MBINU ZA KIKOMPYUTA ZA KUCHOPEKA TAARIFA ZA KIISIMU NA KUCHOPOA DATA YA LUGHA KATIKA KONGOO-MATINI

S. S. Sewangi

Ikisiri

Makala hii inatalii mbinu za kuingiza taarifa za kiisimu na kuchopoa data ya lugha kwenye kongoo-matini kwa msaada wa kompyuta. Makala inaanza kwa kubainisha mambo ya msingi kwa ajili ya matumizi ya kompyuta katika ufafanuzi na uchopoaji wa data ya lugha. Mambo hayo ni kama vile kufanya marekebisho katika kongoo ili kuendana na programu husika na kuingiza taarifa za kiisimu katika kongoo-matini. Kisha makala inafafanua vipengele muhimu katika uingizaji wa taarifa za kiisimu kwenye kongoo-matini kwa msaada wa kompyuta. Pia, makala inagusia uchopoaji wa data ya lugha katika kongoo-matini.

1.0 Utangulizi

Katika miaka ya hivi karibuni, matumizi ya kompyuta katika taaluma ya utafiti wa lugha yameshika kasi kutokana na maendeleo ya kikompyuta katika uhifadhi na ufafanuzi wa data ya lugha ya binadamu. Matumizi hayo yamepelekea kuibuka kwa mkabala mpya wa utafiti wa lugha unaoitwa isimukompyuta. Lengo la mkabala huu ni kurahisisha ufafanuzi na upatikanaji wa data ya utafiti wa lugha kwa kutumia kompyuta. Msingi wa matumizi ya kompyuta katika ufafanuzi na upatikanaji wa data ya lugha upo katika mambo matatu. Jambo la kwanza ni upatikanaji wa programu za kufafanua kongoo-matini na za kukusanya data ya lugha. Programu za kuchambua kongoo huundwa kulingana na mikabala ya aina mbili: mkabala wa kanuni yumkinifu na wa kanuni sarufi ya lugha ya binadamu. Aidha programu hizo huundwa kwa kuzingatia viwango mahususi vya ufafanuzi wa lugha, kama vile, mofolojia, sintaksia na semantiki. Programu za kukusanya data huundwa kwa kutumia ruwaza za kanuni za miundo mbalimbali ya vijenzi lugha. Jambo la pili ni upatikanaji wa kongoo-matini, yaani lugha iliyoteuliwa na kuhifadhiwa katika kompyuta kwa ajili ya kukidhi malengo mahususi ya utafiti. Jambo la tatu ni kuwepo kwa mbinu za kuchambua na kukusanya data kwenye kongoo-matini kulingana na programu zilizopo. Makala hii inamakinikia mbinu za kuchambua na kukusanya data kwenye kongoo-matini kwa kutumia kompyuta.

2.0 Ufafanuzi wa kongoo-matini

Kulingana na muktadha wa isimukompyuta, kongoo-matini ni kusanyo la lugha kwenye muundo wa kikompyuta lililoteuliwa kulingana na malengo mahususi ya utafiti wa lugha (taz. McEnery na wenzake1996). Kwa hali hiyo, kusanyo la lugha ya kikompyuta lisilotangamana na malengo mahususi ya utafiti sio kongoo-matini bali hifadhi ya kugha katika kompyuta. Hivyo basi, ukusanyaji wa kongoo-matini unapaswa uzingatie uwakilishi wa lugha lengwa. Uzingatizi huu ni wa lazima kwa vile mahitimisho ya kitafiti yanayotokana na data ya kongoo hujumuisha aina ya lugha inayowakilishwa katika upana wake (taz. Barnbrook 1996). Ukusanyaji wa kongoo-

matini unaweza kulenga katika kuwakilisha lugha katika ujumla wake, kwa mfano kongoo-matini ya lugha ya Kiswahili. Aidha kongoo-matini yaweza kukusanywa kwa lengo la kuwakilisha aina fulani tu ya lugha, kama vile kongoo ya Kiswahili cha uwanja mahususi wa kitaalamu. Kongoo iliyokusanywa kuwakilisha lugha katika ujumla wake huwa kubwa na huendelea kukua kadri matini mpya yanavyoongezeka. Mfano mzuri wa kongoo ya namna hii ni Kongoo ya Taifa la Uingereza ambayo imekusanywa kwa lengo la kuwakilisha Kiingereza katika ujumla wake. Hadi sasa kongoo hii ina jumla ya maneno milioni 100 (taz.tovuti http://www.hcu.ox.ac.uk/BNC/). Kongoo iliyokusanywa kwa ajili ya kuwakilisha lugha maalumu huwa na ukubwa na ukomo fulani na huwa haiongezeki.

Uundaji wa kongoo-matini hutegemea vyanzo vitatu ambavyo ni: matini zanazopatikana katika mtandao wa Intaneti, matini zilizoko katika maandishi ya kawaida na lugha ya mazungumzo. Upatikanaji wa matini katika mtandao wa Intaneti umerahisisha sana kazi ya kuunda kongoo-matini za baadhi ya lugha ambazo zina matumizi mapana katika mtandao huo. Hata hivyo, lugha nyingi hasa za Kiafrika bado hazitumiki kabisa au zina matumizi madogo sana katika mawasiliano ya mtandao wa Intaneti. Hivyo, uundaji wa kongoo za lugha hizo inabidi utegemee ukusanyaji wa matini za kawaida na wa lugha ya mazungumzo. Katika hali hii, uingizaji wa matini za lugha katika muundo wa kikompyuta hufanywa kwa ama uchapaji wa kawaida au kwa kutumia skana. Uingizaji wa matini za kawaida katika kompyuta kwa kutumia skana hurahisisha kazi na pia huokoa muda. Hata hivyo njia hii ni ghali kwani hutumia kifaa cha skana pamoja na programu yake ambavyo ni ghali. Aidha njia hii huhitaji matini zenye ubora wa hali ya juu kimaandishi, vinginevyo skana huingiza matini katika kompyuta ikiwa na makosa mengi ya othografia, jambo ambalo huhitaji muda mrefu wa uhariri pamoja na programu ya kukagua othografia. Njia ya uchapaji wa kawaida hutumika katika kuingiza lugha ya mazungumzo kwenye maandishi ya kikompyuta.

Kongoo-matini huundwa ili itumike katika shughuli mbalimbali za utafiti wa lugha. Kongoo huweza kutumika katika utafiti wa lugha ama ikiwa bila taarifa zozote za ufafanuzi wa kiisimu (kongoo chasili) au ikiwa imeingizwa taarifa hizo (kongoo fafanuzi). Taarifa za kiisimu zinazoingizwa katika kongoo-matini ni taarifa za kisarufi zilizofafanuliwa kulingana na viwango mbalimbali vya ufafanuzi, kama vile viwango vya ufafanuzi wa mofolojia, sintaksia au semantiki. Fafanuzi hizo huingizwa kwenye kongoo kwa kutumia seti ya alama zilizoteuliwa kuwasilisha taarifa mbalimbali za kisarufi, kwa mfano, taarifa za kategoria za maneno na za viambishi katika ufafanuzi wa mofolojia.

Kongoo chasili yaweza kuwa chanzo cha data ya lugha katika maeneo kadhaa ya utafiti kama vile, utafiti wa kiwango cha urudio wa neno au kirai katika matini na utafiti wa mazingira ya matumizi ya neno au kirai katika matini. Data ya utafiti wa urudio wa maneno au virai katika matini hupatikana kwa kutumia programu maalum ambayo hutambua kila neno katika matini kwa upekee wa umbo lake na kuhesabu idadi ya urudio wake katika matini. Programu hii hutoa orodha ya maneno yote

pamoja na idadi ya urudio wa kila neno kwenye matini (taz. Barnbrook 1996:43-64). Data ya utafiti wa mazingira ya matumizi ya maneno au virai katika matini hupatikana kwa kutumia programu iitwayo konkodansi (taz. Barnbrook 1996:65-85). Konkodansi hutumia ruwaza za mfuatano wa vijenzi vya maneno au virai katika kutambua mazingira ya matumizi ya maneno au virai husika katika kongoo-matini.

2.1 Programu za kufafanua kongoo-matini

Uingizaji wa taarifa za kiisimu kwenye kongoo-matini hufanywa kwa kutumia programu ziliyoundwa mahususi kwa ajili ya kazi hiyo. Kabla ya uingizaji wa taarifa za kiisimu, kongoo hutayarishwa kwa kutumia programu ya kuandaa kongoo-matini kwa ufafanuzi husika. Programu hii hufanya marekebisho mbalimbali katika kongoo, kama vile:

- kubadilisha herufi kubwa kuwa ndogo kwa kuziandika upya kama mfuatano wa alama * na herufi ndogo kama vile, Kilimanjaro→ *kilimanjaro;
- kutenganisha maneno na alama mbalimbali zinazotumika katika maandishi kwa mfano, maandishi, → maandishi;
- kuashiria miisho ya sentensi kwa kuingiza alama maalum, kama vile alama S katika kila mwisho wa sentensi kwenye kongoo;
- kuunganisha maneno virai kama vile, mara kwa mara → mara_kwa_mara, moja kwa moja → moja_kwa_moja; na
- kubadilisha mfuatano mlalo wa maneno katika matini kuwa mfuatano wima ambapo kila neno hukalia msitari moja.

Baada ya kuandaliwa, kongoo-matini huingizwa kwenye programu ya kuingiza taarifa za kiisimu. Programu za kuingiza taarifa za kiisimu kwenye kongoo-

matini huainishwa kulingana na kiwango cha taarifa zinazoingizwa kwenye kongoo kama ifuatavyo:

- Programu ya kuainisha kategoria za maneno katika kongoo-matini. Programu hii hufafanua maneno kwa huchopeka alama ya kategoria ya kila neno kwenye kongoo-matini. Programu hii haiingizi taarifa za kimofolojia, kama vile kategoria za viambishi-ambatizi na viambishi-nyambulishi. Mfano mmoja wa programu ya aina hii ni *CLAWS* (*the Constituent Likelihood Automatic Word-tagging System*) iliyoundwa huko katika Chuo Kikuu cha Lancaster (Garside *na wenzake* 1987).
- Programu ya kufafanua mofolojia kamili ya maneno katika kongoo-matini. Programu hii ina uwezo wa kuingiza alama za kikategoria za maneno na za viambishi. Aidha inaweza kuingiza taarifa za ziada za maneno kama vile taarifa za etimolojia na za matumizi ya neno pamoja na baadhi ya taarifa za kisintaksia. Programu ya aina hii inayojulikana sana kwa sasa ni *Two-level Morphological analyser* (*twol-l*) iliyoundwa huko katika Chuo kikuu cha Helsinki, Finland. Programu hii imeundwa kulingana na kiunzi cha nadharia ijulikanayo kama Mofolojia ya ngazi mbili kilichobuniwa na Koskenniemi (1983).

Programu ya kufafanua kategoria za kisintaksia za maneno kulingana na yalivyotumika katika sentensi kwenye matini. Kimsingi, programu hii hufanya kazi kwa kutumia taarifa za kategoria za maneno zilizokwishafafanuliwa kwenye kongoo-matini. Hivyo, kazi ya programu ya aina hii kwa kiasi fulani hutegemea kazi ya programu za kufafanua kategoria na mofolojia ya maneno.

Kwa kipindi kirefu utafiti wa uchopekaji taarifa za kiisimu kwenye kongoo-matini ulijikita katika ufafanuzi wa kategoria za maneno kwenye kongoo-matini. Programu za kufafanua kategoria za maneno, kwa mfano programu ya CLAWS, ziliundwa kwa kutumia mkabala wa kanuni yumkinifu. Kwa hali hiyo, hazikuwa na uwezo wa kufafanua vipengele vya mofolojia ya lugha. Hata hivyo, kutokana na umuhimu wa taarifa za mofolojia katika data ya baadhi ya lugha, kama vile Kiswahili na Kifini, hivi karibuni kumekuwa na juhudi kubwa za uundaji wa programu zenye uwezo wa kuchambua mofolojia kamili ya kila neno kwenye kongoo-matini. Juhudi hizi zimelenga katika

kutumia mkabala wa kanuni za kisarufi kama msingi wa kuunda programu zenye uwezo wa kuelewa kanuni sarufi ya lugha na kuzitumia katika kuchambua kongoo-matini. Mfano mmoja wa matokeo ya juhudi hizo ni kuundwa kwa programu ya *twol-l* ambayo ina uwezo wa kuchambua mofolojia kamili ya lugha.

Uundaji wa programu za kufafanua kongoo-matini kwa kutumia mkabala wa kanuni za kisarufi huhitaji juhudi za pamoja za wataalamu wa isimu na wataalamu wa programu za kompyuta. Wataalamu wa isimu huandaa kiunzi cha nadharia ya ufafanuzi wa sarufi kwa kuzingatia kanuni za jumla la lugha ya binadamu. Wataalamu wa programu za kompyuta hutumia kiunzi kilichoundwa na wanaisimu kama msingi wa kuunda programu za kuchambua lugha. Kwa kawaida, wataalamu hao hutumia kanuni za kihisabati kuandaa programu hizi katika mtindo wa mashine hatua-ukomo. Hizi si mashine halisi bali za kuwazika ambazo huiwezesha kompyuta kutambua ruwaza za miundo mbalimbali ya vijenzi lugha. Uundaji wa programu hizi huzingatia ruwaza za muundo wa vijenzi lugha kama zilizofafanuliwa katika kiunzi cha nadharia ya sarufi husika. Programu ya mtindo wa mashine hatua-ukomo hufanya kazi kwa kugeuza (kukompaili) kanuni za kawaida za kisarufi, zilizofafanuliwa kulingana na kiunzi husika cha ufafanuzi wa sarufi, kuwa katika muundo wa lugha ya mashine ambao hutumiwa na kompyuta katika kutambua na kufafanua lugha. Programu hizi hufahamika kama kompaila. Kwa hali hiyo, matumizi ya kompyuta katika ufafanuzi wa kongoo-matini hutegemea upatikanaji wa vitu viwili ambavyo ni kompaila na kanuni sarufi ya lugha zilizofafanuliwa kulingana na mahitaji ya kompaila husika. Kazi ya kufafanua sarufi lugha ya kiwango husika cha ufafanuzi, kama vile sarufi ya mofolojia ya Kiswahili, kwa ajili ya ufafanuzi wa kongoo hufanywa na wataalamu wa lugha husika.

2.2 Ufafanuzi wa lugha kwa ajili ya programu ya kufafanua kongoo-matini

Programu zinazofafanua matini kwa kutumia kanuni za kisarufi hufanya hivyo kwa kutumia taarifa za aina mbili za ufafanuzi wa lugha. Hizi ni taarifa za kamusi au leksikoni ya kikompyuta na taarifa za kanuni za kisarufi zilizofafanuliwa kulingana na kiwango husika cha uchambuzi. Taarifa hizi mbili huandaliwa katika mafaili mawili tofauti ya kikompyta. Kazi ya ufafanuzi hufanywa na wataalamu wa lugha husika kulingana na kiunzi cha ufafanuzi kilichotumika katika uundaji wa programu husika. Ufafanuzi hufanywa kulingana na kanuni sarufi ya lugha husika. Mtaalamu wa ufafanuzi hana budi kuwa mweledi mahiri wa sarufi ya lugha husika pamoja na muundo na mahitaji ya proramu ya uchambuzi. Mtaalamu, kwa kuzingatia mahitaji ya programu, huteua alama mbalimbali zitakazotumika katika ufafanuzi, kama vile alama za kuwakilisha kategoria za maneno na za viambishi katika lugha husika. Aidha, mtaalamu huainisha hatua mbalimbali zitakazohusika katika ufafanuzi wa sarufi katika kiwango kinachohusika. Mambo haya huandaliwa kama mpango wa ufafanuzi wa kongoo-matini ya lugha husika (Leech *na wenzake* 1997). Mpango wa ufafanuzi wa kongoo-matini ya lugha huandaliwa kulingana na mahitaji ya programu husika, malengo ya ufafanuzi wa kongoo, na kiwango cha utaalamu wa sarufi ya lugha husika cha muandaaji wa mpango. Kwa hali hiyo, ufafanuzi wa kongoo ya lugha moja unaweza kufanywa kwa kutumia mipango tofauti ya ufafanuzi wa kongoo iliyoandaliwa kwa lengo la kukidhi malengo tofauti ya ufafanuzi. Tofatuti baina ya mipango ya ufafanuzi wa kongoo-matini ya lugha moja yaweza kuwa katika uteuzi wa alama mbalimbali za ufafanuzi wa kisarufi au katika kina cha ufafanuzi. Kwa vile kazi ya uandaaji wa mpango wa ufafanuzi wa kongoo huhitaji muda mwingi, ni vizuri zaidi kama mpango wa ufafanuzi wa kongoo ya lugha moja utaandaliwa katika hali ya kuuwezesha ujumuishe malengo mbalimbali ya ufafanuzi na utumie alama za ufafanuzi zenye ukubalifu mpana miongoni mwa wataalamu wa sarufi ya lugha husika. Maandalizi ya mpango wa namna hii hayana budi yahusishe wataalamu mbalimbali wa sarufi ya lugha husika. Hatua na alama zilizobainishwa kwenye mpango wa ufafanuzi wa kongoo hutumika katika kuandaa faili la leksikoni ya kikompyuta na la kanuni za kisarufi za lugha, mafaili ambayo hutumiwa na proramu katika kuchambua kongoo-matini. Wakati wa ufafanuzi wa kongoo-matini, kompaila hupewa mafaili hayo na kisha huzigeuza fafanuzi za leksikoni na za kanuni za kisarufi zilizomo ndani ya mafaili hayo kuwa katika mtindo wa mashine hatua-ukomo na kuzitumia katika ufafanuzi wa matini. Kwa hali hiyo, programu huingiza taarifa za ufafanuzi kwenye matini kulingana na jinsi taarifa hizo zilivyofafanuliwa kwenye mafaili hayo mawili.

Mfano mmoja wa kazi ya ufafanuzi wa lugha kwa ajili ya programu ya kuchambua matini ni ufafanuzi wa mofolojia ya Kiswahili kwa ajili ya programu ya *twol-l*. Ufafanuzi huo uliofanywa huko katika Chuo Kikuu cha Helsinki, Finland, unafahamika kama SWATWOL '***Swahili Two-Level***' (Hurskainen 1992). Mpango wa ufafanuzi wa SWATWOL umeandaliwa kulingana na kiunzi cha programu ya *twol-l* ambacho kina sehemu mbili: sehemu ya leksikoni na sehemu ya kanuni za kisarufi za kiwango cha mofofonolojia. Alama mbalimbali za ufafanuzi wa mofolojia

ya Kiswahili zimebainishwa katika mpango wa ufafafanuzi wa SWATWOL (taz. Sewangi 2001: 41-47). Alama hizo zimetumika katika ufafanuzi wa taarifa za kimofolojia za aina tatu kama ifuatavyo:

- Alama za ufafanuzi wa kategoria za maneno, kwa mfano, N (jina '*noun*'), V (kitenzi '*verb*'), ADJ (kivumishi '*adjective*'), CC (kiunganishi '*coordinating conjuction*') na GEN-CON (kimilikishi '*genitive connector*'). Idadi kubwa ya alama hizi zimeteuliwa kutoka katika kazi mbalimbali za ufafanuzi wa mofolojia ya Kiswahili zilizoandikwa katika lugha ya Kiingereza.
- Alama za ufafanuzi wa kategoria za viambishi na za vinyambulishi, kwa mfano, 1/2SG (ngeli ya kwanza na pili ya nomino, umoja), REL (alama ya urejeshi kwanye kitenzi '*relative marker in verb*'), OBJ (kiambishi cha shamirisho '*object prefix*'), CAUS:ish (alama ya umbo sababishi '*causative marker*)
- Alama za ziada ambazo hufafanua taarifa nyingine za neno, kama vile taarifa za kietimolojia, kiteminolojia, na baadhi ya taarifa za kisintaksia.

Leksikoni ya kompyuta katika SWATWOL imefafanuliwa kulingana na kanuni za leksikoni ya programu ya *twol-l* kama zilivyobuniwa na Kimmo Koskenniemi (1983). Kulingana na kanuni hizo leksikoni hufafanuliwa kwa kuvunjavunja maneno katika viumbo-mofimu; viumbo-mofimu vyenye sifa moja ya mfuatano huingizwa katika kamusi kama seti ya msamiati. Ufafanuzi wa seti za msamiati hufanywa kwa kuorodhesha vitomeo vya kila seti. Viumbo-mofimu katika seti mbalimbali za msamiati huunganishwa kwa kutumia mbinu ya seti ya viambishi fuatishi. Mbinu hii hutumika wakati wa kufafanua seti za msamiati ambapo kila kitomeo huwa na sehemu kuu mbili: sehemu inayokaliwa na kiumbo-mofimu na sehemu anayokaliwa na seti ya viambishi fuatishi, yaani viambishi ambavyo vinaweza kufuatana na kiumbo mofimu kilichoorodheshwa kwenye kitomeo. Kitomeo kinaweza kuwa na sehemu ya tatu ambayo hukaliwa na fasili ya kiumbo mofimu kilichoorodheshwa. Katika ufafanuzi wa seti za msamiati, sehemu inayokaliwa na kiumbo mofimu yaweza kuachwa wazi lakini sehemu inayokaliwa na seti ya viambishi fuatishi ni lazima ijazwe. Alama '#' hutumika kwenye sehemu ya seti ya viambishi fuatishi kuashiria mwisho wa ufafanuzi wa neno. Ufafanuzi huu umefanywa kwa kuzingatia kanuni za mfuatano wa viumbo-mofimu za maneno ya kategoria mbalimbali, kama vile kategoria ya nomino, au vitenzi, ambapo maneno ya kila kategoria hufafanuliwa yenyewe. Taarifa zinazoingizwa katika sehemu ya kufasili kiumbo mofimu hutolewa na programu kama ufafanuzi wa mofimu mbalimbali kwenye ufafanuzi wa matini. Programu huchambua maneno kwenye matini kwa kutumia mifuatano mbalimbali ya seti za msamiati kama ilivyofafanuliwa katika leksikoni.

Ufafanuzi wa kanuni za kisarufi katika SWATWOL umefanywa kulingana na dhana ya ngazi mbili za neno ambazo zimeainishwa katika kiunzi cha nadharia ya mofolojia ya ngazi mbili kama umbo la neno katika leksikoni (umbo la ndani) na umbo la neno katika othografia (umbo la nje). Katika programu ya *twol-l* kanuni za kisarufi zina jukumu la kuoanisha daraja hizi mbili na kuipa uwezo mashine wa kutambua

na kuchambua maneno katika matini kulingana na ufafanuzi wa mofolojia kwenye leksikoni ya kompyuta. Aidha kanuni hizi huwezesha mashine kusawazisha tofauti yoyote ya kiumbo baina ya daraja hizi mbili za neno. Kwa mfano, 'mu+ana' (katika leksikoni) na 'mwana' (katika maandishi).

Wakati wa ufafanuzi wa matini, faili la leksikoni na la kanuni za kisarufi huingizwa kwenye kompaila ya mtindo wa mashine hatua-ukomo ambayo huzifasiri taarifa katika mafaili hayo katika lugha ya mashine hatua-ukomo na kuzitumia katika kuchambua mofolojia ya maneno katika matini kama inavyoonekana katika ufafanuzi wa matini hii ndogo ifuatayo:

'Baadhi ya magonjwa yanayotokana na kinyesi ni kipindupindu, kuhara, kuhara damu, homa ya matumbo, polio na homa ya ini'

```
"<*baadhi>"
        "baadhi" 9/10-0-SG N AR
        "baadhi" 9/10-0-PL N AR

"<ya>"
        "ya" 3/4-PL GEN-CON
        "ya" 9/10-SG GEN-CON
        "ya" 5/6-PL GEN-CON
        "ya" 5/6-PL

"<magonjwa>"
        "gonjwa" 5a/6-PL N
        "ugonjwa" 11/6-PL N HC

"<yanayotokana>"
        "tokana" 5/6-PL-SP VFIN PR:na 3/4-PL REL V SV SVO STAT REC
        "tokana" 5/6-PL-SP VFIN PR:na 5/6-PL REL V SV SVO STAT REC
        "tokana" 5/6-PL-SP VFIN PR:na 9/10-SG REL V SV SVO STAT REC

"<na>"
        "na" CC @CC

"<kinyesi>"
        "kinyesi" 7/8-SG N HC
        "kinyesi" 9/10-0-SG N
        "kinyesi" 9/10-0-PL N

"<ni>"
        "ni" ADV:ni
        "ni" SG1
"<kipindupindu>"
        "kipindupindu" 9/10-0-SG N HC
        "kipindupindu" 9/10-0-PL N HC
```

```
"<,>"

121
"<kuhara>"
        "hara" INF V AR SV HC

"<,>"

"<kuhara>"
        "hara" INF V AR SV HC

"<damu>"
        "damu" 9/10-0-SG N AR HC
        "damu" 9/10-0-PL N AR HC

"<,>"

"<homa>"
        "homa" 9/10-0-SG N AR HC
        "homa" 9/10-0-PL N AR HC
"<ya>"
        "ya" 3/4-PL GEN-CON
        "ya" 9/10-SG GEN-CON
        "ya" 5/6-PL GEN-CON
        "ya" 5/6-PL

"<matumbo>"
        "tumbo" 5a/6-PL N HC

"<,>"

"<polio>"
        "polio" 9/10-0-SG N HC
        "polio" 9/10-0-PL N HC

"<na>"
        "na" CC @CC

"<homa>"
        "homa" 9/10-0-SG N AR HC
        "homa" 9/10-0-PL N AR HC

"<ya>"
        "ya" 3/4-PL GEN-CON
        "ya" 9/10-SG GEN-CON
        "ya" 5/6-PL GEN-CON
        "ya" 5/6-PL

"<ini>"
```

"ini" 5a/6-SG N HC
"ini" 9/10-0-SG N
"ini" 9/10-0-PL N

"<.S>"

Kama inavyoonekana katika ufafanuzi huu, maneno yaliyofafanuliwa yamezungushiwa alama <> ambapo chini ya kila neno kuna ufafanuzi unaoanza na shina la neno na kufuatiwa na alama mbalimbali. Ifuatayo ni fasili ya alama hizo:

9/10-0-SG = ngeli ya 9/10 umoja
- 9/10-0-PL = nglei ya 9/10 wingi
- 5a/6-SG = ngeli ya 5a/6 umoja
- N = kategoria nomino
- HC = istilahi ya afya
- GEN-CON = kiunganishi kimilikishi
- AR = neno lenye asili ya kiarabu
- CC = kiunganishi
- @CC = kiunganishi katika muktadha wa sintaksia
- INF = alama siukomo
- V = kitenzi
- SV = kitenzi kisokielekezi
- SVO = kitenzi kielekezi
- 5/6-PL-SP = kipatanishi cha ngeli ya 5/6 wingi
- VFIN = kitenzi ukomo
- REL = kirejeshi
- STAT = hali ya kutendeka
- REC = hali ya kutendana
- PR:na = wakati uliopo : na

Karibu maneno yote yana ufafanuzi zaidi ya mmoja, hali ambayo hufahamika kama utata wa kimofolojia. Sababu ya utata wa kimofolojia ni kwamba programu ya *twol-l* huchambua neno kama lilivyofafanuliwa katika leksikoni ya kikompyuta bila kuzingatia mazingira ya matumizi ya neno katika matini. Utata wa namna hii huondolewa kwa kutumia programu nyingine ambayo huondoa fafanuzi zote za neno zilizotolewa na programu ya *twol-l* ambazo hazilingani na mazingira ya matumizi ya neno katika matini. Programu hiyo imeundwa kwa kutumia kiunzi cha nadharia ya sarufi kikwazo iliyobuniwa na Karlson (Karlson *na wenzake* 1995). Kanuni za kisarufi na alama zinazotumiwa na programu hii katika ufafanuzi wa Kiswahili zimeandikwa pia na Hurskainen na zinajulikana kama SWACGP (Swahili *Constraint Grammar Parser*). Kwa mfano, baada ya programu hiyo kuondoa utata katika ufafanuzi wa hapo juu tunapata ufafanuzi ufuatao:

"<*baadhi>"
"baadhi" 9/10-0-SG N AR

"<ya>"
	"ya" 9/10-SG GEN-CON
"<magonjwa>"
	"ugonjwa" 11/6-PL N HC
"<yanayotokana>"
	"tokana" 5/6-PL-SP VFIN PR:na 9/10-SG REL V SV SVO STAT REC
"<na>"
	"na" CC @CC
"<kinyesi>"
	"kinyesi" 7/8-SG N HC
"<ni>"
	"ni" ADV:ni
"<kipindupindu>"
	"kipindupindu" 9/10-0-SG N HC
"<,>"
"<kuhara>"
	"hara" INF V AR SV HC
"<,>"
"<kuhara>"
	"hara" INF V AR SV HC
"<damu>"
	"damu" 9/10-0-SG N AR HC
"<,>"
"<homa>"
	"homa" 9/10-0-SG N AR HC
"<ya>"
	"ya" 9/10-SG GEN-CON
"<matumbo>"
	"tumbo" 5a/6-PL N HC
"<,>"
"<pol'o>"
	"polio" 9/10-0-SG N HC
"<na>"
	"na" CC @CC
"<homa>"
	"homa" 9/10-0-SG N AR HC
"<ya>"
	"ya" 9/10-SG GEN-CON
"<ini>"
	"ini" 5a/6-SG N HC

"<.S>"

Kama inavyoonekana katika ufafanuzi huu, kila neno lina ufafanuzi mmoja tu ambao unaanza na shina la neno na kufuatiwa na alama mbalimbali. Alama zinazofuata baada ya shina la neno ni zile zilizotculiwa katika mpango wa ufafanuzi wa kongoo

wa SWATWOL. Alama hizo zimeingizwa kwenye kongoo kama zilivyofafanuliwa katika leksikoni ya SWATWOL. Alama zilizotumika katika ufafanuzi wa kongoo-matini huwakilisha sifa za aina mbalimbali katika kiwango husika cha ufafanuzi. Kwa mfano katika mfano wa uchambuzi wa mofolojia wa hapo juu, alama 'N' (*Noun*) inawakilisha sifa ya kategoria **Nomino**, alama 'V'(*Verb*) sifa ya kategoria **Kitenzi**, alama 'HC'

(*Health Care*) sifa ya **Istilahi ya Afya**, na alama 9/10-SG inawakilisha sifa ya kiambishi cha nomino za Ngeli ya **9** na **10** katika **Umoja.** Kwa hali hii, seti ya vijenzi lugha katika kongoo iliyofafanuliwa ambavyo vinachangia sifa moja huunda aina moja ya data ya lugha. Kwa mfano, vijenzi lugha vyote ambavyo vinachangia sifa ya kategoria Nomino huunda data ya maneno ya kategoria ya Nomino yaliyoko katika matini. Ni wazi kwamba uainishaji wa data ni rahisi katika kongoo iliyofafanuliwa kuliko katika kongoo chasili. Kazi ya uchopoaji wa data ya lugha katika kongoo-matini hutumia alama zilizotumika katika ufafanuzi wa matini na programu zilizoandaliwa mahususi kwa kazi hiyo.

3. 0 Programu za kuchopoa taarifa kwenye kongoo-matini

Programu za kuchopoa data ya lugha kwenye matini hufanya kazi kwa kutegemea mambo mawili:

- Taarifa zilizofafanuliwa kwenye kongoo-matini. Kama tulivyokwisha sema, taarifa hizi hufafanuliwa kwa kutumia alama zilizoteuliwa kuwakilisha sifa mbalimbali katika kiwango husika cha ufafanuzi wa lugha. Kila alama katika ufafanuzi wa kongoo inapaswa kuwakilisha aina moja tu ya sifa ya ufafanuzi.
- Ruwaza zilizojengwa kwa alama zinazowakilisha sifa za data iliyolengwa kuchopolewa ndani ya kongoo-matini.. Kwa mfano, kama lengo ni kuchopoa data ya seti ya maneno yote ya kategoria Nomino, basi alama 'N' itatumika kama ruwaza ya kuchopolea data ya Nomino kutoka kwenye matini husika. Ruwaza za kuchopolea taarifa zaweza kuundwa kwa kutumia mlolongo wa alama mbili au zaidi zinazowakilisha sifa tofauti katika ufafanuzi wa kongoo. Kwa mfano kama lengo ni kuchopoa data ya vijenzi-lugha vilivyoko katika mfuatano wa kategoria Nomino na kategoria Kitenzi, ruwaza itakayoundwa kwa ajili ya kazi hii itahusisha muungano wa alama za sifa hizi mbili. Uunganishaji hufanywa kwa kutumia alama ya kujumlisha '+'. Kulingana na ufafanuzi wa hapo juu, ruwaza hiyo itakuwa 'N + V'.

Ruwaza za kuchopolea data ya lugha huandaliwa kwenye faili la kikompyuta na kuingizwa kwenye Programu ya kuchopoa data kwenye kongoo-matini. Programu hukompaili ruwaza husika katika lugha ya mashine hatua-ukomo na kuzitumia katika kuchopoa data kulingana na ufafanuzi uliopo katika kongoo-matini. Kwa hali hiyo, programu haiwezi kuchopoa data ambayo sifa zake hazikuwakilishwa kwenye faili la ruwaza. Aidha programu hawezi kuchopoa data kwa kutumia ruwaza yenye alama

yoyote ambayo haikutumika katika ufafanuzi wa kongoo-matini. Kwa mfano, kama katika ufafanuzi wa kongoo-matini alama 'V' imetumika kufafanua sifa ya kategoria kitenzi, halafu alama 'T' ikatumika kwenye ruwaza ya kuchopolea data, basi programu haitachopoa chochote kwenye matini kwani itatafuta maneno yaliyofafanuliwa kwa alama hiyo na kuyakosa. Hivyo basi, uundaji wa ruwaza hauna budi kuzingatia kwa makini aina ya alama zilizotumika katika ufafanuzi wa kongoo.

4.0 Hitimisho

Katika makala hii tumejadili mbinu za kuingiza taarifa za isimu kwenye kongoo-matini na kuzitumia taarifa hizo kama msingi wa kukusanya data ya lugha kutoka kwenye kongoo-matini. Lengo la msingi la mbinu hizi ni kurahisisha ufafanuzi na upatikanaji wa data ya lugha. Mbinu hizi zinajumuisha utaalamu wa kompyuta na wa isimu na zinahitaji maandalizi ya kutosha kwa ajili ya upatikanaji wa kompyuta zenye kasi na uwezo wa kuhifadhi na kuchambua lugha, kongoo-matini kulingana na malengo husika, programu za ufafanuzi wa matini na za uchopoaji wa data, na fafanuzi za taarifa za kisarufi ajili ya programu hizo. Japo mkabala wa kutumia kompyuta katika utafiti wa lugha umekuwa ukishika kasi katika siku za hivi karibuni, mkabala huo bado haujawa na mashiko katika tafiti za lugha za Kiafrika. Hii inatokana na ukosefu wa mambo muhimu yanayohitajika katika mkabala huo kama yalivyobainishwa katika makala hii. Hata hivyo, tayari juhudi zimeanza za kujenga mazingira yatakayoingiza matumizi ya mkabala huu mpya katika utafiti wa lugha hizo. Kama makala hii ilivyobainisha, moja ya juhudi hizo ni kuundwa kwa programu ya ufafanuzi wa mofolojia ya Kiswahili. Programu hii ni muhimu, sio tu katika kurahisisha ufafanuzi wa mofolojia ya Kiswahili bali pia katika kurahisisha uchopoaji wa data ya Kiswahili kwa ajili ya tafiti mbalimbali kama vile tafiti za kamusi na istilahi.

Nyongeza 1: Maelezo ya istilahi

Alama - tag

Hifadhi ya lugha katika kompyuta - Computer text archieve

Isimu kompyuta – computational linguistics

Kiunzi cha nadharia – theoretical framework

Kompaila(programu inayogeuza lugha ya kawaida kuwa lugha ya mashine)- compiler

Kongoo – corpus

Kongoo matini – text corpus

Kongoo ya Taifa la Uingereza - The British National Corpus (BNC)
Kongoo ya lugha maalumu - sample corpus
Kongoo ya lugha jumui - monotor corpus
Kongoo chasili -- raw corpus
Kongoo fafanuzi- annotated corpus

Kuchopoa data – data extraction

Kuchopeka taarifa za kiisimu- marking linguistic information
Leksikoni ya kompyuta – computer lexicon
Mkabala – approach

Mofolojia ya ngazi-mbili – Two-level morpology

Mpango wa ufafanuzi wa kongoo-matini- corpus annotation scheme

Mashine hatua-ukomo - Finite state machine

Programu ya kuandaa kongoo matini – pre-processor
Programu ya kufafanua kategoria za maneno – tagger
Programu ya kufafanua kategoria za kisintaksia - syntactic parser
Programu ya kufafanua mofolojia - morphological analyser
Programu ya kukagua othografia - spell checker
Programu ya kutambua urudio wa maneno – frequency list
Programu ya kutambua mazingira ya matumizi ya maneno/virai -concordance

Programu ya kuchambua kongoo - Corpus annnotation program

Sarufi kikwazo - Constraint grammar

Ruwaza - Pattern/Template

Ruwaza ya kuchopolea data – data extraction pattern/ template
Faili la ruwaza – pattern/template file

Ufafanuzi wa kongoo mtaini- corpus annotation

Uyumkinifu - Probability

Mkabala yumkinifu- probability approach

Marejeo

Banbrook, G. 1996. *Language and Computers*. Edinburgh: Edinburgh University Press.

Frantzi, K. 1998. *Automatic recognition of multi-word terms*. Tasnifu ya Ph.D., Chuo Kikuu cha Manchester Metropolitan, England.

Garside, R., Leech, G., and Sampson, G. 1987. (wahariri), *The computational analysis of English: A corpus-based approach*. London: Longman.

Garside, R., Leech, G., and McEnry, T. (wahariri), 1997. *Corpus Annotation*. London & New York: Longman.

Greenbaum,S., and Yobin, N. 1994. "Tagging the British ICE Corpus". Katika Oostdijk, N. and Haan, P. (wahariri), *Corpus-Based Research into Language, In Honour of Jan Aarts*. Amsterdam-Altlanta, GA, uk. 33-46.

Hurskainen, A. 1992a. "A two-level computer formalism for the analysis of Bantu Morphology: An application to Swahili" Katika *Nordic Journal of African Studies*.1(1): 87-122.

1992b. "Computer Archives of Swahili Language and Folklore – What is it?" Katika *Nordic Journal of African Studies*. 1(1): 123-127.

1995. "Manual for the Computer Archives of Swahili" Muswada.

Karlson, F., Voutilainen, A., Heikkila, J. and Anttila, A. (Wahariri.), 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text.* Berlin & New York: Mounton de Gruyter.

Koskenniemi, K. 1983. *Two-level Morphology: A General Computational Model for Word-Form Recognition and Analysis.* (Machapisho ya Idara ya Isimu, Chuo Kikuu cha Helsinki, No. 11.) Helsinki: University of Helsinki, Department of General Linguistics.

Leech, G. 1993. "Corpus Annotation Schemes." Katika *Literary and Linguistic Computing*, 8(4). Oxford: Oxford University Press, 275-281.

1997a. "Introducing corpus annotation". Katika Garside, R. na wenzake (wahariri.), 1-18.

Leech, G., Garside, R., and Bryant, M. 1994. " The large-scale grammatical tagging of text: Experience with the British National Corpus". Katika Oostdijk, N. and Haan, P. (wahariri), *Corpus-Based Research into Language, In Honour of Jan Aarts.* Amsterdam-Altlanta, GA, uk. 47-64

Lynch, L. 1997. "Medical Terminology Management." Katika Wright, S. E and Budid, G. (wahariri) 1997. *Handbook of Terminology Management, Vol.1, Basic Aspects of Terminology Management.* Amsterdam & Philadelphia: Benjamins Publishing Company.

McEnery,T and Wilson, A. 1996. *Corpus Linguistics.* Edinburgh University Press.

Sager, J.C.1990. *Practical Course in Terminology Processing.* Amsterdam & Philadelphia: Benjamins Publishing Company.

Sewangi, S. S. 2000. "Tapping the Neglected Resource in Kiswahili Terminology: Automatic Compilation of the Domain-Terms From Corpus." Katika *Nordic Journal of African Studies.* 9(2): 60-84.

_____2001. *Computer-Assisted Extraction of Terms in Specific Domains: The Case of Swahili.* Tasnifu ya Ph.D, Chuo Kikuu cha Helsinki.

Sinclair, J. M. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.