*Nordic Journal of African Studies 11(2): 283-301 (2002)*

# Tathmini ya Kamusi Tano za Kiswahili

ARVI HURSKAINEN
*University of Helsinki, Finland*

## UFUPISHO

Makala hii inaeleza mbinu za kikompyuta za kutathmini kamusi, na inatoa pia matokeo ya tathmini ya kamusi tano za Kiswahili. Kamusi hizo ni: Kamusi ya Kiswahili Sanifu (TUKI), Kamusi ya Maana na Matumizi (OUP), Modern Swahili - Modern English Dictionary (MS-tryck), Kamusi ya Kiswahili - Kiingereza (TUKI), na Swahili - Suomi - Swahili -sanakirja (SKS). SALAMA (Swahili Language Manager), ambayo ilitumiwa katika tathimini hii, huweza kuainisha maneno ya Kiswahili na kuteua lemma za maneno katika matini. Ufanisi wa kila kamusi ulitathminiwa kwa kutumia aina tatu za matini, na matokeo yametolewa kwa njia ya matarakimu na jedwali. Tathmini inaonyesha uzuri na upungufu wa kila kamusi na kuleta mwongozo juu ya kurekebisha kamusi hizo.

*Maneno muhimu: leksikografia, Kiswahili, mbinu za kikompyuta*

## UTANGULIZI

Kila mmoja aliyewahi kushiriki katika kazi ya kutunga kamusi ya lugha anajua kwamba si kazi rahisi. Wataalamu wanaohakiki kamusi hizo huweza kutofautiana sana katika maoni yao. Mmoja anaona kwamba kamusi ni nzuri tu ingawa hajachunguza ya kutosha kasoro za kamusi. Mwingine ametafiti kamusi kwa undani zaidi na huweza kuonyesha kosa au kasoro moja moja hapa na pale. Kwa kweli, mpaka sasa hatujawa na mbinu za kuridhisha kwa kutathmini kamusi kwa namna inayotakiwa. Tunaweza kusema kwamba badala ya kudadisi na kutathmini kamusi hizo kwa njia sahihi, wataalamu na wengine wanaohusika wametoa maoni yao tu.

Sababu kubwa ya kushindwa katika kazi hii ni kwamba uzuri au ubaya wa kamusi unaonekana tu baada ya mtu kuitumia kwa muda mrefu. Na pia ni kweli kwamba mtaalamu wa kutathmini kamusi sio yule anayejua lugha ile vizuri, bali ni yule anayejifunza lugha ile kuanzia mwanzo kabisa. Kwake ni lazima kuangalia kila neno na matumizi yake kwenye kamusi, naye atakabiliana na kasoro zile ana kwa ana, kwa sababu hawezi kutegemea ujuzi wake wa awali wa lugha ile kwa kujazia mapengo ya kamusi.

Nyakati hizi tumeanza kupata mbinu mpya za kutathmini kamusi, tukitumia kompyuta katika kazi hii. Programu za kuainisha lugha zaweza kusaidia sana katika kazi hii. Uzuri wa mbinu za namna hii ni kwamba tathmini ya kamusi nzima yawezekana, na kwamba matini za aina mbalimbali zaweza kutumiwa katika kufanya tathimini hii. Faida nyingine ni kwamba mbinu hizo zaweza kuonyesha

hata maneno yale, ambayo yanayokosekana katika kamusi lakini yanapatikana katika matini. Kwa kutumia mbinu za aina hii nimekwishatathmini Kamusi ya Kiswahili Sanifu (KKS) (Hurskainen 1994) na Kamusi ya Maana na Matumizi (KMM) (Hurskainen 1999). Nimeendelea na kazi hii na nimetathmini kamusi tatu zaidi, yaani Swahili - English Dictionary (SE) na Feeley, Kamusi ya Kiswahili - Kiingereza (KKK), na Swahili - Suomi - Swahili -sanakirja (SSS) .

Kamusi hizi zote tano ni kamusi za lugha ya kawaida, ingawa lengo la kila kamusi si sawa. KKS ni kamusi ya lugha moja na hii haiwezi kutumiwa na mtu anayeanzia kujifunza lugha. Hata hivyo, isipokuwa Kamusi ya Maana na Matumizi, ambayo lengo lake kubwa ni kutoa mifano ya matumizi ya maneno, kamusi hizo hufanana kwa kadiri ya kutuwezesha kuzilinganisha na kuangalia jinsi kila kamusi inavyotimiza kusudi lake.

Saizi halisi ya kamusi hizo ni kama ifuatavyo:

KKS vidahizo 14,288 (kamusi yenyewe: vidahizo 20,000 na maneno 50,000)
KMM vidahizo 8,057 (kamusi yenyewe: vidahizo 9,000 na mifano ya matumizi 12,000)
SE vidahizo 10,461 (kamusi yenyewe: maneno 13,300 na mifano ya matumizi 3,000)
KKK vidahizo 14,533 (kamusi yenyewe: zaidi ya vidahizo 30,000)
SSS vidahizo 11,500 (kamusi yenyewe: kama vidahizo 10,000 hivi)

Basi, kwanza tuorodheshe masharti kadhaa, ambazo tunatazamia kamusi ya lugha ya kawaida lazima itimize.

- Kamusi iwe na maneno yale yote kama vidahizo ambazo zinapatikana katika matini.
- Kila kidahizo kiwe na mwongozo wa kutosha, ili mtumiaji wa kamusi ajue jinsi ya kutumia neno lile. Kwa mfano, kila nomino iwe na ngeli yake au zake, ikiwa ni zaidi ya moja, na mwongozo wa kunyambua ikiwa haifuati sheria zile za kawaida. Kasoro inayopatikana katika kamusi nyingi ni kwamba ngeli haionyeshwi kwa maneno ambayo hunyambuliwa kwa ngeli ya 9/10 au 5/6, au hata 9/6.
- Kwa kivumishi ionyeshwe ikiwa kinanyambua au la, au kama inaweza kufanya zote mbili.
- Kamusi isiwe na vidahizo, ambavyo havina maelezo yo yote.

Hayo ni mambo kadhaa ambayo tunatazamia kamusi ya lugha ya kawaida itimize.

## 1. MBINU ZA KUTATHMINI KAMUSI

Tathmini ya kamusi huweza kufanywa kwa namna ya kuchagua sehemu ndogo ya kamusi na kuitafiti kwa undani. Au pengine sehemu kadhaa zaweza kutumiwa katika kazi hii. Lakini kwa vyovyote tathmini ya aina hii huleta picha ya jumla tu kuhusu kamusi ile. Mbinu za kikompyuta hutofautiana sana na mbinu zile za zamani, na hapa kwanza napenda kueleza jinsi mbinu hizo zilivyoundwa na kutumiwa.

## 1.1 MBINU ZA KUAINISHA LUGHA

Katika kutathmini kamusi za Kiswahili nimetumia SALAMA[1] (Swahili Language Manager), ambayo ni mazingira ya kikompyuta ya kufanya shughuli za aina nyingi kwa lugha ya Kiswahili. Miongoni mwa matumizi ya SALAMA ni: (1) kusahihisha lugha (spelling checker); (2) kuainisha lugha ya maandishi kwa namna zote za kisarifu, kama mofolojia, sintaksi, na semantiki; (3) kuainisha na kupanga data kwa kurahisisha utungaji wa kamusi; (4) kusaidia katika kutafuta habari kutoka matini kwa ufasaha kubwa, n.k. Lakini SALAMA inatuwezesha kutathmini kamusi vilevile, ingawa hii haikuwa lengo la kutunga SALAMA.

Katika kutathmini kamusi nimetumia hasa programu moja ya SALAMA, yaani SWATWOL, ambayo ni programu ya kuainisha lugha (Hurskainen 1992). Programu hii inapitia matini neno kwa neno na kujaribu kuona ikiwa neno ni Kiswahili. Ikiwa inaamua kwamba ni Kiswahili, basi inakubali na kutoa maelezo yote yanayotakiwa. Kwa mfano kwa nomino inatoa ngeli yake na kuonyesha ikiwa ni umoja au wingi. Kwa vitenzi huonyesha viambishi vyote vilivyopo pamoja na maelezo ya kila kiambishi, na pia lemma ya neno huonyeshwa.

Zaidi ya kuainisha neno, SALAMA yaweza kutumiwa kwa kuchagua maana halisi (disambiguate), ikiwa neno lina maana zaidi ya moja (Tapanainen 1996; Hurskainen 1996). Kwa mfano neno 'na' yaweza kutumiwa kama kiunganishi (k.mf. baba na mama) au kuonyesha mtendaji katika mifumo ya kauli ya kutendwa (k.mf. alisaidiwa na mama). Kwa hivyo, SALAMA haigundui na kuainisha maneno tu, bali pia hufanya uchaguzi kati ya matumizi mbalimbali ya maneno.

Kasoro mojawapo iliyopo katika tathmini hii ni kwamba SALAMA inatambua neno moja moja tu. Dhana inayowakilishwa na maneno zaidi ya moja (multi-word concepts) haionyeshwi katika tathmini hii. Kwa hiyo tunakazania hasa vidahizo vya neno moja. Pia hatushughuliki na maana za maneno zilizotolewa na kuamua kama ni sawa au la katika sentensi hiyo. Tatizo lingine ni kwamba ingawa programu inaweza kuainisha lemma ya neno, itakuwa vigumu kuhesabu idadi halisi ya lemma. Kuna njia mbili za kuzihesabu. Njia moja inahesabu lemma ya neno bila kujali ikiwa ni nomino, kitenzi, kivumishi, n.k., na hivyo idadi ya lemma itakuwa ndogo kuliko inavyotakiwa. Njia nyingine inatofautisha pia aina ya maneno kwa kadiri ya nomino, kitenzi, n.k., lakini inahesabu pia kama maneno ya pekee maneno yale, ambayo 'disambiguation' yake ni tofauti. Hivyo idadi ya lemma itakuwa kubwa kuliko inavyotakiwa. Baada ya majaribio na namna zote mbili nimeamua kutumia njia ile ya pili, ingawa idadi ya lemma itakuwa kubwa

---

1 Katika kuunda SALAMA, programu za makampuni kadhaa zimetumiwa. Napenda kushukuru makampuni yafuatayo kwa kutoa ruhusa ya kutumia mbinu zao: Rank Xerox Corporation kwa kutumia Two-Level rule compiler, Lingsoft kwa kutumia TWO-L analyser, na Connexor kwa kutumia CG-2 disambiguator. Pia napenda kuwashukuru Kimmo Koskenniemi (1983), Fred Karlsson (1995) na Pasi Tapanainen (1996) kwa mchango wao mkubwa katika kufanikisha SALAMA.

zaidi. Kwa hiyo msomaji asistaajabu ingawa idadi ya lemma zilizotambuliwa na kamusi huweza kuzidi idadi ya vidahizo katika kamusi. Napenda kusisitiza, kwamba hitilafu hiyo haidhuru ulinganifu baina ya kamusi, maana hii itaathiri kwa namna moja kila kamusi.

## 1.2 MATINI ZILIZOTUMIWA

Kwa kutathmini kamusi inatubidi kujua kamusi ile imekusudiwa kwa kukidhi mahitaji ya aina gani. Kwa sababu kamusi hizo tano zimekusudiwa kukidhi haja za matumizi ya lugha ya kawaida, nimechagua matini ya lugha ya namna hii kwa kufanyia tathmini hii. Matini hizo zimegawanywa katika vikundi vitatu, na vikundi hivyo vitaitwa Korpus 1, Korpus 2 na Korpus 3. Matini ni hizo:

- Korpus ya magazeti yanayopatikana katika Internet, pamoja na habari za Deutsche Welle kwa Kiswahili (Korpus 1). Data hizo ni kutoka miaka 1998 - 2001, na jumla ya maneno ni 2.484.852.
- Korpus ya magazeti ya zamani zaidi (miaka 1988-1994) pamoja na vitabu mbalimbali (jumla ya vitabu kama 45 hivi, Korpus 2). Korpus hii inayo maneno 1,190,489 .
- Korpus ya magazeti ya kisasa (mwaka 2002) pamoja na vitabu kadhaa (Korpus 3), jumla ya maneno ni 552.021.

## 2. LENGO LA TATHMINI

Katika kazi hii nimejaribu kutafiti hasa mambo yale ambayo yamewezeshwa na mbinu zilizotumiwa. Maana yake mbinu za kikompyuta huwezesha aina ya utafiti ambayo ni mpya kabisa kwa kulinganisha na njia zile za zamani za utafiti. Muhimu ni hasa kwamba tunaweza kutafiti kamusi nzima, kila kidahizo pamoja na maelezo yake. Pia tunaweza kutumia matini za aina mbalimbali katika kutafiti vidahizo hivyo. Na kisha tunaweza kutafiti jinsi maneno yaliyotumiwa katika maelezo ya vidahizo hupatikana pia kama vidahizo katika kamusi. Kwa hivyo nimeweka malengo yafuatayo matano kwa utafiti huu:

(1) Kutafuta maneno yale ambayo yanapatikana katika matini, lakini hayapatikani kama vidahizo katika kamusi.

(2) Kutafuta maneno yale ambayo yako kama vidahizo katika kamusi, lakini hayatumiwi katika matini.

(3) Kulinganisha kamusi zote tano katika mambo hayo mawili yaliyotajwa katika 1 na 2 hapo juu na kuonyesha uzuri na upungufu wa kila kamusi.

(4) Kulinganisha Kamusi ya Kiswahili Sanifu na Kamusi ya Kiswahili-Kiingereza kwa kuona maendeleo yaliyofanywa kati ya miaka 1981 na 2001 katika kuchagua vidahizo.

## 3. MANENO YAKOSEKANAYO KATIKA KAMUSI

Kila moja katika kamusi hizo tano ina kasoro zake, ingawa tofauti pia zipo. Jambo la kwanza ilikuwa ni kutafiti jinsi vidahizo ya kila kamusi vinavyokidhi mahitaji ya matini. Kwa kutafiti jambo hilo, SWATWOL ilitakiwa irekebishwe ili ifanane na kila kamusi. Kwa hiyo ilitakiwa tutengeneze aina tano za SWATWOL, ambazo kila moja ilikuwa na kasoro zake, kwa kulingana na kamusi yenyewe. Kwa maneno mengine, kila moja kati ya nakala hizo mpya za SWATWOL ilikuwa kama taswira ya kamusi yake.

Kwa kuonyesha jinsi kazi hii ilivyofanyika naeleza hapo chini hatua zilizochukuliwa kwa kupata matokeo yaliyotakiwa. Katika maelezo hayo natumia mfano wa Kamusi ya Kiswahili Sanifu (KKS). Hatua zifuatazo zilichukuliwa kwa kutathmini swala hilo.

(1) Tafuta katika matini maneno yale, ambayo hayamo katika KKS kama vidahizo. Kwa kazi hii nilitumia SWATWOL-KKS, ambayo inatambua maneno yale tu ambayo KKS inayo kama vidahizo.

(2) Baada ya kupata orodha ya maneno hayo, pitisha orodha hii kwa SWATWOL, ambayo imekamilika na kuwa na uwezo wa kutambua na kuainisha kila neno halisi katika matini tuliyotumia kwa utafiti huu.

(3) Kutokana na matokeo ya SWATWOL, chukua lemma ya kila neno na orodhesha lemma hizo ili kila lemma ipatikane mara moja tu katika orodha hii. Tunga programu inayotafuta lemma hizo.

(4) Kwa kutumia SWATWOL, tengeneza orodha ya lemma za maneno yote yanayopatikana katika matini. Panga lemma hizo katika orodha ya alfabeti na hesabu kila lemma.

(5) Sasa tumia programu ile tuliyotayarisha (3 hapo juu) na kutafuta laini zile ambazo zimekuwa na maneno yale ambayo yapo pia katika ile orodha kubwa (4 hapo juu).

Matokeo ya hatua hizo tano ni orodha ya maneno halisi ya Kiswahili ambayo yamo katika matini lakini hayapatikani katika KKS kama vidahizo. Maneno hayo yako katika mwundo wa lemma, na pia frikvensi (idadi ya kutokea) ya maneno inaonekana. Kwa sababu ya ukosefu wa nafasi, orodha hizo hazitaonyeshwa hapa. Badala yake nimetumia jedwali na matarakimu, ili tuone uzuri wa kamusi kwa kulinganisha na kamusi nyinginezo zilizotathminiwa.

# 3. 1 MATOKEO YA KORPUS 1

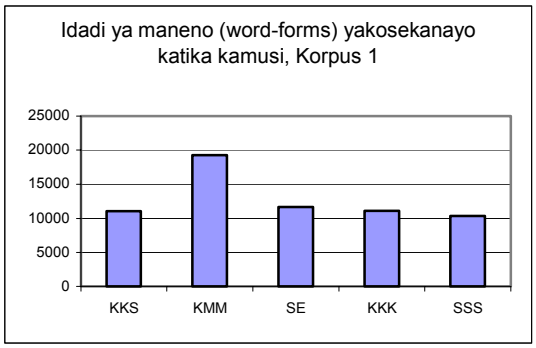
## 3.1.1 Tathmini kwa kutumia mwundo wa maneno

Ni muhimu sana kufanya tofauti kati ya mwundo wa neno (word-form) na lemma yake. Kwa mwundo wa neno tunamaanisha mwundo ule ambao neno linao katika matini (k.mf. kusoma, anasoma, anayesoma, anayesomewa n.k.). Kwa lemma tunamaanisha mwundo ule wa msingi wa neno, ambao tunakuta kama kidahizo katika kamusi (k.mf. soma). Katika tathmini hii nimetumia maana zote mbili za neno. Hasa nimetumia lemma za maneno katika kulinganisha matarakimu. Nikimaanisha lemma katika tathmini hii, nimetumia neno lemma. Nikimaanisha mwundo wa maneno nimetumia usemi huohuo, yaani mwundo wa maneno.

Baada ya kupitishwa hatua zile tano zilizoonyeshwa katika (3.) hapo juu, matokeo yafuatayo yalipatikana. Hapo kwanza tunaangalia maneno katika mwundo ule uliopo katika matini.

**Jedwali 1. Idadi ya maneno yakosekanayo katika makamusi, Korpus 1.**

| | |
|---|---|
| KKS | 11054 |
| KMM | 19245 |
| SE | 11680 |
| KKK | 11101 |
| SSS | 10364 |



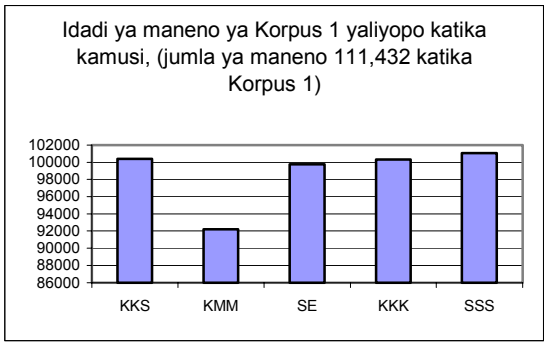

Tunaona katika Jedwali 1 kwamba kati ya kamusi hizo hakuna tofauti kubwa, ila KMM ina kasoro kubwa zaidi. Kamusi hii haikutungwa kwa watu wanaojifunza lugha, bali kwa watu wanaojua lugha tayari. Huenda hii ni sababu kwa kukosa maneno mengi zaidi kuliko kamusi nyingine.

Tunaona pia, kwamba ukosefu wa maneno haulingani na ukubwa wa kamusi. Ingawa KKS na KKK zina vidahizo vingi kuliko kamusi nyingine, upungufu wa maneno ni mkubwa.

Tunaweza kuonyesha mambo hayohayo pia kwa namna nyingine, yaani kwa njia ya kuonyesha jinsi kamusi zilivyofanikiwa kuingiza maneno ya Korpus 1. Hii imeonyeshwa katika Jedwali 2.

**Jedwali 2. Jumla ya miundo ya maneno yaliyogunduliwa na makamusi katika Korpus 1.**

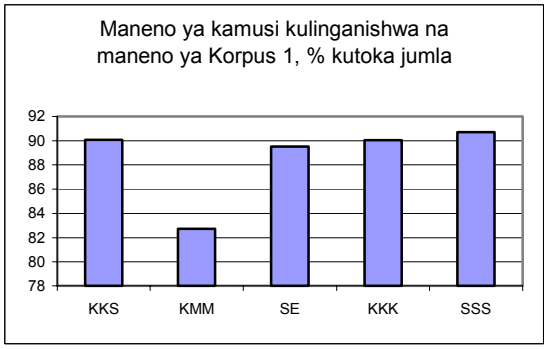
| | |
|---|---|
| KKS | 100378 |
| KMM | 92187 |
| SE | 99752 |
| KKK | 100331 |
| SSS | 101068 |




Tunaweza pia kuonyesha ufanisi wa kamusi hizo kwa njia ya kulinganisha maneno hayo yaliyogunduliwa na maneno yote ya Korpus 1. Matokeo yameonyeshwa kwa njia ya asilimia katika Jedwali 3.

**Jedwali 3. Maneno ya kamusi kulinganishwa na maneno ya Korpus 1, % kutoka jumla.**

| | |
|---|---|
| KKS | 90,08 |
| KMM | 82,73 |
| SE | 89,52 |
| KKK | 90,04 |
| SSS | 90,7 |
| TOTAL | 100 |

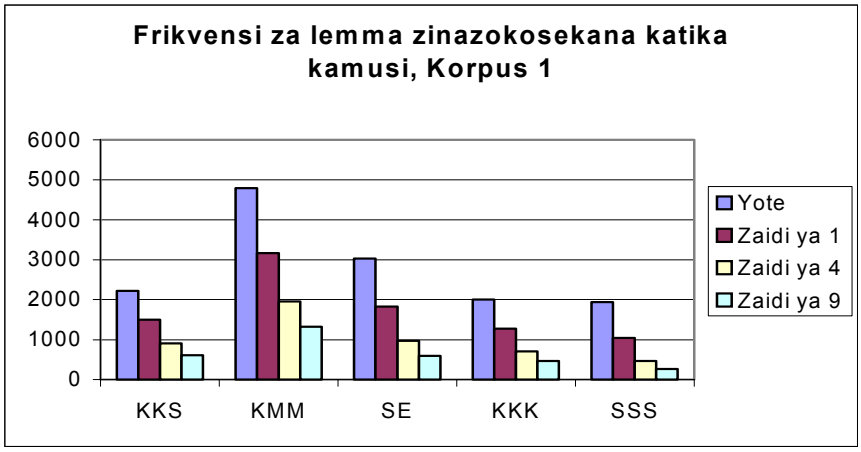



## 3.1.2 Tathmini kwa kutumia lemma

Njia ya pili, na muhimu zaidi, ni kuonyesha matarakimu kwa kutumia lemma ya maneno na sio mwundo wa maneno kamili kama tulivyoonyesha hapo juu. Hivyo tutapata kujua kwa ukamilifu ni maneno gani hasa yanayopungua katika kamusi. Pia orodha ya lemma hizo itapatikana, pamoja na frikvensi yake katika Korpus 1. Katika Jedwali 4 tunaona jumla ya lemma zilizopo katika Korpus 1 lakini zinazokosekana katika kamusi zote tano. Pia inaonekana kuna lemma ngapi zilizotokea zaidi ya mara moja, mara nne, na mara tisa katika Korpus 1. Matarakimu haya yanaonyesha kwamba maneno yanayokosekana katika kamusi hizo si maneno nadra tu katika matini bali mengi ni maneno ya kawaida.

**Jedwali 4. Frikvensi za maneno yakosekanayo katika kamusi, Korpus 1**

| | Yote | Zaidi ya 1 | Zaidi ya 4 | Zaidi ya 9 |
|---|---|---|---|---|
| KKS | 2223 | 1505 | 911 | 611 |
| KMM | 4792 | 3164 | 1952 | 1324 |
| SE | 3029 | 1828 | 968 | 596 |
| KKK | 1999 | 1278 | 704 | 467 |
| SSS | 1940 | 1047 | 463 | 264 |

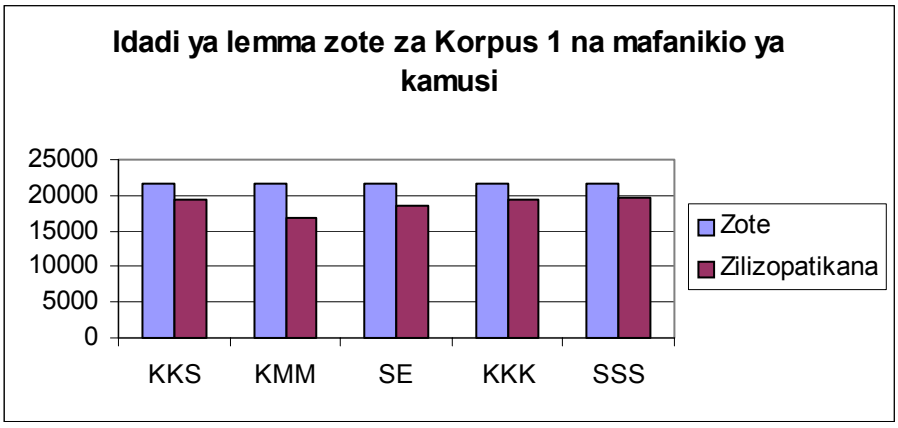



Jedwali 4 inaonyesha idadi ya maneno katika mwundo wa lemma, ambayo inafanana na kidahizo. Kwa hiyo jedwali hii inaonyesha kwa ufasaha upungufu halisi wa kamusi hizo tano. Kwanza tunaona kwamba, tukihesabu lemma zote za Korpus 1, KKK na SSS ni nzuri kuliko kamusi nyingine. Idadi ya lemma zinazokosekana katika kamusi hizo mbili ni kama 1,999 na 1,940. KKS ina upungufu zaidi (jumla 2,223), lakini si nyingi kama KMM (4,792) na SE (3,029). Matokeo haya ni karibu kama ilivyotazamiwa, ingawa KKK ilitazamiwa kuwa nzuri zaidi, kwa sababu ni kamusi mpya na inayo vidahizo vya kutosha kwa kukidhi mahitaji ya Korpus 1.

Ili kuonyesha kwamba upungufu wa kamusi hizo unahusu pia baadhi ya maneno ya kawaida, nimehesabu maneno pia kwa kadiri ya frikvensi. Jumla kubwa ya maneno yakosekanayo katika kila kamusi ni maneno yaliyotokea mara mbili au zaidi. Hasa KMM lakini pia SE zilikosa lemma nyingi zilizotokea mara mbili au zaidi, na hata zile zilizotokea mara tano au zaidi. Katika kikundi cha mara kumi au zaidi, KKS, SE na KKK zilikaribiana kwa kiasi kikubwa. SSS imetokea kama mshindi katika kila kikundi, lakini hasa katika kikundi cha maneno yenye frikvensi kubwa. Hii inathibitisha kwamba SSS imefanikiwa vizuri kuliko kamusi nyingine kuingiza maneno ya kisasa kwenye kamusi.

**Jedwali 5. Idadi ya lemma zote za Korpus 1 na mafanikio ya kamusi.**

| | Zote | Zilizopatikana | % |
|---|---|---|---|
| KKS | 21509 | 19286 | 89.7 |
| KMM | 21509 | 16717 | 77.7 |
| SE | 21509 | 18480 | 85.9 |
| KKK | 21509 | 19510 | 90.7 |
| SSS | 21509 | 19569 | 91.0 |



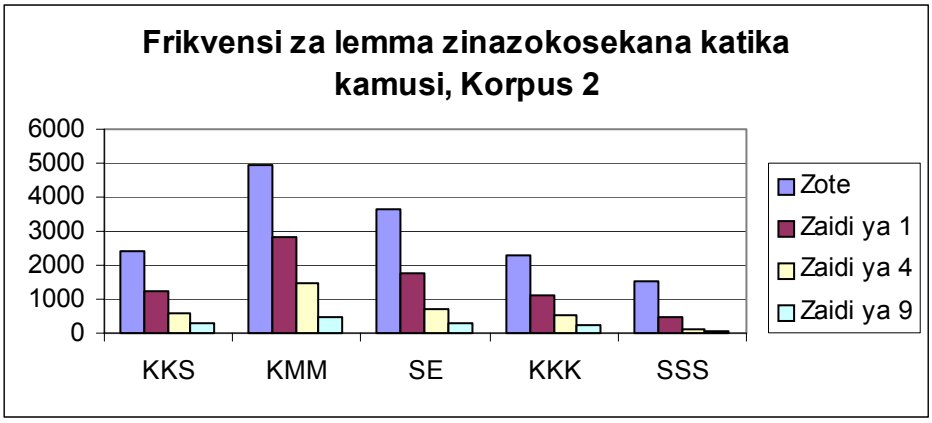

Jedwali 5 inadhihirisha kwamba KKK na SSS tu zimefanikiwa kuingiza zaidi ya asilimia 90 ya lemma za Korpus 1. KKK ni nzuri zaidi kuliko KKS, lakini tofauti si kubwa. SSS ni nzuri kuliko kamusi nyingine, lakini tofauti na KKK si kubwa.

## 3.2 MATOKEO NA KORPUS 2

Korpus 2 inayo matini za aina mbili. Jumla kubwa ni vitabu vya riwaya, lakini pia vitabu fulani vya fasihi vimo. Sehemu nyingine kubwa katika korpus hii ni maandishi ya magazeti kutoka miaka 1988 - 1995. Ni muhimu kujua pia kwamba katika kuunda SSS matini hizi zilitumiwa kwa kuchagua vidahizo, ingawa maneno yote hayakuingizwa katika kamusi. Jumla ya maneno katika Korpus 2 si kubwa kama katika Korpus 1, lakini hata hivyo ni kubwa ya kutosha kwa kuonyesha mafanikio ya kamusi katika matini za aina hiyo.

**Jedwali 6. Frikvensi za lemma zinazokosekana katika kamusi, Korpus 2**

| | Zote | Zaidi ya 1 | Zaidi ya 4 | Zaidi ya 9 |
|---|---|---|---|---|
| KKS | 2426 | 1256 | 567 | 309 |
| KMM | 4924 | 2798 | 1442 | 482 |
| SE | 3647 | 1758 | 694 | 296 |
| KKK | 2311 | 1146 | 503 | 258 |
| SSS | 1512 | 490 | 146 | 88 |

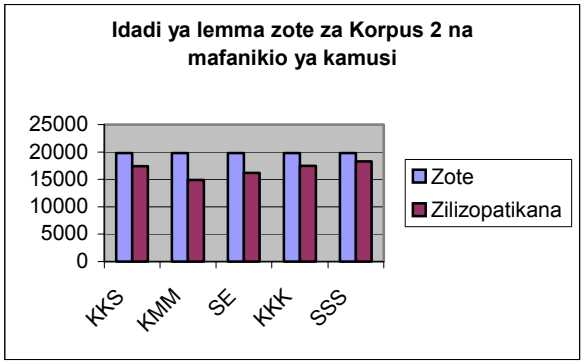



Katika Jedwali 6 tunaona kwamba matokeo kwa jumla ni sawa na Korpus 1. KMM na SE ni kati ya kamusi zilizo na upungufu zaidi. Tena tofauti ipo kati ya KKS na KKK kama ilivyotazamiwa. Lakini jambo la kushangaza ni kwamba ingawa jumla

ya maneno yanayokosekana ni kubwa kuliko katika Korpus 1, tukiangalia matarakimu ya maneno ya kawaida tunaona kwamba matokeo ni karibu sawa katika korpus zote mbili.

**Jedwali 7. Idadi ya lemma zote za Korpus 2 na mafanikio ya kamusi.**

| | Zote | Zilizopatikana | % |
|---|---|---|---|
| KKS | 19799 | 17379 | 87.8 |
| KMM | 19799 | 14875 | 75.1 |
| SE | 19799 | 16152 | 81.6 |
| KKK | 19799 | 17488 | 88.3 |
| SSS | 19799 | 18287 | 92.4 |




Kama Jedwali 7 inavyoonyesha, Korpus 2 ina lemma 19,799. Idadi hii kubwa inatokana na matini za aina nyingi zilizopo katika korpus hii. Sababu nyingine ni kwamba, ingawa majina na namba zimeondolewa, maneno kadhaa yametokea zaidi ya mara moja, hasa ikiwa maneno yale yamefanyiwa 'disambiguation' kwa namna tofauti. Hitilafu hii inahusu pia lemma zilizopatikana kwa kila kamusi. Kwa hiyo namba zote katika jedwali hii ni kubwa kiasi, lakini hii haiathiri sana asilimia, ambazo zinaonyesha hali halisi. Na hasa katika kulinganisha kamusi hitilafu hii hiana maana kabisa.

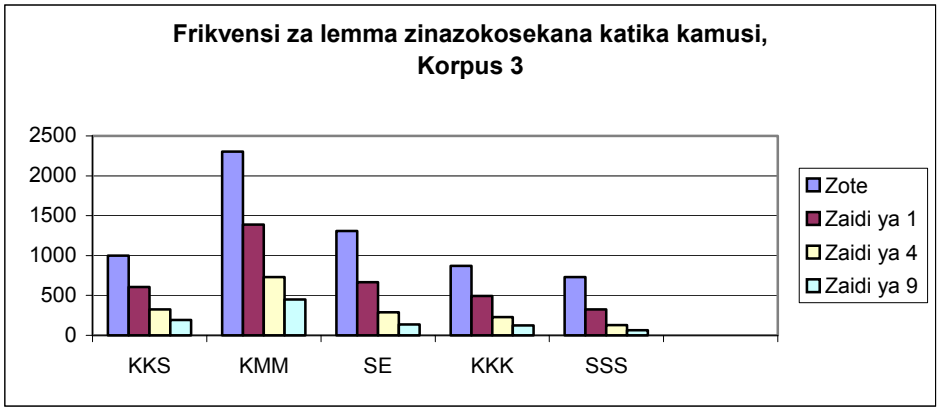
Ingawa Korpus 1 ni kubwa zaidi, idadi ya lemma zake ni ndogo zaidi. Pia mafanikio ya kamusi mbalimbali katika kugundua maneno hayo ni ya wastani tu, isipokuwa SSS, ambayo, kama nilivyosema, imejengwa kwa kiasi kikubwa juu ya korpus hii. Mafanikio yake mazuri yanatokana na hali hii kwa kiasi.

## 3.3 MATOKEO NA KORPUS 3

Korpus 3 ni ndogo kuliko nyingine zilizotumiwa katika tathmini hii, na maneno yake hutokana zaidi na magazeti ya mwaka 2002. Vitabu fulani vya riwaya pia vimeingizwa. Korpus hii haina uzito sana. Sababu yake ya kuingiza hapa ni kuona ikiwa maandishi ya sasa hivi hugunduliwa na makamusi. Tunaonyesha matokeo haya kwa kifupi hapo chini.

**Jedwali 8. Frikvensi za lemma zinazokosekana katika kamusi, Korpus 3.**

| | Zote | Zaidi ya 1 | Zaidi ya 4 | Zaidi ya 9 |
|---|---|---|---|---|
| KKS | 1001 | 605 | 325 | 193 |
| KMM | 2304 | 1387 | 731 | 449 |
| SE | 1309 | 665 | 287 | 136 |
| KKK | 871 | 494 | 229 | 125 |
| SSS | 731 | 326 | 128 | 64 |

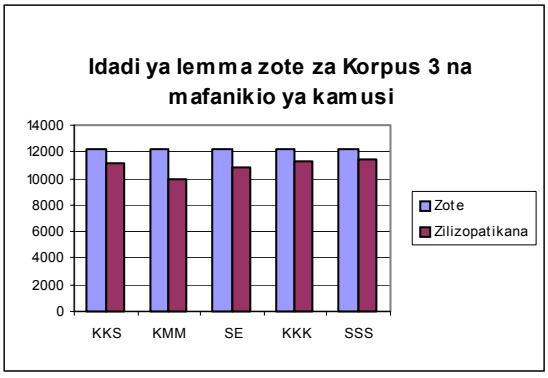



Hapo tunaona kwamba SSS imeshinda kamusi zote tukifanya tathmini na matini hizi za kisasa za Korpus 3. Tofauti ni kubwa zaidi kati ya maneno yaliyotokea katika korpus mara nyingi. Kitika kikundi cha maneno yaliyotokea mara kumi au zaidi katika korpus, SSS ilikuwa na maneno 64 tu ambayo hayako katika kamusi. Kamusi iliyochukua nafasi ya pili ni ni KKK, ambayo ilikuwa na maneno 125 katika kikundi hiki. Hapa SE imeshinda KKS, jambo ambalo si kawaida. Labda sababu yake ni kwamba KKS inayo maneno ya zamani zaidi, pamoja na maneno yanayotumiwa mara chache sana, na SE imefanikiwa kuingiza maneno ya kisasa zaidi. KKS imefanikiwa vizuri zaidi katika maneno nadra, na KMM ni ya mwisho katia kila kikundi.

Jumla ya maneno yanayokosekana hapa kwa jumla ni ndogo kuliko katika Korpus 1 na 2, na sababu yake ni kwamba korpus hii ni ndogo zaidi. Kwa hiyo pia jumla ya lemma za aina mbalimbali ni ndogo. Kwa hiyo haifai kulinganisha matokeo ya korpus mbalimbali moja kwa moja. Njia halisi ya kutathmini ni kuona mafanikio ya kila kamusi kwa kutumia kila korpus peke yake, kama tulivyofanya hapa.

**Jedwali 9. Idadi ya lemma zote za Korpus 3 na mafanikio ya kamusi.**

| | Zote | Zilizopatikana | % |
|---|---|---|---|
| KKS | 12209 | 11208 | 91.8 |
| KMM | 12209 | 9905 | 81.1 |
| SE | 12209 | 10900 | 89.3 |
| KKK | 12209 | 11338 | 92.9 |
| SSS | 12209 | 11478 | 94.0 |



Jedwali 9 inadhihirisha kwamba kamusi tatu zimefanikiwa kuingiza zaidi ya asilimia 90 ya maneno ya Korpus 3. SE pia imefanikiwa kufikia kwenye asilimia 90, lakini KMM imebaki kwenye asilimia 81.1.

## 4. MANENO YA 'ZIADA' KATIKA KAMUSI

Pengine wazo linaweza kutolewa kwamba kamusi haina maneno ya ziada. Ni kweli kwamba maneno ambayo hayatumiwi mara kwa mara yanaweza kuwemo katika kamusi ikiwa nafasi ipo. Wale wanaojua vizuri lugha huhitaji kamusi kwa kutafuta maneno hayo hasa. Lakini ikiwa kamusi imekusudiwa kwa watu wanaohitaji kutafuta maneno ya kawaida, na ikiwa mipaka imewekwa kwa saizi ya kamusi, uchaguzi lazima ufanywe. Kwa kufikiria jinsi ya kupunguza vidahizo, orodha ya maneno ya matini itasaidia.

Sehemu hii ya utathmini ni ngumu, kwa sababu si rahisi kuonyesha ni vidahizo gani hasa katika kamusi ambavyo havikupatikana katika matini hata mara moja. Sababu mojawapo kwa ugumu ni kwamba maneno katika SWATWOL si katika mwundo wa lemma, hasa vitenzi na baadhi ya nomino. Kwa hivyo si rahisi kulinganisha maneno yenye mwundo wa lemma na maneno yenye mwundo wa SWATWOL. Hata hivyo, kazi hii imefanikiwa, na kwa kifupi hatua ni zifuatazo:

(1) Ainisha matini kwa SWATWOL-KKS.
(2) Panga lemma za maneno yote yaliyotambuliwa kwenye vikundi kwa kadiri ya aina yake, yaani vitenzi, nomino za kila ngeli, vivumishi, n.k.
(3) Badilisha maneno katika SWATWOL-KKS yawe sawa na lemma za maneno haya.
(4) Toa sasa maneno ya SWATWOL-KKS na panga katika vikundi kama katika (2) juu.
(5) Kutokana na vikundi vya maneno (4 hapo juu) ondoa kila laini ambayo ina neno ambalo linapatikana pia katika orodha ya (2) hapo juu. Hivyo tutaondoa kutoka SWATWOL-KKS maneno yote yaliyogunduliwa na programu hii. Yaliyobaki kwenye SWATWOL-KKS ni maneno yale, ambayo yamo katika KKS lakini hayakupatikana katika matini.

Kazi hii inatakiwa kufanywa kwa kila kikundi cha maneno pekee, ili tusilinganishe maneno ya vikundi tofauti, na hivyo kuharibu tathmini. Kazi hii pia imefanyiwa kwa kila kamusi, na matokeo ya tathmini yameonyeshwa. Kwa sababu ya matatizo katika kuainisha maneno ya kamusi kadhaa nimeamua kufanya utafiti huu kuhusu vitenzi na nomino za ngeli ya 1/2, 3/4, 5/6, 7/8, na 9/10. Maneno ya ngeli ya 11 na pia vivumishi, vielezi na kadhalika nimeacha nje. Katika kutathmini swala hili nimetumia Korpus 1 tu.

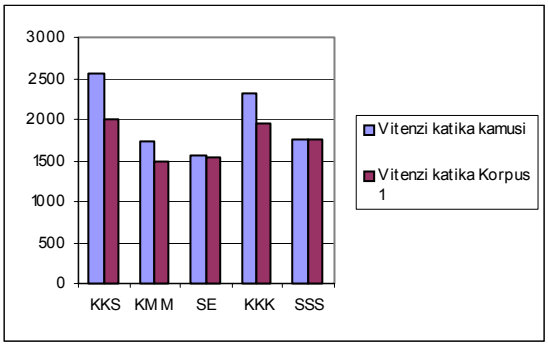
## 4.1 VITENZI

Napenda kusisitiza kwamba katika kuhesabu vitenzi vinyambuo vya vitenzi havikuhesabiwa, isipokuwa pale tu, ambapo kinyambuo kina maana ya pekee katika lugha (k. mf. enda, endesha, endelea). Kwa hiyo, katika kuchambua maneno SWATWOL inarudisha lemma ya msingi kwa vinyambuo vya vitenzi. Ndipo programu hii inahesabu vitenzi kama vilivyo katika vidahizo vikuu vya kamusi.

Tuangalie kwanza jinsi vitenzi vya kamusi na vitenzi katika korpus vilivyokutana. Tunaona katika Jedwali 10, kwamba KKS inayo vitenzi zaidi kuliko kamusi nyingine. Pia KKK inayo vitenzi vingi kuliko kamusi nyingine tatu. Inaonekana kwamba KKS ilifanikiwa kuingiza vitenzi vingi vya korpus (1,991) kuliko kamusi nyingine Hata hivyo, wakati uleule KKS ilikuwa na vitenzi 569, ambavyo havikupatana hata mara moja katika korpus. KKK ilikuwa na vitenzi 1,948 vya korpus, na vitenzi vyake vya 'ziada' vilikuwa 357. Kwa hiyo KKK haikuwa na vitenzi vingi vya ziada kama KKS. Lakini kamusi nyingine tatu, hasa SE na SSS, zilifanikiwa vizuri sana. KMM ilikuwa na hitilafu zaidi, maana kati ya vitenzi vyake (1,742) jumla ya vitenzi 1,500 tu ilipatikana katika korpus. SE ilifanikiwa vizuri akuingiza vitenzi ambavyo vilitumiwa katika korpus (1,557 kwa 1,540). Lakini SSS ilifanikiwa katika kazi yake vizuri kuliko kamusi nyingine zote. Karibu vitenzi vyake vyote vilitumiwa katika korpus (1,751 kwa 1,748).

**Jedwali 10. Vitenzi katika kamusi na katika Korpus 1.**

| | Vitenzi katika kamusi | Vitenzi katika Korpus 1 |
|---|---|---|
| KKS | 2560 | 1991 |
| KMM | 1742 | 1500 |
| SE | 1557 | 1540 |
| KKK | 2305 | 1948 |
| SSS | 1751 | 1748 |

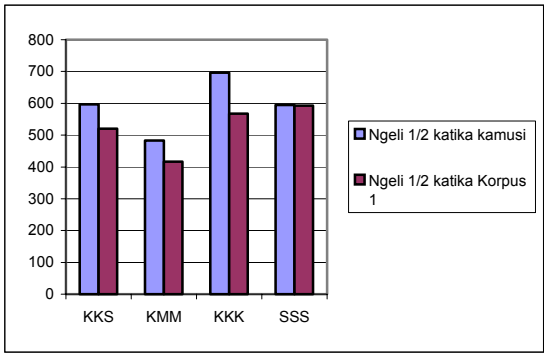



## 4.2 NOMINO

Katika tathmini ya nomino imenibidi kuacha SE nje kabisa, kwa sababu njia ya kamusi hii kuonyesha ngeli haikuwa nzuri ya kutosha kwa kutathmini swala hili kwa njia ya kuaminika. Kwa hiyo, ili kuepukana na makosa katika tathmini hii niliamua kuiacha kabisa.

**Jedwali 11. Ngeli 1/2 katika kamusi na katika Korpus 1.**

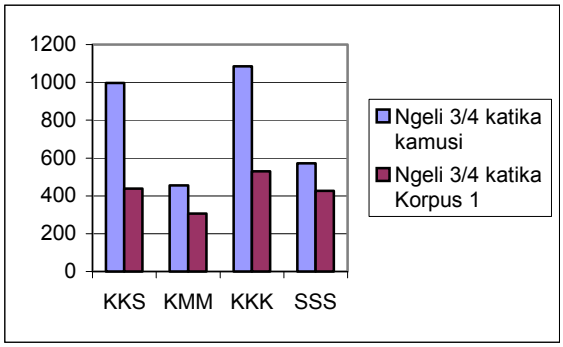
| | Ngeli 1/2 katika kamusi | Ngeli ½ katika Korpus 1 |
|---|---|---|
| KKS | 597 | 520 |
| KMM | 483 | 417 |
| KKK | 696 | 568 |
| SSS | 595 | 593 |




Jedwali 11 inatuonyesha kwamba SSS imefanikiwa vizuri kuliko nyingine tatu. Idadi ya nomino za ngeli ya 1/2 zilizopatikana katika korpus ni kubwa, na karibu kila nomino ya kamusi imetumiwa (595 kwa 593). KKK inayo nomino za ngeli hii zaidi kuliko kamusi nyingine, lakini nomino zake hazikutani sawa na korpus (696 kwa 568). KKS na KMM zina hitilafu ya aina ile pia, lakini nomino za 'ziada' katika kamusi hizo si nyingi.

**Jedwali 12. Ngeli 3/4 katika kamusi na katika Korpus 1**.

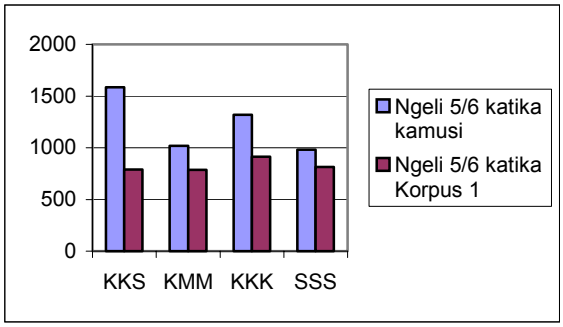
| | Ngeli 3/4 katika kamusi | Ngeli 3/4 katika Korpus 1 |
|---|---|---|
| KKS | 997 | 439 |
| KMM | 455 | 306 |
| KKK | 1085 | 529 |
| SSS | 572 | 427 |




Maneno ya ngeli ya 3/4 yanaonyesha tofauti kubwa kati ya kamusi. KKS na KKK zinafanana hata katika swala hilo. Idadi ya maneno ya ngeli hii ni karibu mara mbili kwa kulinganisha na maneno yaliyopo katika korpus. Kwa hiyo maneno ya 'ziada' ya ngeli hii ni nyingi. KMM haina maneno ya ziada mengi, na kwa jumla kamusi hii ina kasoro kubwa ya maneno. SSS imefanikiwa vizuri pia katika kikundi hiki cha maneno, ingawa hata kamusi hii ina maneno ya ziada.

**Jedwali 13. Ngeli 5/6 katika kamusi na katika Korpus 1**

| | Ngeli 5/6 katika kamusi | Ngeli 5/6 katika Korpus 1 |
|---|---|---|
| KKS | 1584 | 788 |
| KMM | 1020 | 786 |
| KKK | 1319 | 914 |
| SSS | 981 | 815 |

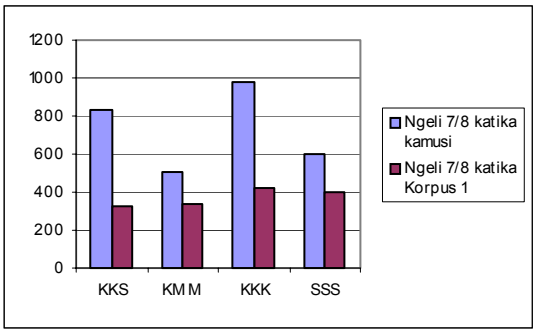



Katika kikundi cha ngeli ya 5/6 KKS inayo maneno mengi kuliko kamusi nyingine, lakini nusu tu ya maneno yake yametumiwa katika korpus. KKK

inayo maneno ya ngeli hii kidogo zaidi, lakini maneno yake mengi zaidi hupatikana pia katika korpus. SSS ina maneno ya ngeli hii kidogo kuliko kamusi nyingine, lakini maneno yake mengi hupatikana pia katika korpus.

**Jedwali 14. Ngeli 7/8 katika kamusi na katika Korpus 1**

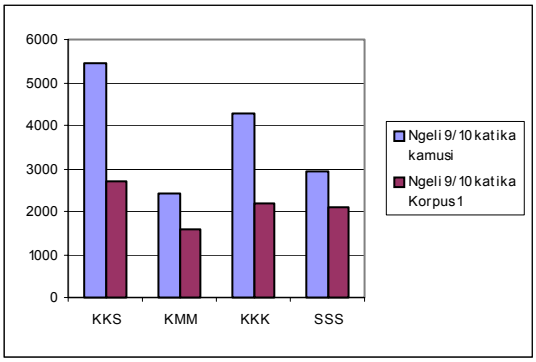
| | Ngeli 7/8 katika kamusi | Ngeli 7/8 katika Korpus 1 |
|---|---|---|
| KKS | 836 | 324 |
| KMM | 501 | 336 |
| KKK | 984 | 420 |
| SSS | 600 | 401 |




Hapa tena tunaona kwamba KKS na KKK zimeingiza maneno mengi ya ngeli hii, lakini chini ya nusu ya maneno hupatikana katika korpus. Hapo KKS ni hafifu kuliko kamusi nyingine, maana ingawa jumla ya maneno ni 836, kati ya hayo 324 tu yametumiwa katika korpus. Hali kadhalika na KKK, ambayo imeingiza maneno zaidi (984), lakini 420 tu yametumiwa katika korpus. Kamusi nyingine mbili pia zina maneno ya ziada zaidi katika kikundi hiki kuliko katika vikundi tulivyowahi kuangalia.

**Jedwali 15. Maneno ya ngeli 9/10 katika kamusi na katika Korpus 1**

| | Ngeli 9/10 katika kamusi | Ngeli 9/10 katika Korpus 1 |
|---|---|---|
| KKS | 5448 | 2713 |
| KMM | 2441 | 1576 |
| KKK | 4295 | 2192 |
| SSS | 2940 | 2106 |




Idadi ya maneno katika ngeli ya 9/10 ni kubwa kuliko katika ngeli nyingine. Hapa pia tunaona kwamba KKS na KKK zimeingiza maneno mengi, lakini kama nusu tu ya maneno yametumiwa katika korpus. Kamusi nyingine mbili zina maneno kidogo zaidi, lakini SSS imefanikiwa vizuri zaidi kukidhi haja za korpus.

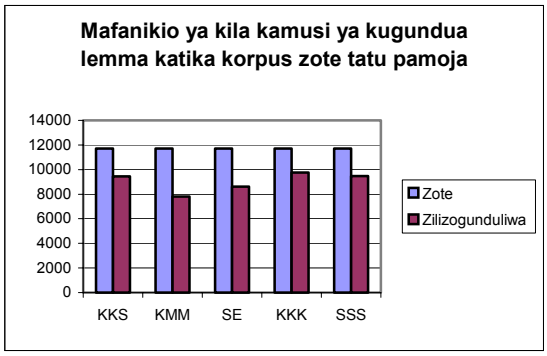
## 5. TATHMINI YA KORPUS ZOTE TATU PAMOJA

Tunaweza pia kuangalia jinsi kamusi hizo tano zilivyofanikiwa kugundua maneno ya korpus hizi tatu pamoja. Hapa nimehesabu maneno yale tu ambayo ni maneno halisi ya Kiswahili kwa uhakika. Nilitumia kwanza SALAMA na kupata lemma 11,713. Halafu nilitumia programu ya kila kamusi kwa zamu, na matokeo yake yanaonekana katika Jedwali 16. Inaonekana kwamba KKK

ilifanikiwa vizuri kuliko kamusi nyingine, lakini KKS na SSS zilifanikiwa vizuri pia.

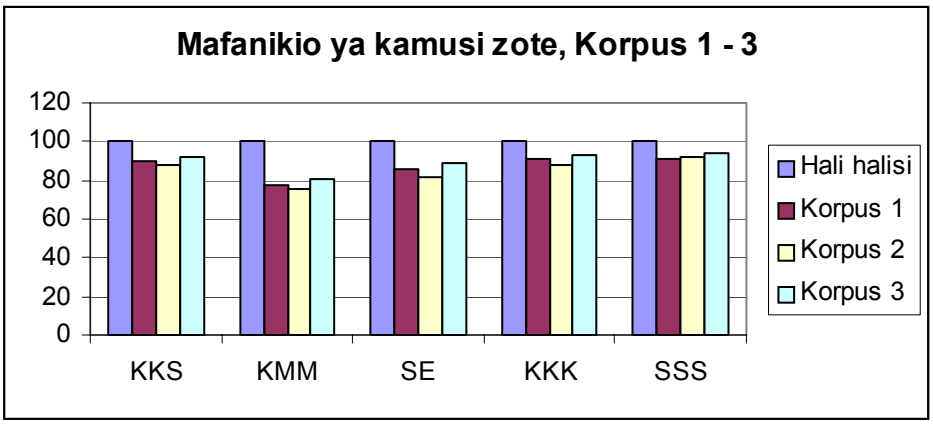
**Jedwali 16. Mafanikio ya kila kamusi katika kugundua lemma katika korpus zote tatu pamoja.**

| | Zote | Zilizogunduliwa | % |
|---|---|---|---|
| KKS | 11713 | 9441 | 80.6 |
| KMM | 11713 | 7808 | 66.7 |
| SE | 11713 | 8613 | 73.5 |
| KKK | 11713 | 9746 | 83.2 |
| SSS | 11713 | 9475 | 80.9 |




**Jedwali 17. Mafanikio ya kamusi zote, Korpus 1-3.**

| | Hali halisi | Korpus 1 | Korpus 2 | Korpus 3 |
|---|---|---|---|---|
| KKS | 100 | 89,7 | 87,7 | 91,8 |
| KMM | 100 | 77,7 | 75,1 | 81,1 |
| SE | 100 | 85,9 | 81,6 | 89,3 |
| KKK | 100 | 90,7 | 88,3 | 92,9 |
| SSS | 100 | 91 | 92,4 | 94 |




Tunaona katika Jedwali 17 kwamba Korpus 2 ilikuwa ngumu kuliko korpus nyingine kwa kamusi nyingine, isipokuwa kwa SSS, ambayo, kama nimekwishasema, imefaidi kutokana na korpus hiyo wakati wa kutungwa. Pia tunaona kwamba Korpus 3 imekuwa rahisi kwa kila kamusi. Sababu mojawapo ni kwamba korpus hii ni ndogo tukilinganisha na korpus

## 6. KULINGANISHA KKS NA KKK

Kamusi ya Kiswahili Sanifu (KKS) ilitolewa mara ya kwanza mwaka 1981. Tangu wakati ule Kiswahili kimeendelea kukua na maneno mapya yameingia kwenye lugha. Nilipoanza utafiti huu, nilikuwa na nadharia tete kwamba Kamusi ya Kiswahili-Kiingereza, ambayo ilitolewa mara ya kwanza mwaka 2001, ingekuwa tofauti na KKS katika mambo mawili: (a) upungufu wa maneno

uliopo katika KKS ungalirekebishwa katika KKK, na (b) maneno mapya ya kisasa yangaliongezwa kwa kukidhi mahitaji ya leo.

Kwa kutumia Korpus 1 kwa kupima swala hilo, utafiti ulionyesha kwamba tatu kwa nne ya maneno yaliyopungua ni yaleyale katika kamusi zote mbili (1.583). Jumla ya maneno 509 ni tofauti katika KKS, na 338 katika KKK. Kwa hiyo KKK ni nzuri zaidi kidogo, lakini tofauti si kubwa. Pia inaonekana kwamba nadharia tete yangu haikupata uthibitisho kamili. Makosa ya KKS hayakurekebishwa ya kutosha katika KKK, na maneno mengi mapya yanayotumiwa katika magazeti ya kisasa yanakosekana katika KKK.

## HITIMISHO

Kwa msingi wa tathmini hii tunaweza kusema yafuatayo:
(1) Kamusi zote tano zina upungufu katika kuingiza vidahizo kwa kukidhi mahitaji ya matini tuliyotumia katika tathmini hii.
(2) Kuna tofauti za maana kati ya kamusi hizo. KMM hasa inatofautiana na kamusi nyinginezo na matokeo yake hafifu yanatakiwa kuangaliwa kwa msingi huo.
(3) Nadharia tete ile, kwamba kamusi yenye maneno zaidi ndiyo inayofanikiwa vizuri, haikuthibitishwa na tathmini hii. KKS na KKK ndizo zilizo na vidahizo zaidi kuliko kamusi nyingine, lakini hata hivyo SSS iliyo na vidahizo 11.000 tu ilizishinda kamusi zote katika kila aina ya tathmini.
(4) KKS na KKK ni kamusi ambazo zina zaidi ya kamusi nyingine maneno ya aina ile, ambayo hayakutumiwa katika matini hata mara moja.
(5) Nadharia tete ile kwamba KKK ni nzuri kuliko KKS ilithibitishwa kwa kiasi. Kamusi hizo mbili zilikuwa karibu sawa katika kutambua miundo ya maneno ya magazeti, lakini katika tathmini ya lemma KKK ilikuwa nzuri zaidi, lakini si kwa kiasi kilichotazamiwa..
(6) Mafanikio mazuri ya SSS yanaonyesha kwamba mbinu za kikompyuta zinaweza kuboresha kamusi kwa kiasi kikubwa bila ya saizi yake kukua mno.
(7) Mbinu za kikompyuta zinaweza kutumiwa kwa kuonyesha kila neno linalopungua katika kamusi, na pia kuonyesha kila kidahizo ambacho hakitumiwi katika matini. Pia frikvensi za maneno hayo zinapatikana. Kwa hiyo mbinu hizo ni njia bora ya kusaidia katika kutunga kamusi.

## MAREJEO

Hurskainen, A. 1992.
*A Two-Level Computer Formalism for the Analysis of Bantu Morphology: An Application to Swahili*. **Nordic Journal of African Studies** 1(1): 87-122.

1994 *Kamusi ya Kiswahili Sanifu in test: A computer system for analyzing dictionaries and for retrieving lexical data*. **Afrikanistische Arbeitspapiere** 37 (Swahili Forum I): 169-179.

1996 Disambiguation of morphological analysis in Bantu languages. In *COLING-96, Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, August 5-9, 1996. Pp. 568-573.

1999 *Salim K. Bakhressa, Kamusi ya Maana na Matumizi*. Nairobi: Oxford University Press. Book review. **Journal of African Languages and Linguistics** 20. Book review.

Karlsson, F. 1995.
Designing a parser for unrestricted text. In Karlsson et al (eds.), *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin. Pp. 1-40.

Koskenniemi. K. 1983
*Two-level morphology: A general computational model for word-form recognition and production*. Publications No. 11. Department of General Linguistics, University of Helsinki.

Tapanainen, P. 1996.
*The Constraint Grammar Parser CG-2.* Publications No. 27. Department of General Linguistics, University of Helsinki.

## KAMUSI ZILIZOTATHMINIWA

Abdulla, A., Halme, R., Harjula, L. and Pesari-Pajunen, M. 2002.
*Swahili - Suomi - Swahili -sanakirja*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Bakhressa, Salim K. 1992.
*Kamusi ya Maana na Matumizi.* Nairobi: Oxford University Press.

Feeley, Gerald, 1994.
*Modern Swahili - Modern English Dictionary* (Revised and enlarged second edition). Denmark: MS-tryk.

*Kamusi ya Kiswahili - Kiingereza* (Swahili - English Dictionary), 2001. Dar-es-Salaam: Taasisi ya Uchunguzi wa Kiswahili.

*Kamusi ya Kiswahili Sanifu*, 1981. Dar-es-Salaam: Oxford University Press.