilgio bei glaudinimo laipsnio, naudoto tekstams. Tam tikroje dalyje santraukų (ir lietuviškų, ir angliškų) žodžių skaičius atitinkamo stiliaus tekste yra panašus ir skiriasi vidutiniškai 30 žodžių. Tačiau santraukos, kurios yra ilgesnės, dažniausiai yra suprantamesnės ir prasmingesnės. Esant mažesniam 30% glaudinimui sistemoms *Shvoong Summarizer* ir *Pertinence Summarizer* santraukų ilgiai skiriasi 46 žodžiais. *Pertinence Summmarizer* pateikiamos santraukos ilgis žodžiais ir kitos sistemos, tokios kaip *Intellexer Summarizer* žodžių skirtumas yra palyginti labai mažas (51 žodis, o palyginus su *Automatic Text Summarizer* dar mažesnis – 41 žodis).

Skirtumai tarp to paties glaustumo teksto santraukoms svyruoja priklausomai nuo to, kokius santraukų rengimo būdus jos naudoja. Sistemos, kurios santraukos rengimui naudoja sakinių skaičių automatiškai pateikia didesnį žodžių kiekį, nes dažniausiai į santrauką įtraukiami ilgiausi teksto sakiniai. Nors *Automatic Text Summarizer* ir *Subject Search Summarizer* glaudinimui pasirinktas vienodas sakinių skaičius, atitinkantis 25% viso teksto, tačiau šių sistemų pateiktos santraukos skiriasi 234 žodžiais. Toks skirtumas sistemoms naudojančioms tą patį glaudinimo būdą yra pakankamai didelis. Tačiau, esant ir kitokam glaudinimo lygiui santraukose kartais pastebimi gana dideli žodžių skaičiaus skirtumai. Eksperimentai parodė, kad trumpiausias santraukas žodžiais pateikia *QuickJist Summarizer*, kuris, kaip ir likusios sistemos santrauką rengia naudodamas 25% glaudinimo laipsnį. Vidutiniškai nuo kitų santraukų *QuickJist Summarizer* parengta santrauka skiriasi ~260 žodžiais. Tai, palyginus su kitomis sistemomis ir kitokiais glaudinimo lygiais yra didelis skirtumas, nes vidutinis skirtumas yra ~27,33 žodžiai. Kaip ir žodžių kiekis santraukoje, taip ir vidutinis jos ilgis kiekvienu atveju yra skirtingas. Skirtumai, kaip jau minėta, dažniausiai priklauso nuo glaustumo laipsnio, santraukos ilgio ir sistemos galimybių. Skirtumai tarp vidutinio santraukos sakinių ilgio žodžiais nėra dideli – kai kuriais atvejais mažesni nei naudojant 25% ar 30% glaudinimo lygį. Naudojant 50% glaudinimo laipsnį vidutinis sakinio ilgis skiriasi nuo 2 iki 10 dešimties žodžių sakiniui. Tai nėra daug, nors bendras sakinių skaičius santraukose skiriasi. Iš gautų duomenų galima teigti, jog SARS santraukose

ilgiausius sakinius pateikia, priklausomai nuo teksto stiliaus, ilgio bei kalbos. Tendencija, jog trumpiausias 50% suspaudimas santraukas pateikia *QuickJist Summarizer*, išlieka beveik visose lietuvių ir anglų kalbų santraukose. *QuickJist Summarizer* pateikiamos santraukos yra trumpesnės nei kitos, tačiau dėl to santrauka nepraranda prasmingumo ir suprantamumo.

Iš pateiktų duomenų matome, jog vidutiniškai ilgiausius sakinius santraukoje pateikia *Shvoong Summarizer*. Kaip jau minėta anksčiau, taip yra todėl, kad šiuo atveju *Shvoong Summarizer* naudoja 30% glaudinimą, ir ima didesnę teksto dalį, nei kitos santraukų sistemos. Nors *Pertinence Summarizer* taip pat naudoja 30% glaudinimą, tačiau šios sistemos pateikiama santrauka nuo 25% teksto santraukos skiriasi nežymiai. *Automatic Text Summarizer* ir *Subject Search Summarizer* vidutinis sakinio ilgis skiriasi 8,67 žodžio (plg. su *Pertinence summarizer* ir *Shoong Summarizer*, kur skirtumas lygus 24,84 žodžiams).

Kiekvienas tekstas, kaip ir kiekviena santrauka turi prasmę, kuri koduojama žodžiais, sakiniais, pastraipomis. Tačiau dažnai iškyla klausimas, ar pasinaudojus SARS iš gautosios santraukos bus įmanoma adekvačiai nustatyti teksto turinį, neperskaičius originalaus teksto. Iš gautų rezultatų, galima teigti, jog teksto prasmė santraukose priklauso nuo glaustumo laipsnio. 50% glaustumo santraukos yra aiškesnės, prasmingesnės, nors ir atitinkamai ilgesnės. Naudojant 25%, 10%, 5% glaustumą, gaunamos santraukos yra trumpesnės, tačiau ne visada informatyvios, prasmingos ir suprantamos. Dažnai vyrauja padriki sakiniai, bendras kontekstas sunkiau suprantamas. Ypač tai pastebima grožinio žanro kūriniuose, kur sakiniai ištraukiami iš bendro konteksto ir pateikiami kaip santraukos dalis. Mokslinio žanro santraukoje sakinių išsiderinimas yra mažiau pastebimas, bet, atidžiau skaitant, jį galima įžvelgti. *Subject Search Summarizer* mokslinio teksto santraukoje pateikia keletą vienodų sakinių, kuriuos kitos SARS santraukoje pateikia tik vieną kartą.

Naudojant mažesnę nei 25% glaustumą tekstas tampa sunkiau suprantamas, o naudojant žemiausią 5% glaustumą gaunamas tekstas yra beprasmis ir mažai adekvatus originalui. Iš to galima teigti, jog geriausiai teksto prasmę atspindi nuo 25% iki 50% glaudintos santraukos. Tai patvirtina išvadą, jog kuo daugiau spaudžiamas tekstas, tuo mažiau prasmės lieka santraukoje.

Teksto santraukos suprantamumas (nesigilinant į prasmę, o tik į sakinių išdėstymą). Santraukos suprantamumas yra vienas iš svarbiausių kokybės kriterijų, kaip esminis reikalavimas gerai santraukai. Deja tobulą santrauką gauti nėra lengva. Tiek lietuviškuose, tiek angliškuose santraukų tekstuose pasitaiko įvairių klaidų. Dažnai SARS programose pateikiamų santraukų sakiniai būna išplėšti iš konteksto, prasideda mažosiomis raidėmis, pasitaikė keli atvejai, kai buvo sujungiami žodžiai, romėniški skaičiai painiojami su raidėmis. Tokie dalykai pastebimi angliškame grožiniame tekste, kuriame romėnišku skaitmeniu žymimas skaičius *„1“* supainiojamas su angliškuoju raidės *„I“* simboliu. Pavyzdžiui, *QuickJist Summarizer* grožinio teksto santraukoje yra sakinys *I HAD just been looking long and sadly at Holbein's ploughman, and was walking through the fields, musing on rustic life and the destiny of the husbandman*. Sistemos, kurios nepainiojo skaičiaus vienas su žodžiu *„I“* yra *Automatic Text Summarizer* ir *Shvoong Summarizer*. Kitos sistemos *„I“* priskiria prie šalia esančio sakinio, kaip įvardį „aš“, dar kitos – „suklijuoja“ su prieš tai einančiu žodžiu, taip sukurdamos nesamą žodį. Žodžių suliejimų pasitaiko ir *Subject Search Summarizer*, *Automatic Text Summarizer*.

Dalis SARS sakinius tiesiog paima iš teksto, neskirdamos nei mažųjų, nei didžiųjų raidžių, taip sudarydamos įspūdį, jog sakinys ne tik, kad nebaigtas, bet ir trūksta jo dalies, kuri atskleistų kokią nors prasmę, pvz., *Subject Search Summarizer* lietuviško grožinio teksto santraukos sakinys: *jis jau matė daug pilių ir daug moterų (bet nė viena negali prilygti tai, kuri jo laukia po dviejų dienų).* Pasitaiko atvejų, kai keli žodžiai sujungiami į vieną, pavyzdžiui, romėniškas skaičius *I* sujungiamas su *HAD* ir santraukoje pateikiama kaip vienas žodis *IHAD*, lietuviško teksto santraukoje randamas toks žodžių sujungimo pavyzdys: *Pirmasis skyriusJo*. Žinoma, tokie atvejai nesukelia daug nepatogumų skaitant, bet anglų kalboje painiojant romėniškai rašomą skaičių *„I“* su angliškuu *„aš“* kartais iškreipiama prasmė. Panašių atvejų pasitaiko beveik visose SARS, todėl negalima tvirtinti, jog kuri nors iš jų pateikė santrauką be trūkumų.

Santraukos įskaitomumas. Dar vienas svarbus santraukos rodiklis yra santraukos įskaitomumas. Net jei santrauka būtų ir ideali, ypač svarbu kaip įgyvendinta galimybė perskaityti santrauką. Santraukų įskaitomumo vertinimas – ar yra neatpažintų ženklų, ar pakeičiamos raidės ir tai, kas dar gali turėti įtakos į santraukos įskaitomumą. Iš gautų rezultatų galima teigti, kad geriausiai įskaitomos tiek lietuvių, tiek anglų kalbomis yra santraukos parengtos *MS Word AutoSummarize, Automatic Text Summarizer, Pertinence Summarizer, Shvoong Summarizer, Subject Search Summarizer*. Gerai įskaitomas angliškas santraukas pateikia ir *Copernic Summarizer, Intellexer Summarizer* bei *QuickJist Summarizer*, tačiau šių sistemų lietuviškų tekstų santraukų įskaitomumas gerokai nusileidžia kitoms sistemoms. *QuickJist Summarizer* sunkiau įskaitomos santraukos yra tada, kai santrauka skaitoma tos sistemos lange. Ją nukopijavus į MS Word dokumentą, santrauka yra įskaitoma, joje nebūna neatpažintų ženklų, kitaip nei programos lange. *Copernic Summarizer* ir *Intellexer Summarizer* parengtos santraukos (net ir į MS Word dokumento kopijos), yra sunkiai įskaitomos. *Copernic Summarizer* lietuviškoje santraukoje panaikina lietuviškus rašmenis – *š* virsta *s, ž –z; ą – a; ū, ų – u; į – i; ė ,ę –e;* ši sistema lietuviškus rašmenis padaro lotyniškais. *Intellexer Summarizer* lietuviškose santraukose lietuviški rašmenys paverčiami neįskaitomais, tenka spėti, kokia tai raidė. Pvz., *Copernic Summarizer*: *iatobulejus – ištobulėjus; dideliu pakitimu –*

didelių pakitimų; aakoje – šakoje; tekstynu – tekstynų; pa~velgti – pažvelgti; atskleid~ia – atskleidžia; ~mones – žmones, ~vaig~des - žvaigdždes ir t.t. Intellexer Summarizer: *dÄ—l – dėl; sÅ«nÅ³ - sūnų; u¸tvara – užtvarą; gra¸u – gražu; triusiancios – triūsiančios; panaÅ□Å«s – panašūs; io – šio.*

3 lentelė. Lietuvių kalbos teksto santraukos įskaitomumo įvertinimas[5]

Programa	**Grožinis tekstas**	**Mokslinis tekstas**	**Publicistinis tekstas**
Automatic Text Summarizer	5	5	5
Copernic Summarizer	2	2	2
Intellexer Summarizer	2	2	2
MS Word AutoSummarize	5	5	5
Pertinence Summarizer	5	5	5
QuickJist Summarizer	3	3	3
Shvoong Summarizer	5	-	5
Subject Search Summarizer	5	5	5

Atlikti tyrimai parodė, kad visos sistemos gali rengti puikiai įskaitomas angliškų tekstų santraukas. Tuo tarpu geriausiai įskaitomas lietuviškas santraukas rengia tik *MS Word AutoSummarize, Automatic text summarizer, Subject Search Summarizer, Pertinence summarizer, Shvoong summarizer* sistemos. Prasčiau įvertinamos *QuickJist summarizer* gautos lietuviškų tekstų santraukos, kadangi šios programos lange lietuviškos santraukos nėra lengvai įskaitomos, atsiranda neįskaitomų ženklų, tačiau pasinaudojus funkcija *Copy Clipboard* nukopijuotas tekstas yra taip pat gerai įskaitomas, kaip ir santrauka gauta *MS Word AutoSummarize* sistema. Blogiausiai įvertinamas *Intellexer summarizer* ir *Copernic Summarizer* sistemomis parengtų santraukų įskaitomumas, kadangi tekstuose tenka spėlioti raides ar žodžius.

Santraukų kokybė. Santraukų kokybė yra vienas svarbiausių geros santraukos rodiklių. Ją sudaro santraukos apimtis, perteikiamo teksto prasmė, santraukos įskaitomumas, jos naudingumas, aiškumas, konkretumas, pagrindinių faktų pateikimas ir prasmingumas. Bendras integralinis vertinimas pagal santraukos įskaitomumą, prasmingumą, jos apimtį, gaunamas kai santrauką skaitęs asmuo įvertina ją penkiabalėje skalėje.

Pagal gautus tyrimo rezultatus nustatyta, kad blogiausiai vertinamos *Subject Search Summarizer* parengtos santraukos. Šios sistemos pateiktos santraukos pagal jas vertinusius asmenis yra blogiausios. Geriausiomis santraukų sistemomis būtų galima vadinti *MS Word AutoSummarize* bei *Automatic Text Summarizer*. Jos pagal visus rodiklius veikia geriausiai ir jų pateikiamos santraukos yra pakankamai geros. Šių sistemų geresnį įvertinimą lėmė ir tai, kad lietuviškos santraukos yra įskaitomos ir pakankamai prasmingos, naudingos ir informatyvios. *QuickJist Summarizer* taip pat gali būti priskiriamas prie geriau santraukas rengiančių sistemų, nors lietuviškų tekstų santraukose pastebimi keli trūkumai. *Copernic Summarizer* ir *Intellexer Summarizer* pateikia pakankamai geras santraukas anglų kalbos tekstams, tačiau lietuvių kalbos santraukos šiose sistemose yra prastesnės kokybės nei angliškos santraukos, todėl joms buvo skirtas žemesnis balas.

Apibendrinant rezultatus galima teigti, jog geriausios santraukų sistemos, tinkančios lietuvių kalbai yra *Automatic Text Summarizer, MS Word AutoSummarize* ir *QuickJist Summarizer*, kurios pateikia geriausios kokybės lietuviškas santraukas. Blogiausias lietuviškas santraukas pateikia *Subject Search Summarizer, Pertinence Summarizer, Copernic Summarizer, Intellexer Summarizer,* nors anglų kalbos santraukos gautos šiomis sistemomis (išskyrus *Subject Search Summarizer*) yra pakankamai geros. Vidutiniškos kokybės lietuviškas santraukas pateikia *Shvoong Summarizer*.

4. Išvados

SARS tyrimui naudota angliški ir lietuviški mokslinio, publicistinio ir grožinio stiliaus tekstai su aštuoniomis santraukų rengimo sistemomis: *Automatic text summarizer, Copernic summarizer, Intellexer summarizer, MS Word AutoSummarize, Pertinence summarizer, Subject Seearch Summarizer, QuickJist summarizer, Shvoong summarizer.* Nustatyta, kad automatinės teksto santraukų sistemos pakankamai gerai veikia su bet kokio stiliaus ar žanro anglų kalbos tekstais. Analizuojant lietuvių kalbos tekstus iškyla problemos dėl pačio teksto nuskaitymo ir jo programinės interpretacijos, gautos santraukos įskaitomumo ir suprantamumo bei prasmės perteikimo galimybių. Naudojant didesnį teksto glaudinimą, žymi dalis SARS nesugeba santraukose adekvačiai perteikti pagrindinio teksto turinio, neatspindi jo prasmės. Todėl, norėdami gauti pakankamai gerą santrauką rurėtume naudoti ne didesnį 50% glaustumo laipsnį, atskirais atvejais – 25%-30%, bet ne didesnius. Pastebėta, kad *Copernic summarizer* ir *Intellexer summarizer* sistemų parengtas santraukas lietuvių kalbos tekstui yra sunku įskaityti, žodžių reikšmė dažnai gali būti nuspėjama tik pagal bendrą sakinio kontekstą. Šios santraukų sistemos neatpažįsta lietuviškų raidžių, tačiau puikiai veikia su tekstais parašytais lotyniško alfabetu. Lietuvių kalbai tinkančios santraukų sistemos yra *MS Word*

[5] Įskaitomumas buvo vertinamas penkiabalėje sistemoje, kur 5 –labai gerai, 4 – gerai, 3 – vidutiniškai, 2 – blogai, 1 – labai blogai.

AutoSummarize, *Automatic Text Summarizer* ir *QuickJist Summarizer*. Jos pateikia santraukas, kurios yra geresnės kokybės nei kitų santraukų sistemų, išlaiko informatyvumą ir prasmingumą.

Nustatyta, kad prastos kokybės lietuviškas santraukas pateikia *Subject Search Summarizer*, *Pertinence Summarizer*, *Copernic Summarizer*, *Intellexer Summarizer* sistemos, nors anglų kalbos santraukos, gautos šiomis sistemomis (išskyrus *Subject Search Summarizer*) yra vertinamos pakankamai gerai. Vidutiniškos kokybės lietuviškas santraukas pateikia *Shvoong Summarizer*. Reikia pripažinti, kad automatinės teksto santraukų sistemos yra nepaprastai naudingos, nes sumažina laiko sąnaudas sutrumpindamos tekstą pagal vartotojo pageidavimus. Šias sistemas pritaikius lietuvių kalbai, ar sukūrus lietuvišką automatinę skaitmeninio teksto santraukų rengimo sistemą, palengvėtų ir taptų spartesnis tekstų skaitomumas gimtąją kalba, jos leistų efektyviau atrinkti informaciją bei taupyti didelės apimties tekstų peržvalgos trukmę.

Literatūros sąrašas

1. Mani, I. (2001), Automatic Summarization. Philadephia, John Benjamins Publishing Company. 285 p. ISBN 9027249865
2. Mitkov, R., (2005). The Oxford Handbook of Computational Linguistics. Oxford, Oxford University Press. 583-598 p. 786 p. ISBN 019927634X,

Šaltiniai

10. Automatic text summarizer. Search and Information Extraction Research Lab. Language technologies research center International institute of information technology. Interaktyvus. Žiūrėta [2009-05-01]. Prieiga per internetą: http://search.iiit.ac.in/~jags/summarizer/index.cgi
11. Copernic summarizer. Copernic Inc. Interaktyvus. Žiūrėta [2009-05-01]. Prieiga per internetą: http://www.copernic.com/en/products/summarizer/
12. Intellexer summarizer. Intellexer. Interaktyvus. Žiūrėta [2009-05-01]. Prieiga per internetą: http://summarizer.intellexer.com/
13. Pertinence summarizer. Pertinence Mining. Interaktyvus. Žiūrėta [2009-05-01]. Prieiga per internetą: http://www.pertinence.net/ps/main.jsp?ui.lang=
14. QuickJist summarizer. Vitaly Demin. Interaktyvus. Žiūrėta [2009-05-01]. Prieiga per internetą: http://www.quickjist.com/
15. Shvoong summarizer. Shvoong Ltd. USA Interaktyvus. Žiūrėta [2009-05-01]. Prieiga per internetą: http://www.shvoong.com/summarizer/
16. Subject Search Summarizer. Kryloff technologies, Inc. Interaktyvus. Žiūrėta [2009-05-01]. Prieiga per internetą: http://www.kryltech.com/summarizer.htm
17. MS Word AutoSummarize – Word Help. Microsoft Corpeoration. Interaktyvus. Žiūrėta [2009-05-01]. Prieiga per internetą: http://www.microsoft.com/education/autosummarize.mspx

Jurgita Lasytė, Bronius Tamulynas

Summarizers – similarities, differences and applications in Lithuanian language

Summary

Summarizer is useful system, which allows compressing text and represents it in a shorter way. In Lithuania summarizers are new technology and information about them is poor, although these systems are created and developed by computational linguistics specialists in other countries. In summarizers' research English and Lithuanian texts were used, which allowed seeing differences between received summaries. In the English texts, summaries using all summarizers were quite informative, useful and did not main information from original texts, except cases when a high compression rate was used. While using for summarization Lithuanian texts, some summarizers confront with language recognition problem – summaries in Lithuanian are with unrecognizable symbols, or in individual cases these symbols are missing. Summaries in Lithuanian language are not of the same quality as summaries in English. It was also noticed, that Lithuanian summaries with higher compression rate are loosing original text meaning, information and their usefulness. To sum up results it was noticed that commercial summarizers give better summaries in English language, but in Lithuanian language better summaries are given by noncommercial summarizers. Furthermore, summaries received using high compression rate are loosing their information and meaning. Analyzed summarizers are best fitted for English language, therefore using summarizer for other language than English stipulates various problems – from language recognition to rendering of text meaning. It is very important to adapt different languages to summarizers, as this would significantly improve summarizers. Findings of this paper can be used in creating Lithuanian summarizer or improving other summarizers.

Apie autorius

Jurgita Lasytė, Kauno technologijos universiteto magistrantė. *El. paštas:* jurgita.lasyte@gmail.com

Bronius Tamulynas, technikos mokslų daktaras, Kauno technologijos universiteto Kompiuterių tinklų katedros docentas,
Mokslinės veiklos sritys: Kompiuterinio kalbų vertimo technologijos, Kompiuterinė lingvistika, Intelektualiųjų sistemų modeliavimas, Lingvistinių duomenų struktūros
Adresas: Kauno technologijos universitetas, Informatikos fakultetas, Studentų 50-414, 51368, Kaunas, *El. paštas:* bronius.tamulynas@ktu.lt.

2 priedas. Konferencijos „Kalbos teorija ir praktika“ pranešimo tezės.

Santraukų automatinio rengimo sistemų (summarizer) dalinio tyrimo rezultatai

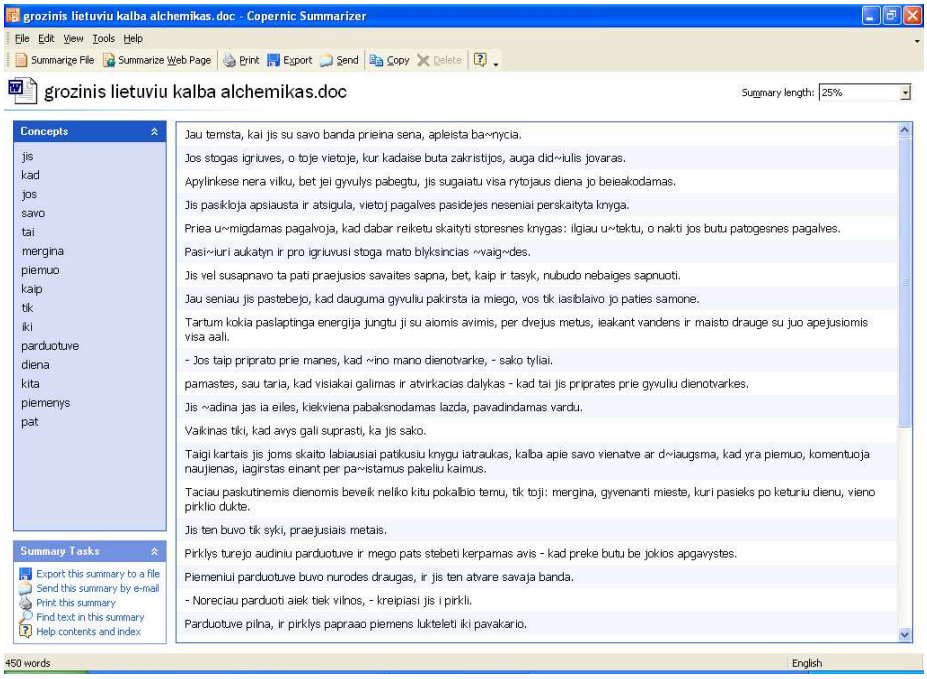
Jurgita Lasytė, KTU magistrantė, KTU Informatikos fakultetas,
Studentų 50, Kaunas, jurgita.lasyte@gmail.com
Bronius Tamulynas dr., doc. KTU kompiuterių tinklų katedra,
Studentų 50-414, Kaunas, bronius.tamulynas@ktu.lt.

Santraukų automatinio rengimo sistemos – tai kompiuterizuotos programos, galinčios iš skaitmeninių tekstų parengti santraukas pagal vartotojo pageidavimus. Santraukų automatinio rengimo sistemos (SARS) neapsiriboja skaitmeniniu tekstu, bet siekia apimti įvairialypę informaciją – tekstą, vaizdinę ir garsinę medžiagą, paveikslėlius. Atrinkę ir susipažinę su keliomis SARS ir technologijomis (*summarizers*), atskleisime skaitmeninio **teksto** santraukų rengimo skirtumus, atsirandančius kuriant skirtingų kalbų ar žanrų tekstų santraukas skirtingomis sistemomis, palyginsime gautus praktinio taikymo rezultatus su angliškais ir lietuviškais skirtingų žanrų tekstais.

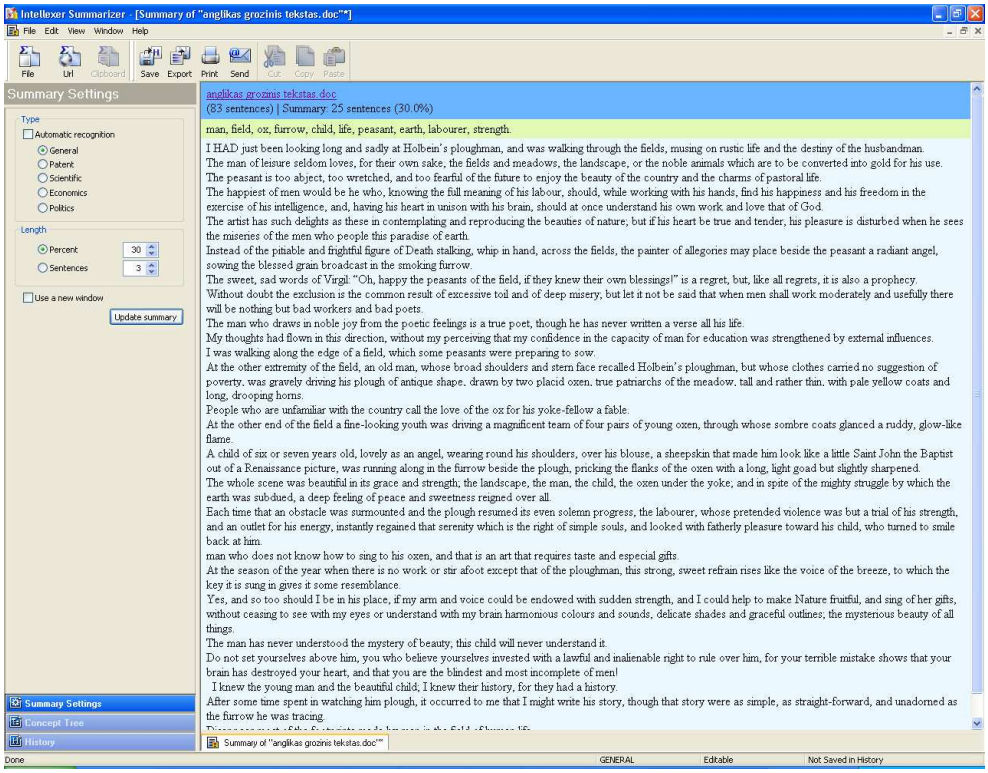
SARS tyrimui atlikti buvo pasirinktos santraukų sistemos, kurios santraukas atlieka tekstams: *Automatic Text Summarizer, Copernic Summarizer, Intellexer Summarizer, Pertinence Summarizer, QuickJist Summarizer, Shvoong Summarizer, Subject Search Summarizer, Microsoft Office MS Word 2003 AutoSummarize.* Esminis klausimas yra ar pasinaudojus SARS iš gautosios santraukos bus įmanoma nustatyti apie ką kalbama. Iš gautų įvertinimų galima teigti, kad prasčiausiai teksto prasmė perteikiama beveik visose grožinių kūrinių santraukose, blogiausias šiuo atžvilgiu SARS yra *Subject Search Summarizer*. Geriausiai prasmę santraukoje perteikia *MS Word AutoSummarize*. Nedaug prastesnes - *Copernic Summarizer* bei *Intellexer Summarizer* sistemos. Tyrimo rezultatai patvirtina - teksto prasmė santraukose priklauso nuo kompresijos lygio. Santraukos, kurios gaunamos panaudojant 50% kompresiją yra aiškesnės, prasmingesnės. Naudojant žemesnę kompresiją gaunamos prastesnės santraukos. Galima daryti išvadą, jog kuo didesne kompresija spaudžiamas tekstas, tuo mažiau prasmės yra gaunamoje santraukoje.

Atlikta SARS analizė parodė, kad su lietuvių kalbos tekstais iškyla problemos dėl teksto nuskaitymo, gautos santraukos įskaitomumo, suprantamumo bei prasmės perteikimo. Nustatyta, kad naudojant gilesnį teksto suglaudinimą, didesnė dalis SARS nesugeba gautose santraukose perteikti pagrindinį teksto turinį, neatspindi teksto prasmės. Esant aukštam suglaudinimui, gaunama santrauka netenka prasmės lyginant su teksto originalu, t.y. informacija perteikiama neinformatyviai. Todėl norint gauti pakankamai gerą santrauką geriausia naudoti 50% suglaudinimo lygį. Lietuvių kalbai tinkančios SARS yra *MS Word AutoSummarize, Automatic Text Summarizer* ir *QuickJist Summarizer,* nes jos pateikia santraukas, kurios yra geresnės kokybės, kurios yra pakankamai informatyvios ir prasmingos. SARS pritaikius lietuvių kalbai, ar sukūrus lietuvišką automatinę skaitmeninio teksto santraukų rengimo sistemą, palengvėtų ir taptų spartesniu tekstų skaitymas gimtąją kalba, jos leistų efektyviau atrinkti informaciją.

3 priedas Copernic Summarizer programos langas



4 priedas Intellexer summarizer programos langas



5 priedas Anglų kalbos grožinio teksto santraukos, naudojant 50% kompresiją

Copernic summarizer parengta anglų kalbos grožinio teksto santrauka

Concepts: plough, labourer, peasant, life, furrow, pleasure, strength, child, voice, animals, earth, heart, happiness, beautiful, country.

Summary: I HAD just been looking long and sadly at Holbein's ploughman, and was walking through the fields, musing on rustic life and the destiny of the husbandman.

It is certainly tragic for him to spend his days and his strength delving in the jealous earth, that so reluctantly yields up her rich treasures when a morsel of coarse black bread, at the end of the day's work, is the sole reward and profit to be reaped from such arduous toil.

The wealth of the soil, the harvests, the fruits, the splendid cattle that grow sleek and fat in the luxuriant grass, are the property of the few, and but instruments of the drudgery and slavery of the many.

The man of leisure seldom loves, for their own sake, the fields and meadows, the landscape, or the noble animals which are to be converted into gold for his use.

He comes to the country for his health or for change of air, but goes back to town to spend the fruit of his vassal's labour.

On the other hand, the peasant is too abject, too wretched, and too fearful of the future to enjoy the beauty of the country and the charms of pastoral life.

To him, also, the yellow harvest-fields, the rich meadows, the fine cattle represent bags of gold; but he knows that only an infinitesimal part of their contents, insufficient for his daily needs, will ever fall to his share.

Yet year by year he must fill those accursed bags, to please his master and buy the right of living on his land in sordid wretchedness.

Yet nature is eternally young, beautiful, and generous, She pours forth poetry and beauty on all creatures and all plants that are allowed free development.

She owns the secret of happiness, of which no one has ever robbed her.

The happiest of men would be he who, knowing the full meaning of his labour, should, while working with his hands, find his happiness and his freedom in the exercise of his intelligence, and, having his heart in unison with his brain, should at once understand his own work and love that of God.

The artist has such delights as these in contemplating and reproducing the beauties of nature; but if his heart be true and tender, his pleasure is disturbed when he sees the miseries of the men who people this paradise of earth.

True happiness will be theirs when mind, heart, and hand shall work in concert in the sight of Heaven, and there shall be a sacred harmony between God's goodness and the joys of his creatures.

Then, instead of the pitiable and frightful figure of Death stalking, whip in hand, across the fields, the painter of allegories may place beside the peasant a radiant angel, sowing the blessed grain broadcast in the smoking furrow.

The dream of a serene, free, poetic, laborious, and simple life for the tiller of the soil is not so impossible that we should banish it as a chimera.

The sweet, sad words of Virgil: "Oh, happy the peasants of the field, if they knew their own blessings!"

is a regret, but, like all regrets, it is also a prophecy.

The day will come when the labourer too may be an artist, and may at least feel what is beautiful, if he cannot express it---a matter of far less importance.

Do not we know that this mysterious poetic intuition is already his, in the form of instinct and vague reverie?

Among those peasants who possess some of the comforts of life, and whose moral and intellectual development is not entirely stifled by extreme wretchedness, pure happiness that can be felt and appreciated exists in the elementary stage; and, moreover, since poets have already raised their voices out of the lap of pain and of weariness, why should we say that the labour of the hands excludes the working of the soul?

Without doubt this exclusion is the common result of excessive toil and of deep misery; but let it not be said that when men shall work moderately and usefully there will be nothing but bad workers and bad poets.

The man who draws in noble joy from the poetic feelings is a true poet, though he has never written a verse all his life.

My thoughts had flown in this direction, without my perceiving that my confidence in the capacity of man for education was strengthened by external influences.

At the other extremity of the field, an old man, whose broad shoulders and stern face recalled Holbein's ploughman, but whose clothes carried no suggestion of poverty, was gravely driving his plough of antique shape, drawn by two placid oxen, true patriarchs of the meadow, tall and rather thin, with pale yellow coats and long, drooping horns.

They had the short, curly heads that belong to the wild bull, the same large, fierce eyes and jerky movements; they worked in an abrupt, nervous way that showed how they still rebelled against the yoke and goad and trembled with anger as they obeyed the authority so recently imposed.

A child of six or seven years old, lovely as an angel, wearing round his shoulders, over his blouse, a sheepskin that made him look like a little Saint John the Baptist out of a Renaissance picture, was running along in the furrow beside the plough, pricking the flanks of the oxen with a long, light goad but slightly sharpened.

The spirited animals quivered under the child's light touch, making their yokes and head-bands creak, and shaking the pole violently, whenever a root stopped the advance of the ploughshare, the labourer would call every animal by name in his powerful voice, trying to calm rather than to excite them; for the oxen, irritated by the sudden resistance, bounded, pawed the ground with their great cloven hoofs, and would have jumped aside and dragged the plough across the fields, if the young man had not kept the first four in order with his voice and goad, while the child controlled the four others.

This song, which was probably sacred in its origin, and to which mysterious influences must once have been attributed, is still thought to possess the virtue of putting animals on their mettle, allaying their irritation, and of beguiling the weariness of their long, hard toil.

Its irregular form and its intonations that violate all the rules of musical art make it impossible to describe.

But it is none the less a noble song, and so appropriate is it to the nature of the work it accompanies, to the gait of the oxen, to the peace of the fields, and to the simplicity of the men who sing it, that no genius unfamiliar with the tillage of the earth, and no man except an accomplished labourer of our part of the country, could repeat it.

this man has never understood the mystery of beauty; this child will never understand it.

I knew the young man and the beautiful child; I knew their history, for they had a history.

Everybody has his own, and could make the romance of his life interesting, if he could but understand it.

Although but a peasant and a labourer, Germain had always been aware of his duties and affections.

He had related them to me clearly and ingenuously, and I had listened with interest.
After some time spent in watching him plough, it occurred to me that I might write his story, though that story were as simple, as straight-forward, and unadorned as the furrow he was tracing.
Next year that furrow will be filled and covered by a fresh one.
Thus disappear most of the footprints made by man in the field of human life.
A little earth obliterates them, and the furrows we have dug succeed one another like graves in a cemetery.
Is not the furrow of the labourer of as much value as that of the idler, even if that idler, by some absurd chance, have made a little noise in the world, and left behind him an abiding name?
I mean, if possible, to save from oblivion the furrow of Germain, the skilled husbandman.
He will never know nor care, but I shall take pleasure in my task.

Automatic text summarizer anglų kalbos grožinio teksto santrauka

I. The Tillage of the SoilI HAD just been looking long and sadly at Holbein's ploughman, and was walking through the fields, musing on rustic life and the destiny of the husbandman.

It is certainly tragic for him to spend his days and his strength delving in the jealous earth, that so reluctantly yields up her rich treasures when a morsel of coarse black bread, at the end of the day's work, is the sole reward and profit to be reaped from such arduous toil.

The wealth of the soil, the harvests, the fruits, the splendid cattle that grow sleek and fat in the luxuriant grass, are the property of the few, and but instruments of the drudgery and slavery of the many.

The man of leisure seldom loves, for their own sake, the fields and meadows, the landscape, or the noble animals which are to be converted into gold for his use.

He comes to the country for his health or for change of air, but goes back to town to spend the fruit of his vassal's labour.

To him, also, the yellow harvest-fields, the rich meadows, the fine cattle represent bags of gold; but he knows that only an infinitesimal part of their contents, insufficient for his daily needs, will ever fall to his share.

Yet nature is eternally young, beautiful, and generous, She pours forth poetry and beauty on all creatures and all plants that are allowed free development.

The happiest of men would be he who, knowing the full meaning of his labour, should, while working with his hands, find his happiness and his freedom in the exercise of his intelligence, and, having his heart in unison with his brain, should at once understand his own work and love that of God.

The artist has such delights as these in contemplating and reproducing the beauties of nature; but if his heart be true and tender, his pleasure is disturbed when he sees the miseries of the men who people this paradise of earth.

True happiness will be theirs when mind, heart, and hand shall work in concert in the sight of Heaven, and there shall be a sacred harmony between God's goodness and the joys of his creatures.

Then, instead of the pitiable and frightful figure of Death stalking, whip in hand, across the fields, the painter of allegories may place beside the peasant a radiant angel, sowing the blessed grain broadcast in the smoking furrow.

Among those peasants who possess some of the comforts of life, and whose moral and intellectual development is not entirely stifled by extreme wretchedness, pure happiness that can be felt and appreciated exists in the elementary stage; and, moreover, since poets have already raised their voices out of the lap of pain and of weariness, why should we say that the labour of the hands excludes the working of the soul?

Without doubt this exclusion is the common result of excessive toil and of deep misery; but let it not be said that when men shall work moderately and usefully there will be nothing but bad workers and bad poets.

The man who draws in noble joy from the poetic feelings is a true poet, though he has never written a verse all his life.

I was walking along the edge of a field, which some peasants were preparing to sow.

The space was vast as that in Holbein's picture; the landscape, too, was vast and framed in a great sweep of green, slightly reddened by the approach of autumn.

Here and there in the great russet field, slender rivulets of water left in the furrows by the late rains sparkled in the sunlight like silver threads.

At the other extremity of the field, an old man, whose broad shoulders and stern face recalled Holbein's ploughman, but whose clothes carried no suggestion of poverty, was gravely driving his plough of antique shape, drawn by two placid oxen, true patriarchs of the meadow, tall and rather thin, with pale yellow coats and long, drooping horns.

They were those old workers who, through long habit, have grown to be brothers, as they are called in our country, and who, when one loses the other, refuse to work with a new comrade, and pine away with grief.

People who are unfamiliar with the country call the love of the ox for his yoke-fellow a fable.

Let them come and see in the corner of the stable one of these poor beasts, thin and wasted, restlessly lashing his lean flanks with his tail, violently breathing with mingled terror and disdain on the food offered him, his eyes always turned toward the door, scratching with his hoof the empty place at his side, sniffing the yokes and chains which his fellow used to wear, and incessantly calling him with melancholy lowings.

" The old labourer worked slowly, silently, and without waste of effort.

My attention was next caught by a fine spectacle, a truly noble subject for a painter.
The man who drove them had to clear a corner of the field that had formerly been given up to pasture, and was filled with old tree-stumps; and his youth and energy, and his eight half-broken animals, hardly sufficed for the Herculean task.
A child of six or seven years old, lovely as an angel, wearing round his shoulders, over his blouse, a sheepskin that made him look like a little Saint John the Baptist out of a Renaissance picture, was running along in the furrow beside the plough, pricking the flanks of the oxen with a long, light goad but slightly sharpened.
The spirited animals quivered under the child's light touch, making their yokes and head-bands creak, and shaking the pole violently, whenever a root stopped the advance of the ploughshare, the labourer would call every animal by name in his powerful voice, trying to calm rather than to excite them; for the oxen, irritated by the sudden resistance, bounded, pawed the ground with their great cloven hoofs, and would have jumped aside and dragged the plough across the fields, if the young man had not kept the first four in order with his voice and goad, while the child controlled the four others.
The whole scene was beautiful in its grace and strength; the landscape, the man, the child, the oxen under the yoke; and in spite of the mighty struggle by which the earth was subdued, a deep feeling of peace and sweetness reigned over all.
Each time that an obstacle was surmounted and the plough resumed its even solemn progress, the labourer, whose pretended violence was but a trial of his strength, and an outlet for his energy, instantly regained that serenity which is the right of simple souls, and looked with fatherly pleasure toward his child, who turned to smile back at him.
Its irregular form and its intonations that violate all the rules of musical art make it impossible to describe.
Each phrase ends with a long trill, the final note of which is held with incredible strength of breath, and rises a quarter of a tone, sharping systematically.
Instead of a wretched old man, a young and active one; instead of a team of weary and emaciated horses, four yoke of robust and fiery oxen; instead of death, a beautiful child; instead of despair and destruction, energy and the possibility of happiness.
Then the old French verse, "A la sueur de ton visaige," etc., and Virgil's "O fortunatos… agricolas," returned to my mind, and seeing this lovely child and his father, under such poetic conditions, and with so much grace and strength, accomplish a task full of such grand and solemn suggestions, I was conscious of deep pity and involuntary respect.
Yes, and so too should I be in his place, if my arm and voice could be endowed with sudden strength, and I could help to make Nature fruitful, and sing of her gifts, without ceasing to see with my eyes or understand with my brain harmonious colours and sounds, delicate shades and graceful outlines; in short, the mysterious beauty of all things.
And above all, if my heart continued to beat in concert with the divine sentiment that presided over the immortal sublimity of creation.
The proof that they feel this is that they cannot be exiled with impunity, that they love the soil they have watered with their tears, and that the true peasant dies of homesickness under the arms of a soldier far from his native field.
He lacks the consciousness of his sentiment.
Although but a peasant and a labourer, Germain had always been aware of his duties and affections.
After some time spent in watching him plough, it occurred to me that I might write his story, though that story were as simple, as straight-forward, and unadorned as the furrow he was tracing.
Next year that furrow will be filled and covered by a fresh one.
I mean, if possible, to save from oblivion the furrow of Germain, the skilled husbandman.

6 priedas Anglų kalbos mokslinio teksto santraukos, naudojant 50% kompresiją

Tools 4 Noobs summarizer anglų kalbos mokslinio teksto santrauka

In this paper, we empirically characterize human-written summaries provided in a widely used summarization corpus by attempting to answer the questions: Can multi-document summaries that are written by humans be characterized as extractive or generative? (133)
In this paper, we seek to empirically characterize ideal multi-document summaries in part by attempting to answer the questions: Can multi-document summaries that are written by humans be characterized as extractive or generative? (127)
While this sheds light on how much gain can be achieved by optimizing sentence extraction methods for single-document summarization, to our knowledge, no one has assessed the potential for extraction-based systems when attempting to summarize multiple documents. 3 Using N-gram Sequences to Characterize Summaries Our approach to characterizing summaries is much simpler than what Jing has described and is based on the following idea: if

human-written summaries are extractive, then we should expect to see long spans of text that have been lifted from the source documents to form a summary. (230)
Since we wished to collectively compare single-document summaries against multi-document summaries, we used the 100-word multi-document summaries for our analysis. (137)

Intellexer summarizer anglų kalbos mokslinio teksto santrauka

Concepts:
summary, document understanding conference, document, summary abstract, sentence combination, sentence reduction, sentence, summarization, multi-document, document clustering.
Summary:
Using N-Grams to Understand the Nature of Summaries
Abstract
Although single-document summarization is a well-studied task, the nature of multidocument summarization is only beginning to be studied in detail.
While close attention has been paid to what technologies are necessary when movint from single to multi-document summarization, the properties of humanwritten multi-document summaries have not been quantified.
Our results suggest that extraction-based techniques which have been successful for single-document summarization may not be sufficient when summarizing multiple documents.
Document clustering and multi-document summarization technologies working in tandem promise to ease some of the burden on users when browsing related documents.
Summarizing a set of documents brings about challenges that are not present when summarizing a single document.
A useful multi-document summary may also indicate the presence of new or distinct information contained within a set of documents describing the same topic (McKeown et.
To meet the expectations, a multi-document summary is required to generalize, condense and merge information coming from multiple sources.
Although single-document summarization is a wellstudied
task (see Mani and Maybury, 1999 for an overview), multi-document summarization is only recently being studied closely (Marcu & Gerber 2001).
While close attention has been paid to multi-document summarization technologies (Barzilay et al. 2002, Goldstein et al 2000), the inherent properties of humanwritten multi-document summaries have not been quantified.
Our aim in answering these questions is to discover how the nature of multi-document summaries will impact our system requirements.
Section 3 describes a new approach for assessing the degree to which a summary can be described as extractive, and reports our findings for both single and multiple document summarization tasks.
Based on a manual inspection of 15 humanwritten summaries, she proposes that for the task of single-document summarization, human summarizers use a “cut-and-paste” approach in which six main operations are performed: sentence reduction, sentence combination, syntactic transformation, reordering, lexical paraphrasing, and generalization or specification.
According to the model, 81% of summary sentences contained in a corpus of 300 human-written summaries of news articles on telecommunications were found to fit the cut-and-paste method, with the rest believed to have been composed from scratch.
3 Using N-gram Sequences to
Characterize Summaries
Our approach to characterizing summaries is much simpler than what Jing has described and is based on the following idea: if human-written summaries are extractive, then we should expect to see long spans of text that have been lifted from the source documents to form a summary.
[The V-chip will give parents a] [device to block out programs they don’t want their children to see.] Syntactic Transformation: Document sentence: Since annoy.com enables visitors to send unvarnished opinions to political and other figures in the news, the company was concerned that its activities would be banned by the statute.
[Annoy.com enables visitors to send unvarnished opinions to political and other figures in the news] and feared the law could put them out of business.
Sentence Combination:
Document sentence 1: But it also raises serious questions about the privacy of such highly personal information wafting about the digital world.

Document sentence 2: The issue fits squarely into the broader debate about privacy and security on the Internet, whether it involves protecting credit card numbers or keeping children from offensive information.
As part of the DUC 2001 summarization corpus, NIST also provides four hand-written summaries of different lengths for every document cluster, as well as 100-word summaries of each document.
Since we wished to collectively compare single-document summaries against multi-document summaries, we used the 100-word multi-document summaries for our analysis.
It is important to note that for each cluster, all summaries (50, 100, 200 and 400-word multi-document and 100-word per-document) have been written by the same author.
NIST used a total of ten authors, each providing summaries for 3 of the 30 topics.
To compare the text of human-authored multidocument summaries to the full-text documents describing the events, we automatically broke the documents into sentences, and constructed a minimal tiling of each summary sentence.
For each sentence in the summary, we searched for all n-grams that are present in both the summary and the documents, placing no restrictions on the potential size of an n-gram.
We then covered each summary sentence with the ngrams, optimizing to use as few n-grams as possible (i.e. favoring n-grams that are longer in length).

Pertinence summarizer anglų kalbos mokslinio teksto santrauka

Using N-Grams to Understand the Nature of Summaries Abstract Although single-document summarization is a well-studied task, the nature of multidocument summarization is only beginning to be studied in detail.
While close attention has been paid to what technologies are necessary when movint from single to multi-document summarization, the properties of humanwritten multi-document summaries have not been quantified.
In this paper, we empirically characterize human-written summaries provided in a widely used summarization corpus by attempting to answer the questions: Can multi-document summaries that are written by humans be characterized as extractive or generative?
Our results suggest that extraction-based techniques which have been successful for single-document summarization may not be sufficient when summarizing multiple documents.
The increased accessibility of worldwide online news sources and the continually expanding size of the worldwide web place demands on users attempting to wade through vast amounts of text.
One might expect that a good multidocument summary will present a synthesis of multiple views of the event being described over different documents, or present a high-level view of an event that is not explicitly reflected in any single document.
A useful multi-document summary may also indicate the presence of new or distinct information contained within a set of documents describing the same topic (McKeown et.
To meet these expectations, a multi-document summary is required to generalize, condense and merge information coming from multiple sources.
In this paper, we seek to empirically characterize ideal multi-document summaries in part by attempting to answer the questions: Can multi-document summaries that are written by humans be characterized as extractive or generative?
Our aim in answering these questions is to discover how the nature of multi-document summaries will impact our system requirements.
While we recognize that other summarization corpora may exhibit different properties than what we report, the data prepared for DUC evaluations is widely used, and continues to be a powerful force in shaping directions in summarization research and evaluation.
Section 3 describes a new approach for assessing the degree to which a summary can be described as extractive, and reports our findings for both single and multiple document summarization tasks.
We conclude with a discussion of our findings in Section 4. 2 Related Work Jing (2002) previously examined the degree to which single-document summaries can be characterized as extractive.
According to this model, 81% of summary sentences contained in a corpus of 300 human-written summaries of news articles on telecommunications were found to fit the cut-and-paste method, with the rest believed to have been composed from scratch.
By computing a performance upper-bound for pure sentence extraction, they found that state-of-the-art extraction-based systems are still 15%-24%2 away from this limit, and 10% away from average human performance.
While this sheds light on how much gain can be achieved by optimizing sentence extraction methods for single-document summarization, to our knowledge, no one has assessed the potential for extraction-based systems when

attempting to summarize multiple documents.
Summary sentence: [Annoy.com enables visitors to send unvarnished opinions to political and other figures in the news] and feared the law could put them out of business.
As part of the DUC 2001 summarization corpus, NIST also provides four hand-written summaries of different lengths for every document cluster, as well as 100-word summaries of each document.
Since we wished to collectively compare single-document summaries against multi-document summaries, we used the 100-word multi-document summaries for our analysis.
It is important to note that for each cluster, all summaries (50, 100, 200 and 400-word multi-document and 100-word per-document) have been written by the same author.
The instructions provided did not differ per task; in both single and multi-document scenarios, the authors were directed to use complete sentences and told to feel free to use their own words (Over, 2004).
To compare the text of human-authored multidocument summaries to the full-text documents describing the events, we automatically broke the documents into sentences, and constructed a minimal tiling of each summary sentence. Specifically, for each sentence in the summary, we searched for all n-grams that are present in both the summary and the documents, placing no restrictions on the potential size of an n-gram.

7 priedas Anglų kalbos publicistinio teksto santraukos, naudojant 50% kompresiją

Tools 4 Noobs summarizer anglų kalbos publicistinio teksto santrauka

President Joao Bernardo Vieira and General Tagme Na Waie died in separate incidents only hours apart. (25)
An army spokesman claimed responsibility for the President's death saying it was in reprisal for the earlier assassination of the army chief. "President Vieira was killed by the army as he tried to flee his house," said Zamora Induta. (27)
He said that the President was "taken down by bullets fired by … soldiers. " Mr Induta alleged that Mr Vieira was "one of the main people responsible for the death of [General Na Waie]. " Mr Vieira ruled Guinea-Bissau from 1980 to 1999 before being deposed in a military coup. (46)
This militia was, however, partly disarmed by the army after its gunmen were accused of shooting at General Na Waie's convoy in January in an incident that underlined the extent of the hostility between the President and his top military man. (31)
Drugs enforcement officials have complained that Mr Vieira failed to crackdown on the lucrative trade in which an estimated 50 tonnes of cocaine transit the region destined for Europe every year. (32)
But analysts say that the blowing up of General Na Waie bears the hallmarks of an attack by drugs cartels rather than the result of power struggles within the military. "There is no mutiny aspect to the bombing. (31)
There are fears that the instability might spread beyond Guinea-Bissau's own borders. "This is a very bad situation," said Richard Moncrieff, West Africa project director at International Crisis Group. "There is a power vacuum, people are not coming out onto the streets and there is still shooting going on. "There are many factions within the armed forces, the fear is that the army could fracture further," he warned. (31)

Shvoong summarizer anglų kalbos publicistinio teksto santrauka

President Joao Bernardo Vieira of Guinea-Bissau assassinated by army The President and head of the army in Guinea-Bissau were killed in tit-for-tat murders that have plunged the West African "narco-state" into crisis.
President Joao Bernardo Vieira and General Tagme Na Waie died in separate incidents only hours apart. General Na Waie was killed in a bomb blast at army headquarters. Hours later Mr Vieira died in a hail of bullets as he tried to flee his home in the capital Bissau. An army spokesman claimed responsibility for the President's death saying it was in reprisal for the earlier assassination of the army chief. "President Vieira was killed by the army as he tried to flee his house," said Zamora Induta. He said that the President was "taken down by bullets fired by … soldiers." Mr Induta alleged that Mr Vieira was "one of the main people responsible for the death of [General Na Waie]." Mr Vieira ruled Guinea-Bissau from 1980 to 1999 before being deposed in a military coup.
But tensions between the President and the army have remained high.
Days after parliamentary elections in November gunmen attacked the presidential palace leading Mr Vieira to establish an elite unit of personal bodyguards. This militia was, however, partly disarmed by the army after its

gunmen were accused of shooting at General Na Waie's convoy in January in an incident that underlined the extent of the hostility between the President and his top military man.
But analysts say that the blowing up of General Na Waie bears the hallmarks of an attack by drugs cartels rather than the result of power struggles within the military.
The murder of Mr Vieira was a revenge attack by General Na Waie's army loyalists.
"This is a very bad situation," said Richard Moncrieff, West Africa project director at International Crisis Group.

Word AutoSummarize anglų kalbos publicistinio teksto santrauka

President Joao Bernardo Vieira of Guinea-Bissau assassinated by army
The President and head of the army in Guinea-Bissau were killed in tit-for-tat murders that have plunged the West African "narco-state" into crisis.
President Joao Bernardo Vieira and General Tagme Na Waie died in separate incidents only hours apart. General Na Waie was killed in a bomb blast at army headquarters. An army spokesman claimed responsibility for the President's death saying it was in reprisal for the earlier assassination of the army chief.
"President Vieira was killed by the army as he tried to flee his house," said Zamora Induta. Mr Induta alleged that Mr Vieira was "one of the main people responsible for the death of [General Na Waie]."
Mr Vieira ruled Guinea-Bissau from 1980 to 1999 before being deposed in a military coup. Days after parliamentary elections in November gunmen attacked the presidential palace leading Mr Vieira to establish an elite unit of personal bodyguards.
This militia was, however, partly disarmed by the army after its gunmen were accused of shooting at General Na Waie's convoy in January in an incident that underlined the extent of the hostility between the President and his top military man.
Diplomats have accused General Na Waie of involvement in the growing cocaine trade through West Africa.
Drugs enforcement officials have complained that Mr Vieira failed to crackdown on the lucrative trade in which an estimated 50 tonnes of cocaine transit the region destined for Europe every year.
The murder of Mr Vieira was a revenge attack by General Na Waie's army loyalists.
There are fears that the instability might spread beyond Guinea-Bissau's own borders. "This is a very bad situation," said Richard Moncrieff, West Africa project director at International Crisis Group.

8 priedas Anglų kalbos grožinio teksto santrauka, naudojant 25% kompresiją

Word AutoSummarize anglų kalbos grožinio teksto santrauka

I HAD just been looking long and sadly at Holbein's ploughman, and was walking through the fields, musing on rustic life and the destiny of the husbandman. The man of leisure seldom loves, for their own sake, the fields and meadows, the landscape, or the noble animals which are to be converted into gold for his use. The happiest of men would be he who, knowing the full meaning of his labour, should, while working with his hands, find his happiness and his freedom in the exercise of his intelligence, and, having his heart in unison with his brain, should at once understand his own work and love that of God. The artist has such delights as these in contemplating and reproducing the beauties of nature; but if his heart be true and tender, his pleasure is disturbed when he sees the miseries of the men who people this paradise of earth. Then, instead of the pitiable and frightful figure of Death stalking, whip in hand, across the fields, the painter of allegories may place beside the peasant a radiant angel, sowing the blessed grain broadcast in the smoking furrow.
The sweet, sad words of Virgil: "Oh, happy the peasants of the field, if they knew their own blessings!" is a regret, but, like all regrets, it is also a prophecy. The man who draws in noble joy from the poetic feelings is a true poet, though he has never written a verse all his life.
At the other extremity of the field, an old man, whose broad shoulders and stern face recalled Holbein's ploughman, but whose clothes carried no suggestion of poverty, was gravely driving his plough of antique shape, drawn by two placid oxen, true patriarchs of the meadow, tall and rather thin, with pale yellow coats and long, drooping horns. The ox-herd will say: "There is a pair of oxen gone; this one will work no more, for his brother is dead. The old labourer worked slowly, silently, and without waste of effort. At the other end of the field a fine-looking youth was driving a magnificent team of four pairs of young oxen, through whose sombre coats glanced a ruddy, glow-like flame. The spirited animals quivered under the child's light touch, making their yokes and head-bands creak, and shaking the pole violently, whenever a root stopped the advance of the ploughshare, the labourer would call every animal by name in his powerful voice, trying to calm rather than to excite them; for the oxen, irritated by the sudden resistance, bounded, pawed the ground with their great cloven hoofs, and would have jumped aside and dragged the plough

across the fields, if the young man had not kept the first four in order with his voice and goad, while the child controlled the four others. The whole scene was beautiful in its grace and strength; the landscape, the man, the child, the oxen under the yoke; and in spite of the mighty struggle by which the earth was subdued, a deep feeling of peace and sweetness reigned over all. It is not enough to guide them skilfully, to trace a perfectly straight furrow, and to lighten their labour by raising the ploughshare or driving it into the earth; no man can be a consummate husband. man who does not know how to sing to his oxen, and that is an art that requires taste and especial gifts.
Instead of a wretched old man, a young and active one; instead of a team of weary and emaciated horses, four yoke of robust and fiery oxen; instead of death, a beautiful child; instead of despair and destruction, energy and the possibility of happiness.
Happy the peasant of the fields! this man has never understood the mystery of beauty; this child will never understand it. I knew the young man and the beautiful child; I knew their history, for they had a history. I mean, if possible, to save from oblivion the furrow of Germain, the skilled husbandman.

Intellexer summarizer anglų kalbos grožinio teksto santrauka

Concepts:
man, field, ox, furrow, child, life, peasant, earth, labourer, strength.
Summary:
I HAD just been looking long and sadly at Holbein's ploughman, and was walking through the fields, musing on rustic life and the destiny of the husbandman.
The man of leisure seldom loves, for their own sake, the fields and meadows, the landscape, or the noble animals which are to be converted into gold for his use.
The peasant is too abject, too wretched, and too fearful of the future to enjoy the beauty of the country and the charms of pastoral life.
The happiest of men would be he who, knowing the full meaning of his labour, should, while working with his hands, find his happiness and his freedom in the exercise of his intelligence, and, having his heart in unison with his brain, should at once understand his own work and love that of God.
The artist has such delights as these in contemplating and reproducing the beauties of nature; but if his heart be true and tender, his pleasure is disturbed when he sees the miseries of the men who people this paradise of earth.
Instead of the pitiable and frightful figure of Death stalking, whip in hand, across the fields, the painter of allegories may place beside the peasant a radiant angel, sowing the blessed grain broadcast in the smoking furrow.
The sweet, sad words of Virgil: "Oh, happy the peasants of the field, if they knew their own blessings!" is a regret, but, like all regrets, it is also a prophecy.
Without doubt the exclusion is the common result of excessive toil and of deep misery; but let it not be said that when men shall work moderately and usefully there will be nothing but bad workers and bad poets.
The man who draws in noble joy from the poetic feelings is a true poet, though he has never written a verse all his life.
My thoughts had flown in this direction, without my perceiving that my confidence in the capacity of man for education was strengthened by external influences.
At the other extremity of the field, an old man, whose broad shoulders and stern face recalled Holbein's ploughman, but whose clothes carried no suggestion of poverty, was gravely driving his plough of antique shape, drawn by two placid oxen, true patriarchs of the meadow, tall and rather thin, with pale yellow coats and long, drooping horns.
At the other end of the field a fine-looking youth was driving a magnificent team of four pairs of young oxen, through whose sombre coats glanced a ruddy, glow-like flame.
A child of six or seven years old, lovely as an angel, wearing round his shoulders, over his blouse, a sheepskin that made him look like a little Saint John the Baptist out of a Renaissance picture, was running along in the furrow beside the plough, pricking the flanks of the oxen with a long, light goad but slightly sharpened.
The whole scene was beautiful in its grace and strength; the landscape, the man, the child, the oxen under the yoke; and in spite of the mighty struggle by which the earth was subdued, a deep feeling of peace and sweetness reigned over all.
Each time that an obstacle was surmounted and the plough resumed its even solemn progress, the labourer, whose pretended violence was but a trial of his strength, and an outlet for his energy, instantly regained that serenity which is the right of simple souls, and looked with fatherly pleasure toward his child, who turned to smile back at him.

Subject search summarizer anglų kalbos grožinio teksto santrauka

At the other extremity of the field, an old man, whose broad shoulders and stern face recalled Holbein's ploughman, but whose clothes carried no suggestion of poverty, was gravely driving his plough of antique shape, drawn by two placid oxen, true patriarchs of the meadow, tall and rather thin, with pale yellow coats and long, drooping horns.

The Tillage of the Soil

I HAD just been looking long and sadly at Holbein's ploughman, and was walking through the fields, musing on rustic life and the destiny of the husbandman.

At the season of the year when there is no work or stir afoot except that of the ploughman, this strong, sweet refrain rises like the voice of the breeze, to which the key it is sung in gives it some resemblance.

Then the young father would raise his manly voice in the solemn and melancholy chant that ancient tradition transmits, not indeed to all ploughmen indiscriminately, but to those who are most perfect in the art of exciting and sustaining the spirit of cattle while at work.

Each time that an obstacle was surmounted and the plough resumed its even solemn progress, the labourer, whose pretended violence was but a trial of his strength, and an outlet for his energy, instantly regained that serenity which is the right of simple souls, and looked with fatherly pleasure toward his child, who turned to smile back at him.

Let them come and see in the corner of the stable one of these poor beasts, thin and wasted, restlessly lashing his lean flanks with his tail, violently breathing with mingled terror and disdain on the food offered him, his eyes always turned toward the door, scratching with his hoof the empty place at his side, sniffing the yokes and chains which his fellow used to wear, and incessantly calling him with melancholy lowings.

The spirited animals quivered under the child's light touch, making their yokes and head-bands creak, and shaking the pole violently, whenever a root stopped the advance of the ploughshare, the labourer would call every animal by name in his powerful voice, trying to calm rather than to excite them;

This song, which was probably sacred in its origin, and to which mysterious influences must once have been attributed, is still thought to possess the virtue of putting animals on their mettle, allaying their irritation, and of beguiling the weariness of their long, hard toil.

for the oxen, irritated by the sudden resistance, bounded, pawed the ground with their great cloven hoofs, and would have jumped aside and dragged the plough across the fields, if the young man had not kept the first four in order with his voice and goad, while the child controlled the four others.

At the other end of the field a fine-looking youth was driving a magnificent team of four pairs of young oxen, through whose sombre coats glanced a ruddy, glow-like flame.

but, thanks to the undistracted steadiness of his toil and the judicious expenditure of his strength, his furrow was as soon ploughed as that of his son, who was driving, at some distance from him, four less vigorous oxen through a more stubborn and stony piece of ground.

But it is none the less a noble song, and so appropriate is it to the nature of the work it accompanies, to the gait of the oxen, to the peace of the fields, and to the simplicity of the men who sing it, that no genius unfamiliar with the tillage of the earth, and no man except an accomplished labourer of our part of the country, could repeat it.

The happiest of men would be he who, knowing the full meaning of his labour, should, while working with his hands, find his happiness and his freedom in the exercise of his intelligence, and, having his heart in unison with his brain, should at once understand his own work and love that of God.

and, moreover, since poets have already raised their voices out of the lap of pain and of weariness, why should we say that the labour of the hands excludes the working of the soul?

and if I had to narrate the story of his life, the pleasure I should take in bringing out the tender and touching side of it would be greater than your merit in painting the degradation and contempt into which he is cast by your social code.

Then, instead of the pitiable and frightful figure of Death stalking, whip in hand, across the fields, the painter of allegories may place beside the peasant a radiant angel, sowing the blessed grain broadcast in the smoking furrow.

A child of six or seven years old, lovely as an angel, wearing round his shoulders, over his blouse, a sheepskin that made him look like a little Saint John the Baptist out of a Renaissance picture, was running along in the furrow beside the plough, pricking the flanks of the oxen with a long, light goad but slightly sharpened.

Is not the furrow of the labourer of as much value as that of the idler, even if that idler, by some absurd chance, have made a little noise in the world, and left behind him an abiding name?

It is certainly tragic for him to spend his days and his strength delving in the jealous earth, that so reluctantly yields up her rich treasures when a morsel of coarse black bread, at the end of the day's work, is the sole reward and profit to be reaped from such arduous toil.

Among those peasants who possess some of the comforts of life, and whose moral and intellectual development is not entirely stifled by extreme wretchedness, pure happiness that can be felt and appreciated exists in the elementary stage;

9 priedas Lietuvių kalbos grožinio teksto santraukos, naudojant 50% kompresiją

Tools 4 Noobs lietuvių kalbos grožinio teksto santrauka

Tai jauna mergina, tipiška andalūzietė, ilgais juodais plaukais ir primenančiomis senuosius užkariautojus maurus akimis. - Avys išmoko daugiau nei knygos, - atsako piemuo. (37)
Jis laimingas, turėdamas kitokį pašnekovą nei jo avys. - Kaip tu išmokai skaityti? - klausia mergina. - Kaip visi, - atsako jis. - Mokykloje. - Bet jei moki skaityti, kodėl esi tik piemuo? (36)
Ir kad piemenys, kaip ir jūrininkai, kaip ir komivojažieriai, visi turi vieną miestą, kuriame gyvena būtybė, sugebanti juos priversti užmiršti malonumą laisvai klajoti po pasaulį. * Kuomet pasirodo pirmoji ryto žara, piemuo ima raginti savo avis tekančios saulės kryptimi. (43)
Nesvarbu, kad visos jų dienos, susidedančios iš ilgų, tarp saulėtekio ir saulėlydžio išsirikiavusių valandų, panašios viena į kitą, kad per savo trumpą gyvenimą jos neperskaitė nė vienos knygos ir nesupranta žmonių kalbos, kuria jis joms pasakoja, kas dedasi kaimuose. (43)
Dvejus metus praleidęs klajodamas po Andalūzijos lygumas, jis mintinai žino visus krašto miestelius, ir prasmę jo gyvenimui teikia būtent tai: keliavimas. Šį sykį jis ketina paaiškinti merginai, kodėl paprastas piemuo moka skaityti: iki šešiolikos metų jis lankė seminariją. (52)
Nesvarbu, ar šviesiaplaukiai, ar tamsiaodžiai, jie panašus į mūsų kaimo žmones. - Bet aš nesu matęs tų kraštų, iš kur atkeliavo šie žmonės, nei jų pilių, - nesutinka jaunuolis. - Šie žmonės, matydami mūsų laukus ir moteris, sako, kad jie norėtų čia gyventi, - toliau kalba tėvas. - Aš noriu pažinti jų moteris ir jų kraštus, - tuomet sako sūnus. - Nes šie žmonės niekuomet pas mus nepasilieka. - Jie turi pilnas kišenes pinigų. (72)
Nusipirk bandą ir eik per pasaulį, iki suprasi, kad mūsų pilis svarbiausia ir mūsų moterys gražiausios. Ir jis palaimina sūnų. Žiburėliai tėvo akyse išduoda norą taip pat leistis klajoti po pasaulį. (38)
Ten galės iškeisti į kitą, storesnę, savo knygą, prisipildyti vyno butelį, nusiskusti ir apsikirpti: jis turi būti visiškai pasiruošęs susitikti su mergina ir nenori net pagalvoti apie tai, kad koks kitas piemuo su daugiau avių galėjo ateiti ir paprašyti jos rankos anksčiau už jį. , - pamano jis, vėl žvilgteldamas į dangų ir paspartindamas žingsnį. (71)

Copernic Summarizer lietuvių kalbos grožinio teksto santrauka

Concepts:
jis, kad, jos, savo, tai, mergina, piemuo, kaip, tik, iki, parduotuve, diena, kita, piemenys, pat.
Summary:
Jau temsta, kai jis su savo banda prieina sena, apleista ba~nycia.
Jos stogas igriuves, o toje vietoje, kur kadaise buta zakristijos, auga did~iulis jovaras.
Apylinkese nera vilku, bet jei gyvulys pabegtu, jis sugaiatu visa rytojaus diena jo beieakodamas.
Jis pasikloja apsiausta ir atsigula, vietoj pagalves pasidejes neseniai perskaityta knyga.
Priea u~migdamas pagalvoja, kad dabar reiketu skaityti storesnes knygas: ilgiau u~tektu, o nakti jos butu patogesnes pagalves.
Pasi~iuri aukatyn ir pro igriuvusi stoga mato blyksincias ~vaig~des.
Jis vel susapnavo ta pati praejusios savaites sapna, bet, kaip ir tasyk, nubudo nebaiges sapnuoti.
Jau seniau jis pastebejo, kad dauguma gyvuliu pakirsta ia miego, vos tik iasiblaivo jo paties samone.
Tartum kokia paslaptinga energija jungtu ji su aiomis avimis, per dvejus metus, ieakant vandens ir maisto drauge su juo apejusiomis visa aali.
- Jos taip priprato prie manes, kad ~ino mano dienotvarke, - sako tyliai.
pamastes, sau taria, kad visiakai galimas ir atvirkacias dalykas - kad tai jis priprates prie gyvuliu dienotvarkes.
Jis ~adina jas ia eiles, kiekviena pabaksnodamas lazda, pavadindamas vardu.
Vaikinas tiki, kad avys gali suprasti, ka jis sako.
Taigi kartais jis joms skaito labiausiai patikusiu knygu iatraukas, kalba apie savo vienatve ar d~iaugsma, kad yra piemuo, komentuoja naujienas, iagirstas einant per pa~istamus pakeliu kaimus.
Taciau paskutinemis dienomis beveik neliko kitu pokalbio temu, tik toji: mergina, gyvenanti mieste, kuri pasieks po keturiu dienu, vieno pirklio dukte.
Jis ten buvo tik syki, praejusiais metais.
Pirklys turejo audiniu parduotuve ir mego pats stebeti kerpamas avis - kad preke butu be jokios apgavystes.
Piemeniui parduotuve buvo nurodes draugas, ir jis ten atvare savaja banda.
- Noreciau parduoti aiek tiek vilnos, - kreipiasi jis i pirkli.
Parduotuve pilna, ir pirklys papraao piemens lukteleti iki pavakario.

Tuomet ais atsiseda ant aaligatvio prieaais parduotuve ir iasitraukia knyga.
- Aa ir ne~inojau, kad piemenys skaito knygas, - suskamba aalia moters balsas.
Tai jauna mergina, tipiaka andaluziete, ilgais juodais plaukais ir primenanciomis senuosius u~kariautojus maurus akimis.
- Avys iamoko daugiau nei knygos, - atsako piemuo.
Ji pasisako esanti pirklio dukte ir kalba apie miestelio gyvenima, kur kiekviena diena panaai i praejusia.
O piemuo pasakoja apie Andaluzijos pievas, apie aplankytus miestelius ir paskutines ju naujienas.
Jis laimingas, turedamas kitoki paanekova nei jo avys.
- Bet jei moki skaityti, kodel esi tik piemuo?
Todel tesia savo kelioniu istorijas, o ma~os mauriakos akutes ia nustebimo ar iagascio tai iasiplecia, tai visai u~simerkia.
Laikas bega, ir kuo toliau, tuo labiau jaunuolis trokata, kad ai diena niekuomet nepasibaigtu, kad merginos tevas ilgai butu u~imtas ir papraaytu jo palaukti dar tris dienas.
Staiga jis suvokia pajutes ka~ka, ko niekad iki aiol nebuvo jautes: nora likti cia visam laikui.
Su aia juodaplauke mergina jo dienos niekada nebutu panaaios viena i kita.
Bet ateina pirklys ir papraao nukirpti keturias avis.
Po to sumoka, kiek skolingas, ir pakviecia u~sukti kitais metais.
Iki to miestuko liko ne daugiau kaip keturios dienos.
Jis susijaudines ir kartu sunerimes: juk gali buti, kad mergina ji pamirao.
- Tai ne taip svarbu, - sako jis, kalbedamasis su savo avimis.
- Aa pa~istu ir kitu merginu kituose miesteliuose.
Bet giliai airdyje jis ~ino, kad tai toli gra~u nera nesvarbu.
Ir kad piemenys, kaip ir jurininkai, kaip ir komivoja~ieriai, visi turi viena miesta, kuriame gyvena butybe, sugebanti juos priversti u~mirati malonuma laisvai klajoti po pasauli.
Kuomet pasirodo pirmoji ryto ~ara, piemuo ima raginti savo avis tekancios saules kryptimi.
"Joms niekada nereikia nieko spresti, - galvoja jis.
- Gal todel jos visad laikosi greta manes".
Vienintelis aviu poreikis - vanduo ir ~ole.
Ir tol, kol piemuo ~ino geriausias Andaluzijos ganyklas, jos bus jo drauges.
Nesvarbu, kad visos ju dienos, susidedancios ia ilgu, tarp sauletekio ir saulelyd~io iasirikiavusiu valandu, panaaios viena i kita, kad per savo trumpa gyvenima jos neperskaite ne vienos knygos ir nesupranta ~moniu kalbos, kuria jis joms pasakoja, kas dedasi kaimuose.
Jos tenkinasi ~ole ir vandeniu, ir to joms visiakai gana.
Mainais jos dosniai atsilygina savaja vilna, draugyste, o kartais ir savo mesa.
"Jei retkarciais pasiversciau pabaisa ir imciau jas viena po kitos galabyti, jos nesuprastu aito, iki nebutu ianaikinta visa banda, - masto jis.
- Nes pasitiki manim ir nebesivadovauja instinktais.
Ir tik todel, kad tai aa jas vedu i ganyklas".
Staiga jaunuolis suvokia, kokios keistos jo mintys.
O gal toji ba~nycia su viduje auganciu jovaru buvo u~keikta?
Ir del to jis vel susapnavo ta sapna, o dabar jaucia ka~koki pykti visad iatikimoms savo biciulems avims?
Jaunuolis gurkateli likusio nuo vakarienes vyno ir tvirciau susisupa apsiaustu.
Jis ~ino, kad po keleto valandu, saulei pakilus i zenita, bus taip karata, jog jis nebegales vesti bandos.
Vasaros metu aia valanda visa Ispanija miega.
Bus karata iki pat nakties, ir visa ta laika apsiausta teks neati rankose.
Kai jam u~eina noras padusauti del neaulio, jis sau primena, kad to neaulio deka jam nebuvo aalta ryta.
"Mes turime buti pasiruoae visokiems orams", - tuomet pagalvoja jis ir pajunta dekinguma apsiausto svoriui.
Apsiaustas turi prie~asti egzistuoti, lygiai kaip ir jis pats.
Dvejus metus praleides klajodamas po Andaluzijos lygumas, jis mintinai ~ino visus kraato miestelius, ir prasme jo gyvenimui teikia butent tai: keliavimas.
`i syki jis ketina paaiakinti merginai, kodel paprastas piemuo moka skaityti: iki aeaiolikos metu jis lanke seminarija.
Tevai norejo, kad jis taptu kunigu - kukliõs, triusiancios tik del valgio ir gerimo, kaip tos jo avys, valstieciu aeimos pasidid~iavimu.
Bet jis nuo pat vaikystes svajojo pa~inti pasauli, ir tai jam buvo daug svarbiau nei pa~inti Dieva ar ~moniu nuodemes.
- Per ai kaima perejo ~moniu ia viso pasaulio, sunau.
Jie ateina ieakodami ka~ko naujo, taciau visuomet lieka tokie patys.

Intellexer summarizer lietuvių kalbos grožinio teksto santrauka

Concepts:
ir, jis, kad, ir kad piemenys, ir dÄ—l, ir tol, ir jis palaimina sÅ«nÅ³, jis prisimena, jis tikras, lygiai kaip ir jis pat.
Summary:
Jau temsta, kai jis su savo banda prieina sena, apleista ba˛nycia.
Suvaro avis i griuvesius ir, kad ios per nakti
nei silakstytu, i keliu lentu padaro u˛tvara.
Apylinkese nera vilku, bet jei gyvulys pabegtu, jis sugai tu visa rytojaus diena jo beieškodamas.
Jis pasikloja apsiausta ir atsigula, vietoj pagalves pasidejes neseniai perskaityta knyga.
Jis vel susapnavo ta pati praejusios savaites sapna, bet, kaip ir tasyk, nubudo nebaiges sapnuoti.
Jau seniau jis pastebejo, kad dauguma gyvuliu pakirsta iš miego, vos tik išsiblaivo jo paties samone.
- Jos taip priprato prie manes, kad ˛ino mano dienotvarke, - sako tyliai.
Paskui, akimirka
pamastes, sau taria, kad visi kai galimas ir atvirk cias dalykas - kad tai jis priprates prie gyvuliu dienotvarkes.
Jis ˛adina jas i eiles, kiekviena pabaksnodamas lazda, pavadindamas vardu.
Vaikinas tiki, kad avys gali suprasti, ka jis sako.
Taigi kartais jis joms skaito labiausiai patikusiu knygu i traukas, kalba apie savo vienatve ar d˛iaugsma, kad yra piemuo, komentuoja naujienas, i girstas einant per pa˛istamus pakeliu kaimus.
Jis ten buvo tik syki, praejusiais metais.
Pirklys turejo audiniu parduotuve ir mego pats stebeti kerpamas avis - kad preke butu be jokios apgavystes.
Piemeniui parduotuve buvo nurodes draugas, ir jis ten atvare savaja banda.
*
- Noreciau parduoti šiek tiek vilnos, - kreipiasi jis i pirkli.
- A ir ne˛inojau, kad piemenys skaito knygas, - suskamba alia moters balsas.
Tai jauna mergina, tipi ka andaluziete, ilgais juodais plaukais ir primenanciomis senuosius u˛kariautojus maurus akimis.
O piemuo pasakoja apie Andaluzijos pievas, apie aplankytus miestelius ir paskutines ju naujienas.
Jis laimingas, turedamas kitoki pašnekova nei jo avys.
- Kaip visi, - atsako jis.
Jis tikras, kad mergina jo nesuprastu.
Laikas bega, ir kuo toliau, tuo labiau jaunuolis trokšta, kad ši diena niekuomet nepasibaigtu, kad merginos tevas ilgai butu u˛imtas ir papra ytu jo palaukti dar tris dienas.
Staiga jis suvokia pajutes ka˛ka, ko niekad iki iol nebuvo jautes: nora likti cia visam laikui.
Bet ateina pirklys ir papra o nukirpti keturias avis.
Jis susijaudines ir kartu sunerimes: juk gali buti, kad mergina ji pamir o.
- Tai ne taip svarbu, - sako jis, kalbedamasis su savo avimis.
Bet giliai irdyje jis ˛ino, kad tai toli gra˛u nera nesvarbu.
Ir kad piemenys, kaip ir jurininkai, kaip ir komivoja˛ieriai, visi turi viena miesta, kuriame gyvena butybe, sugebanti juos priversti u˛mir ti malonuma laisvai klajoti po pasauli.
Joms niekada nereikia nieko spresti, - galvoja jis.
Vienintelis aviu poreikis - vanduo ir ˛ole.
Ir tol, kol piemuo ˛ino geriausias Andaluzijos ganyklas, jos bus jo drauges.
Nesvarbu, kad visos ju dienos, susidedancios i ilgu, tarp sauletekio ir saulelyd˛io i sirikiavusiu valandu, pana ios viena i kita, kad per savo trumpa gyvenima jos neperskaite ne vienos knygos ir nesupranta ˛moniu kalbos, kuria jis joms pasakoja, kas dedasi kaimuose.
Jos tenkinasi ˛ole ir vandeniu, ir to joms visi kai gana.
Mainais jos dosniai atsilygina savaja vilna, draugyste, o kartais ir savo mesa.
"Jei retkarciais pasiversciau pabaisa ir imciau jas viena po kitos galabyti, jos nesuprastu ito, iki nebutu i naikinta visa banda, - masto jis.
Ir tik todel, kad tai a jas vedu i ganyklas".
Jis ˛ino, kad po keleto valandu, saulei pakilus i zenita, bus taip kar ta, jog jis nebegales vesti bandos.
Kai jam u˛eina noras padusauti del ne ulio, jis sau primena, kad to ne ulio deka jam nebuvo alta ryta.
"Mes turime buti pasiruo e visokiems orams", - tuomet pagalvoja jis ir pajunta dekinguma apsiausto svoriui.
Apsiaustas turi prie˛asti egzistuoti, lygiai kaip ir jis pats.
Dvejus metus praleides klajodamas po Andaluzijos lygumas, jis mintinai ˛ino visus kra to miestelius, ir prasme jo gyvenimui teikia butent tai: keliavimas.
i syki jis ketina paai kinti merginai, kodel paprastas piemuo moka skaityti: iki e iolikos metu jis lanke seminarija.

Tevai norejo, kad jis taptu kunigu - kukliõs, triusiancios tik del valgio ir gerimo, kaip tos jo avys, valstieciu eimos pasidid¸iavimu.
Bet jis nuo pat vaikystes svajojo pa¸inti pasauli, ir tai jam buvo daug svarbiau nei pa¸inti Dieva ar ¸moniu nuodemes.
Viena vakara, atejes aplankyti eimos, jis sukaupe drasa ir pasake tevui, kad
nenori buti kunigu.
U¸kopia kalvon, aplanko pili, o po to sau taria, kad praeitis vertesne u¸ dabarti.
- ie ¸mones, matydami musu laukus ir moteris, sako, kad jie noretu cia gyventi, - toliau kalba tevas.
Nusipirk banda ir eik per pasauli, iki suprasi, kad musu pilis svarbiausia ir musu moterys gra¸iausios.
Ir jis palaimina sunu.
Visad jo irdyje slypejusi nora, nepaisant de imciu metu, per kuriuos jis stengesi ji u¸slopinti, kas ryta toje pacioje vietoje keldamasis ir guldamas, gerdamas ir valgydamas.
Jaunuolis prisimena pokalbi su tevu ir jauciasi laimingas; jis jau mate daug piliu ir daug moteru (bet ne viena negali prilygti tai, kuri jo laukia po dvieju dienu).
Jis turi apsiausta, knyga, kuria gali iškeisti i kita, turi aviu banda.
Taciau, svarbiausia, kiekviena diena jis igyvendina savo did¸iaja svajone keliauti.
Kai pavares nuo Andaluzijos lygumu, gales parduoti savo avis ir tapti jurininku.
- nusistebi jis, ¸iuredamas i tekancia saule.
Jos nepastebi, kad keiciasi ganyklos, metu laikai.
Nes vienintelis ju rupestis - ¸ole ir vanduo".
"Gal taip yra ir mums visiems,- masto piemuo.
- Juk ir a nuo tada, kai susipa¸inau su pirklio dukra, nebegalvoju apie kitas moteris".
Ten gales iškeisti i kita, storesne, savo knyga, prisipildyti vyno buteli, nusiskusti ir apsikirpti: jis turi buti visiškai pasiruošes susitikti su mergina ir nenori net pagalvoti apie tai, kad koks kitas piemuo su daugiau aviu galejo ateiti ir paprašyti jos rankos anksciau u¸ ji.
"Galimybe igyvendinti svajone - tai kas gyvenima daro idomu", - pamano jis, vel ¸vilgteldamas i dangu ir paspartindamas ¸ingsni.
Jis prisimena, kad Tarifoje gyvena sena moteris, mokanti ai kinti sapnus.

Word AutoSummarize lietuvių kalbos grožinio teksto santrauka

Jau temsta, kai jis su savo banda prieina seną, apleistą bažnyčią. Suvaro avis į griuvėsius ir, kad šios per naktį
Jis pasikloja apsiaustą ir atsigula, vietoj pagalvės pasidėjęs neseniai perskaitytą knygą. Dar nepražvitus jis nubunda. Jis vėl susapnavo tą patį praėjusios savaitės sapną, bet, kaip ir tąsyk, nubudo nebaigęs sapnuoti.
Vaikinas atsikelia ir išgeria gurkšnį vyno. Tuomet stveria botagą ir pradeda žadinti dar
Jau seniau jis pastebėjo, kad dauguma gyvulių pakirsta iš miego, vos tik išsiblaivo jo paties sąmonė. pamąstęs, sau taria, kad visiškai galimas ir atvirkščias dalykas - kad tai jis pripratęs prie gyvulių dienotvarkės.
Vaikinas tiki, kad avys gali suprasti, ka jis sako. Jis ten buvo tik sykį, praėjusiais metais. Piemeniui parduotuvę buvo nurodęs draugas, ir jis ten atvarė savąja bandą.
*
- Norėčiau parduoti šiek tiek vilnos, - kreipiasi jis į pirklį.
Parduotuvė pilna, ir pirklys paprašo piemens luktelėti iki pavakario. - Aš ir nežinojau, kad piemenys skaito knygas, - suskamba šalia moters balsas.
O piemuo pasakoja apie Andalūzijos pievas, apie aplankytus miestelius ir paskutines jų naujienas. Jis laimingas, turėdamas kitokį pašnekovą nei jo avys.
- Kaip visi, - atsako jis. Jis tikras, kad mergina jo nesuprastų. Laikas bėga, ir kuo toliau, tuo labiau jaunuolis trokšta, kad ši diena niekuomet nepasibaigtų, kad merginos tėvas ilgai būtų užimtas ir paprašytų jo palaukti dar tris dienas. Bet ateina pirklys ir paprašo nukirpti keturias avis. Jis susijaudinęs ir kartu sunerimęs: juk gali būti, kad mergina ji pamiršo. - Tai ne taip svarbu, - sako jis, kalbėdamasis su savo avimis. - Aš pažįstu ir kitu merginu kituose miesteliuose.
Bet giliai širdyje jis žino, kad tai toli gražu nėra nesvarbu. Ir kad piemenys, kaip ir jūrininkai, kaip ir komivojažieriai, visi turi vieną miestą, kuriame gyvena būtybė, sugebanti juos priversti užmiršti malonumą laisvai klajoti po pasaulį.
*
"Joms niekada nereikia nieko spręsti, - galvoja jis. Vienintelis avių poreikis - vanduo ir žolė. Ir tol, kol piemuo žino geriausias Andalūzijos ganyklas, jos bus jo draugės.
Nesvarbu, kad visos jų dienos, susidedančios iš ilgų, tarp saulėtekio ir saulėlydžio išsirikiavusių valandų, panašios viena į kitą, kad per savo trumpą gyvenimą jos neperskaitė nė vienos knygos ir nesupranta žmonių kalbos, kuria jis joms pasakoja, kas dedasi kaimuose. Jos tenkinasi žole ir vandeniu, ir to joms visiškai gana. "Jei retkarčiais pasiverščiau pabaisa ir imčiau jas viena po kitos galabyti, jos nesuprastų šito, iki nebūtų išnaikinta visa banda, - mąsto jis. - Nes pasitiki manim ir nebesivadovauja instinktais. Ir tik todėl, kad tai aš jas vedu į ganyklas".

Jis žino, kad po keleto valandų, saulei pakilus į zenitą, bus taip karšta, jog jis nebegalės vesti bandos. "Mes turime būti pasiruošę visokiems orams", - tuomet pagalvoja jis ir pajunta dėkingumą apsiausto svoriui.
Apsiaustas turi priežastį egzistuoti, lygiai kaip ir jis pats. Šį sykį jis ketina paaiškinti merginai, kodėl paprastas piemuo moka skaityti: iki šešiolikos metų jis lankė seminariją. Tėvai norėjo, kad jis taptų kunigu - kukliõs, triūsiančios tik dėl valgio ir gėrimo, kaip tos jo avys, valstiečių šeimos pasididžiavimu. Bet jis nuo pat vaikystės svajojo pažinti pasaulį, ir tai jam buvo daug svarbiau nei pažinti Dievą ar žmonių nuodėmes. Vieną vakarą, atėjęs aplankyti šeimos, jis sukaupė drąsa ir pasakė tėvui, kad
Kad nori keliauti.
*
- Šie žmonės, matydami mūsų laukus ir moteris, sako, kad jie norėtų čia gyventi, - toliau kalba tėvas.
- Aš noriu pažinti jų moteris ir jų kraštus, - tuomet sako sūnus. Nusipirk bandą ir eik per pasaulį, iki suprasi, kad mūsų pilis svarbiausia ir mūsų moterys gražiausios.
Ir jis palaimina sūnų. Visad jo širdyje slypėjusį norą, nepaisant dešimčių metų, per kuriuos jis stengėsi jį užslopinti, kas rytą toje pačioje vietoje keldamasis ir guldamas, gerdamas ir valgydamas.
*
Jaunuolis prisimena pokalbį su tėvu ir jaučiasi laimingas; jis jau matė daug pilių ir daug moterų (bet nė viena negali prilygti tai, kuri jo laukia po dviejų dienų). Jis turi apsiaustą, knygą, kurią gali iškeisti į kitą, turi avių bandą. Tačiau, svarbiausia, kiekvieną dieną jis įgyvendina savo didžiąją svajonę keliauti. Kai pavaręs nuo Andalūzijos lygumų, galės parduoti savo avis ir tapti jūrininku. - nusistebi jis, žiūrėdamas į tekančią saulę. Nes vienintelis jų rūpestis - žolė ir vanduo".
"Gal taip yra ir mums visiems,- mąsto piemuo. Ten galės iškeisti į kitą, storesnę, savo knygą, prisipildyti vyno butelį, nusiskusti ir apsikirpti: jis turi būti visiškai pasiruošęs susitikti su mergina ir nenori net pagalvoti apie tai, kad koks kitas piemuo su daugiau avių galėjo ateiti ir paprašyti jos rankos anksčiau už jį.
Jis prisimena, kad Tarifoje gyvena sena moteris, mokanti aiškinti sapnus.

10 priedas Lietuvių kalbos mokslinio teksto santraukos, naudojant 50% kompresiją

Tools 4 Noobs lietuvių kalbos mokslinio teksto santrauka

LIETUVIŲ KALBOS TEKSTYNO MORFOLOGINĖS ANALIZĖS AUTOMATIZAVIMAS Erika Rimkutė Vytauto Didžiojo universitetas, Humanitarinis fakultetas Daukanto 28, LT-3000 Kaunas, Lietuva Šiame pranešime bus paaiškinta, kas yra tekstynas, kaip jis gali būti žymimas, kaip gramatiškai analizuojami kompiuteriniai tekstai. (67)
Bus pristatyti šios programos privalumai ir trūkumai, problemų sprendimo būdai. 1 Įvadas XX a. pabaigoje ištobulėjus kompiuteriams didelių pakitimų įvyko ir gana stabilioje mokslo šakoje – kalbotyroje, kurioje tekstynų lingvistika, galima sakyti, padarė perversmą: leido visiškai kitaip pažvelgti į kalbą, ją analizuoti objektyviai, pateikti tikslesnius duomenis, kadangi atsiribojama nuo subjektyvios analizės, dažniausiai pagrįstos kalbos jausmu ir intuicija. (79)
Tekstynų lingvistika parodė visai kitą kalbos vaizdą, nes didelis ir ganėtinai reprezentatyvus kompiuteriu tvarkomas ir nuolat papildomas tekstynas atskleidžia kur kas didesnę kalbos vienetų vartojimo įvairovę, o statistika leidžia nustatyti būdingiausius, dažniausiai vartojamus atvejus ir atskirti juos nuo retesnių [2]. (68)
Informatikų sukurtos programos gali anotuoti tekstus, bet be lingvistų kalbinės analizės ir taisymų tokie tekstai nesuteikia daug informacijos, nelabai padeda gramatinei analizei, kadangi bent jau kol kas neįmanoma automatiškai panaikinti daugiaprasmybių, kurios trukdo analizei. (54)
Todėl ir norisi, remiantis savo darbo patirtimi su morfologiškai anotuotu tekstu, parodyti, kokių iškyla problemų ir kaip jos galėtų būti sprendžiamos. 2 Tekstyno samprata Tekstynas – bet koks, kad ir pats mažiausias, elektroninę formą turinčių tekstų rinkinys. (62)
Galima skirti net tris šio žodžio reikšmes: pati bendriausia tekstyno reikšmė yra „bet koks tekstų rinkinys”, dažniausia – „elektroninių tekstų rinkinys”, o griežčiausia, terminologiškiausia - „baigtinis elektroninių tekstų rinkinys, sudarytas taip, kad kuo geriau atspindėtų kalbą ar jos atmainą [3]. (66)
Taigi esminė teksto ir tekstyno skirtybė ne sandara, ne kalbų kiekis, net ne dydis ar reprezantatyvumas, bet su kompiliacine teksto prigimtimi susijusi jo tyrimo metodologija [3, 5]. (64)
Vytauto Didžiojo universiteto tekstynas buvo sumanytas kaip didelis neanotuotas ir nekoduotas, į skaitytoją orientuotas, daugiausiai ištisų periodikos ir knygų tekstų, daugiau bendro pobūdžio nei specialus, didelės temų ir kitokios įvairovės autentiškos rašytinės lietuvių kalbos tekstynas, kuris dabar jau yra pasiekęs 60 mln. žodžių apimties ribą ir toliau didinamas [3]. (90)

Dėl tekstynų kodavimo ir anotavimo kyla daug lingvistų nesutarimų: ar teksto pažymos padeda analizuoti, ar trukdo.
3 Tekstyno žymėjimas Tekstynas gali būti pažymimas ir anotuojamas teksto formatavimo ar analizės labui. (96)
Teksto žymėjimas arba kodavimas atsirado tekstynų gyvavimo pradžioje dėl to, kad senieji kompiuteriai nesugebėjo apdoroti teksto kaip teksto, bet jie mokėjo atpažinti pažymas. (77)
Koduojant pažymimi tokie formalieji teksto struktūros elementai kaip nuorodos apie rašytinio teksto citatas, sąrašus, vardo raidžių ar kitokias santrumpas, akronimus, knygų įžangas ir apendiksus, skyrius, atskirų teksto dalių ir kitokius pavadinimus, tikrinius vardus. (83)
Dar žymimi tikriniai pavadinimai, pvz. , Juodoji jūra, Ivanas Rūstusis, Lietuvos ir Lenkijos karalystė ir t. t. , kadangi Lemuokliui išskaidžius tokius junginius negalima suprasti, kad tai tikrinis pavadinimas, pvz. , junginį Juodoji jūra sulemavus į antraštines formas būdvardį juodas ir daiktavardį jūra, neparodoma, jog tai tikrinis pavadinimas, todėl šio junginio vientisumas tvarkomame sulemuotame tekste atrodo taip: Juodoji jūra Taigi kodavimas daugiau susijęs su struktūriniais teksto elementais, tačiau jis apima ir interpretacijos atvejus, kai pažymos vietą ir pobūdį lemia subjektyvi tyrėjo nuomonė. (74)
Tekstyno žymėjimą galima suskirstyti į tris etapus: pirmiausia grynas tekstas lemuojamas – pateikiama tekstyne pavartoto žodžio antraštinė forma, t. y. lema, pvz. , forma namą sulemuojama į antraštinį pavidalą namas, t. y. vienaskaitos vardininko linksnį. (49)
Antrojo etapo metu gali būti pateikti žodžių formų gramatiniai apibūdinimai, reiškiami gramatinėmis kategorijomis, pvz. , nurodoma, kad forma namą - tai daiktavardis, vyriškoji giminė, vienaskaitos galininkas. Šiuos žymėjimus gali atlikti ir jau veikianti morfologinės analizės programa Lmuoklis. (49)
Sudėtingiausias teksto apdorojimo etapas po anotavimo – sintaksinė analizė, kurią sudaro sintaksinius ryšius vaizduojančio medžio kūrimas kiekvienam teksto sakiniui. (53)

QuickJist lietuvių kalbos mokslinio teksto santrauka

Vytauto Didžiojo universiteto dabartinės lietuvių kalbos tekstynas suartino gana tolimų specialybių žmones – informatikus ir lingvistus.
Glaudus informatikų ir lingvistų bendradarbiavimas ypač akivaizdus nagrinėjant automatiškai morfologiškai sužymėtą tekstą.
2 Tekstyno samprata
Tekstynas – bet koks, kad ir pats mažiausias, elektroninę formą turinčių tekstų rinkinys.
Tekstyną su tekstu sieja tai, kad jie sudaryti iš tekstų; gali sutapti teksto ir tekstyno apimtys.
Tačiau tekstas analizuojamas ištisai, jis turi struktūrą: pradžią, vidurį, pabaigą, yra daugiau ar mažiau rišlus ir vientisas, o tekstynas neturi struktūros, tik sandarą.
Jo neverta ir neįmanoma tirti tiesiogiai, skaityti taip, kaip teksto, o tik su programinėmis priemonėmis, įvairiais įrankiais.
Taigi esminė teksto ir tekstyno skirtybė ne sandara, ne kalbų kiekis, net ne dydis ar reprezantatyvumas, bet su kompiliacine teksto prigimtimi susijusi jo tyrimo metodologija [3, 5].
Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centras ketina ne tik didinti dabar esantį tekstyną, bet ir sukurti kelis mažesnius, iš kurių vienas bus morfologiškai anotuotas.
Dėl tekstynų kodavimo ir anotavimo kyla daug lingvistų nesutarimų: ar teksto pažymos padeda analizuoti, ar trukdo.
3 Tekstyno žymėjimas
Tekstynas gali būti pažymimas ir anotuojamas teksto formatavimo ar analizės labui.
Teksto žymėjimas arba kodavimas atsirado tekstynų gyvavimo pradžioje dėl to, kad senieji kompiuteriai nesugebėjo apdoroti teksto kaip teksto, bet jie mokėjo atpažinti pažymas.
Dabartinės technologijos leidžia apdoroti tekstus jau be pažymų, bet senoji žymėjimo praktika išliko.
Kartais tokios pažymos gali padėti analizei, bet kur kas dažniau jai gali pakenkti.
Be to, pažymos pažeidžia teksto vientisumą, jos yra žmogaus intervencija į tekstą.
Tik per pažymas žvelgiama į tekstyno kalbą, o tai, kas nesužymėta, yra prarandama.
Anotuoto teksto privalumas yra tas, kad jei anotacijos yra neklaidinančios, jos labai palengvina paiešką ir bet kokią lingvistinę tekstyno analizę [3, 4].
Tekstyno žymėjimą galima suskirstyti į tris etapus: pirmiausia grynas tekstas lemuojamas – pateikiama tekstyne pavartoto žodžio antraštinė forma, t. y. lema, pvz., forma namą sulemuojama į antraštinį pavidalą namas, t. y. vienaskaitos vardininko linksnį.
Sudėtingiausias teksto apdorojimo etapas po anotavimo – sintaksinė analizė, kurią sudaro sintaksinius ryšius vaizduojančio medžio kūrimas kiekvienam teksto sakiniui.
Eama ir dar sudėtingesnių darbo su tekstu etapų ir juos atitinkančių priemonių: teksto generavimo, vertimo, santraukų automatinio rengimo sistemų, bet jos – jau nebe tekstynų, o kompiuterinės lingvistikos sritis [3].

Automatic summarizer lietuvių kalbos mokslinio teksto santrauka

LIETUVIŲ KALBOS TEKSTYNOMORFOLOGINĖS ANALIZĖS AUTOMATIZAVIMASErika RimkutėVytauto Didžiojo universitetas, Humanitarinis fakultetasDaukanto 28, LT-3000 Kaunas, LietuvaŠiame pranešime bus paaiškinta, kas yra tekstynas, kaip jis gali būti žymimas, kaip gramatiškai analizuojamikompiuteriniai tekstai.

Didžiausias dėmesys skiriamas jau veikiančiai automatinei morfologinei analizei,kurią atlieka kompiuterinė programa Lemuoklis.

Bus pristatyti šios programos privalumai ir trūkumai,problemų sprendimo būdai.

1 ĮvadasXX a. pabaigoje ištobulėjus kompiuteriams didelių pakitimų įvyko ir gana stabilioje mokslo šakoje – kalbotyroje, kurioje tekstynų lingvistika, galima sakyti, padarė perversmą: leido visiškai kitaip pažvelgti į kalbą, ją analizuoti objektyviai, pateikti tikslesnius duomenis, kadangi atsiribojama nuo subjektyvios analizės, dažniausiai pagrįstos kalbos jausmu ir intuicija.

Tekstynų lingvistika parodė visai kitą kalbos vaizdą, nes didelis ir ganėtinai reprezentatyvus kompiuteriu tvarkomas ir nuolat papildomas tekstynas atskleidžia kur kas didesnę kalbos vienetų vartojimo įvairovę, o statistika leidžia nustatyti būdingiausius, dažniausiai vartojamus atvejus ir atskirti juos nuo retesnių [2].

Vytauto Didžiojo universiteto dabartinės lietuvių kalbos tekstynas suartino gana tolimų specialybių žmones – informatikus ir lingvistus.

Akivaizdu, kad tekstynų lingvistikos pažanga yra tiesiogiai susijusi su informatikos mokslo raida ir su efektyvesniu informatikų ir lingvistų bendradarbiavimu,- juk norint geriau aprėpti ir išanalizuoti didėjantį informacijos kiekį reikia naudoti vis pažangesnes lingvistines kompiuterines programas [6].

Glaudus informatikų ir lingvistų bendradarbiavimas ypač akivaizdus nagrinėjant automatiškai morfologiškai sužymėtą tekstą.

Informatikų sukurtos programos gali anotuoti tekstus, bet be lingvistų kalbinės analizės ir taisymų tokie tekstai nesuteikia daug informacijos, nelabai padeda gramatinei analizei, kadangi bent jau kol kas neįmanoma automatiškai panaikinti daugiaprasmybių, kurios trukdo analizei.

Darbas su morfologiškai tekstą pažyminčia kompiuterine programa paskatino parodyti, kad mūsų kalba nėra nedviprasmiška, lengvai suklasifikuojama į kalbos dalis.

Toks sunkiai sugrupuojamų kalbos vienetų vaizdas ypač išryškėja tada, kai kompiuterinė programa, nesuvokdama konteksto, ryšių tarp žodžių, pateikia daugybę hipotetinių variantų.

Todėl ir norisi, remiantis savo darbo patirtimi su morfologiškai anotuotu tekstu, parodyti, kokių iškyla problemų ir kaip jos galėtų būti sprendžiamos.

Tačiau vienas žymiausių tekstynų lingvistikos atstovų J. Sinclairis tekstynu siūlo vadinti tik tokį tekstų rinkinį, kuris yra pakankamai didelis, matuojant pagal šių dienų kompiuterinių technologijų galimybes, ir sudarytas ne dėl kokio specialaus tyrimo, bet nepriklausomai nuo jo panaudojimo tikslų.

Galima skirti net tris šio žodžio reikšmes: pati bendriausia tekstyno reikšmė yra ,,bet koks tekstų rinkinys", dažniausia – ,,elektroninių tekstų rinkinys", o griežčiausia, terminologiškiausia - ,,baigtinis elektroninių tekstų rinkinys, sudarytas taip, kad kuo geriau atspindėtų kalbą ar jos atmainą [3].

Tekstyną su tekstu sieja tai, kad jie sudaryti iš tekstų; gali sutapti teksto ir tekstyno apimtys.

Tačiau tekstas analizuojamas ištisai, jis turi struktūrą: pradžią, vidurį, pabaigą, yra daugiau ar mažiau rišlus ir vientisas, o tekstynas neturi struktūros, tik sandarą.

Jo neverta ir neįmanoma tirti tiesiogiai, skaityti taip, kaip teksto, o tik su programinėmis priemonėmis, įvairiais įrankiais.

Vytauto Didžiojo universiteto tekstynas buvo sumanytas kaip didelis neanotuotas ir nekoduotas, į skaitytoją orientuotas, daugiausiai ištisų periodikos ir knygų tekstų, daugiau bendro pobūdžio nei specialus, didelės temų ir kitokios įvairovės autentiškos rašytinės lietuvių kalbos tekstynas, kuris dabar jau yra pasiekęs 60 mln.

Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centras ketina ne tik didinti dabar esantį tekstyną, bet ir sukurti kelis mažesnius, iš kurių vienas bus morfologiškai anotuotas.

Dėl tekstynų kodavimo ir anotavimo kyla daug lingvistų nesutarimų: ar teksto pažymos padeda analizuoti, ar trukdo.

Teksto žymėjimas arba kodavimas atsirado tekstynų gyvavimo pradžioje dėl to, kad senieji kompiuteriai nesugebėjo apdoroti teksto kaip teksto, bet jie mokėjo atpažinti pažymas.

Pavyzdžiui, Vytauto Didžiojo universiteto tekstyne yra pažymimimi užsienio kalbų žodžiai, pvz.

Taip pat žymimi Lemuoklio išskaidyti, nors prasmės atžvilgiu neskaidomi junginiai tokie kaip be abejo, iš tikrųjų, kas nors, tam tikras, taip pat ir panašūs. Taigi šie žymėjimai tvarkomame tekste atrodo taip:<phr type=įterp>
<phr type=prvks>be abejo arba: kada nors</phr>
</phr>Dar žymimi tikriniai pavadinimai, pvz.

Dar daugiau interpretacijos esama gramatinėse anotacijose, nes paprastai jos remiasi tradicine, daugeliu atvejų subjektyvia nuomone ir susitarimu paremta gramatika (consensus grammar).
Gavęs gramatiškai anotuotą tekstą, kompiuteris netikrina gramatinų kategorijų, bet priima jas tokias, kokios jos yra, taigi ir analizės rezultatai daugeliu atvejų yra iš anksto nulemti, nes kompiuteris dirba su pažymomis ir ignoruoja pačią kalbą.
Kartais tokios pažymos gali padėti analizei, bet kur kas dažniau jai gali pakenkti.
Dar viena anotacijų yda – tekstas perkraunamas pažymomis, jį tampa sunkiau perdirbti, nes apimtis patrigubėja ar padidėja dar daugiau kartų.
Anotuoto teksto privalumas yra tas, kad jei anotacijos yra neklaidinančios, jos labai palengvina paiešką ir bet kokią lingvistinę tekstyno analizę [3, 4].
, forma namą sulemuojama į antraštinį pavidalą namas, t. y. vienaskaitos vardininko linksnį.
Antrojo etapo metu gali būti pateikti žodžių formų gramatiniai apibūdinimai, reiškiami gramatinėmis kategorijomis, pvz.
, nurodoma, kad forma namą - tai daiktavardis, vyriškoji giminė, vienaskaitos galininkas.
Tam tikslui reikalinga speciali programa, kurios pagrindinė funkcija būtų dviprasmybių panaikinimas (angliškas terminas – ambiguity resolution).
Tekstyno anotavimas neapsiriboja vien morfologija, todėl dar esama sintaksinių, semantinių, nusakančių skirtingas daugiareikšmio žodžio reikšmes, ir net tekstinių ar diskursinių žymių.
Sudėtingiausias teksto apdorojimo etapas po anotavimo – sintaksinė analizė, kurią sudaro sintaksinius ryšius vaizduojančio medžio kūrimas kiekvienam teksto sakiniui.
Eama ir dar sudėtingesnių darbo su tekstu etapų ir juos atitinkančių priemonių: teksto generavimo, vertimo, santraukų automatinio rengimo sistemų, bet jos – jau nebe tekstynų, o kompiuterinės lingvistikos sritis [3].

Subject Search Summarizer lietuvių kalbos mokslinio teksto santrauka

3 Tekstyno žymėjimas
Tekstynas gali būti pažymimas ir anotuojamas teksto formatavimo ar analizės labui.
Tekstyno žymėjimą galima suskirstyti į tris etapus: pirmiausia grynas tekstas lemuojamas – pateikiama tekstyne pavartoto žodžio antraštinė forma, t.
Teksto žymėjimas arba kodavimas atsirado tekstynų gyvavimo pradžioje dėl to, kad senieji kompiuteriai nesugebėjo apdoroti teksto kaip teksto, bet jie mokėjo atpažinti pažymas.
Tekstyną su tekstu sieja tai, kad jie sudaryti iš tekstų;
2 Tekstyno samprata
Tekstynas – bet koks, kad ir pats mažiausias, elektroninę formą turinčių tekstų rinkinys.
Galima skirti net tris šio žodžio reikšmes: pati bendriausia tekstyno reikšmė yra ,,bet koks tekstų rinkinys”, dažniausia – ,,elektroninių tekstų rinkinys”, o griežčiausia, terminologiškiausia - ,,baigtinis elektroninių tekstų rinkinys, sudarytas taip, kad kuo geriau atspindėtų kalbą ar jos atmainą [3].
Sinclairis tekstynu siūlo vadinti tik tokį tekstų rinkinį, kuris yra pakankamai didelis, matuojant pagal šių dienų kompiuterinių technologijų galimybes, ir sudarytas ne dėl kokio specialaus tyrimo, bet nepriklausomai nuo jo panaudojimo tikslų.
Taigi esminė teksto ir tekstyno skirtybė ne sandara, ne kalbų kiekis, net ne dydis ar reprezantatyvumas, bet su kompiliacine teksto prigimtimi susijusi jo tyrimo metodologija [3, 5].
Tekstynų lingvistika parodė visai kitą kalbos vaizdą, nes didelis ir ganėtinai reprezentatyvus kompiuteriu tvarkomas ir nuolat papildomas tekstynas atskleidžia kur kas didesnę kalbos vienetų vartojimo įvairovę, o statistika leidžia nustatyti būdingiausius, dažniausiai vartojamus atvejus ir atskirti juos nuo retesnių [2].
Eama ir dar sudėtingesnių darbo su tekstu etapų ir juos atitinkančių priemonių: teksto generavimo, vertimo, santraukų automatinio rengimo sistemų, bet jos – jau nebe tekstynų, o kompiuterinės lingvistikos sritis [3].
Anotuoto teksto privalumas yra tas, kad jei anotacijos yra neklaidinančios, jos labai palengvina paiešką ir bet kokią lingvistinę tekstyno analizę [3, 4].
Tačiau vienas žymiausių tekstynų lingvistikos atstovų J.
Dėl tekstynų kodavimo ir anotavimo kyla daug lingvistų nesutarimų: ar teksto pažymos padeda analizuoti, ar trukdo.
Informatikų sukurtos programos gali anotuoti tekstus, bet be lingvistų kalbinės analizės ir taisymų tokie tekstai nesuteikia daug informacijos, nelabai padeda gramatinei analizei, kadangi bent jau kol kas neįmanoma automatiškai panaikinti daugiaprasmybių, kurios trukdo analizei.
Akivaizdu, kad tekstynų lingvistikos pažanga yra tiesiogiai susijusi su informatikos mokslo raida ir su efektyvesniu informatikų ir lingvistų bendradarbiavimu,- juk norint geriau aprėpti ir išanalizuoti didėjantį informacijos kiekį reikia naudoti vis pažangesnes lingvistines kompiuterines programas [6].
Glaudus informatikų ir lingvistų bendradarbiavimas ypač akivaizdus nagrinėjant automatiškai morfologiškai sužymėtą tekstą.

Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centras ketina ne tik didinti dabar esantį tekstyną, bet ir sukurti kelis mažesnius, iš kurių vienas bus morfologiškai anotuotas.
Darbas su morfologiškai tekstą pažyminčia kompiuterine programa paskatino parodyti, kad mūsų kalba nėra nedviprasmiška, lengvai suklasifikuojama į kalbos dalis.
Tekstyno anotavimas neapsiriboja vien morfologija, todėl dar esama sintaksinių, semantinių, nusakančių skirtingas daugiareikšmio žodžio reikšmes, ir net tekstinių ar diskursinių žymių.
LIETUVIŲ KALBOS TEKSTYNO
MORFOLOGINĖS ANALIZĖS AUTOMATIZAVIMAS
Erika Rimkutė
Vytauto Didžiojo universitetas, Humanitarinis fakultetas
Daukanto 28, LT-3000 Kaunas, Lietuva
Šiame pranešime bus paaiškinta, kas yra tekstynas, kaip jis gali būti žymimas, kaip gramatiškai analizuojami kompiuteriniai tekstai.
Vytauto Didžiojo universiteto dabartinės lietuvių kalbos tekstynas suartino gana tolimų specialybių žmones – informatikus ir lingvistus.
Vytauto Didžiojo universiteto tekstynas buvo sumanytas kaip didelis neanotuotas ir nekoduotas, į skaitytoją orientuotas, daugiausiai ištisų periodikos ir knygų tekstų, daugiau bendro pobūdžio nei specialus, didelės temų ir kitokios įvairovės autentiškos rašytinės lietuvių kalbos tekstynas, kuris dabar jau yra pasiekęs 60 mln.
Koduojant pažymimi tokie formalieji teksto struktūros elementai kaip nuorodos apie rašytinio teksto citatas, sąrašus, vardo raidžių ar kitokias santrumpas, akronimus, knygų įžangas ir apendiksus, skyrius, atskirų teksto dalių ir kitokius pavadinimus, tikrinius vardus.
Be to, pažymos pažeidžia teksto vientisumą, jos yra žmogaus intervencija į tekstą.
Gavęs gramatiškai anotuotą tekstą, kompiuteris netikrina gramatinų kategorijų, bet priima jas tokias, kokios jos yra, taigi ir analizės rezultatai daugeliu atvejų yra iš anksto nulemti, nes kompiuteris dirba su pažymomis ir ignoruoja pačią kalbą.
Pavyzdžiui, Vytauto Didžiojo universiteto tekstyne yra pažymimimi užsienio kalbų žodžiai, pvz.
Taigi šie žymėjimai tvarkomame tekste atrodo taip:
<phr type=įterp> <phr type=prvks>
be abejo arba: kada nors
</phr> </phr>
Dar žymimi tikriniai pavadinimai, pvz.
gali sutapti teksto ir tekstyno apimtys.

11 priedas Lietuvių kalbos publicistinio teksto santraukos, naudojant 50% kompresiją

Tools 4 Noobs lietuvių kalbos publicistinio teksto santrauka

NASA erdvėlaivis „Discovery“ atsiskyrė nuo Tarptautinės kosminės stoties ir grįžta į Žemę AFP-BNS ir lrytas. lt inf. 2009-03-26 11:25 JAV erdvėlaivis „Discovery“ trečiadienį atsiskyrė nuo Tarptautinės kosminės stoties (TKS) ir greitai grįš į Žemę, pranešė Nacionalinė aeronautikos ir kosmoso administracija (NASA). (26)
Prieš aštuonias dienas prie TKS prisijungęs „Discovery“ atsiskyrė kaip planuota 19 val. 53 min. Grinvičo (21 val. 53 min. (15)
Prieš nuskriedami tolyn įgulos nariai trumpam sustabdė erdvėlaivį už 200 metrų nuo TKS stoties, kad ją nufotografuotų ir nufilmuotų. „Discovery“ prisijungė prie TKS kovo 17 dieną, praėjus dviem dienoms po starto. (16)
Viena iš šios misijos svarbiausių užduočių - atgabenti ir sumontuoti paskutinį kosminės stoties saulės baterijų komplektą. (17)
Kosminės stoties „jėgainę“ sudaro 32 800 saulės baterijų elementai, kuriuose šviesa paverčiama elektros energija. (23)
Sumontavus paskutiniąsias saulės baterijas TKS jėgainės galingumas padidėjo iki 120 kilovatų, todėl stočiai pakaks energijos visiems planuojamiems moksliniams bandymams, kurie bus atliekami laboratorijų moduliuose „Kibo“ ir „Columbus“. (13)

QuickJist Summarizer lietuvių kalbos publicistinio teksto santrauka

Prieš aštuonias dienas prie TKS prisijungęs „Discovery“ atsiskyrė kaip planuota 19 val. 53 min. Grinvičo (21 val. 53 min. Lietuvos) laiku, kai orbitinė stotis skriejo 350 kilometrų aukštyje virš Indijos vandenyno.
„Discovery“ prisijungė prie TKS kovo 17 dieną, praėjus dviem dienoms po starto.
Viena iš šios misijos svarbiausių užduočių - atgabenti ir sumontuoti paskutinį kosminės stoties saulės baterijų komplektą.
Dabar TKS sumontuotos keturios saulės baterijų poros.
Kosminės stoties „jėgainę“ sudaro 32 800 saulės baterijų elementai, kuriuose šviesa paverčiama elektros energija.

Copernic Summarizer lietuvių kalbos publicistinio teksto santrauka

Concepts:
TKS, saules bateriju, kosmines stoties, Discovery, atsiskyre, erdvelaivis, sumontuoti paskutini, atsiskyre nuo Tarptautines, NASA, energija, jegaine, aio, igulos, vienas, laiku.
Summary:
JAV erdvelaivis,,Discovery" treciadieni atsiskyre nuo Tarptautines kosmines stoties (TKS) ir greitai gria ieme, praneae Nacionaline aeronautikos ir kosmoso administracija (NASA).
Priea aatuonias dienas prie TKS prisijunges,,Discovery" atsiskyre kaip planuota 19 val.
,,Hjustonai, mes fiziakai atsiskyreme", - praneae vienas erdvelaivio igulos narys per aio manevro tiesiogine transliacija televizijoje,,NASA TV".
,,Discovery" ir jo septyniu astronautu igula turetu nutupti Floridoje esanciame Kennedy kosminiame centre aeatadieni 1 val.
Priea nuskriedami tolyn igulos nariai trumpam sustabde erdvelaivi u~ 200 metru nuo TKS stoties, kad ja nufotografuotu ir nufilmuotu.
Viena ia aios misijos svarbiausiu u~duociu - atgabenti ir sumontuoti paskutini kosmines stoties saules bateriju komplekta.
Dabar TKS sumontuotos keturios saules bateriju poros.
Kosmines stoties,,jegaine" sudaro 32 800 saules bateriju elementai, kuriuose aviesa paverciama elektros energija.
Sumontavus paskutiniasias saules baterijas TKS jegaines galingumas padidejo iki 120 kilovatu, todel stociai pakaks energijos visiems planuojamiems moksliniams bandymams, kurie bus atliekami laboratoriju moduliuose,,Kibo" ir,,Columbus".

Intellexer Summarizer lietuvių kalbos publicistinio teksto santrauka

Concepts:
min. lietuvos laiku, discovery, stoties, nasa, ir greitai grÄÆÅ□, kibo ir columbus, atgabenti ir sumontuoti paskutinÄÆ kosminÄ—s stoties saulÄ—s baterijÅ³ komplektÄ…, discovery prisijungÄ— prie tks kovo dienÄ…, min. grinviÄ¨o, min. vietos.
Summary:
NASA erdvelaivis „Discovery“ atsiskyre nuo Tarptautines kosmines stoties ir gri¸ta i ˇeme
AFP-BNS ir lrytas.lt inf.
2009-03-26 11:25
JAV erdvelaivis „Discovery“ treciadieni atsiskyre nuo Tarptautines kosmines stoties (TKS) ir greitai gri i ˇeme, prane e Nacionaline aeronautikos ir kosmoso administracija (NASA).
Prie a tuonias dienas prie TKS prisijunges „Discovery“ atsiskyre kaip planuota 19 val.
53 min. Lietuvos) laiku, kai orbitine stotis skriejo 350 kilometru auk tyje vir Indijos vandenyno.
43 min. Lietuvos) laiku, jeigu bazeje i silaikys palankus orai.
Viena i ios misijos svarbiausiu u¸duociu - atgabenti ir sumontuoti paskutini kosmines stoties saules bateriju komplekta.
Kosmines stoties „jegaine“ sudaro 32 800 saules bateriju elementai, kuriuose viesa paverciama elektros energija.
Sumontavus paskutiniasias saules baterijas TKS jegaines galingumas padidejo iki 120 kilovatu, todel stociai pakaks energijos visiems planuojamiems moksliniams bandymams, kurie bus atliekami laboratoriju moduliuose „Kibo“ ir „Columbus“.

12 priedas Lietuvių kalbos tekstų santraukos, naudojant 25% kompresiją

Tools 4 Noobs lietuvių kalbos grožinio teksto santrauka

Nesvarbu, ar šviesiaplaukiai, ar tamsiaodžiai, jie panašus į mūsų kaimo žmones. - Bet aš nesu matęs tų kraštų, iš kur atkeliavo šie žmonės, nei jų pilių, - nesutinka jaunuolis. - Šie žmonės, matydami mūsų laukus ir moteris, sako, kad jie norėtų čia gyventi, - toliau kalba tėvas. - Aš noriu pažinti jų moteris ir jų kraštus, - tuomet sako sūnus. - Nes šie žmonės niekuomet pas mus nepasilieka. - Jie turi pilnas kišenes pinigų. (72)
Ten galės iškeisti į kitą, storesnę, savo knygą, prisipildyti vyno butelį, nusiskusti ir apsikirpti: jis turi būti visiškai pasiruošęs susitikti su mergina ir nenori net pagalvoti apie tai, kad koks kitas piemuo su daugiau avių galėjo ateiti ir paprašyti jos rankos anksčiau už jį. , - pamano jis, vėl žvilgteldamas į dangų ir paspartindamas žingsnį. (71)

Copernic Summarizer lietuvių kalbos grožinio teksto santrauka

Concepts:
jis, kad, jos, savo, tai, mergina, piemuo, kaip, tik, iki, parduotuve, diena, kita, piemenys, pat.
Summary:
Jau temsta, kai jis su savo banda prieina sena, apleista ba~nycia.
Jos stogas igriuves, o toje vietoje, kur kadaise buta zakristijos, auga did~iulis jovaras.
Apylinkese nera vilku, bet jei gyvulys pabegtu, jis sugaiatu visa rytojaus diena jo beieakodamas.
Jis pasikloja apsiausta ir atsigula, vietoj pagalves pasidejes neseniai perskaityta knyga.
Priea u~migdamas pagalvoja, kad dabar reiketu skaityti storesnes knygas: ilgiau u~tektu, o nakti jos butu patogesnes pagalves.
Pasi~iuri aukatyn ir pro igriuvusi stoga mato blyksincias ~vaig~des.
Jis vel susapnavo ta pati praejusios savaites sapna, bet, kaip ir tasyk, nubudo nebaiges sapnuoti.
Jau seniau jis pastebejo, kad dauguma gyvuliu pakirsta ia miego, vos tik iasiblaivo jo paties samone.
Tartum kokia paslaptinga energija jungtu ji su aiomis avimis, per dvejus metus, ieakant vandens ir maisto drauge su juo apejusiomis visa aali.
- Jos taip priprato prie manes, kad ~ino mano dienotvarke, - sako tyliai.
pamastes, sau taria, kad visiakai galimas ir atvirkacias dalykas - kad tai jis priprates prie gyvuliu dienotvarkes.
Jis ~adina jas ia eiles, kiekviena pabaksnodamas lazda, pavadindamas vardu.
Vaikinas tiki, kad avys gali suprasti, ka jis sako
Taigi kartais jis joms skaito labiausiai patikusiu knygu iatraukas, kalba apie savo vienatve ar d~iaugsma, kad yra piemuo, komentuoja naujienas, iagirstas einant per pa~istamus pakeliu kaimus.
Taciau paskutinemis dienomis beveik neliko kitu pokalbio temu, tik toji: mergina, gyvenanti mieste, kuri pasieks po keturiu dienu, vieno pirklio dukte.
Jis ten buvo tik syki, praejusiais metais.
Pirklys turejo audiniu parduotuve ir mego pats stebeti kerpamas avis - kad preke butu be jokios apgavystes.
Piemeniui parduotuve buvo nurodes draugas, ir jis ten atvare savaja banda.
- Noreciau parduoti aiek tiek vilnos, - kreipiasi jis i pirkli.
Parduotuve pilna, ir pirklys papraao piemens lukteleti iki pavakario.
Tuomet ais atsiseda ant aaligatvio prieaais parduotuve ir iasitraukia knyga.
Tai jauna mergina, tipiaka andaluziete, ilgais juodais plaukais ir primenanciomis senuosius u~kariautojus maurus akimis.
- Avys iamoko daugiau nei knygos, - atsako piemuo.
Ji pasisako esanti pirklio dukte ir kalba apie miestelio gyvenima, kur kiekviena diena panaai i praejusia.
O piemuo pasakoja apie Andaluzijos pievas, apie aplankytus miestelius ir paskutines ju naujienas.
Laikas bega, ir kuo toliau, tuo labiau jaunuolis trokata, kad ai diena niekuomet nepasibaigtu, kad merginos tevas ilgai butu u~imtas ir papraaytu jo palaukti dar tris dienas.
Su aia juodaplauke mergina jo dienos niekada nebutu panaaios viena i kita.
- Tai ne taip svarbu, - sako jis, kalbedamasis su savo avimis.
Ir kad piemenys, kaip ir jurininkai, kaip ir komivoja~ieriai, visi turi viena miesta, kuriame gyvena butybe, sugebanti juos priversti u~mirati malonuma laisvai klajoti po pasauli.
Vasaros metu aia valanda visa Ispanija miega.
Bus karata iki pat nakties, ir visa ta laika apsiausta teks neati rankose.
"Mes turime buti pasiruoae visokiems orams", - tuomet pagalvoja jis ir pajunta dekinguma apsiausto svoriui.
Apsiaustas turi prie~asti egzistuoti, lygiai kaip ir jis pats.

QuickJist Summarizer lietuvių kalbos grožinio teksto santrauka

Jau temsta, kai jis su savo banda prieina seną, apleistą bažnyčią.

Jis vėl susapnavo tą patį praėjusios savaitės sapną, bet, kaip ir tąsyk, nubudo nebaigęs sapnuoti.
Jau seniau jis pastebėjo, kad dauguma gyvulių pakirsta iš miego, vos tik išsiblaivo jo paties sąmonė.
Paskui, akimirką
pamąstęs, sau taria, kad visiškai galimas ir atvirkščias dalykas - kad tai jis pripratęs prie gyvulių dienotvarkės.
Vaikinas tiki, kad avys gali suprasti, ka jis sako.
Taigi kartais jis joms skaito labiausiai patikusių knygų ištraukas, kalba apie savo vienatvę ar džiaugsmą, kad yra piemuo, komentuoja naujienas, išgirstas einant per pažįstamus pakelių kaimus.
Jis ten buvo tik sykį, praėjusiais metais.
- Aš ir nežinojau, kad piemenys skaito knygas, - suskamba šalia moters balsas.
Jis susijaudinęs ir kartu sunerimęs: juk gali būti, kad mergina ji pamiršo.
Bet giliai širdyje jis žino, kad tai toli gražu nėra nesvarbu.
Ir kad piemenys, kaip ir jūrininkai, kaip ir komivojažieriai, visi turi vieną miestą, kuriame gyvena būtybė, sugebanti juos priversti užmiršti malonumą laisvai klajoti po pasaulį.
Ir tik todėl, kad tai aš jas vedu į ganyklas".
Ir dėl to jis vėl susapnavo ta sapną, o dabar jaučia kažkokį pyktį visad ištikimoms savo bičiulėms avims?
Jis žino, kad po keleto valandų, saulei pakilus į zenitą, bus taip karšta, jog jis nebegalės vesti bandos.
Kai jam užeina noras padūsauti dėl nešulio, jis sau primena, kad to nešulio dėka jam nebuvo šalta ryta.
Tėvai norėjo, kad jis taptų kunigu - kukliõs, triūsiančios tik dėl valgio ir gėrimo, kaip tos jo avys, valstiečių šeimos pasididžiavimu.
Vieną vakarą, atėjęs aplankyti šeimos, jis sukaupė drąsa ir pasakė tėvui, kad
nenori būti kunigu.
- Šie žmonės, matydami mūsų laukus ir moteris, sako, kad jie norėtų čia gyventi, - toliau kalba tėvas.
- Aš noriu pažinti jų moteris ir jų kraštus, - tuomet sako sūnus.
Nusipirk bandą ir eik per pasaulį, iki suprasi, kad mūsų pilis svarbiausia ir mūsų moterys gražiausios.
Jaunuolis prisimena pokalbį su tėvu ir jaučiasi laimingas; jis jau matė daug pilių ir daug moterų (bet nė viena negali prilygti tai, kuri jo laukia po dviejų dienų).
Jis turi apsiaustą, knygą, kurią gali iškeisti į kitą, turi avių bandą.
"Aviu bėda, kad jos nesuvokia kasdien einančios naujais keliais.
Jos nepastebi, kad keičiasi ganyklos, metų laikai.
Jis prisimena, kad Tarifoje gyvena sena moteris, mokanti aiškinti sapnus.
Programos lange nerodomos nosinės raidės. Kiti lietuviški rašmenys rodomi. NUkopijavus tekstą, lietuviški rasmenys yra, tik rodoma, jog tai anglų kalbos tekstas.

Intellexer Summarizer lietuvių kalbos mokslinio teksto santrauka

Concepts:
kad ir, teksto, yra, todÄ—l ir norisi, pauzes ir pan, informatikus ir lingvistus, bet, gali sutapti teksto ir tekstyno apimtys, taip pat ir panaÅ Å«s, bet jos.
Summary:
Informatiku sukurtos programos gali anotuoti tekstus, bet be lingvistu kalbines analizes ir taisymu tokie tekstai nesuteikia daug informacijos, nelabai padeda gramatinei analizei, kadangi bent jau kol kas neimanoma automati kai panaikinti daugiaprasmybiu, kurios trukdo analizei.
Todel ir norisi, remiantis savo darbo patirtimi su morfologi kai anotuotu tekstu, parodyti, kokiu i kyla problemu ir kaip jos galetu buti sprend¸iamos.
2 Tekstyno samprata
Tekstynas – bet koks, kad ir pats ma¸iausias, elektronine forma turinciu tekstu rinkinys.
Tekstyna su tekstu sieja tai, kad jie sudaryti i tekstu; gali sutapti teksto ir tekstyno apimtys.
Jo neverta ir neimanoma tirti tiesiogiai, skaityti taip, kaip teksto, o tik su programinemis priemonemis, ivairiais irankiais.
Vytauto Did¸iojo universiteto tekstynas buvo sumanytas kaip didelis neanotuotas ir nekoduotas, i skaitytoja orientuotas, daugiausiai i tisu periodikos ir knygu tekstu, daugiau bendro pobud¸io nei specialus, dideles temu ir kitokios ivairoves autenti kos ra ytines lietuviu kalbos tekstynas, kuris dabar jau yra pasiekes 60 mln.
Dabartines technologijos leid¸ia apdoroti tekstus jau be pa¸ymu, bet senoji ¸ymejimo praktika i liko.
Koduojant pa¸ymimi tokie formalieji teksto strukturos elementai kaip nuorodos apie ra ytinio teksto citatas, sara us, vardo raid¸iu ar kitokias santrumpas, akronimus, knygu i¸angas ir apendiksus, skyrius, atskiru teksto daliu ir kitokius pavadinimus, tikrinius vardus.
Taip pat ¸ymimi Lemuoklio i skaidyti, nors prasmes at¸vilgiu neskaidomi junginiai tokie kaip be abejo, i tikruju, kas nors, tam tikras, taip pat ir pana us.
</phr>

Dar ˛ymimi tikriniai pavadinimai, pvz., Juodoji jura, Ivanas Rustusis, Lietuvos ir Lenkijos karalyste ir t. t., kadangi Lemuokliui i skaid˛ius tokius junginius negalima suprasti, kad tai tikrinis pavadinimas, pvz., jungini Juodoji jura sulemavus i antra tines formas budvardi juodas ir daiktavardi jura, neparodoma, jog tai tikrinis pavadinimas, todel io junginio vientisumas tvarkomame sulemuotame tekste atrodo taip:
<name type=place>
Juodoji jura
</name>
Taigi kodavimas daugiau susijes su strukturiniais teksto elementais, taciau jis apima ir interpretacijos atvejus, kai pa˛ymos vieta ir pobudi lemia subjektyvi tyrejo nuomone.
Gaves gramati kai anotuota teksta, kompiuteris netikrina gramatinu kategoriju, bet priima jas tokias, kokios jos yra, taigi ir analizes rezultatai daugeliu atveju yra i anksto nulemti, nes kompiuteris dirba su pa˛ymomis ir ignoruoja pacia kalba.
Tik per pa˛ymas ˛velgiama i tekstyno kalba, o tai, kas nesu˛ymeta, yra prarandama.
Dar viena anotaciju yda – tekstas perkraunamas pa˛ymomis, ji tampa sunkiau perdirbti, nes apimtis patrigubeja ar padideja dar daugiau kartu.
Toliau turetu buti i ai kinami keleriopas lemas turintys atvejai, nes, pvz., forma laimes gali buti sulemuota ir kaip laime ir kaip laimeti.

Word AutoSummarize lietuvių kalbos publicistinio teksto santrauka

NASA erdvėlaivis „Discovery“ atsiskyrė nuo Tarptautinės kosminės stoties ir grįžta į Žemę
JAV erdvėlaivis „Discovery“ trečiadienį atsiskyrė nuo Tarptautinės kosminės stoties (TKS) ir greitai grįš į Žemę, pranešė Nacionalinė aeronautikos ir kosmoso administracija (NASA).
Dabar TKS sumontuotos keturios saulės baterijų poros.

13 priedas Lietuvių kalbos tekstų santraukos, naudojant 10% kompresiją

Intellexer Summarizer lietuvių kalbos grožinio teksto santrauka

Concepts:
ir, jis, kad, ir kad piemenys, ir dÄ—l, ir tol, ir jis palaimina sÅ«nÅ³, jis prisimena, jis tikras, lygiai kaip ir jis pat.
Summary:
Vaikinas tiki, kad avys gali suprasti, ka jis sako.
Pirklys turėjo audinių parduotuvė ir mėgo pats stebėti kerpamas avis - kad prekė būtų be jokios apgavystės.
- Kaip visi, - atsako jis.
Jis tikras, kad mergina jo nesuprastų.
Bet ateina pirklys ir paprašo nukirpti keturias avis.
- Tai ne taip svarbu, - sako jis, kalbėdamasis su savo avimis.
Bet giliai širdyje jis žino, kad tai toli gražu nėra nesvarbu.
Ir kad piemenys, kaip ir jūrininkai, kaip ir komivojažieriai, visi turi vieną miestą, kuriame gyvena būtybė, sugebanti juos priversti užmiršti malonumą laisvai klajoti po pasaulį.
Ir tol, kol piemuo žino geriausias Andalūzijos ganyklas, jos bus jo draugės.
Jos tenkinasi žole ir vandeniu, ir to joms visiškai gana.
Jis žino, kad po keleto valandų, saulei pakilus į zenitą, bus taip karšta, jog jis nebegalės vesti bandos.
Kai jam užeina noras padūsauti dėl nešulio, jis sau primena, kad to nešulio dėka jam nebuvo šalta ryta.
- nusistebi jis, žiūrėdamas į tekančią saulę.

Copernic Summarizer lietuvių kalbos grožinio teksto santrauka

Concepts:
plough, labourer, peasant, life, furrow, pleasure, strength, child, voice, animals, earth, heart, happiness, beautiful, country.
Summary:

It is certainly tragic for him to spend his days and his strength delving in the jealous earth, that so reluctantly yields up her rich treasures when a morsel of coarse black bread, at the end of the day's work, is the sole reward and profit to be reaped from such arduous toil.
On the other hand, the peasant is too abject, too wretched, and too fearful of the future to enjoy the beauty of the country and the charms of pastoral life.
The day will come when the labourer too may be an artist, and may at least feel what is beautiful, if he cannot express it---a matter of far less importance.
At the other extremity of the field, an old man, whose broad shoulders and stern face recalled Holbein's ploughman, but whose clothes carried no suggestion of poverty, was gravely driving his plough of antique shape, drawn by two placid oxen, true patriarchs of the meadow, tall and rather thin, with pale yellow coats and long, drooping horns.
Everybody has his own, and could make the romance of his life interesting, if he could but understand it.
Although but a peasant and a labourer, Germain had always been aware of his duties and affections.
After some time spent in watching him plough, it occurred to me that I might write his story, though that story were as simple, as straight-forward, and unadorned as the furrow he was tracing.
Is not the furrow of the labourer of as much value as that of the idler, even if that idler, by some absurd chance, have made a little noise in the world, and left behind him an abiding name?

Copernic Summarizer lietuvių kalbos mokslinio teksto santrauka

Concepts:
tekstyno, teksto, yra, lingvistikos, kad, kaip, kalbos, programa, kompiuteris, gali, daugiau, analizei, teksts, jos, informatiks.
Summary:
Akivaizdu, kad tekstyns lingvistikos pa~anga yra tiesiogiai susijusi su informatikos mokslo raida ir su efektyvesniu informatiks ir lingvists bendradarbiavimu,- juk norint geriau apr_pti ir iaanalizuoti did_jant/ informacijos kiek/ reikia naudoti vis pa~angesnes lingvistines kompiuterines programas [6].
Informatiks sukurtos programos gali anotuoti tekstus, bet be lingvists kalbin_s analiz_s ir taisyms tokie tekstai nesuteikia daug informacijos, nelabai padeda gramatinei analizei, kadangi bent jau kol kas ne/manoma automatiakai panaikinti daugiaprasmybis, kurios trukdo analizei.
Darbas su morfologiakai tekst_ pa~ymin ia kompiuterine programa paskatino parodyti, kad mkss kalba n_ra nedviprasmiaka, lengvai suklasifikuojama / kalbos dalis.
Tekstynas -- bet koks, kad ir pats ma~iausias, elektronin_ form_ turin is teksts rinkinys.
Ta iau vienas ~ymiausis tekstyns lingvistikos atstovs J. Sinclairis tekstynu siklo vadinti tik tok/ teksts rinkin/, kuris yra pakankamai didelis, matuojant pagal ais diens kompiuterinis technologijs galimybes, ir sudarytas ne d_l kokio specialaus tyrimo, bet nepriklausomai nuo jo panaudojimo tiksls.
Jo neverta ir ne/manoma tirti tiesiogiai, skaityti taip, kaip teksto, o tik su programin_mis priemon_mis, /vairiais /rankiais.
Vytauto Did~iojo universiteto tekstynas buvo sumanytas kaip didelis neanotuotas ir nekoduotas, / skaitytoj_ orientuotas, daugiausiai iatiss periodikos ir knygs teksts, daugiau bendro pobkd~io nei specialus, didel_s tems ir kitokios /vairov_s autentiakos raaytin_s lietuvis kalbos tekstynas, kuris dabar jau yra pasiek_s 60 mln.
Gav_s gramatiakai anotuot_ tekst_, kompiuteris netikrina gramatins kategorijs, bet priima jas tokias, kokios jos yra, taigi ir analiz_s rezultatai daugeliu atvejs yra ia anksto nulemti, nes kompiuteris dirba su pa~ymomis ir ignoruoja pa i_ kalb_.

Pertinence Summarizer lietuvių kalbos mokslinio teksto santrauka

Informatikų sukurtos programos gali anotuoti tekstus, bet be lingvistų kalbinės analizės ir taisymų tokie tekstai nesuteikia daug informacijos, nelabai padeda gramatinei analizei, kadangi bent jau kol kas neįmanoma automatiškai panaikinti daugiaprasmybių, kurios trukdo analizei.
Dar daugiau interpretacijos esama gramatinėse anotacijose, nes paprastai jos remiasi tradicine, daugeliu atvejų subjektyvia nuomone ir susitarimu paremta gramatika (consensus grammar).
Tam tikslui reikalinga speciali programa, kurios pagrindinė funkcija būtų dviprasmybių panaikinimas (angliškas terminas – ambiguity resolution).

Shvoong Summarizer lietuvių kalbos publicistinio teksto santrauka

NASA erdvėlaivis „Discovery“ atsiskyrė nuo Tarptautinės kosminės stoties ir grįžta į Žemę AFP-BNS ir lrytas.lt inf. 2009-03-26 11:25 JAV erdvėlaivis „Discovery“ trečiadienį atsiskyrė nuo Tarptautinės kosminės stoties (TKS) ir greitai grįš į Žemę, pranešė Nacionalinė aeronautikos ir kosmoso administracija (NASA).

Copernic Summarizer lietuvių kalbos publicistinio teksto santrauka

Concepts:
TKS, baterijs, stoties, Discovery, sumontuoti, nuo, NASA, energija, aio, vienas, laiku, Lietuvos, planuota, prie TKS, Priea.
Summary:
JAV erdv_laivis,,Discovery" tre iadien/ atsiskyr_ nuo Tarptautin_s kosmin_s stoties (TKS) ir greitai gr/a /em_, pranea_ Nacionalin_ aeronautikos ir kosmoso administracija (NASA).

14 priedas Lietuvių kalbos mokslinio teksto santrauka, naudojant 5% kompresiją

Intellexer Summarizer lietuvių kalbos mokslinio teksto santrauka

Concepts:
kad ir, teksto, yra, todÄ—l ir norisi, pauzes ir pan, informatikus ir lingvistus, bet, gali sutapti teksto ir tekstyno apimtys, taip pat ir panaÅ Å«s, bet jos.
Summary:
Informatikų sukurtos programos gali anotuoti tekstus, bet be lingvistų kalbinės analizės ir taisymų tokie tekstai nesuteikia daug informacijos, nelabai padeda gramatinei analizei, kadangi bent jau kol kas neįmanoma automatiškai panaikinti daugiaprasmybių, kurios trukdo analizei.
2 Tekstyno samprata
Tekstynas – bet koks, kad ir pats mažiausias, elektroninę formą turinčių tekstų rinkinys.
Taip pat žymimi Lemuoklio išskaidyti, nors prasmės atžvilgiu neskaidomi junginiai tokie kaip be abejo, iš tikrųjų, kas nors, tam tikras, taip pat ir panašūs.