

LIETUVIŠKIEJI INTERNETO PUSLAPIAI TEKSTINĖS ANALIZĖS ASPEKTU

Agnė Vinickaitė

*Vytauto Didžiojo universitetas, Humanitarinis fakultetas
Daukanto 28, LT-3000 Kaunas, Lietuva*

Šiame straipsnyje kalbama apie internetinio tekstyno kūrimą bei kilusias problemas renkant ir siunčiant tekstus iš interneto. Taip pat pateikiamas lietuviškųjų interneto svetainių skirstymas į tipus pagal tekstų pobūdį, į kokias žmonių grupes jos orientuotos.

1 Įvadas

Populiarėjant informacinėms technologijoms gana svarbią vietą visuomenėje užėmė internetas, kaip informacijos šaltinis ir bendravimo priemonė. Nekontroliuojamai plaukiančios informacijos gausa bei įvairūs jos pobūdys atskleidžia gana plačią, beveik neaprepiamą interneto tematiką. Pamažu vis labiau susidomima lietuviškaisiais interneto puslapiais. Atsiranda įvairiausių tyrinėjimų ir darbų šia tema.

Vytauto Didžiojo universitete Kompiuterinės lingvistikos centre parengus bei dar toliau pildant bendrojo pobūdžio Dabartinės lietuvių kalbos tekstyną (šiuo metu apimantį 60 milijonų žodžių) taip pat paruošus kelis specialaus pobūdžio tekstynus, iškilio būtinybė parengti internetinį tekstyną, kuris būtų pagrindinis šaltinis interneto tekstų tyrinėjimui. Šiuo metu kaupiama internetinių tekstų duomenų bazė pasiekė apytiksliai 18,6 milijono žodžių skaičių ir dar toliau pildoma. Prognozuojama, jog internetinis tekstynas turėtų apimti apie 20 milijonų žodžių. Kadangi jis yra specialaus pobūdžio, tai tokia apimtis yra pakankama. Tekstynas turėtų būti pakankamai reprezentatyvus, kad “tekstynė lyg veidrodyje atspindėtų kalbą” [36]. Tai yra viena iš tekstynų kūrimo sąlygų. Todėl tekstų atranka šiam tekstynui nėra oporūnistiinė, o paremta subalansuotu atrankos principu, [36] t. y. tekstai atrenkami taip, kad atspindėtų interneto tekstų tipiškumą.

Šis darbas atliekamas keliais etapais. Visų pirma, reikia susirasti nemažą kiekį internetinių svetainių adresų. Lietuvoje kol kas nėra parengto didelio adresų žinyno, kuriame būtų suregistruoti bent jau dauguma adresų, ir to, matyt, neįmanoma padaryti. Galima būtų paminėti Lietuviškojo interneto katalogą [1] bei interneto svetainę www.online.lt, kur pateikiama nemažai adresų, suskirstytų pagal veiklos sritis.

Kitas darbo etapas – iš interneto svetainių išrinkti tekstus, kurie būtų bent jau pusės puslapio dydžio. Tekstai imami kuo įvairiausių sričių, kad kuo labiau atspindėtų konkrečios srities leksiką bei sintaksinę sandarą, internetinių tekstų pobūdį bei žanrinę įvairovę. Į šį etapą įeina tų tekstų parsiuntimas, neretai sukeltantis įvairių problemų bei parsiųstų failų tvarkymas. Šį darbą atlikus, tekstai jungiami skirstant juos pagal veiklos sritis ir pobūdį.

Peržvelgus daugelį lietuviškojo interneto puslapių ir žiūrint į juos kaip į visumą, galima daryti nemažai apibendrinimų ir iškelti klausimų: kokią informaciją teikia adresas, kokio tipo vyrauja svetainės, kiek informatyvūs tekstai, kodėl apskritai domimasi tektais ir ar jie internete tokie svarbūs.

2 Internetinių a dresų informatyvumas

Vienas pagrindinių uždavinių – kaip iš gausybės adresų išrinkti, kurios svetainės galėtų būti tinkamiausios kuriant tekstyną. Juk adresą dažniausiai sudaro firmos, įstaigos išsamus pavadinimas, pvz.: <http://www.ambergallery.lt> – Gintaro muziejus-galerija; <http://muziejai.mch.mii.lt> – Lietuvos muziejai arba tik žodžių pirmųjų raidžių santrumpos, pvz.: <http://www.lnm.lt> – Lietuvos nacionalinis muziejus. Todėl, jei dar iš išsamiai nurodyto pavadinimo galima pasakyti kieno tie puslapiai, tai iš santrumpų gana sunku spręsti apie ką ta svetainė.

Nemažai adresų netgi nėra susiję su puslapių kūrėju ar įstaiga, pvz.: <http://www.elnet.lt/inema/> – periodinis leidinys “Iš pirmo žvilgsnio”; <http://www.iti.lt/~piketas/landyne/> – elektroninis “Šluotos” variantas; <http://freehosting2.at.webjump.com/> – Vidmanto Karoso puslapis; <http://mp3.ku.lt/bilas/> – anekdotai apie Bilą Klintoną; <http://www.muza.lt/index.html> – Kultūros ministerija; <http://www.is.lt/karma/> – Antandrijos veislynas.

Tai apsunkina vartotoją, nes adresai nėra informatyvūs ir daug sunkiau nežinant adreso surasti reikiama informacija. Lietuviškojo interneto kataloge šiek tiek padeda skirstymas pagal veiklos sritis.

3 Internetinių svetainių tipai

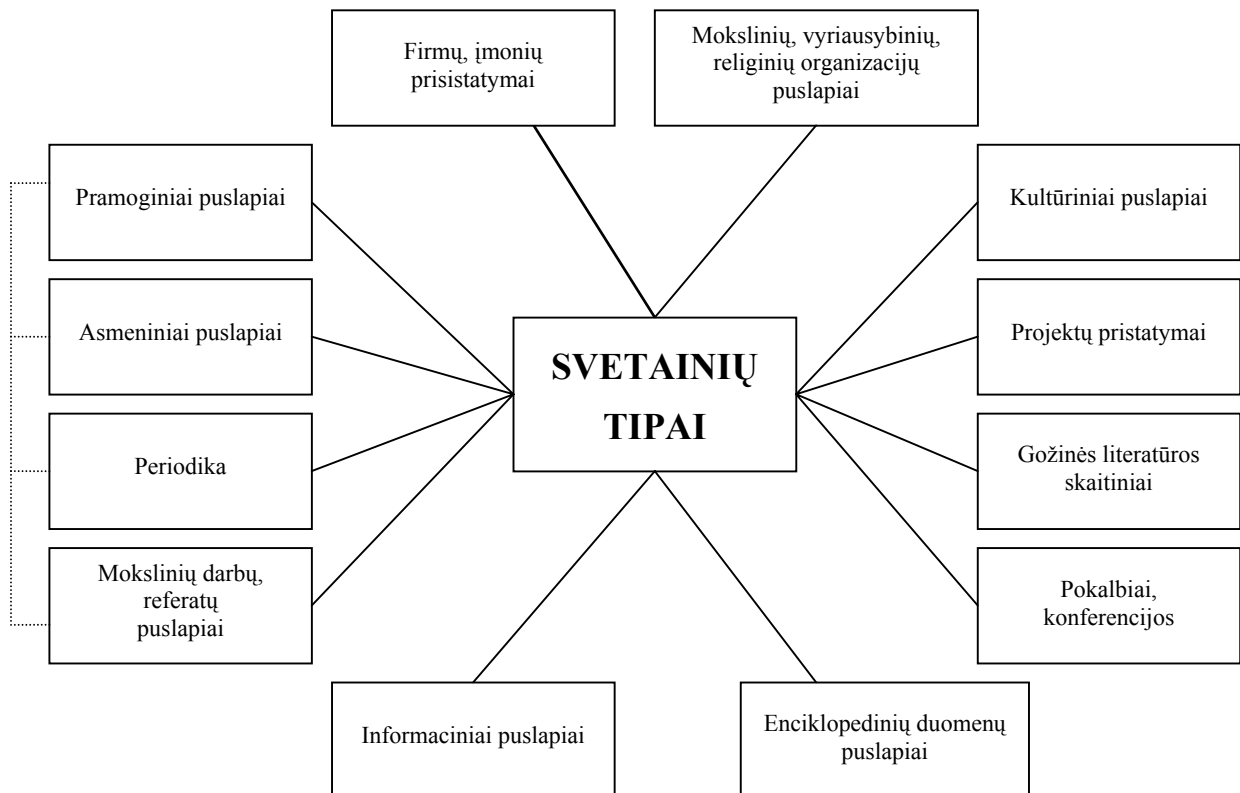
3.1 Svetainių tipai pagal informacijos pobūdį

Kadangi objektas yra tekstai, pagal tai internetines svetaines galima skirstyti į du tipus: tekstines ir netekstines. Tekstinių svetainių pagrindas – informacijos pateikimas tekstu: aprašomos naujienos, įvykiai, supažindinama su firmos veikla ar pan. Neretai iliustruojama nuotraukomis. Tokios svetainės dažniausiai pasižymi nedidele nuorodų gausa, aiškia struktūra, informacija pateikiama gana išsamiai.

Netekstinėse svetainėse paprastai teksto nedaug, dominuoja vaizdinė medžiaga: nuotraukos, piešiniai, žemėlapiai, diagramos, įvairios schemos bei lentelės. Jos pasižymi nuorodų gausa, taip pat pernelyg didele dizaino spalvų gama. Tokios svetainės patraukia interneto naršytojų dėmesį, bet jos labai painios, neturi aiškios struktūros ir, žinoma, tokią svetainę peržiūrėti užima gana daug laiko.

Teksto svarba internete gana nemaža. Kad ir kaip populiarėtų grafinis vaizdas, vis tik interneto puslapyje informacija dažniausiai perduodama tekstu, todėl didžioji dauguma svetainių yra tekstinio pobūdžio. Tekstas vis dar išlieka kaip pats patogiausias informacijos perdavimo būdas.

Peržiūrėjus daugybę svetainių išryškėja bendros jų tendencijos, panašumai ir skirtumai. Labai skiriasi svetainių pobūdis, todėl pagal pateiktą informaciją galima išskirti kelis svetainių tipus (žr.1 pav):



1 pav. Svetainių tipai.

1. Bene didžiausią dalį internete užima firmų, įmonių prisistatymai, pvz.: saldinių fabrikas “Rūta” [32], “Narbutas ir Ko”[26], prekybos namai “Piramidė”[21] ir kt.

Peržvelgus tokias svetaines matyti, kad jos neišsiskiria originalumu. Ryškėja bendra šių interneto svetainių struktūra. Puslapyje galima pastebėti maždaug septynias nuorodas, pasikartojančias beveik visose svetainėse: *apie mus*, *naujienos*, *veikla*, *produktai*, *paslaugos*, *kontaktai*, *nuorodos*. Pirmosios keturios nuorodos beveik be išimčių pasižymi tekstine medžiaga, kuri šiuo atveju labiausiai ir reikalinga.

- *Apie mus* skyrelyje pristatoma firma, papasakojama jos įkūrimo ir gyvavimo istorija;
- *Naujienose* aptariamos kai kurios firmos naujovės, įvykiai, nauji produktai;
- *Veikla* – veiklos apibūdinimas, teisinis veiklos pagrindas, kur pateikiami įstatymai, nutarimai ar jų nuorodos ir kt.

- *Produktai* – firmos gaminamos ar parduodamos prekės, neretai pateikiamos ištisos kolekcijos ar parodos su aprašymais nuotraukomis. Šis nuorodų skyrius dažniausiai būna bene išsamiausias ir reprezentatyviausias.

- *Paslaugos* – dažniausiai papunkčiui išvardijamos firmoje teikiamos paslaugos, klientų aptarnavimas.
- *Kontaktai* – nurodomos panašios veiklos kryptys ir firmos.
- *Nuorodų* skyrelyje pateikiamos kitų svetainių nuorodos panašia tema, dažniausiai užsienio svetainių.

2. Mokslinių įstaigų, institutų, draugijų, religinių bendruomenių, vyriausybės puslapių (pvz.: „Lietuvos Raudonasis kryžius“ [31], Vegetarų draugija [5], Etinių farmacijos kompanijų atstovybių asociacija [11], Vilniaus krašto totorių bendruomenė [13]) struktūra panaši į firmų puslapių, tik neretai pridedami įvairūs projektai, agitaciniai straipsniai, ilgi veiklos aprašymai, periodiniai leidiniai.

3. Kultūriniai puslapiai, pvz.: Po Kamanų rezervatą [15], Kuršių Nerijos nacionalinis parkas [27]. Šie puslapiai išsiskiria tuo, kad yra daugiausia pažintinio pobūdžio. Jie nustebino informacijos gausa, plačiais aprašymais, moksliniais tyrinėjamojo pobūdžio tekstais. Bene daugiausia puslapių gamtos, geografijos, zoologijos tema, nemažai parodų, istorinių ekspozicijų pristatymų. Gana daug informacijos apie įvairiuose miestuose vykstančius kultūrinius renginius.

4. Pramoginiai puslapiai, pavyzdžiui, anekdotai, laisvalaikio pasiskaitymai, hobi skirti laisvalaikiui ir yra dažniausiai humoristinio pobūdžio. Prie šio tipo puslapių galima priskirti taip pat dauguma asmeninių puslapių, kurie skiriami vien tik pramogai ar hobi aprašymui. Tekstų kalba pasižymi šnekamosios kalbos atspalviu, neretai iškraipomas lietuviškasis alfabetas. Nemažai šia tema yra periodinių leidinių, tokių, kaip elektroninis „Šluotos“ variantas [16], žurnalai „Šuns pasaulis“ [6], „Ore“ [17]. Atskiros svetainės anekdotų, horoskopų, virtuvės receptų, televizijos laidų.

5. Mokslinių darbų, referatų puslapiai dažniausiai skirti atskirų profesijų specialistams. Prie šio tipo svetainių galima priskirti ir asmenines, kuriose taip pat neretai publikuojami pačių parašyti moksliniai darbai. Galima būtų paminėti svetainę Mokslo centras [24], kur publikuojami pagal temines sritis suskirstyti įvairūs referatai ir moksliniai darbai.

6. Enciklopediniai duomenys svetainėse taip pat nereti. Pateikiami įvairūs žinynai, pvz.: ginklu, astronomijos, fizikos, vaistų, Lietuvos bei kitų šalių ir miestų enciklopediniai duomenys.

7. Projektų pristatymai, pvz.: projektas „Aš galiu“, „Sniego gniūžtė“, bankininkystės studentų projektai, nusikalstamumo prevencija Lietuvoje.

8. Informaciniai puslapiai (kas kur kada), pvz.: Lithuania Online, Lietuvos – Vokietijos informacinė sistema [10], Lietuvos nacionalinis informacijos centras „Eurika“ [12], kino filmai, Lietuvos avialinijos [18] ir dar daug kitų. Šie puslapiai dažniausiai skirti informacijos paieškai.

9. Periodiką internete galima skirstyti į pramoginę (žurnalai „Moteris“ [25], „Panelė“ [29], „Bombonešis“ [8]), bendrąją („Kauno diena“ [3], „Lietuvos rytas“ [22], „Verslo žinios“ [34]) ir specialiąją („Mokslo Lietuva“ [2], „Geras skonis“ [23], „Bažnyčios žinios“ [19], „Keturi ratai“ [30], „Septynios meno dienos“ [9]). Dauguma šių žurnalų išleidžiami ne tik elektronine forma. Jų elektroninis variantas šiek tiek skiriasi nuo popierinio varianto.

10. Grožinės literatūros skaitiniai, pvz.: Klasikinė lietuvių literatūra-antologija [1]. Čia įeitų ne tik literatūros kūriniai, bet ir pačių skaitytojų pasakojimai bei kūryba, pvz.: Polar page [20]– poliarinės kelionės, Vilniaus autostopo klubas [7], puslapis apie NSO [28].

11. Asmeniniai puslapiai (hobi, savęs pristatymai). Tai individualūs puslapiai, neretai ir garsių žmonių, kurių tikslas papasakoti apie save, pristatyti savo pomėgius, hobi, kūrybą ir noras užmegzti kontaktą su panašių pomėgių žmonėmis. Šiuose puslapiuose dominuoja nuotraukos, paveikslėliai, teksto gana nedaug, dominuoja šnekamosios kalbos frazės.

12. Pokalbiai, konferencijos, seimo posėdžiai, pvz.: interviu su įvairiais žmonėmis, elektroninių konferencijų svetainės. Šioms svetainėms būdingas dialogiškumas [37], klausimo-atsakymo forma.

3.2 Svetainių tipai pagal adresatą

Beveik kiekviena interneto svetainė orientuota į tam tikrą žmonių grupę, besiskiriančią savo amžiumi, išsilavinimu, su skirtingais tikslais, turinčią savų pomėgių, interesų. Internetas yra toks platus informacijos šaltinis, kad jame smalsumą gali patenkinti įvairiausių poreikių turintys individai. Taigi pagal tai, į kokią žmonių grupę orientuota svetainė, puslapius galima skirstyti į kelias grupes:

1. Puslapiai smalsiam interneto naršytojui, mėgstančiam sensacijas (pvz.: raganų klubas, skaičių magija, burtai);

2. Eiliniam interneto vartotojui (firmų reklaminiai puslapiai, naujienos);
3. Specialaus pobūdžio puslapiai specialiam žmonių ratui (įvairių profesijų, religijų, organizacijų, bendruomenių atstovams);
4. Specialaus pobūdžio puslapiai plačiajai visuomenei (įvairių ligų gydymas, grožinė literatūra, kultūriniai puslapiai);
5. Puslapiai vaikams.

Apskritai internetinis tekstas skiriasi nuo knyginio teksto. Dėl elektroninės formos ir nuorodų naudojimo tekstui būdingas fragmentiškumas, informacijos paskirstymas dalimis. Suskaidytą į daug dalių tekstą žymiai lengviau skaityti, matoma aiški struktūra. Nors, kita vertus, daugybė nuorodų klaidina. Viena pagrindinių internetinio teksto funkcijų – sudominti ir patraukti vartotoją savo informacijos turiniu. Todėl tekstas kai kuriais aspektais primena reklaminį, ir neretai taip yra dėl komercinių paskatų. Dėl tokio pobūdžio, tekstams charakterizuoti nebetinka įprastiniai funkciniai stiliai ir jų žanrai, kadangi internetinis tekstas balansuoja tarp rašytinės ir sakytinės kalbos. Vienas iš tokių bruožų – grįžtamasis ryšys, arba dialogiškumas, kada svetainėje paliekamas atskiras skyrelis nuomonėms ir komentarams.

Taigi toks internetinių puslapių skirstymas atsižvelgiant į puslapiuose pateikiamą tekstinę informaciją, leidžia geriau suvokti bendrą svetainių vaizdą ir, žinoma, yra svarbus internetinio teksto sudarymui. Tokiu ar panašiu skirstymu bus paremtas internetinių tekstų grupavimas bei jų jungimas. Tačiau prieš tai minėtinas dar vienas internetinio teksto kūrimo etapas – tekstų parsisiuntimas bei jų tvarkymas.

4 Tekstų siuntimas ir jų tvarkymo problematika

Kadangi tekstų analizei bei internetinio teksto kūrimui reikalinga tekstų visuma, tai tekstai turi būti parsiųsti ir sugrupuoti pasirenkant tam tikrus kriterijus. Tai daroma pasitelkiant Vido Daudaravičiaus (Kompiuterinės lingvistikos centras) parengtą programą “Kolektorių”, kuri skirta tekstų parsisiuntimui, jų apdorojimui ir kt. Be šios programos dar dirbama su “Webzip”, kuri papildomai naudojama tik tekstų parsisiuntimui.

Pagrindinė problema, su kuria susiduriama parsisiuntus tekstus – tai didelė disproporcija tarp interneto puslapyje matomų tekstų ir parsiųstų failų. Paprastai atsiunčiama daugiau tekstų nei matoma svetainėje. Taip atsitinka dėl to, kad svetainėse yra beveik nepastebimų nuorodų, kurios dažnai lokalizuotos puslapio pakraštyje ir išryškėja tik bakstelėjus pele. Tai apsunkina darbą kuriant tekstyną, nes be jau peržiūrėtų tekstų tenka papildomai tikrinti ar parsiųstieji tekstai atitinka nusistatytus kriterijus, t. y. ar parsiųstame faile pakankamai teksto ir kokio jis pobūdžio.

Parsiųsti failai su tekstais yra peržiūrimi, nustatomas teksto tipas ir žanras. Tekstai nėra išskaidomi ir sumaišomi. Jie lieka konkrečios interneto svetainės kontekste ir tuo būdu atspindi svetainės tipą ir pobūdį.

5 Išvados

Aptarus internetinio teksto kūrimo etapus ir apibūdinus teksto kūrimo problematiką galima teigti, jog gana nemažai problemų kelia internetinių svetainių adresai, kurie yra neinformatyvūs ir apsunkina vartotoją, taip pat tekstų siuntimas.

Atlikta gana subjektyvi interneto svetainių klasifikacija teksto aspektu leidžia suvokti bendrą lietuviškojo interneto svetainių tekstų vaizdą, tematiką, išskirti kai kuriuos tekstų žanrus ir svetainių pobūdį bei tipus. Ši klasifikacija taip pat atspindi internetinio teksto tekstų visumą bei žanrinę įvairovę.

Kadangi iškilo būtinybė tyrinėti internetą, tai internetinis tekstynas turėtų pasitarnauti kaip pagalbiniė priemonė interneto analizei.

Literatūros sąrašas

- [1] <http://anthology.lms.lt/index.html>
- [2] <http://ic.lms.lt/ml.html>
- [3] <http://kaunodiena.lt>
- [4] <http://receptai.w3.lt/>
- [5] <http://sveikata.elnet.lt/vegetarai/>
- [6] <http://www.5ci.net/dogs/>
- [7] <http://www.autostop.lt/pirmas.html>
- [8] <http://www.coolzone.lt/music/bomboneshis.htm>

- [9] <http://www.culture.lt/7menodienos>
- [10] <http://www.deutschland.lt/lt/default.htm>
- [11] <http://www.efa.lt/>
- [12] <http://www.eureka-cost.lt/>
- [13] <http://www.gaumina.lt/totoriai/>
- [14] <http://www.humoras.lt/>
- [15] <http://www.is.lt/venta/mokiniu/kamanai/>
- [16] <http://www.iti.lt/~piketas/landyne/>
- [17] <http://www.laisvalaikis.lt/ore.htm>
- [18] <http://www.lal.lt/>
- [19] <http://www.lcn.lt/bzinios/>
- [20] <http://www.lgt.lt/polar/>
- [21] <http://www.lithill.lt/piramide/>
- [22] <http://www.lrytas.lt>
- [23] <http://www.meniu.lt/>
- [24] <http://www.mokslo.centras.lt/main.php3>
- [25] <http://www.moteris.lt/>
- [26] http://www.narbutas.lt/biuro_baldai.phtml
- [27] <http://www.nerija.lt/>
- [28] <http://www.nso.lt/index.html>
- [29] <http://www.panele.lt/>
- [30] <http://www.ratai.lt/4ratai/>
- [31] <http://www.redcross.lt/veikla.html>
- [32] <http://www.ruta.lt/>
- [33] <http://www.teatras.lt/main.htm>
- [34] <http://www.vz.lt/>
- [35] Lietuviškojo interneto katalogas 2000. *UAB Penki kontinentai*. 2001. NR. 1.
- [36] **Marcinkevičienė R.** Tekstynų lingvistika (Teorija ir praktika) // Darbai ir dienos. *VDU leidykla*. 2001. Nr.24.P. 7 – 64.
- [37] **Ryklienė A.** Bendravimas internetu: Kalbėjimas rašant // Darbai ir dienos. *VDU leidykla*. 2001. Nr.24.P. 99 – 107.

Summary

The Lithuanian Internet Pages in Aspect of Textual Analysis

Nowadays the Internet is an important source of information and a means of communication. More people are interested in Lithuanian Internet pages. There are various studies and works in this subject. At Vytautas Magnus university in the Computer Linguistics Centre the corpus of present day Lithuanian is prepared and is further filling and this centre wants to prepare Internet corpus too which would be the main source of internet text studies.

This work consist of several stages. First of all it is necessary to find a big amount of internet pages addresses. These addresses are not informative. The next stage of internet corpus preparation is to pick out text of internet pages. The texts are taken from various spheres in order to reflect the vocabulary of concrete sphere, the character of internet texts and variety of genres. Because the object is texts, so according to this, internet pages are divided in several types.

After selection and sending of text from the Internet the main problem is that there is a big disproportion between texts, references which you can see in the internet page and the texts which are already sent.