Esperantologio / Esperanto Studies **6** (2013), 45–57

Malpliigi pretervidojn kaj misavertojn de esperanta literumilo per eraromodelo

Edmund Grimley Evans

Enhavo

Resumo. Estas priskribita metodo por objektive kaj kvante pritaksi literumilon. Pluraj esperantaj literumiloj estas pritaksataj per la sama procedo, kaj la rezultoj estas komparataj ankaŭ kun literumiloj por tri aliaj lingvoj. Montriĝas ke la nunaj literumiloj por esperanto ne taŭgas por kontroli longan tekston kun malmultaj eraroj, ĉar ili donas tro da misavertoj: inter la veraj eraroj ili listigas tro da ĝustaj, sed nerekonitaj vortoj. La literumiloj por kelkaj aliaj lingvoj funkcias multe pli bone en tiu rilato. Per akceptado de arbitraj kunmetaĵoj oni povas draste malaltigi la frekvencon de misavertoj, sed koste de pli alta frekvenco de pretervidoj (netrovitaj eraroj). Pluraj prototipoj por nova literumilo estas prezentitaj kaj pritaksataj, kaj nova aliro estas priskribita, la aplikado de eraromodelo, kiu kondukis al prototipo kiu funkcias signife pli bone ol aliaj literumiloj en la pritaksado. Per la nova aliro eblas draste malaltigi la frekvencon de misavertoj kaj samtempe malaltigi la frekvencon de pretervidoj kompare kun literumilo vaste uzata hodiaŭ. Fine estas proponitaj pluraj direktoj en kiuj oni povus plibonigi la prototipon.

Abstract. *Reducing the frequency of false negatives and false positives from an Esperanto spelling checker by using an error model.*
A method is described for objectively and quantitatively evaluating a spelling checker. Several Esperanto spelling checkers are evaluated using the same procedure and the results are also compared with spelling checkers for three other languages. It is shown that current spelling checkers for Esperanto are not suitable for checking a long text with only a few errors because they give too many false positives: among the real errors they list too many correct but unrecognised words. The spelling checkers for some other languages

work much better in this respect. By accepting arbitrary compounds it is possible to drastically reduce the frequency of false positives, but at the cost of a higher frequency of false negatives (undetected errors). Several prototypes for a new spelling checker are presented and evaluated and a new approach is described, the use of an error model, which leads to a prototype that works significantly better than other spelling checkers under evaluation. With the new approach it is possible to drastically reduce the frequency of false positives and at the same time reduce the frequency of false negatives compared with a spelling checker widely used today. Finally several directions are suggested for how the prototype might be improved.

1. Enkonduko

Laŭ *La Nova Plena Ilustrita Vortaro* (2002) literumilo estas "Programo aŭ parto de programo por kontroli ortografion per vortlisto de la koncerna lingvo". Ĝi povas do esti aparta programo, kiel la uniksa komando `ispell`, aŭ parto de redaktilo, retpoŝtilo aŭ alia programo en kiu oni entajpas tekston. Ekzemple, kaj la oficeja programaro OpenOffice.org kaj la TTT-legilo Firefox enhavas literumilon por multaj lingvoj, inkluzive de esperanto. Literumilo atentigas la uzanton pri vortoj kiuj ŝajnas eraraj, ekzemple substrekante ilin, kaj ĝi povas ankaŭ proponi korekton. Tipa literumilo proponas ankaŭ la eblon aldoni nerekonitan vorton al privata vortaro, por ke poste ne plu estu avertoj pri ĝi.

Laŭ la citita difino literumilo funkcias "per vortlisto de la koncerna lingvo". Tiu parto de la difino servas por plipreciziĝi aŭ pliklarigi la manieron en kiu tia programo tipe funkcias. Certajn erarojn, ekzemple neeblajn sinsekvojn de literoj, aŭ literon el malĝusta alfabeto, eblus rekoni sen vortlisto, sed literumilo sen vortlisto ne estus tre kapabla. Aliajn erarojn ne eblas rekoni sen analizo de la kunteksto, sed programo kun tia kapablo oni pli verŝajne nomus gramatikilo ol literumilo.

Vortlisto necesas, sed por tipa lingvo ĝi ne sufiĉas. Multaj lingvoj havas gramatikajn regulojn kiuj ebligas nefinian nombron da eblaj vortoj. Eĉ se en iu lingvo praktike eblas, por la celo de literumilo, trakti la aron de eblaj vortoj kiel finian, se la lingvo havas iujn often aplikatajn derivregulojn, ekzemple por formi pluralon, oni verŝajne volus ke literumilo apliku ilin, por ke, se nova vorto estas aldonita al la privata vortaro en ĝia baza formo, estu rekonataj ankaŭ formoj derivitaj el ĝi.

Literumilo plej ofte estas uzata de aŭtoro por kontroli la tekston kiun li mem entajpis, aŭ poste de redaktoro, sed kelkfoje oni aplikas literumilon al la eligaĵo de tekstlegilo (optika signorekono). Tial oni povas distingi tri kategoriojn de eraroj kiujn oni esperas trovi per literumilo:

1. tajperaroj, ekzemple *repsondis* anstataŭ *respondis*;
2. lingvaj eraroj, ekzemple *rebelo* anstataŭ *ribelo*;
3. legeraroj, ekzemple *etema* anstataŭ *eterna*.

Decidante, ĉu vorto estas erara, literumilo povas mem erari en du manieroj:

1. **misaverto**: la literumilo avertas pri vorto kiu ne estas erara;
2. **pretervido**: la literumilo ne avertas pri eraro en la teksto.

Ĉe statistikaj kaj medicinaj testoj oni aplikas la terminojn *falsa pozitivo* kaj *falsa negativo*. Same kiel en tiuj kampoj oni celas minimumigi la frekvencojn de ambaŭ specoj de eraro, sed ĝenerale oni devas kompromisi. Ĉiam eblas malplioftigi unu specon de eraro, se oni akceptas plioftigi la alian specon.

El la vidpunkto de uzanto la du specoj de eraro malsimilas. Ĉe misaverto la uzanto ricevas averton kiun li devas ignori, aŭ al kiu li devas iel reagi, depende de la funkcireĝimo de la literumilo. Ĉe pretervido la uzanto ricevas nenian averton, kaj povas esti ke la eraro restos nekorektita kaj li neniam ekscios pri ĝi. Do la uzanto tuj observas misavertojn, sed ne tuj observas pretervidojn. Tamen, pli videbla por la uzanto ol la ofteco de misavertoj estas la proporcio de misavertoj el ĉiuj avertoj: kia proporcio el avertoj estas misa. Se granda plimulto el la avertoj donataj de la literumilo estas misaj, tiam la uzanto baldaŭ ne plu tre atentos la avertojn. Evidente la proporcio de misavertoj dependas de la nombro de eraroj en la teksto; en senerara teksto, ĉiuj avertoj estos misaj. Tial oni akceptas pli altan frekvencon de misavertoj, kiam oni kontrolas tekston en kiu estas multaj eraroj.

2. Metodo por pritaksi literumilon

En la ĝisnuna literaturo pri esperantaj literumiloj mankas priskriboj de objektiva pritaksado de literumiloj, krom kelkaj tre simplaj eksperimentoj, en kiuj oni mezuris kia proporcio de la vortoj en unu tekstaro estis rekonita.

Por mezuri la frekvencon de misavertoj estis elektitaj por ĉi tiu studo dek unu tekstoj en esperanto. Ili estis de diversaj aŭtoroj, de diversaj ĝenroj, kaj de diversaj fontoj, escepte ke estis du el Monato (1980–). La plej mallonga teksto havis ĉirkaŭ 500 vortojn. Ĉar la celo estas mezuri misavertojn, la tekstoj devus esti seneraraj. Tamen, kvankam ĉiuj tekstoj estis jam bone redaktitaj, dum la eksperimentado malkaŝiĝis kelkaj malmultaj eraroj. Ĉar en ĉiu tia okazo la intenco estis klara, tiuj eraroj estis korektitaj antaŭ ol kalkuli la finajn rezultojn. Apartaj listoj estis faritaj de la neesperantaj vortoj, inkluzive proprajn nomojn kiuj aperis en la tekstoj, por ke tiuj estu

ignorataj en la kalkuloj. (La pritraktado en literumilo de propraj nomoj kaj alilingvaj citaĵoj estas interesa temo, sed ne traktita en ĉi tiu studo.)

En la rezultoj estas prezentitaj la 1-a, 2-a kaj 3-a kvariloj de la elcento de nerekonitaj vortoj en la 11 tekstoj. Tiel estas taksitaj ne nur la mezuma frekvenco de nerekonitaj vortoj, sed ankaŭ ia mezuro kiel varias tiu frekvenco laŭ la aŭtoro kaj ĝenro. (La n-a kvarilo de aro da nombroj estas valoro tia ke n kvaronoj de la nombroj estas malpli granda ol tiu valoro. La 2-a kvarilo egalas al la duilo, nomata ankaŭ *mediano*, speco de mezumo.)

Por mezuri la frekvencon de pretervidoj estis kolektita listo de 113 veraj eraroj: 49 devenis de tekstoj kiujn la aŭtoro de ĉi tiu artikolo reviziis por Monato; 33 devenis de aliaj tekstoj kiujn li reviziis por kolegoj; kaj 31 devenis de tekstoj transskribitaj per tekstlegilo, do konsistas el legeraroj.

Principe estus interese dividi la erarojn en diversajn kategoriojn kaj pritrakti ilin aparte, sed 113 ne estas sufiĉe granda specimeno por elteni disdividon, kaj praktike ne estis facile kolekti eĉ nur tiom da vere okazintaj eraroj, do en la rezultoj estas prezentitaj nur la elcento de vortoj en la listo de 113 eraroj kiuj ne estis rekonitaj kiel eraraj.

Estas ĝenerala principo en tiaj studoj ke oni uzu apartajn tekstarojn por evoluigo kaj pritakso: oni ne pritaksu programon per la sama materialo per kiu oni evoluigis aŭ trejnis la programon. En iuj rilatoj tio estas tre evidenta. Ekzemple, se oni aldonus la mankantajn vortojn al la vortaro uzata de literumilo, poste la literumilo funkcius preskaŭ perfekte, kontrolante tiun saman tekston, sed evidente tiu rezulto ne plu estus bona mezuro kiel la literumilo funkcius ĉe novaj tekstoj. Sed la principo validas ankaŭ en pli subtilaj situacioj. Ekzemple, se oni modifus literumilon por pli bone pritrakti certajn specojn de kunmetaĵoj kiujn oni renkontis dum provado ĉe certaj tekstoj, tiam oni ne plu povus uzi tiujn samajn tekstojn por pritraksi la literumilon: en novaj tekstoj eble aperos pli ofte ne tiuj, sed aliaj specoj de kunmetaĵoj.

3. Pritaksado de jamaj literumiloj

3.1. Esperantaj literumiloj

Kvar esperantaj literumiloj estis pritaksitaj en la maniero priskribita en la antaŭa sekcio. Ili estis:

1. ispell: versio 3.1.20.0-4.5ubuntu1 kun la vortaro iesperanto (versio 2.1.2000.02.25-40);
2. aspell: versio 0.60.6-2 kun la vortaro aspell-eo (versio 2.1.2000.02.25-40);
3. hunspell (malnova): versio 1.2.8-4ubuntu2 kun la vortaro myspell-eo (versio 2.1.2000.02.25-40);

4. hunspell (nova): versio 1.2.8-4ubuntu2 kun nova vortaro de Marek Blahuš (versio 1.0 de 2009 de aldonaĵo por OpenOffice.org).

Escepte la vortaron por 4, ĉiuj literumiloj kaj vortaroj estis instalitaj kun Ubuntu 9.10 (mastruma sistemo bazita sur Linux/Linukso). La vortaroj por 1–3 devenas de la sama vortlisto kreita por ispell de Sergio Pokrovskij en 1998. Ili devenas de la sama fontopako, eo-spell, kaj tial ili havas la saman versinumeron. La vortaro por 4 estis prenita de
`http://extensions.services.openoffice.org/en/project/literumilo`

La rezultoj estis:

Literumilo	Misavertoj (%)			Pretervidoj
	1-a kvarilo	2-a kvarilo	3-a kvarilo	(%)
ispell	2,57	3,89	4,45	15,9
aspell	3,57	5,07	5,73	10,6
hunspell (malnova)	3,54	5,07	6,52	10,6
hunspell (nova)	1,44	1,80	2,54	41,6

Ĉar ili uzas baze la saman vortaron, ne estas surprize ke ispell, aspell kaj la malnova hunspell donas tre similajn rezultojn. Ŝajne la vortaro de Pokrovskij plej bone funkcias kun ispell, la literumilo por kiu ĝi estis destinita. Ekzamenado de la listoj de nerekonitaj vortoj sugestas ke aspell ne akceptas certajn kunmetaĵojn kiuj funkcias ĉe ispell, eble pro misadapto de la vortaro al la nova literumilo. La etaj diferencoj inter aspell kaj la malnova hunspell rilatas al la traktado de apostrofoj kaj dividstrekoj: ĉe la malnova hunspell vorto kun dividstreko estas traktata kiel du apartaj vortoj.

Rilate misavertojn la nova hunspell montras gravan pliboniĝon: ili duoniĝis kompare kun ispell. Tamen, rilate pretervidojn estas malpliboniĝo: la nova hunspell trovis nur ĉirkaŭ 70 % el la eraroj kiujn trovis ispell.

3.2. Literumiloj por aliaj lingvoj

Por havi ian komparon inter esperantaj kaj alilingvaj literumiloj estis prenitaj po tri tekstoj en la lingvoj angla, franca kaj germana: unu teknika, unu beletra, kaj unu ĵurnala pri politiko. Ĉiu estis pritraktita per literumilo, ispell aŭ hunspell, kun vortaro de la koncerna lingvo, kaj la listoj de trovitaj eraroj estis filtritaj por forigi fremdaĵojn kaj proprajn nomojn. La rezultoj estis:

Lingvo	Misavertoj (%)		
	Teknika	Beletra	Ĵurnala
angla	0,1	0,1	0,2–0,8
franca	0,1	0,1	0,1
germana	3–6	1,4	1,1–1,7

La gamoj indikas malcertecon ĉu specifaj vortoj estu kalkulitaj kiel fremdaĵoj aŭ propraj nomoj.

Estas tre malgranda specimeno, la ciferoj estas malprecizaj, kaj komparo inter la lingvoj estas dubinda pro la malsamaj stiloj kaj temoj de la tekstoj, sed la sekvaj konkludoj ŝajnas kredindaj:

1. La literumiloj en la angla kaj franca donas multe malpli da misavertoj ol la plej bona literumilo en esperanto: ĉirkaŭ dekono.

2. La literumilo en la germana donas multe pli da misavertoj ol la literumiloj en la angla kaj franca, sed verŝajne malpli ol la esperantaj literumiloj. (La teknika artikolo en la germana, plena je derivaĵoj kaj kunmetaĵoj kiel *Parallelisierbarkeit* ‘paraleligeblo’ kaj *Arbeitsalphabetgröße* ‘laboralfabetgrando’, ŝajnas escepta.)

3.3. Diskuto

Kio estas akcepteblaj frekvencoj de misavertoj kaj pretervidoj ĉe literumilo?

Kiel jam notite, tio dependas de la nombro de eraroj en la kontrolata teksto, kaj cetere la frekvencoj mem dependos de la stilo kaj la temo. Tamen, kiel konkretan ekzemplon ni konsideru la kontroladon de 50-mil-vorta teksto, en kiu estas dek eraroj, kaj ni supozu ke la frekvencoj de misavertoj kaj pretervidoj estas 1,8 % kaj 42 %, kiel mezurite ĉe la nova hunspell. Tiam la literumilo proponus liston de 900 eraroj, inter kiuj aperus 6 el la 10 veraj eraroj. Trastudi tiun liston estus pena laboro. (Fakte la situacio povus esti pli favora ĉe la kontrolado de tiel longa teksto, pro la ripetado de nerekonitaj vortoj.)

4. Prototipoj de nova literumilo

Estus bone havi por esperanto literumilon kiu donas malaltan frekvencon de misavertoj kiel la nacilingvaj literumiloj pritaksitaj supre. Ĉu eblus tion atingi per literumilo kiu funkcias per grandega vortaro, ne analizante la strukturon de esperantaj vortoj (derivaĵoj kaj kunmetaĵoj)? Esplori tiun demandon estas celo de la unua prototipo, priskribita sube, kaj la eksperimentaj rezultoj sugestas, ne surprize, la respondon *ne*.

Sekva demando estas: kiel, en literumilo, oni analizu la strukturon de esperantaj vortoj? Ĉu oni akceptu arbitrajn kunmetaĵojn el esperantaj vorteroj (radikoj kaj aliaj elementoj)? Danĝero de tia aliro estas ilustrata de la tri ekzemploj jam uzitaj en la enkonduko: *repsondis* anstataŭ *respondis*, *rebelo* anstataŭ *ribelo*, kaj *etema* anstataŭ *eterna*. La ekzemploj estis elpensitaj, sed per retserĉilo eblas konfirmi ke tiuj eraroj ankaŭ vere okazis. Notu, kiel ĉiu el tiuj vortoj estus analizebla kiel kunmetaĵo: *rep-sond-is, re-bel-o, et-em-a*. Povas esti, do, ke akceptado de arbitraj kunmetaĵoj solvus la problemon de troaj misavertoj koste de tre alta, eble tro alta, frekvenco de pretervidoj. Tiun demandon esploras la dua prototipo, sube, kaj la rezultoj montras ke la frekvenco de pretervidoj ĉe tia literumilo ja estas tre alta.

Evidenta reago al tiu problemo estas iel limigi la kunmetadon de vorteroj, ekzemple per sintaksa aŭ semantika klasado de la vorteroj kaj pli precizaj reguloj por ilia kunmetado. Tian aliron sekvis Blahuš (2008), kun la aldonaĵo ke la klasoj de radikoj estis aŭtomate derivitaj el tekstaro. En ĉi tiu artikolo estas proponita alternativa aliro, la aplikado de eraromodelo, kaj la valoro de tiu propono estas pritaksita de la tria prototipo, priskribita sube.

Ĉiu el la tri prototipoj uzas unu aŭ ambaŭ el jenaj du fontoj de leksikaj informoj:

1. La kapvortoj de NPIV (*La Nova Plena Ilustrita Vortaro*, 2002). Estas 47 162 kapvortoj, se oni kalkulas aparte la alternativojn montritajn interalie per komo aŭ parentezo en la presita originalo kaj apartigitajn per punktokomo en la komputila dosiero, sed uzataj estis nur la *radikoj*, tio estas la artikolaj kapvortoj kiuj havas gramatikan finaĵon */a*, */aj*, */e*, */i*, */o* aŭ */oj* kaj krom tio konsistas nur el literoj, plus la artikolaj kapvortoj kiuj konsistas nur el literoj kaj havas derivaĵon kun la formo $\sim a$, $\sim e$, $\sim i$ aŭ $\sim o$. Tiel estis eltiritaj 15 707 radikoj.

2. Ĉiuj tekstoj elŝuteblaj de tekstaro.com. Se oni difinas vorton kiel maksimuman nemalplenan sinsekvon el literoj, ciferoj kaj dividstrekoj, estis sume 4 353 793 vortoj, kaj 182 397 diversaj formoj, se ignori majuskl-econ. La tekstoj ĉe tekstaro.com venis de diversaj fontoj, el kiuj unu el la plej riĉaj, pro la amplekso kaj pro la diverseco de aŭtoroj kaj temoj, estas la magazino Monato. La tekstoj estas ĝenerale bone redaktitaj kaj reviziitaj, sed ne tute sen eraroj. Ĉiuj tekstoj ĉe tekstaro.com estis pli malnovaj ol la tekstoj elektitaj por pritaksi la literumilojn.

Estis uzataj nur tiuj du fontoj, ĉar la celo estas esplori pri diversaj algoritmoj por literumilo. Ĉiun literumilon eblus plibonigi per ŝanĝado de la vortaro, sed en studo pri algoritmoj uzado de diversaj vortaroj nur konfuzus la rezultojn.

4.1. Provo 1: granda vortaro sen derivreguloj

La unua prototipo uzas por siaj leksikaj informoj nur la tekstojn elŝuteblajn de tekstaro.com.

Utila statistiko pri teksto estas la proporcio de vortoj kiuj aperas nur unufoje en la teksto (kelkfoje oni nomas ilin *hapaksoj*). En la uzata tekstaro la proporcio de tiaj vortoj estis 2,1 %, se oni konsideras nur la vortojn konsistantajn nur el esperantaj literoj kaj dividstrekoj. Tio sugestas ke la frekvenco de pretervidoj estus proksimume tiom, se oni aplikus la tekstaron en literumilo rekte, sen gramatika scio, akceptante nur la vortojn kiuj aperis en la tekstaro, kaj se la kontrolataj tekstoj venus de la sama populacio kiel la tekstaro.

Pli kredinda aliro estas apliki scion almenaŭ de la gramatikaj finaĵoj: oni akceptas vorton kiu aperas en la tekstaro ankaŭ kun alia gramatika finaĵo, do oni akceptas ekzemple *homas*, se en la tekstaro troviĝas *homo*. Pli precize, antaŭ ol kompari vortojn, oni reduktas ilin per la sekva regulesprima substituo (en la programlingvo Perlo):

```
s!([aeiou].*)([ao]j?n?|en?|[aiou]s|[iu])$!$1/!;
```

Tiu esprimo rekonas 16 kombinaĵojn de finaĵoj, sed nur ĉe vorto, en kiu troviĝas ankaŭ alia vokalo.

Post tia redukto restas 92 148 diversaj formoj de la vortoj konsistantaj nur el esperantaj literoj kaj dividstrekoj, kaj la proporcio de “hapaksoj” el la vortoj estas pli esperiga: 1,1 %.

Estis konstruita literumilo kiu funkcias laŭ tiu aliro. Majusklecon kaj dividstrekojn ĝi simple ignoras, sed estis pli speciala pritraktado de du signoj kiuj estas praktike uzataj kelkfoje kiel apostrofo kaj kelkfoje kiel citilo. La literumilo estis pritaksita per la sama procedo kiel antaŭe. Jen la rezultoj, kun la ciferoj por la nova hunspell ripetitaj por komparo:

Literumilo	Misavertoj (%)			Pretervidoj
	1-a kvarilo	2-a kvarilo	3-a kvarilo	(%)
hunspell (nova)	1,44	1,80	2,54	41,6
prototipo 1	1,43	1,78	3,31	24,8

Estas rimarkinde ke la simpla kaj naiva aliro de prototipo 1 donas rezultojn similajn al la multe pli kompleksa aliro de la nova hunspell.

Eblus aldoni parametron al prototipo 1 por ke ĝi akceptu nur la vortojn kiuj aperas en la tekstaro almenaŭ n fojojn, sed la frekvenco de misavertoj kun $n = 1$ estas jam tro alta por multaj aplikoj, kiel jam diskutite.

La 28 pretervidoj de prototipo 1 estis ekzamenitaj kaj dividitaj en jenajn du kategoriojn:

15 vortoj kiuj sen kunteksto ŝajnas eraraj;
13 vortoj kiuj sen kunteksto ŝajnas ĝustaj.

Pretervido povus deveni de gramatika misanalizo, ekzemple la vorto *tamen* en la tekstaro akceptigus la vorton *tamas* en kontrolata teksto, sed vera ekzemplo de tiu fenomeno ne estis renkontita.

El tiu analizo eblas konkludi ke oni povus proksimume duonigi la frekvencon de pretervidoj, se oni forigus ĉiujn erarojn el la vortlisto eltirita de la tekstaro.

Por analizi la misavertojn estis prenita hazarda specimeno de 100 el la nerekonataj vortoj en la provotekstoj. Tiuj estis ekzamenitaj kaj dividitaj en jenajn kvar kategoriojn:

5 esperantaj nomoj;
7 maloftaj vortoj aŭ neologismoj;
23 vortoj formitaj el konataj vortoj per oficialaj afiksoj;
65 vortoj formitaj alimaniere per kunmetado de konataj vortoj.

Kvankam kelkaj el la 65 alimaniere formitaj kunmetaĵoj eble troviĝus en pli granda tekstaro, multaj el ili estis verŝajne tute novaj, neniam antaŭe uzitaj. Reale ni ne disponas multe pli grandan tekstaron en esperanto, speciale se ni postulas tekstaron kiu ne enhavas tro da eraroj. Eblas do konkludi ke apenaŭ eblus malpliigi la frekvencon de misavertoj en normalaj tekstoj sub 1 % per literumilo kiu funkcias sen vortfaraj reguloj.

4.2. Provo 2: libera kunmetado

La dua prototipo uzas ambaŭ fontojn de leksikaj informoj. Ĝi akceptas ĉiujn vortojn troveblajn en la tekstaro, eventuale kun alia gramatika finaĵo, kiel en prototipo 1, sed ankaŭ ĉiujn kunmetaĵojn el radikoj el NPIV. Kunmetaĵoj estu formitaj laŭ jena regula esprimo:

(prefikso* radiko sufikso* mezfinaĵo?)* prefikso* radiko sufikso* finaĵo

Prefikso: *bo-*, *ĉi-*, *de-*; la 15 tabelvortoj kun finaĵo *-a*, *-o* aŭ *-u* (*ĉiu-*, *nenia-* k.t.p.); 8 pronomoj kun *-a*-finaĵo (*mia-*, *via-* k.t.p.).

Radiko: la radikoj de NPIV, eltiritaj kiel priskribite supre, tamen sen la tabelvortaj radikoj (*ki-*, *ti-* k.t.p.), sen la pronomoj (*mi*, *vi* k.t.p.), kaj kun aldono de ĉiuj oficialaj afiksoj escepte *-ĉj-* kaj *-nj-*.

Sufikso: la 6 participaj sufiksoj.

Mezfinaĵo: *a*, *e*, *o*.

Finaĵo: 16 kombinaĵoj, kiel en prototipo 1.

Notu ke la difinoj de *prefikso*, *radiko* kaj *sufikso* uzataj ĉi tie ne tre kongruas kun la normalaj sencoj de tiuj vortoj. La plej multaj oficialaj afiksoj ne devas esti traktataj kiel prefikso aŭ sufikso, ĉar ili funkcias ankaŭ kiel radikoj: el la vidpunkto de ĉi tiu literumilo *eklabori* estas kunmetaĵo pravigita per *eki* kaj *labori*, same kiel *manlaboro* estas kunmetaĵo pravigita per *mano* kaj *labori*.

NPIV traktas la tabelvortojn kun finaĵo *-a* kaj *-o* kiel radikon plus finaĵon. Indas forigi tiujn radikojn el la radikaro por eviti troan akceptemon. Simile estas utile ne trakti la pronomojn kiel radikojn, kvankam ili troviĝas en NPIV kun adjektiva finaĵo. Pronomoj ja aperas en novaj kunmetaĵoj, sed preskaŭ nur kun adjektiva finaĵo (*miaflanke*, *vianome* k.t.p.).

Ne necesas aldoni *en* al la listo de "mezfinaĵoj", ĉar vorto kiel *hejmeniri* estas analizebla kiel kunmetaĵo el tri radikoj: *hejm-en-ir-i*; *en-* estas radiko pro la NPIV-a kapvorto *ene*. Tio estas misanalizo, kompreneble, sed la misanalizo donas ĝustan rezulton.

Rememore la antaŭan komenton pri la apartigo de tekstaroj por evoluigo kaj pritakso, estu notite ke la specifo de kunmetaĵo kaj la difinoj de prefikso, sufikso k.t.p. estis disvolvitaj dum pli fruaj eksperimentoj, antaŭ ol estis kolektitaj la provotekstoj uzataj en la nunaj pritaksoj.

Sama pritakso de la nova literumilo donis jenajn rezultojn, kun la pli fruaj rezultoj por komparo:

Literumilo	Misavertoj (%)			Pretervidoj
	1-a kvarilo	2-a kvarilo	3-a kvarilo	(%)
hunspell (nova)	1,44	1,80	2,54	41,6
prototipo 1	1,43	1,78	3,31	24,8
prototipo 2	0,00	0,10	0,27	54,0

Kompare kun la aliaj esperantaj literumiloj prototipo 2 donas tre bonan frekvencon de misavertoj; ĝi similas al la frekvenco mezurata ĉe literumiloj en la angla kaj la franca. Tamen, la frekvenco de pretervidoj multe kreskis. La kialo estas facile divenebla: eraroj kiel *repsondis* estas analizataj kiel bizaraj kunmetaĵoj. Ekzamenado de la listo de pretervidoj konfirmis tiun divenon.

4.3. Provo 3: libera kunmetado kun eraromodelo

Kiel oni povas rekoni ke *rimdependa*, en recenzo pri poemaro, estas bona kunmetaĵo, kvankam neniam antaŭe renkontita, dum *repsondis* estas ne kunmetaĵo, sed eraro? Homa leganto distingas per la kunteksto kaj la signifo, sed alia maniero, pli facila por komputilo, estas rimarki ke *repsondis* tre similas al alia vorto jam ofte renkontita, nome *respondis*. Estas notinde ke la eraroj trovitaj en tekstoj ne estas ajnaj sinsekvoj de literoj, sed ĝenerale sekvas ian sistemon: ili apartenas al unu el la kategorioj menciitaj en la enkonduko, kaj ofte la erara vorto similas al la celita vorto, aŭ al la vorto al kiu ĝi estu korektita. Verŝajne kelkfoje ankaŭ vere okazas ke kato kuras sur la klavaron dum oni eliras por fari teon, sed tiaj eraroj estas tiel okulfrapaj ke oni apenaŭ bezonas literumilon por trovi ilin.

Tiaj pripensoj kondukis al la algoritmo aplikita en la tria prototipo: ĝi akceptas vortojn trovitajn en la tekstaro, eble kun alia gramatika finaĵo, kiel en prototipo 1, kaj novajn kunmetaĵojn, kiel en prototipo 2, sed ĝi rifuzas kunmetaĵon kiu similas al vorto trovita pli ol n fojojn en la tekstaro, kun tre simpla difino de simileco: ĝi konsideras vortojn kiuj diferencas per aldono de iu ajn litero en iu ajn loko, per forpreno de iu ajn litero, per substituo de iu ajn litero, aŭ per interŝanĝo de du apudaj literoj. (Tiuj ŝanĝetoj, kaj la parametro n, konsistigas la eraromodelon.)

La algoritmo estis rekte realigita: la programo faras liston de la similaj vortoj kaj traktas ilin laŭvice. Ĉar deklitera vorto sen ripetitaj literoj similas al 587 aliaj vortoj, oni povus timi ke la literumilo funkcios tro malrapide. Tamen, strange kaj neatendite, la prototipo, realigita en Perlo, funkciis proksimume 60-oble pli rapide ol la nova hunspell, realigita en C++. Jen la rezultoj, por diversaj valoroj de n, kun pli fruaj rezultoj por komparo:

Literumilo	Misavertoj (%)			Pretervidoj
	1-a kvarilo	2-a kvarilo	3-a kvarilo	(%)
hunspell (nova)	1,44	1,80	2,54	41,6
prototipo 1	1,43	1,78	3,31	24,8
prototipo 2	0,00	0,10	0,27	54,0
prototipo 3 ($n = 1$)	0,35	0,92	1,00	30,1
prototipo 3 ($n = 10$)	0,16	0,39	0,45	34,5
prototipo 3 ($n = 30$)	0,10	0,23	0,45	37,2
prototipo 3 ($n = 100$)	0,10	0,18	0,43	45,1

Kun granda valoro de n, prototipo 3 egalas al prototipo 2. Reduktante n, oni reduktas la frekvencon de pretervidoj, sed la frekvenco de misavertoj kreskas. La rezultoj ĉe $n = 30$ ŝajnas bona kompromiso, sed oni povus preferi divers-

ajn valorojn de n por diversaj aplikoj, do eble oni devus lasi la parametron adaptebla fare de la uzanto.

5. Konkludoj

La ekzistantaj literumiloj por esperanto ne taŭgas por kontroli longajn tekstojn kun malmultaj eraroj: ili donas tro da misavertoj. La literumiloj por kelkaj aliaj lingvoj funkcias multe pli bone en tiu rilato. Per akceptado de arbitraj kunmetaĵoj oni povas draste malaltigi la frekvencon de misavertoj, sed koste de pli alta frekvenco de pretervidoj (netrovitaj eraroj). Per nova aliro, la aplikado de eraromodelo, eblas draste malaltigi la frekvencon de misavertoj kaj samtempe malaltigi la frekvencon de pretervidoj, kompare kun kelkaj literumiloj uzataj hodiaŭ.

Ekzemple, kontrolante 50-mil-vortan tekston en kiu estas dek eraroj, oni per prototipo 3 kun $n = 30$ eble ricevus liston de ĉirkaŭ 115 vortoj, inter kiuj aperus 6 el la dek eraroj, anstataŭ listo de 900 vortoj, kiel estis taksite por la nova hunspell.

6. Plua laboro

Estis montrita la praktika utilo de eraromodelo, sed la koncepto ne estis plu disvolvita. La prototipo konsideris aliajn vortojn kiuj diferencas je unu enŝovo, forviŝo, interŝanĝo aŭ substituo, sed pli adekvata eraromodelo konsiderus aliajn specojn de eraro kaj atribuus al diversaj eraroj diversajn probablojn.

Aliaj specoj de eraro inkluzivas: mislegojn de grupoj de apudaj literoj, ekzemple *rn* mislegita kiel *m* kaj inverse; pli grandajn translokiĝojn de litero, ekzemple *problabe* anstataŭ *probable*, kie la *l* translokiĝis al alia simila kunteksto, post alia *b*; ellason de ripetita sinsekvo, ekzemple *anstaŭ* anstataŭ *anstataŭ*, kie *tata* anstataŭiĝis per *ta*.

La probabloj atribuitaj al diversaj eraroj supozeble dependus interalie de la distancoj inter klavoj kaj la simileco de sonoj, sed ili devus deveni de empiriaj esploroj.

Ideala eraromodelo pritraktus okazojn en kiuj estas pluraj eraroj en la sama vorto, kaj vorto estus komparata ne nur kun vortoj trovitaj en tekstaro, sed ankaŭ kun aliaj kandidataj kunmetaĵoj, do necesus ankaŭ modelo de kunmetado sen eraroj. (La reguloj derivitaj de Blahuš (2008) estas tia modelo, sed verŝajne oni preferus statistikan version.) Pri vorto renkontita en kontrolata teksto eblus demandi, ĉu estas pli kredinde ke temas pri certa kunmetaĵo el konataj elementoj, aŭ pri mistajpita (aŭ mislegita k.t.p.) formo de alia certa kunmetaĵo el konataj elementoj, kaj eblus respondi tiun demandon per la modeloj.

Poste leviĝus la demando, kiel oni praktike apliku tiajn modelojn en programo kiu funkcias ne tro malrapide, eventuale per metodoj el la kampo de markovaj modeloj.

Alia direkto estus esplori ĉu simila aliro helpus ankaŭ en aliaj lingvoj kun multaj kunmetaĵoj, ekzemple la germana. (Povas esti ke esperanto estas iom escepta rilate la facilecon ricevi unu vorton per mistajpo de alia. La rubriko "Spritaj splitoj kaj preskeraroj" de Ertl (2002–) eble ilustras tion.)

La prototipan literumilon eblus evoluigi ankaŭ en simplaj, praktikaj manieroj por krei pli utilan programon. Por maloftigi pretervidojn oni devus forigi erarojn el la vortlisto eltirita el la tekstaro, per homa interveno, per aŭtomataj rimedoj, aŭ plej verŝajne per miksaĵo el ambaŭ.

Verŝajne estus bone aldoni specifajn regulojn por participoj kaj transitiveco. La prototipo jam trovas kelkajn tiajn erarojn. Ekzemple, se la tekstaro enhavas la vorton *estinta*, sed neniun vorton kun la radiko *estit-*, tiam ĝi ne akceptos la vorton *estita*, sed ankaŭ la vorton *estanta* ĝi malakceptus, se en la tekstaro estus neniu vorto kun la radiko *estant-* (*estita* kaj *estanta* diferencas je unu forviŝo kaj unu substituo, respektive, de *estinta*).

Praktika literumilo devus pritrakti majusklecon kaj dividstrekojn, kiujn la prototipo simple ignoras. (Ekzemple, ĝi traktas la vortojn *UN-dungitoj* kaj *undungitoj* kiel la saman vorton.) Oni povus fari ion pli utilan ĉe vortoj kiuj enhavas ciferojn aŭ fremdajn literojn, anstataŭ ignori tiujn vortojn. Oni verŝajne dezirus rimedon por aldoni vortojn al la vortaro, kaj oni eble volus funkciigi la literumilon kun aliaj programoj, ekzemple OpenOffice.org.

Referencoj

Blahuš, Marek. 2008. *A Spell Checker for Esperanto.* Bachelor Thesis, 46 pp. Brno: Masaryk University, Faculty of Informatics.
`http://is.muni.cz/th/172464/fi_b/bc.pdf` (kontrolita 2012-12-13).

Ertl, István. 2002– . Spritaj splitoj kaj preskeraroj. Rubriko, ekde numero 98. **En:** *La Ondo de Esperanto.* Internacia sendependa magazino en Esperanto.

Monato. 1980– . Internacia magazino sendependa pri politiko, ekonomio kaj kulturo. Eldonata en Belgio. ISSN 0772-456X.

La Nova Plena Ilustrita Vortaro de Esperanto. 2002. Parizo: Sennacieca Asocio Tutmonda. ISBN 2-9502432-5-8.

Edmund Grimley Evans laboras kiel esplorinĝeniero pri programaro, pli specife programtradukiloj, ĉe ARM Ltd en Kembriĝo, Britio. Antaŭe li okupiĝis pri parolrekonado kaj komputa lingvoscienco. Li diplomiĝis pri matematiko kaj komputiko.

Retpoŝta adreso: `edmund.grimley.evans@gmail.com`

Ricevita 2011-08-05. Prilaborita versio ricevita 2012-02-22.
Akceptita por publikigo 2012-02-26