SALAMA Dictionary Compiler

Arvi Hurskainen
University of Helsinki

SALAMA Dictionary Compiler takes raw text as input and processes it into a format of a dictionary. The output contains dictionary entries of all words appearing in the input text. The system finds the base form (form needed for a dictionary head-word) of each inflected word in text, analyzes it and finds its correct interpretation. The ability of the system to decide between homonyms (lexical entries with similar form but different meaning) in most cases brings the output close to what it should be in the final dictionary.

The system also identifies various kinds of multi-word expressions (MWE), such as idioms, nominal expressions, adjectival and adverbial expressions, and also proverbs. Currently the system recognizes 2,127 idioms, 2,173 proverbs, and 6,338 multi-word expressions of other types. Some of the MWEs are fixed forms with no inflection, while most of them appear in more than one form. Especially such MWEs that have a verb as a member may have several thousands of forms each. SALAMA is able to handle also such cases.

In dictionary compilation it is not sufficient to list just the head-word with its grammatical information and, as in bilingual dictionaries, with glosses in the target language. Examples of use are important, as well as such special uses of the head-word that cannot be derived directly on the basis of the combination of the words. Idioms are typically such word uses where the dictionary can be of great help. In traditional dictionaries, the treatment of idioms has been generally poor, and if attempted at all, it has not been done in a systematic way. SALAMA makes it possible to find all idioms in text, together with examples of use in context.

For a number of years there have been automatic methods for compiling word-lists on the basis of the base-form of the word. Such lists can be compiled using a morphological analyzer that includes means for disambiguation. These lists have two weaknesses. First, they work only on the word level, and at best include a few frozen word combinations. Second, they do not include automatic and controlled access to the context where they are used.

In SALAMA we have solved both problems. The highly sophisticated system of mastering the description of MWEs solves the first problem. The second problem contains actually several sub-problems. The immediate question is: How do we find examples of use for each head-word, even for rare words? If we have resolved this problem, the next one is facing us: How can we delimit the number of examples, if there are in the corpus for example 16,845 instances where the verb 'acha' occurs? And if we manage to find a way for delimiting them, how can we guarantee that the examples retrieved are good for representing the use of that word?

Now all this is possible, as the extracts below indicate.

So far I have not said anything about frequency information on head-words. In SALAMA system, it is possible, even quite simple, to include such information for each entry, because the frequency count comes automatically when producing the basic head-word list. This information can be retained there all the time or transferred to the dictionary in a later phase. The numerical information can then be rewritten as reader-friendly symbols. The number after each lexical entry in the extracts below indicates how many times the word appears in the corpus.

The Test

The test on the function of the system was made using five fiction books of E. Kezilahabi as a corpus (a total of 196,150 words). These books are part of the Helsinki Corpus of Swahili. The resulting dictionary contains all those words (in base form) that appear in the corpus. Also all examples of use were retrieved from the same corpus. Because the resulting file contains 18,750 text lines, only brief extracts of it are presented here.

Extract 1: General outlay

This extract illustrates the general outlay of the result with default settings. The head-word with required information comes first, surrounded with curly brackets '{' and '}'. If the head-word is a verb, it retains the derived form which it has in text, but also the basic form is given, surrounded by brackets '(' and ')' the

glosses in English are enclosed within curly brackets, with one space on both sides. There may be also some information on derivation (PASS = passive, CS = causative, APPL = applicative, REC = reciprocal form, STAT = stative, etc.). Also information on etymology is indicated (AR = Arabic, PERS = Persian, IND = Indian, ENG = English, PO = Portuguese). In the end there is a number indicating how many times the word appears in the corpus.

The head-word is followed by examples of use in context. The length of the context is a sentence. The head-word is in base form immediately after the inflected form in text, and it is copied also to the beginning of each sentence. Note that the index code, e.g. <GAM>, is also retained in examples. This tells from which book the example was extracted.

In most cases the number of examples is three or less. If the word (in base form) occurs at least three times in the corpus, it has three examples. If it occurs less often, the number of examples decreases accordingly. If the number of examples in text is bigger than three, the rest is cut off. However, this is not the whole truth. There are also head-words with high frequency, and they have examples of common use. This is illustrated in Extract 2 further down.

{ahadi} N 9/10 { promise, pledge, commitment, gage } AR 11
[ahadi] <NAG> Ipo ahadi [ahadi] kwa wanaomtegemea.
[ahadi] <GAM> Ahadi [ahadi] tulizopewa!
[ahadi] <GAM> Kwa nje, watu walikuwa bado wakihudhuria mikutano na maandamano, lakini kwa ndani walitaka Nyerere mwenyewe awahutubie awaambie ni lini watatimiziwa ahadi [ahadi] hizo.

{ahidi} V (ahidi) { promise, engage } AR 5
[ahidi] <GAM> Mwalimu Magafu aliahidi [ahidi] kupita nyumbani kwa Mambosasa baada ya shughuli yake.
[ahidi] <KIC> Mimi nilimwahidi [ahidi] kwamba sitamsahau kamwe.
[ahidi] <KIC> Mwishowe siku ambayo nilikuwa nimeahidi [ahidi] kuwavizia ilifika.

{ahidiana} V (ahidi) { promise, engage } AR REC { each other } 2
[ahidiana] <KIC> Tuliahidiana [ahidiana] kutofarakana hali siku zilikuwa zimebaki chache.
[ahidiana] <ROS> Mwezi waliokuwa wameahidiana [ahidiana] kwenda nyumbani kuwaona wazazi wao ulikuwa umekaribia.

{ahidiwa} V (ahidi) { promise, engage } AR PASS 1
[ahidiwa] <GAM> Miaka mitano ilikuwa imekwishapita, na sasa mioyo yao ambayo ilikuwa ikihesabu kila pigo kwa tumaini la kuona maajabu yaliyoahidiwa [ahidiwa] na Serikali ilikuwa imeanza kuchoka.

{ahirisha} V (ahirika) { postpone, suspend, defer, adjourn, delay, cause to wait, shunt } AR CS 1
[ahirisha] <KIC> Vijana wako tayari kuahirisha [ahirisha] ndoa, lakini wasome, ingawa wana umri wa miaka thelathini au zaidi.

{ahsante} EXCLAM { thank you! } AR 1
[ahsante] <GAM> "Ahsante [ahsante]," walijibu kwa pamoja.

{aibika} V (aibika) { be humiliated } STAT AR 2
[aibika] <MZI> "Hutaaibika [aibika] milele," alisema.
[aibika] <MZI> Hutaaibika [aibika] milele.

{aibisha} V (aibika) { humiliate, embarrass, discredit, disgrace, mortify, put to shame } AR CS 2
[aibisha] <KIC> Usiniaibishe [aibisha] tena.
[aibisha] <MZI> Umeniaibisha [aibisha] mbele ya watu weusi

{aibu} N 9/10 { shame, reproach, scandal, obloquy, compunction, disgrace } AR 28
[aibu] <KIC> " Kwa sababu tulipokuwa mbinguni tuliona kwamba Mungu anatupenda vile kwamba tuliona vigumu kumkosea kwa sababu ya aibu [aibu] kubwa.
[aibu] <KIC> Kwa aibu [aibu] niliona Kalia akimpelekea Baba noti hiyo.
[aibu] <GAM> "Tulifukuzwa kazi," alijibu Mamboleo kwa aibu [aibu].

Extract 2: Frequent contexts

The tag 'fr' attached immediately after the head-word in the example indicates that this is a high-frequency example. These high-frequency examples are located first, and then follow the three examples selected by the default settings.

The head-word {aina} below has several occurrences of frequent context. It appears mostly in the context 'kila aina' in Kezilahabi's text. After frequent contexts, there are three examples randomly selected. It happens that one of them also is 'kila aina'. Other head-words in Extract 2 do not have words with frequent context.

What are the criteria for determining that a word appears often in such a context that this context can be selected as frequent? There are no precise criteria, of course. In this experiment, a context is considered frequent if a sequence of three words before and two words after the key-word appear at least three times in the corpus. The criteria are fulfilled also if there are less than two words after the key-word in the sentence.

Now knowing the criteria, we can look again at the examples of {aina} below. We see that the contexts are in fact 'ya kila aina' and 'za kila aina'.

{aina} N 9/10 { kind, brand, category, genre, order } 67
[aina]fr <GAM> Ananiita majina ya kila aina [aina].
[aina]fr <GAM> Kabla ya Wazungu kufika, nchi hii ilikuwa imejawa na magonjwa ya kila aina [aina].
[aina]fr <GAM> Kulikuwa ngoma za kila aina [aina].
[aina]fr <GAM> Saa tatu na nusu vyungu vya nyama vilikuwa vimeanza kuchemka na maji yalikuwa yakitoa milio ya kila aina [aina].
[aina]fr <GAM> Zamani wanawake walizoea kukoga kisimani au ziwani na kuchekana matako wakati wakitetana na kupiga za kila aina [aina].
[aina]fr <KIC> Niliona kwamba mwanadamu anateremka kama maji kwa kasi sana; na kama maji ya mto yakusanyavyo takataka za kila aina [aina] ndivyo mwanadamu akusanyavyo taka wakati wa maisha yake.
[aina]fr <MZI> Baada ya mwezi mmoja matunda ya kila aina [aina] yalikuwa yameiva.
[aina]fr <MZI> Kiusalama nilitaka iwe ngome ngumu itakayoweza kuhimili mashambulizi ya kila aina [aina] kwa kame nyingi zijazo.
[aina]fr <MZI> Majani ya kila aina [aina] yaliota.
[aina]fr <MZI> Miti ya kila aina [aina] ilianza kuweka majani.
[aina]fr <MZI> Mwili wa ni kiwanda kitegemeacho mali ghafi ya kila aina [aina] ulimwenguni.
[aina]fr <MZI> Wakati huo nilikuwa sifa za kila aina [aina].
[aina]fr <MZI> Watu wengi kijijini walifika kumwomba awapandie michungwa, mpunga, viazi na mazao ya kila aina [aina].
[aina]fr <MZI> Yasemekana aliweza kuponyesha magonjwa ya kila aina [aina].
[aina]fr <NAG> Nilipokuwa narudi nyumbani niliona ngoma za kila aina [aina] zimeanikwa juani.
[aina] <KIC> "Aina [aina] yo yote ya pombe mimi nakunywa.
[aina] <MZI> "Alidai kuwa na uwezo wa kila aina [aina].
[aina] <GAM> "Ukiendelea kumeza vidonge hivi hatutapata aibu ya aina [aina] yo_yote.

{aisee} EXCLAM { hey!, gosh! (I say) } ENG 1
[aisee] <KIC> "Aisee [aisee], hukusikia kwamba kulikuwa watu wengine motoni?

{ajabu} N 5/6 { wonder, uncommon, strangeness, surprising, amazing } AR 50
[ajabu] <GAM> "Hivyo ndiyo ajabu [ajabu].
[ajabu] <GAM> Miaka mitano ilikuwa imekwishapita, na sasa mioyo yao ambayo ilikuwa ikihesabu kila pigo kwa tumaini la kuona maajabu [ajabu] yaliyoahidiwa na Serikali ilikuwa imeanza kuchoka.
[ajabu] <KIC> Ajabu [ajabu] kwangu ilikuwa kuona wingi wa samaki unapita ugali.

{ajali} N 9/10 { accident, fate } 10
[ajali] <MZI> "Ajali [ajali] nyingine haitatokea tena," nilimhakikishia.
[ajali] <MZI> Vipofu na vichaa wataurudisha ulimwengu Katika nyayo zilizofichama gizani Ondoeni mwanga katika bonde la Kupunguza ajali [ajali] na maluweluwe.
[ajali] <NAG> "Ulifika wakati ambapo ajali [ajali] ilikuwa lazima itokee.

{ajiri} V (ajiri) { employ, hire, engage, retain } AR 1
[ajiri] <GAM> "Babu we hukuniajiri [ajiri] wewe.

Extract 3: Multi-word expressions (idioms)

SALAMA is able to catch also multi-word expressions of various types. In Extract 3 we have examples of cases where a verb is a member of the idiom. Examples are from the verb {piga} only. Even in our fairly restricted corpus we find several special uses of this single verb (from a large corpus I have found about 250 in total, and all of them are included in SALAMA). Even these few examples show that it is important to include information and examples of the idiomatic use of words.

{piga} V (piga) { hit, beat } ACT 100
[piga] <KIC> "Wewe ndiye rafiki yangu," alisema huku akinipiga [piga] begani.
[piga] <GAM> Padri Madevu aliendelea mpaka mwisho wa Injili na kisha Wakristo walipiga [piga] shara ya msalaba.
[piga] <GAM> "Kwa nini mlimpiga [piga] mwalimu wa siasa?

{piga_bunduki} V IDIOM-V { shoot } 1
[piga_bunduki] <GAM> Mara moja askari walipiga [piga_bunduki] bunduki na risasi zilipita juu ya vichwa vyao pyuuu pyuuuuu!

{piga_chafya} V ACT IDIOM-V { sneeze } 1
[piga_chafya] <MZI> Masista walipokuwa wamekaribia altare nilipiga [piga_chafya] chafya.

{piga_chenga} V ACT IDIOM-V { dodge, avoid } 2
[piga_chenga] <NAG> Kwa muda wote wa uhai wangu alinipiga [piga_chenga] chenga.
[piga_chenga] <ROS> Stella alipiga [piga_chenga] chenga mbili, tatu huyoo!

{piga_deki} V ACT IDIOM-V { clean a floor } 2
[piga_deki] <KIC> Kesho yake asubuhi nilipokwenda masomoni nilimwacha wangu akipiga [piga_deki] deki nyumba yote.
[piga_deki] <KIC> Mimi nilikuwa sijapata kuipiga [piga_deki] deki hata_siku_moja.

{piga_goti} V ACT IDIOM-V { kneel } 17
[piga_goti] <GAM> Alipiga [piga_goti] goti, akaanza kusali:
[piga_goti] <GAM> Kanisani, watu walikuwa wamejaa Mama Tinda alikuwa amepiga [piga_goti] goti katikati upande_wa wanawake.
[piga_goti] <GAM> Kisha aliinuka, akapiga [piga_goti] goti moja chini na kuhamia katikati kwenye picha ya Uatu Mtakatifu.

{piga_hatua} V ACT IDIOM-V { advance, go ahead } 3
[piga_hatua] <GAM> Baada ya mwezi mmoja vijiji vyote katika wilaya vilikuwa vimepiga [piga_hatua] hatua kubwa katika jitihada ya kuondoa ujinga.
[piga_hatua] <MZI> Nikapiga [piga_hatua] hatua chache kuelekea mlangoni.
[piga_hatua] <MZI> Nilipopiga [piga_hatua] hatua mbili tatu hivi, aligeuka akaanza safari.

{piga_hodi} V ACT IDIOM-V { knock the door } 3
[piga_hodi] <GAM> Mzee Chilongo alikuwa hapigi [piga_hodi] hodi mara mbili bila kuitikiwa.
[piga_hodi] <KIC> Tulipofika karibu na nyumba mimi ndiye nilikuwa wa kupiga [piga_hodi] hodi.
[piga_hodi] <NAG> Nilipofika mlangoni nilipiga [piga_hodi] hodi.

{piga_kelele} V ACT IDIOM-V { make noise } 87
[piga_kelele] "Padri hakumaliza mtu mmoja alipopiga [piga_kelele] kelele.
[piga_kelele] <KIC> Baba alisema, " na kama ukiendelea kupiga [piga_kelele] kelele namna hii nitakushona sasa_hivi kwa mshale
[piga_kelele] <KIC> alipiga [piga_kelele] kelele.

{piga_vigelegele} V ACT IDIOM-V { ululate } 2
[piga_vigelegele] <ROS> Bigeyo alikimbia huku na huko akipiga [piga_vigelegele] vigelegele kwa ulimi.
[piga_vigelegele] <ROS> Walipiga [piga_vigelegele] vigelegele pia.

{piga_kofi} V ACT IDIOM-V { applaud } 10
[piga_kofi] <GAM> Walimu walipiga [piga_kofi] makofi kumsaidia.
[piga_kofi] <GAM> Alianza kufikiri tena jinsi atakavyo wapungia mkono watu wote hao na kelele ambazo zingesikika wakati wakipiga [piga_kofi] makofi.
[piga_kofi] <GAM> Kisha alipiga [piga_kofi] kofi moja kubwa kwa mikono yake miwili kama alama.

{piga_konzi} V ACT IDIOM-V { hit with a fist } 1
[piga_konzi] <MZI> Alinipiga [piga_konzi] konzi kichwani kwa hasira.

{piga_kura} V ACT IDIOM-V { vote } 1
[piga_kura] <GAM> Sasa tupige [piga_kura] kura...

{piga_marufuku} V ACT IDIOM-V { prohibit } 1
[piga_marufuku] <GAM> Badala ya kupiga [piga_marufuku] marufuku vitabu sasa utapigwa wewe

{piga_mguu} V ACT IDIOM-V { walk } 1
[piga_mguu] <ROS> Wao walisimama mbele ya mwanamume waliyempenda na kuchezesha mabega yao pole pole, bila kupiga [piga_mguu] mguu chini, shingo zao upande.

{piga_moyo_konde} V ACT IDIOM-V { be skilful } 1
[piga_moyo_konde] <KIC> Hofu ilinishika lakini nilipiga [piga_moyo_konde] moyo konde.

{piga_mstari} V ACT IDIOM-V { underline } 1
[piga_mstari] <ROS> Rosa alikuwa amepiga [piga_mstari] mstari chini ya maneno yaliyokuwa ya zaidi kwake; mambo aliyokuwa anahitaji wakati huu; na mambo mengine yaliyokuwa yakimsumbua sana moyoni mwake: Tamaa ya mwili; hofu ya kupata mimba; inaondolewa kabisa; furaha ya mwili; staili; na mimba baada ya mimba.

{piga_mswaki} V ACT IDIOM-V { brush the teeth } 5
[piga_mswaki] <GAM> Kisha alipiga [piga_mswaki] mswaki, na alipomaliza alianza mazoezi yake ya kila asubuhi - mazoezi ya kunyoosha viungo.
[piga_mswaki] <KIC> Niliamka nikapiga [piga_mswaki] mswaki haraka haraka.
[piga_mswaki] <KIC> Nilipokuwa nikipiga [piga_mswaki] mswaki niliwaona Kabenga na mkewe wakiongozana kuja nyumbani kwetu.

{piga_mwayo} V ACT IDIOM-V { yawn } 2
[piga_mwayo] <KIC> Stima ilipofika, saa saba kasorobo, nilikuwa nimekwishaanza kupiga [piga_mwayo] mwayo.
[piga_mwayo] <ROS> Tangu siku hiyo mvulana ye yote aliyekuwa akiingia ndani ya chumba chake alitoka nje akipiga [piga_mwayo] mwayo moja kwa moja.

{piga_ngoma} V ACT IDIOM-V { drum } 2
[piga_ngoma] <KIC> Na nitakapokufa ninawaombeni mpige [piga_ngoma] ngoma, mkatangaze kote kijijini kwamba mbwa wa amelamba mchanga
[piga_ngoma] <NAG> Na huo ndio ulikuwa mwanzo wa vikundi vyote vya ngoma kuanza kupiga [piga_ngoma] ngoma zao na kucheza bila kufuata mpango waliouandaa.

{piga_ngumi} V ACT IDIOM-V { box } 2
[piga_ngumi] <GAM> Alimpiga [piga_ngumi] ngumi moja kali na Padri Madevu akaanguka kwa nyuma akiwa bado kwenye kiti chake.
[piga_ngumi] <GAM> Wakati ulipompiga [piga_ngumi] ngumi alijiangusha ili kuchukua chupa hiyo.

{piga_pembe} V ACT IDIOM-V { blow a horn } 1
[piga_pembe] <MZI> Alipoinama kutaka kunipiga [piga_pembe] pembe, nilipiga.

{piga_ripoti} V ACT IDIOM-V { report } 2
[piga_ripoti] <GAM> Akaonywa, na kisha akaambiwa kuwa mara tu turejeapo kutoka hospitali tupige [piga_ripoti] ripoti bomani.

[piga_ripoti] <GAM> Tazama hapa, kuna barua ingine inaulizia lini utapiga [piga_ripoti] ripoti kazini <GAM> Mambosasa alishangaa.

{piga_simu} V ACT IDIOM-V { call } 4
[piga_simu] <GAM> Mkuu wa Wilaya alipiga [piga_simu] simu aletewe faili namba fulani.
[piga_simu] <KIC> Baada ya kupiga [piga_simu] simu nilimwomba mwalimu mkuu wa shule anipe ruhusa ya kupumzika kwa siku hiyo.
[piga_simu] <KIC> Nilirudi shule kupiga [piga_simu] simu.

{piga_soga} V ACT IDIOM-V { pass time by idle talk } 2
[piga_soga] <GAM> Zamani wanawake walizoea kukoga kisimani au ziwani na kuchekana matako wakati wakitetana na kupiga [piga_soga] soga za kila aina.
[piga_soga] <MZI> Siku za Jumamosi wakati tunatoka kanisani kuungama tulipiga [piga_soga] soga hapa na kusimuliana dhambi zetu za kitoto tulizoungama kwa padiri.

{piga_teke} V ACT IDIOM-V { separate, expel, abandon } 3
[piga_teke] <MZI> Aliwapiga [piga_teke] teke na kuangusha watano kabla hajasalimu baada ya kutupiwa kamba shingoni.
[piga_teke] <NAG> Nietzsche atalipiga [piga_teke] teke buku hilo.
[piga_teke] <ROS> Honorata alimrukia Stella na kumpiga [piga_teke] teke.

{piga_mayowe} V ACT IDIOM-V { shout for help } 1
[piga_mayowe] <KIC> Nilikuwa bado sijajifunika niliposikia ndani ya nyumba ya Baba, Rukia na Mama wakipiga [piga_yowe] mayowe: " Tumekufa!

Extract 4: Other multi-word expressions

In Extract 4 there are various other types of MWEs. Especially adjectival expressions are numerous, because Swahili is not very rich in proper adjectives. Expressions requiring qualification are constructed in various ways. Only some of them are exemplified below.

{nchi_endelea} N 9/10 MW-N { developing country } 2
[nchi_endelea] <NAG> "Wengine kutoka nchi [nchi_endelea] zinazoendelea, lakini wengi zaidi kutoka ulimwengu wa.
[nchi_endelea] <GAM> Ninaamini, na ninakubaliana na wataalamu wasemao kuwa msingi wa maendeleo wa nchi [nchi_endelea] zinazoendelea ni kilimo.

{asubuhi_na_mapema} MW>> ADV { early in the morning } 9
[asubuhi_na_mapema] <GAM> Kazi ya kuteka ilianza asubuhi [asubuhi_na_mapema] na_mapema ili kujaza mapipa kumi yaliyokuwa yamewekwa pale shuleni.
[asubuhi_na_mapema] <KIC> Asubuhi [asubuhi_na_mapema] na mapema Kalia alikuja chumbani mwangu.
[asubuhi_na_mapema] <MZI> Alizoea kuwatembelea jirani zake asubuhi [asubuhi_na_mapema] na mapema akimung'unya ubuyu, "Wenye nyumba hamjambo!

{enye_afya} MW> ADJ { bonny } 1
[enye_afya] <ROS> Alikuwa mtoto mwenye [enye_afya] afya nzuri na nyumbani dada zake walimpenda.

{enye_akili} MW> ADJ { clever, cute } 6
[enye_akili] <KIC> Mwanadamu mwenye [enye_akili] akili hufurahi apitishapo moshi puani!
[enye_akili] <KIC> Yote mwanadamu mwenye [enye_akili] akili anafanya kwa ambalo limempumbaza akili.
[enye_akili] Tanzania mnae Rais mwenye sana," alianza Padri Madevu alipokuwa amerudi kutoka chumbani," tena mwenye [enye_akili] akili sana, "alimalizia akikaa" Lakini...

{enye_bidii} MW> ADJ { untiring, unwearied } 3
[enye_bidii] <ROS> Albert alikuwa mtu mwenye [enye_bidii] bidii.
[enye_bidii] <ROS> Kweli alikuwa msichana mwenye [enye_bidii] bidii, kwani kipindi kilipokwisha alikuwa wa katika mtihani.

[enye_bidii] <ROS> Masista walimwita msichana mwema na mwenye [enye_bidii] bidii.

{enye_busara} MW> ADJ { wise } 7
[enye_busara] <ROS> "Uchungaji wenye [enye_busara] busara ni ule unaofikiria hali ya mtu ilivyo - kwamba hali ya mtu haimruhusu akae kwa muda mrefu bila kuzungumza na mtu mwingine asiyefanana naye.
[enye_busara] <GAM> Mzee Magesa ambaye zamani alikuwa akikodisha watu kulima mashamba yake alimtazama mzee Farjalla kwa muda, kisha akasema kwa wasi wasi kidogo, "Ni mtu mwenye [enye_busara] busara.
[enye_busara] Tanzania mnae Rais mwenye [enye_busara] busara sana, " alianza Padri Madevu alipokuwa amerudi kutoka chumbani, "tena mwenye sana," alimalizia akikaa "Lakini...

{enye_chuki} MW> ADJ { angry, uptight, grumpy } 1
[enye_chuki] <GAM> Pikipiki la Padri Madevu lilikwenda kasi likitimua vumbi la baraka yenye [enye_chuki] chuki kati ya wanakijiji kwa Serikali yao.

{enye_furaha} MW> ADJ { glad } 5
[enye_furaha] <GAM> Siku hiyo wanakijiji wengi walikuwa wenye [enye_furaha] furaha kubwa.
[enye_furaha] <KIC> Lakini Vumilia hakuonekana kuwa msichana mwenye [enye_furaha] furaha: alikuwa hajapata kumwona baba yake.
[enye_furaha] <KIC> Nami nilijiona mwenye [enye_furaha] furaha, furaha ambayo nilikuwa sijapata kuwa nayo.

{enye_hasira} MW> ADJ { angry, wroth, cantankerous, fractious, ornery } 3
[enye_hasira] <GAM> Alisema kwa sauti yenye [enye_hasira] hasira kidogo.
[enye_hasira] <GAM> Upande wa wanaume katika viti vya, vijana wawili - Mambosasa na Mamboleo walikuwa wameinamisha vichwa vyao wakisikiliza kwa masikitiko na mioyo yenye [enye_hasira] hasira.
[enye_hasira] <MZI> Alikuwa mtu mwenye [enye_hasira] hasira.

{enye_haya} MW> ADJ { diffident, squeamish, unassuming, demure, coy } 3
[enye_haya] <GAM> Bibi mmoja alisimama bila kupewa ruhusa, akasema kwa sauti yenye [enye_haya] haya :
[enye_haya] <KIC> Rukia ni msichana mwenye [enye_haya] haya sana.
[enye_haya] <NAG> Mgongo wake ulikuwa umejipinda kidogo na macho yenye [enye_haya] haya ambayo yaliingia ndani.

{enye_hekima} MW> ADJ { sententious } 1
[enye_hekima] <KIC> Hata yule mwenye [enye_hekima] hekima hupelekwa maovuni na tamaa.

{enye_huzuni} MW> ADJ { morose, disconsolate, wan } 1
[enye_huzuni] <ROS> Wimbo wenye [enye_huzuni] huzuni pia wanacheza.

{enye_kiburi} MW> ADJ { arrongant, haughty, uppish } 5
[enye_kiburi] <GAM> Hicho ndicho kijiji cha watu wakaidi na wenye [enye_kiburi] kiburi...
[enye_kiburi] <GAM> Leo jua litalainisha vichwa vyenu vyenye [enye_kiburi] kiburi.
[enye_kiburi] <GAM> Ulitua vizuri pale pale nilipoutuma, ukaziba lile tundu lenye [enye_kiburi] kiburi lililokataa kusikia kilio chetu.

{enye_miwani} MW> ADJ { bespectacled } 1
[enye_miwani] <ROS> Huyu hapa mwenye [enye_miwani] miwani ni mzuri lakini ana vidole sita mkononi.

{enye_nguvu} MW> ADJ { strong, upstanding } 12
[enye_nguvu] <GAM> "Yesu mwenye [enye_nguvu] nguvu uliyeponyesha vipofu, wape macho wanaotutendea uovu huu nitakaokwambia.
[enye_nguvu] <GAM> Herufi iliyofuata "c" ilifananishwa na upinde uliovutwa na mtu mwenye [enye_nguvu] nguvu.
[enye_nguvu] <GAM> Huko pwani nilipata kusikia juu ya Fumo Liyongo, mtu mkubwa mwenye
[enye_nguvu] nguvu na aliyeogopwa sana vitani.

{enye_rangi_ingi} MW>> ADJ { versicolour } 1

[enye_rangi_ingi] <NAG> Siku moja ulimfuata kipepeo mwenye [enye_rangi_ingi] rangi nyingi hadi katikati ya msitu.

{enye_sifa} MW> ADJ { preeminent } 1
[enye_sifa] <ROS> Alikuwa akimtafuta mwanamke mwenye [enye_sifa] sifa nzuri kama Rosa.

{enye_sikitiko} MW> ADJ { glum } 4
[enye_sikitiko] " aliniuliza kwa sauti yenye [enye_sikitiko] masikitiko mengi.
[enye_sikitiko] <KIC> "Kazimoto," Tuza aliniita kwa sauti yenye [enye_sikitiko] masikitiko, "mtoto wa...
[enye_sikitiko] <KIC> "Fika tena siku ingine," Matilda alisema kwa sauti yenye [enye_sikitiko] masikitiko.

{enye_sura_zuri} MW>> ADJ { handsome } 1
[enye_sura_zuri] <ROS> Charles alikuwa kijana mwembamba na mrefu kidogo, mweupe na mwenye [enye_sura_zuri] sura nzuri.

{enye_uwezo_kubwa} MW>> ADJ { puissant } 1
[enye_uwezo_kubwa] <NAG> "Binadamu ni kiumbe kilicho katika daraja la kuliko viumbe vyote na chenye [enye_uwezo_kubwa] uwezo mkubwa wa kufikiri.

{kwa_sasa} ADV { currently } 15
[kwa_sasa] <KIC> " Karibu, lakini kwa [kwa_sasa] sasa bado yanaliwa tumbili, " alijibu.
[kwa_sasa] <GAM> Kwa [kwa_sasa] sasa inafaa tufikirie zaidi huduma zinazoweza kutolewa kwa baada ya watu kuhamia vijijini.
[kwa_sasa] <GAM> " Nafikiri jambo hili kwa [kwa_sasa] sasa tunaweza kuliweka pembeni.

{kwa_ufupi} MW> ADV { in brief } 6
[kwa_ufupi] <KIC> Alisema kwa [kwa_ufupi] ufupi, 'Jamani kwa herini, mimi nimechoka nakwenda zangu.
[kwa_ufupi] <KIC> Baada ya kuwaeleza kwa [kwa_ufupi] ufupi walinipeleka hospitali ya Bukonyo.
[kwa_ufupi] <KIC> Kazimoto, kwa [kwa_ufupi] ufupi ni kwamba watu wanaoishi ndani ya nyumba hii hawaoni tena maana ya maisha.

{kwa_uhakika} MW> ADV { certainly } 1
[kwa_uhakika] <KIC> " Haya, nitajaribu, lakini siwezi kusema kwa [kwa_uhakika] uhakika.

{kwa_urahisi} MW> ADV { easily } 13
[kwa_urahisi] <GAM> Kwa_sasa inafaa tufikirie zaidi huduma zinazoweza kutolewa kwa [kwa_urahisi] urahisi baada_ya watu kuhamia vijijini.
[kwa_urahisi] <KIC> " Ujamaa wenyewe ni wa kidogo, na sidhani kwamba utaweza kuuelewa kwa [kwa_urahisi] urahisi hata kama nikikueleza.
[kwa_urahisi] <KIC> Matumaini yetu yote yalifutika kwa [kwa_urahisi] urahisi kama mwalimu afutavyo maandishi yake ubaoni ambayo bado hayajanakiliwa na wanafunzi.

{kwa_utaratibu} MW> ADV { carefully } 1
[kwa_utaratibu] <MZI> Muziki wa kuingilia ulipolia aliingia kwa [kwa_utaratibu] utaratibu bila kucheza.

{kwa_uwazi} MW> ADV { openly } 2
[kwa_uwazi] <MZI> Sauti ilisikika tena, safari hii kwa [kwa_uwazi] uwazi kidogo.
[kwa_uwazi] <NAG> Ng'ambo ya tuliziona nyayo kwa [kwa_uwazi] uwazi zaidi.

{kwa_wasiwasi} MW> ADV { in uncertainty } 11
[kwa_wasiwasi] <GAM> "Ndiyo," Mambosasa alijibu kwa [kwa_wasiwasi] wasiwasi.
[kwa_wasiwasi] <GAM> Aliutazama kwa [kwa_wasiwasi] wasiwasi ule mraba wenye pembe tatu na macho matatu.
[kwa_wasiwasi] <KIC> Kwa [kwa_wasiwasi] wasiwasi tulichukua masanduku yetu na kuielekea ile nyumba.

{kwa_wingi} MW> ADV { in big amounts } 6
[kwa_wingi] <GAM> Makofi yalisikika kwa [kwa_wingi] wingi.
[kwa_wingi] <GAM> Tukitoa misaada maana yake tunataka mlime pamba kwa [kwa_wingi] wingi viwanda vyetu vipate kuendelea!
[kwa_wingi] <KIC> Pombe ya moshi husemekana kupatikana kwa [kwa_wingi] wingi humo milimani.

Extract 5: Proverbs

This extract shows that also proverbs are found if such ones exist in the corpus. Proverbs are given here an explanation in English. There are no equivalent English proverbs. Such a feature could be easily implemented into the system.

{Aliyekutangulia_usimwambie_akupishe} PROVERB>> { The one who went before you do not ask him to let you pass by } 1

{aliyo_nayo} { which he/she has } 1
[aliyo_nayo] <MZI> Wanatuliza kidogo ile hali ya ubwana aliyo_nayo [aliyo_nayo] nayo binadamu.

{Chombo_kilichopikiwa_samaki_hakiachi_kunuka_vumba} PROVERB>>>>> { A vessel that was used for cooking fish does not stop smelling fish } 1

{chomea} V (choma) { burn, roast } APPL 3
[chomea] <GAM> Kuni sasa zilitafutwa - kuni za kuchomea [chomea] nyama.
[chomea] <ROS> '' Kwa_hiyo mlimchomea [chomea] nguo?
[chomea] <ROS> Kama mlimchomea [chomea] nguo zake sioni

{Kipya_kinyemi_ingawa_kidonda} PROVERB>>> { The new one is good although it may be sore } 1

{Mrina_haogopi_nyuki} PROVERB>> { The honey collector does not fear bees } 1

{mrindimo} N 3/4 (rindima) { booming sound, roar } 2
[mrindimo] <NAG> Mrindimo [mrindimo] wa ngoma na kelele za vurumai vilimwanzishia uchungu, akajifungua kitoto cha siku si zake wakati uleule jua lilipopatwa.
[mrindimo] <NAG> Nilisikia tu mrindimo [mrindimo] wake maana sikuusogelea kwa kuogopa nyoka.

{Mvua_inyeshe_leo_na_uyoga_upate_leo} PROVERB>>>>>> { Can it be so that when the rain falls today and you will get mushrooms today? } 1

{mvulana} N 1/2 { boy, bachelor, boyfriend } 72
[mvulana]fr <ROS> ''<ROS> Kweli Rosa, kwa wakati huu, alikuwa bado hataki kuzungumza na mvulana [mvulana] ye_yote.
[mvulana]fr <ROS> Aliwapiga aliwakataza kuzungumza na mvulana [mvulana] ye yote.
[mvulana]fr <ROS> Honorata sasa alichungwa vikali na alizuiwa kuzungumza na mvulana [mvulana] ye_yote.
[mvulana] <GAM> Mbele yao walikuwa wamesimama vijana wawili - mvulana [mvulana] na msichana.
[mvulana] <GAM> Mkutano huu wa ulikuwa wa wazee wote, vijana waliokwishaoa na wavulana [mvulana] wenye umri usiopungua miaka ishirini.
[mvulana] Namna ya kwamba mvulana [mvulana] anakupenda...

Extract 6: Homonyms

A lexical word may have more than one interpretation. These vary from slight shades of meaning of the same part-of-speech category to fundamental differences. How can we make sure that we get examples of use for all these interpretations, provided that there are examples for all types in the corpus? Of course, one solution is to retrieve all examples, but checking them is a tedious job and certainly not practical.

Our experiment shows that by retrieving three examples at maximum, plus frequent contexts if such are found for the word, we only seldom get examples of all meanings. For example, {chungu} has three

interpretations, but all the context are from one interpretation, ‘cooking pot’. In fact only for the word {jua} there are examples for both interpretations, ‘sun’ and ‘know’.

There are three main reasons for not getting examples for all interpretations. First, the retrieval of examples takes place randomly, and only by good luck we get what we want. Second, some interpretations are more common in text than others. Third, less common interpretations are likely not to be represented at all in a small corpus.

{chungu} ADJ A-INFL { bitter, acrid, sour, pungent } 0

{chungu} N 7/8 { cooking pot } 6

{chungu} N 9/10 { common black ant } AN 0
[chungu] <GAM> Saa tatu na nusu vyungu [chungu] vya nyama vilikuwa vimeanza kuchemka na maji yalikuwa yakitoa milio ya kila aina.
[chungu] <GAM> Tinda aliingia jikoni akawasha na kutenga sufuria la maji - chungu [chungu] cha ugali.
[chungu] <KIC> "Mtu unayekula naye chungu [chungu] kimoja anakupita nini?

{chungu_nzima} ADJ A-INFL { a lot, plenty } 2
[chungu_nzima] <GAM> Mifano chungu nzima [chungu_nzima] nzima.
[chungu_nzima] <KIC> Lakini watu hawa ukisema juu ya mtu fulani wataruka pale pale na kusema, 'Huyo ninamfahamu sana, ana madeni chungu [chungu_nzima] nzima'.

{chungwa} N 5/6 { orange } 12

{chungwa} V (chunga) { look after, tend, guard, lead spiritually } PASS 7
[chungwa] <GAM> Aliendelea kuwasifu katika ukuzaji na uzalishaji wa matunda kama machungwa [chungwa] na ndizi.
[chungwa] <GAM> Mashavu yake yalituna kidogo tu na rangi ilifanya kau yake ya kuyapa mviringo wa chungwa [chungwa].
[chungwa] <GAM> Mtoto wao alikuwa amekalishwa juu ya kitenge cha Mwatex, naye alikuwa akila chungwa [chungwa] kwa utulivu.

{dai} N 5/6 { assertion, allegation, plaint, postulant, claim } 2

{dai} V (dai) { claim, demand, assert, state, advocate, profess } AR 49
[dai] <GAM> Mambosasa alifunua yake, akaona kulikuwa kumekaribia kucha, lakini macho yake mazito yalikuwa yakidai [dai] saa moja ya kufunikwa tena, na akili zake zilitaka bado kuelea katika raha ya aliyokuwa nayo dakika chache kabla ya kuamka.
[dai] <KIC> "Kweli, Baba hawezi kudai [dai] hayo yote kama kwamba ananiuza.
[dai] <MZI> "Alidai [dai] kuwa mwanao.

{funza} N 9/10 { jigger, sand flea, larva, maggot, grub } HUM 1

{funza} V (funza) { educate, teach good manners } 10
[funza] <KIC> Itakuwa vigumu kuondoa ubovu huu sasa kwa_sababu kizazi kimoja kinafunza [funza] kizazi kinachofuata.
[funza] <KIC> Nilikuwa nikimfunza [funza] mke wangu kuendesha gari.
[funza] <MZI> Alikiri kuwa maafa hayo yalimfunza [funza] kuwa mtiifu kwa Yeye mwenye uwezo wote, maana hapo zamani alikuwa mtu mwovu.

{jibu} N 5/6 (jibu) { answer, response, reply } 18

{jibu} V (jibu) { answer, respond, reply, react to } AR 260
[jibu]fr <KIC> "Hapana" Tegemea alijibu [jibu]," isipokuwa tunasikia kwamba ni mja mzito sasa .
[jibu]fr <KIC> "Hatukupata, " Tegemea alijibu [jibu] .

[jibu]fr <KIC> Tegemea alijibu [jibu] .
[jibu]fr <KIC> "Hapana " Tegemea alijibu [jibu] .
[jibu]fr <KIC> "Labda mvua, " Tegemea alijibu [jibu] .
[jibu]fr <KIC> "Nzuri," Tegemea alijibu [jibu] .
[jibu]fr <KIC> Tegemea alijibu [jibu] .
[jibu] <KIC> "Ndiyo," nilijibu [jibu].
[jibu] <KIC> "Ndiyo," nilimjibu [jibu].
[jibu] <KIC> "Baba alikwenda kunywa pombe jana, mpaka leo hajarudi," alijibu [jibu].

{jua} N 5/6 { sun } 51

{jua} N 9/10 { sun } 40

{jua} V (jua) { know } 397
[jua]fr <GAM> Kimya ndani ya nyumba , kimya nje ya nyumba ambako paka walikuwa wamelala upenuni na mijusi ukutani wakiota mionzi ya jua [jua] la asubuhi bila habari ya mambo yaliyokuwa yameathiri watu na mazingara yao .
[jua]fr <GAM> Kwa kuwa 'mashujaa' hawa walilala msituni wamejifunika vishuka vyao walihitaji sana mionzi ya jua [jua] la asubuhi asubuhi .
[jua]fr <NAG> Jumba zima lilikuwa sasa liking'ara kwa mionzi ya jua [jua] la asubuhi .
[jua] "Sauti ingine ilisikika kutoka kwenye kitanda kilichokuwa konani - "Unajua [jua], walioleta shida ni wale watekelezaji.
[jua] <GAM> "Nitajuaje [jua] mimi!
[jua] <GAM> "Sijui [jua] D.D.D. mimi sielewi vizuri, lakini kasema ni katika mkumbo huu wa madaraka mikoani.

Extract 7: Homonyms – an advanced solution

Above we retrieved examples using the base form as a key, regardless the meaning of the word. Because any sentence that contained the key-word was accepted as a candidate for retrieving, we retrieved randomly maximum three sentences as examples (plus frequent context sentences if such were available). We can overcome this problem by giving each member of a homonym a separate code and thus treat them as separate head-words.

As we can see below, each meaning of a homonym has at least one example. In this way we can do very fine-grained search of examples. Also this feature in the system saves a lot of very laborious human work.

{chungwa} N 5/6 { orange } 12
[chungwa] <GAM> Aliendelea kuwasifu katika ukuzaji na uzalishaji wa matunda kama machungwa
[chungwa] na ndizi.
[chungwa] <GAM> Mashavu yake yalituna kidogo tu na rangi ilifanya kau yake ya kuyapa mviringo wa chungwa [chungwa].
[chungwa] <GAM> Mtoto wao wa alikuwa amekalishwa juu ya kitenge cha Mwatex, naye alikuwa akila chungwa [chungwa] kwa utulivu.

{chungwa} V (chunga) { look after, tend, guard, lead spiritually } PASS 7
[chungwa] <ROS> Hata kama mkichungwa [chungwa] mimi nafikiri itafaa zaidi kama uchungaji utafanyika kwa busara.
[chungwa] <ROS> Honorata sasa alichungwa [chungwa] vikali na alizuiwa kuzungumza na mvulana ye_yote.
[chungwa] <ROS> Nilichungwa [chungwa] kama wasichana wa Jela.

{funza} N 9/10 { jigger, sand flea, larva, maggot, grub } HUM 1
[funza] <NAG> Kamasi lilianza kujitingisha kama funza [funza] lililopata joto

{funza} V (funza) { educate, teach good manners } 10
[funza] <KIC> Itakuwa vigumu kuondoa ubovu huu sasa kwa sababu kizazi kimoja kinafunza [funza] kizazi kinachofuata.

[funza] <KIC> Nilikuwa nikimfunza [funza] mke wangu kuendesha gari.
[funza] <MZI> Alikiri kuwa maafa hayo yalimfunza [funza] kuwa mtiifu kwa Yeye mwenye uwezo wote, maana hapo zamani alikuwa mtu mwovu.

{jibu} N 5/6 (jibu) { answer, response, reply } 18
[jibu] <MZI> walitazamana na sikupata jibu [jibu].
[jibu] <GAM> "Jibu [jibu] ni rahisi.
[jibu] <GAM> Alipoona hakuna jibu [jibu] alisimama na kutazama nyuma.

{jibu} V (jibu) { answer, respond, reply, react to } AR 260
[jibu]fr <KIC> "Hapana" Tegemea alijibu [jibu]," isipokuwa tunasikia kwamba ni mja_mzito sasa .
[jibu]fr <KIC> "Hatukupata," Tegemea alijibu [jibu] .
[jibu]fr <KIC> Tegemea alijibu [jibu] .
[jibu]fr <KIC> "Hapana ," Tegemea alijibu [jibu] .
[jibu]fr <KIC> "Labda mvua," Tegemea alijibu [jibu] .
[jibu]fr <KIC> "Nzuri," Tegemea alijibu [jibu] .
[jibu]fr <KIC> Tegemea alijibu [jibu] .
[jibu] <KIC> "Ndiyo," nilijibu [jibu].
[jibu] <KIC> "Ndiyo," nilimjibu [jibu].
[jibu] <KIC> "Baba alikwenda kunywa pombe jana, mpaka leo hajarudi," alijibu [jibu].

The size of the source corpus

The above extracts give a rough idea of what SALAMA Dictionary Compiler is able to produce. The extracts are from a fairly small corpus (196,150 words), and as such it is no good for any real dictionary. What is excellent in the system is that the size of the source corpus does not make any difference. The head-word extractor was compiled using a corpus of about 20 million words. In addition, if needed, also words included in SALAMA but not found in the corpus of 20 million words can be included into the extracting module. This done, it is hard to find a word in any Swahili text not found by the system.

It is also obvious that with very large corpora the frequency counts of contexts will become more fine-grained and varied. This helps to find automatically the most representative examples among large masses of example sentences.

Alternative ways of formatting

There are also issues related to the format of the output. In the examples above we have shown one way of formatting the result. Alternative formats are possible. One might think of putting the head-word and all its context examples on the same line - at least temporarily - so that the correct order of entries is retained in sorting. This effect can be achieved also in other ways, of course.

Example:

{mvulana} N 1/2 { boy, bachelor, boyfriend } [mvulana]fr <ROS> Kweli Rosa, kwa wakati huu, alikuwa bado hataki kuzungumza na mvulana [mvulana] ye_yote. [mvulana]fr <ROS> Aliwapiga aliwakataza kuzungumza na mvulana [mvulana] ye_yote. [mvulana]fr <ROS> Honorata sasa alichungwa vikali na alizuiwa kuzungumza na mvulana [mvulana] ye_yote. [mvulana] <GAM> Mbele_yao walikuwa wamesimama vijana wawili - mvulana [mvulana] na msichana. [mvulana] <GAM> Mkutano huu wa ulikuwa wa wazee wote, vijana waliokwishaoa na wavulana [mvulana] wenye umri usiopungua miaka ishirini. [mvulana] Namna ya kwamba mvulana [mvulana] anakupenda...

Or:

{mvulana} N 1/2 { boy, bachelor, boyfriend } [mvulana]fr <ROS> Kweli Rosa, kwa wakati huu, alikuwa bado hataki kuzungumza na mvulana [mvulana] ye yote. --fr <ROS> Aliwapiga aliwakataza kuzungumza na mvulana [mvulana] ye yote. --fr <ROS> Honorata sasa alichungwa vikali na alizuiwa kuzungumza na mvulana [mvulana] ye_yote. -- <GAM> Mbele yao walikuwa wamesimama vijana wawili - mvulana [mvulana]

na msichana. -- <GAM> Mkutano huu wa ulikuwa wa wazee wote, vijana waliokwishaoa na wavulana [mvulana] wenye umri usiopungua miaka ishirini. -- Namna ya kwamba mvulana [mvulana] anakupenda...

Or:

{mvulana} N 1/2 { boy, bachelor, boyfriend }
[mvulana]fr <ROS> Kweli Rosa, kwa wakati huu, alikuwa bado hataki kuzungumza na mvulana [mvulana] ye yote.
--fr <ROS> Aliwapiga aliwakataza kuzungumza na mvulana [mvulana] ye yote.
--fr <ROS> Honorata sasa alichungwa vikali na alizuiwa kuzungumza na mvulana [mvulana] ye yote.
-- <GAM> Mbele yao walikuwa wamesimama vijana wawili - mvulana [mvulana] na msichana.
-- <GAM> Mkutano huu wa ulikuwa wa wazee wote, vijana waliokwishaoa na wavulana [mvulana] wenye umri usiopungua miaka ishirini.
-- Namna ya kwamba mvulana [mvulana] anakupenda...

Length of examples

Also the length of examples should be considered. The sentence length, which we have used in the examples, might be a good approximation and suits in most cases, as far as we are dealing with a dictionary in computer format, where space limitations are not very strict. If we aim at compiling a printed dictionary, the question on the example length becomes important.

Because sentences are sometimes unnecessary long, they should be truncated in some way. This can be done in a number of ways. One method, not an ideal one, is to define the number of words allowed to be on the left and right of the key-word. A better method is to look for natural boundaries for cutting. Such boundaries are at least the semicolon, and perhaps also the colon. The use of the comma as a boundary is not safe, because the comma has so many roles.

There are also sentences consisting of one or two words. Such sentences are obviously not ideal and we could think of removing them. But we have to be careful so that we do not remove the short example of a rare word, in case it occurs only there in the corpus. So we can only remove short examples for words that have a fairly frequent distribution.

Both of these operations - truncating too long examples and removing too short example sentences – have been implemented in this system.

Further considerations

Although SALAMA Dictionary Compiler already does quite a nice job, there is still manual work to be done until the dictionary is ready. Even though the number of context examples can be controlled, and also selection can be done on the basis of frequent contexts, there is no guarantee that the retrieved examples are the best ones. Further checking can be done in two ways, either by increasing the maximum number of examples so big that also the best examples will be found, or by manually going through all sentences where the word occurs.

As we see above, the biggest problem is with homonyms, because each member of the homonym group would need at least one example. Fortunately we have found a method for finding examples for each member of the homonym group. If this fails in some cases, it is due to inaccuracy in the disambiguation system rather than in the retrieving system itself.

A fine-grained treatment of glosses requires an advanced semantic disambiguation system, and this is only partly implemented in SALAMA. In future we hope to be able to select and retrieve examples also on the basis of each gloss. When working with large corpora, this would save a lot of time.

Our experiment was made with a fairly small corpus consisting of five fiction books. The advantages of SALAMA Dictionary Compiler become even more obvious with larger corpora. An undeniable advantage is that the automatic system saves a lot of work done traditionally by human beings. Anybody who has been involved in compiling a dictionary knows this. Another advantage is that a big corpus is likely to contain examples of many meanings of words. Also the list of frequently occurring contexts becomes more fine-grained with a large corpus and it can be used more reliably for retrieving representative examples.

Key to abbreviations:

N = noun
V = verb
ADJ = adjective
EXCLAM = exclamation
9/10 = noun class affiliation of the noun
PASS = passive
S = causative
APPL = applicative
REC = reciprocal form
STAT = stative, neutropassive
AR = Arabic
PERS = Persian
IND = Indian
ENG = English
PO = Portuguese
fr = frequent context
{---} = head-word
(---) = base-form of a verb
[---] = headword and its context sentence
<---> = index code of a book