Technical Reports in Language Technology
Report No 44, 20019
http://www.njas.helsinki.fi/salama

Two methods for accurate information retrieval

Arvi Hurskainen
Department of World Cultures, Box 59
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

This report is an extension to the Technical report No. 36 (2019). It extends the paragraph 7 of that report with new findings in implementing the method. The new invention is the possibility to mark and identify the words found by the system, and make the output more readable by emphasizing the hits. The emphasis can be done in varius ways, such as by surrounding the word with parantheses, or with a special color. The precise marking makes it also possible to make counts of found words, or words with special properties.

Key Words: *information retrieval, morphological analysis, disambiguation.*

1 Introduction

Most current information systems are based on direct string search, where the search target is the normal written text. This search has several limitations, because we normally do not search for surface text form, but rather meaning units. It is often difficult to formulate the search string so that all wanted hits are found, and that the result has only those hits that were wanted. The problem of precision and recall cannot be resolved, if we direct the search to surface text.

The problem is especially big with inflecting languages. The worst cases are such languages, where inflection takes place to both directions, such as Bantu languages.

Problems of precision and recall can be solved, if the text has also a structured representation alongside the surface text. The morphological analysis and syntactic mapping can provide many kinds of structured details of the language. These details, represented as tags, can be integrated into the search system according to need.

In this report, we only use the POS tags as additional features to base forms of the words. This alone will make a huge difference to accuracy of the search.

2 Search method with runtime analysis

Technical Report No. 36 (Hurskainen 2019) describes the search method, where the runtime analyser formulates the entered search word into the format needed for performing the actual search. For example, the user may type the word *perustuslaissakaan*, and the analyser changes it into the form *{perustuslaki_N}*. Any surface form of the word will be rendered into the same base form.

The target of the search is the analysed and disambiguated text, where each surface word is followed by its base form and its POS category, for example *perutuslaissakaan*

{perustuslaki_N}. If the size of the context is defined as a sentence, each word of the sentence gets its own copy of the sentence, and each word in turn gets its base form displayed. This is demonstrated in (1).

(1)
```
2_§_3 Julkisen {julkinen_A} vallan {valta_N} käytön {käyttö_N}
tulee {tulla_V} perustua {perustua_V} lakiin {laki_N}.
```

Each word in (1) is represented as its surface form and base form. Out of this we produce as many copies as needed for each word to have its own analysed representation (2).

(2)
```
2_§_3 Julkisen {julkinen_A} vallan käytön tulee perustua lakiin.
2_§_3 Julkisen vallan {valta_N} käytön tulee perustua lakiin.
2_§_3 Julkisen vallan käytön {käyttö_N} tulee perustua lakiin.
2_§_3 Julkisen vallan käytön tulee {tulla_V} perustua lakiin.
2_§_3 Julkisen vallan käytön tulee perustua {perustua_V} lakiin.
2_§_3 Julkisen vallan käytön tulee perustua lakiin {laki_N}.
```

Note that the runtime analysis uses the same analysis package as was used for producing the analysed target text.

This search method is particularly helpful in languages, where the word may inflect in both ends. Verbs in Bantu languages are particularly problematic, because the verb may have only a single consonant as a verb stem, and all the rest is inflection or derivation. For example, in the verb *wa+li+o+tu+f+i+a*, only *f* is constant and means 'to die'. The structure means *'those who died on our bahalf'*. In this system, if the user enters the word *waliotufia*, the system converts it to the form *{fa_V}* or *{fia_V}*, depending on whether extended verbs are returned into base form or left into extended form. Also the very common verb 'to be' is monosyllabic *w*.

In this search system, the found word is displayed in surface form and in its base form, for example, waliotufia *{fa_V}*, *perustuslaissa {perustuslaki_N}, lailla {laki_N}*. Note that the system requires that compound words such as perustuslaki must be typed in some form of the compound word. Merely the form *laissa* is not enough. It produces only the search word *{laki_N}*.

The system could be implemented so that the form of the short word would also find such strings, where this is the last part of a compund word, but by so doing we would lose accuracy.

The strength of the method is that it makes the use simple. The user does not need to bother what is the correct search form. It is only needed that the search form is unambiguously some form of the searched word.

The disadvantage is that partial words cannot be searched. On the other hand, there is less need for searching with partial strings, because the system is accurate.

Another disadvantage is that the length of the target text will be multiplied by the number of words in each sentence, or any other record if the search unit is not a sentence.

3 Search method without runtime analysis

In this second search method, the text is first analysed as in the system above. The difference is here that we do not need to make a new copy for each word of the sentence. The search is directed to the analysed text, such as in (3).

(3)
```
2_§_3 Julkisen {julkinen_A} vallan {valta_N} käytön {käyttö_N}
tulee {tulla_V} perustua {perustua_V} lakiin {laki_N}.
```

When we do not use the runtime analyser, which always forces the search word into a certain format, we are free to use any kinds of search keys, such as full surface words, partial surface words, full base forms, and partial base forms.

We first retrieve all such lines, where the string occurs. Then we remove the base forms so that only the surface text is left. The result is clean text with all those sentences, where the search string occurs.

The setback in this method is that although the search result can be covering and accurate, the hit is not displayed in any way, and finding hits in result text is not convenient. This used to be the case, until we, together with colleagues at the Helsinki University, Jyrki Niemi and Anssi Yli-Jyrä, found out that it is possible to mark the hit by surrounding the hit with color codes in *egrep*. When the color codes are added to the hit, those codes can then be rewritten and moved to appropriate places for producing desired marking of the hits.

This invention opened almost limitless possibilities for designing search methods.

a. Now it is possible to indicate that you want to target the base form only, by including the underscore into the search string, e.g. *laki_*. Or, in ambiguous cases, if you want to be sure that you will find only the words of the certain POS category, you can add also the POS tag, *laki_N*.

b. With the search key described in (a), you will get also such words as *{perustuslaki_N},* and all nouns ending in *laki*. You can restrict this by marking the beginning of the word, *{laki_N*.

c. If you omit all such features that occur in base forms, such as *{*, *}*, _ or any POS tag, the search is directed to the surface words and base forms alike. In this case it is not always sure that the hit is marked in any way, because there is ambiguity in marking, if both representeations of the word could be marked. Typical cases are words, that are in base form in text. However, this method is useful in cases, where no ambiguity occurs, and the hit will be surrounded with curly braces even in the case of surface forms.

d. The search of partial words is also possible. The search string can be cut with the Kleene star '*'. This facility is implemented so that if you cut the word in the end, the whole beginning part of the word must be typed. In the same way, if you cut the word from the beginning, the whole last part must be typed.

e. Boolean operators AND and OR can be used for searching more than one word. Several words can be typed in the search strings, with one of the operators between each two words. Operators can be mixed in the same search. Even in these combined searches, the hit can be surrounded with curly braces. However, this does not happen in such cases, where there are more than two words combined with the AND operator. The middle hit

will not be marked, because the color code marker marks the beginning and end of the whole search string, and the middle word will be left without color code marking. In using the Boolean operators, it is better to use base form description for ensuring that the relevant words will be marked.

In brief, the search process has the following steps:

a. Find the lines, which match the search string.

b. Mark the hit with color codes, one code before the hit, and another code after the hit.

c. Convert the left code into the desired marker (different from curly braces) and move it to the appropriate place, usually the beginning of the word.

d. Convert the right code into the desired marker and move it to the appropriate place, usually the end of the word.

e. Remove the words surrounded with curly braces. This removes the base forms of all words except for the hit word.

f. Convert the surrounding codes of the hit words into curly braces. Now the output is ready. The hit is surrounded with curly braces. This applies to surface words as well as the base forms alike.

4 Producing statistics

When the hit is marked in text, it can be kept separate from other words. This makes it possible to produce various statistical lists. In (4) are examples of such searches, when the target text is Suomen Perustuslaki (Constitution of Finland).

(4)
Search string: laki_
Output:
```
  4 itsehallintolaki_N
  3 kirkkolaki_N
142 laki_N
  1 maakuntalaki_N
  1 maanhankintalaki_N
 33 perustuslaki_N
```

Search string: hallinto_ **AND** laki_
Output:
```
 3 hallinto_N
 1 itsehallinto_N
 6 laki_V
 1 paikallishallinto_N
 1 valtionhallinto_N
```

Search string: hallinto_ **OR** laki_
Output:
```
12 hallinto_N
 4 itsehallintolaki_N
```

5 itsehallinto_N
1 keskushallinto_N
3 kirkkolaki_N
142 laki_N
1 maakuntalaki_N
1 maanhankintalaki_N
1 paikallishallinto_N
33 perustuslaki_N
2 valtionhallinto_N

Search string: hallinto_ **OR** laki_ **OR** sääntö_
Output:
12 hallinto_N
4 itsehallintolaki_N
5 itsehallinto_N
2 johtosääntö_N
1 keskushallinto_N
3 kirkkolaki_N
142 laki_N
1 maakuntalaki_N
1 maanhankintalaki_N
1 ohjesääntö_N
2 oikeussääntö_N
1 paikallishallinto_N
33 perustuslaki_N
2 valtionhallinto_N
4 valtiosääntö_N

It is also possible to produce statistical lists of various POS categories in text, such as in (5-7).

(5)
Search string: **N**
Output:

1 aihe_N
4 aika_N
4 ajankohta_N
3 ala_N
1 alijäämä_N
1 alku_N
1 alkuperäiskansa_N
1 alkuperä_N
9 aloite_N
1 aloiteoikeus_N
1 aluejaotus_N
4 alue_N
1 ammatti_N
1 antaja_N
4 apulaisoikeusasiamies_N
4 apulaisoikeuskansleri_N
1 armahdus_N
1 arpa_N
2 arviomääräraha_N
1 arvio_N
2 asema_N
11 asetus_N
1 asevelvollisuuden_N
5 asiakirja_N
72 asia_N
2 asianomainen_N
3 asiasisältö_N
1 asuinpaikka_N
1 asukas_N
1 asunto_N
2 edellytys_N
168 eduskunta_N
2 eduskuntaryhmä_N
5 eduskuntatyö_N
13 eduskuntavaali_N
3 edustaja_N
12 edustajantoimi_N
1 edustusto_N
1 ehdokaslista_N
11 ehdokas_N
23 ehdotus_N
2 ehto_N
1 elämä_N
1 elinkeino_N
1 elinkeinotoiminta_N

2 elinympäristö_N
9 enemmistö_N
2 enimmäismäärä_N
1 ennakko_N
2 epäluottamuslause_N
1 erityistuomioistuin_N
4 ero_N
1 erotus_N
1 esittelijä_N
2 esittely_N
12 esitys_N
1 etu_N
1 etusija_N
2 eurooppa-neuvosto_N
1 hakemus_N
1 hallintoalue_N
1 hallintoasia_N
1 hallintolainkäyttöasia_N
12 hallinto_N
9 hallinto-oikeus_N
1 hallintotehtävä_N
1 hallintotuomioistuin_N
1 hallitusmuoto_N
11 hallitus_N
1 hallitusohjelma_N
1 hallitusvalta_N
3 haltu_N
2 häntä_N
1 havainto_N
1 heinäkuu_N
7 henkilö_N
1 henkilöstö_N
1 henkilötieto_N
3 hoito_N
2 hovioikeus_N
2 huolenpito_N
1 huolimattomuus_N
1 huoltaja_N
1 huolto_N
4 huomautus_N
2 huomio_N
1 hyökkäys_N
1 hyvä_N
1 hyvinvointi_N
1 ihminen_N
3 ihmisarvo_N
5 ihmisoikeus_N
1 ihmisoikeussopimus_N
1 ihmisoikeusvelvoite_N
1 ihmisyys_N
1 ikä_N
3 ilma_N
3 ilmoitus_N
3 irtisanominen_N
1 isänmaa_N
3 istunto_N
4 itsehallintolaki_N
5 itsehallinto_N
1 jälkeinen_N
2 jaos_N
3 järjestö_N
2 järjestysmuoto_N
3 järjestys_N
3 jäsenmäärä_N
32 jäsen_N
1 johto_N
2 johtosääntö_N
1 joki_N
1 julkaisemispäivä_N
2 julkisuus_N
3 julkisyhteisö_N
1 käännösapu_N
7 kannanotto_N
1 kansakunta_N
14 kansalainen_N
1 kansalaisaloite_N
1 kansalaiskunto_N
4 kansalaisuus_N
1 kansalliskieli_N
3 kansa_N
5 kansanäänestys_N
34 kansanedustaja_N
1 kansaneläkelaitos_N
1 kansanvalta_N
1 kanslia_N
3 kanta_N
1 käräjäoikeus_N
37 käsittely_N
1 kasvu_N
1 käytävä_N
1 käyttöehto_N
3 käyttö_N
4 käyttötarkoitus_N
1 kehitys_N
3 kerta_N
7 kertomus_N
1 keskushallinto_N
9 keskustelu_N
1 keskuus_N
1 kidutus_N
11 kieli_N
2 kirjelmä_N
1 kirje_N
3 kirkkolaki_N
1 kirkko_N
2 kohde_N
2 kohta_N
1 kohtelu_N
7 koko_N
5 kokoonpano_N
1 kokoontua_N
1 kokoontumispäivä_N
1 kokoontumisvapaus_N
5 kokous_N
9 kolmasosa_N
2 korkein_N
1 korvaus_N
2 koskemattomuus_N
3 kotirauha_N
1 kotiseutualue_N
1 koulutus_N
2 kulttuuri_N
1 kulttuuriperintö_N
1 kumma_N
1 kunnallisvaali_N
1 kunnia_N
1 kunnioitus_N
1 kuntajako_N
8 kunta_N
1 kuolemanrangaistus_N
1 kutsu_N
2 kuudesosa_N
4 kuukausi_N
1 kuvaohjelma_N
2 kyky_N
4 kysymys_N
1 laatu_N
2 laillisuus_N
3 laillisuusvalvonta_N
1 laiminlyönti_N
2 lainanotto_N
1 lainkäyttöelin_N
2 lainkäyttö_N
6 lainmukaisuus_N
3 lainsäädäntö_N
1 lainsäädäntötoimi_N
1 lainsäädäntövalta_N
2 laintuntija_N
1 lainvalmistelu_N
1 lainvastaisuus_N
1 laitos_N
2 lakialoite_N
12 lakiehdotus_N
142 laki_N
1 lakka_N
6 lapsi_N
1 lausuma_N
10 lausunto_N
1 liikekannallepano_N
3 liikelaitos_N
1 lisätalousarvioesitys_N
1 lisätalousarvio_N
1 loukkaamattomuus_N
1 lukumäärä_N
1 luonnos_N
1 luopua_N
7 luottamus_N
2 lupa_N

1 maakuntalaki_N
3 maakunta_N
1 maakuntapäivä_N
1 maaliskuu_N
12 maa_N
1 maanhankintalaki_N
1 maanpetosrikos_N
1 maanpuolustus_N
1 maanpuolustusvelvollisuus_N
3 määräaika_N
1 määräenemmistö_N
1 määräikä_N
7 määrä_N
15 määräraha_N
3 määräys_N
3 määräysvalta_N
1 määrittely_N
5 mahdollisuus_N
1 maksullisuus_N
1 maksu_N
1 marraskuu_N
1 menestys_N
12 menettely_N
1 menettelytapa_N
1 menetys_N
4 meno_N
3 merkitys_N
1 mielenosoitus_N
5 mielipide_N
7 mietintö_N
1 ministeriaika_N
32 ministeri_N
14 ministeriö_N
1 ministerivaliokunta_N
3 momentti_N
1 monimuotoisuus_N
5 muutos_N
3 muuttamaton_N
1 muutto_N
1 neljäsosa_N
1 neuvottelu_N
1 niisi_N
1 nimitysperuste_N
4 noja_N
1 ohjelma_N
1 ohje_N
1 ohjesääntö_N
5 oikeudenkäynti_N
1 oikeudenloukkaus_N
1 oikeudenmukaisuus_N
18 oikeusasiamies_N
21 oikeuskansleri_N
68 oikeus_N
2 oikeussääntö_N
4 oikeusturva_N
3 omaisuus_N
3 omatunto_N
1 omistus_N
3 opetus_N
1 oppilaitos_N
1 oppivelvollisuus_N
1 osakasvalta_N
9 osa_N
3 päällikkö_N
2 pääministeriehdokas_N
18 pääministeri_N
2 pää_N
2 pääte_N
3 päätöksenteko_N
34 päätös_N
1 paikallishallinto_N
11 päivä_N
1 pakkolunastus_N
1 palkkaus_N
2 palvelu_N
2 palvelussuhde_N
1 palvelutavoite_N
3 pankki_N
2 pankkivaltuutettu_N
2 parlamentti_N
1 perhe_N
6 perusoikeus_N
2 perusopetus_N
2 perustelu_N
27 peruste_N
1 perustoimeentulo_N
1 perustuslainmukaisuus_N
33 perustuslaki_N
12 perustuslakivaliokunta_N
1 piiri_N
2 poikkeus_N
2 poikkeusolo_N
3 pöytäkirja_N
2 presidentintoimi_N
69 presidentti_N
2 puheenjohtaja_N
1 puheenvuoro_N
1 puhelu_N
14 puhemies_N
7 puhemiesneuvosto_N
1 puhe_N
4 puoli_N
2 puolue_N
1 puolustus_N
2 puolustusvoima_N
1 puute_N
1 pykälä_N
4 pyyntö_N
3 rahasto_N
1 raja_N
3 rajoitus_N
5 rangaistus_N
3 ratkaisuehdotus_N
1 ratkaisuvalta_N
1 rauha_N
1 rauhanen_N
1 riita_N
1 rikosasia_N
10 rikos_N
2 ristiriita_N
1 romani_N
6 ruotsi_N
1 ryhmä_N
3 säädöskokoelma_N
2 säädös_N
4 saamelainen_N
1 saame_N
10 säännös_N
2 sääntely_N
4 säätämisjärjestys_N
1 sairaus_N
2 salaisuus_N
1 sala_N
3 sananvapaus_N
1 selitys_N
4 selonteko_N
7 selvitys_N
1 seuraamus_N
1 sidonnaisuus_N
3 siirtomääräraha_N
3 siirto_N
6 sijainen_N
1 sisältö_N
4 sopimus_N
1 sosiaali_N
1 sota_N
1 sotilasvirka_N
2 soveltamisala_N
4 suhde_N
2 sukupuoli_N
1 suoja_N
1 suojelu_N
52 suomi_N
9 suostumus_N
2 suuruus_N
2 syntymä_N
14 syy_N
17 syyte_N
1 syyteoikeus_N
2 syyttäjälaitos_N
1 syyttäjä_N
1 tae_N
1 tahto_N
1 taide_N
1 taito_N
1 takuu_N
2 tallenne_N

2 taloudenhoito_N
2 talousarvioaloite_N
5 talousarvioesitys_N
26 talousarvio_N
2 talous_N
1 tammikuu_N
3 tapa_N
2 tapaus_N
3 tarkastusvaliokunta_N
3 tarkastusvirasto_N
7 tarve_N
1 tarvitseva_N
1 tasa-arvo_N
35 tasavalta_N
21 täysistunto_N
2 täysivaltaisuus_N
3 täytäntö_N
1 täytäntöönpaneminen_N
1 täytäntöönpano_N
23 tehtävä_N
1 tekeminen_N
2 tekohetki_N
1 teko_N
1 terveydentila_N
1 terveys_N
1 terveyspalvelu_N
1 tiede_N
3 tiedonanto_N
19 tieto_N
1 tilaisuus_N
1 tila_N
1 tilinpäätös_N
2 toimeentulo_N
1 toimenpidealoite_N
8 toimenpide_N
8 toimiala_N
15 toimielin_N
7 toimikausi_N
6 toimi_N
19 toiminta_N
1 toimintatavoite_N
1 toimituskirja_N
4 toimivalta_N
2 toteutuma_N
1 totuus_N
1 tuki_N
1 tulkitsemis_N
4 tuloarvio_N
3 tulo_N
2 tulos_N
6 tuomari_N
1 tuomioistuinlaitos_N
18 tuomioistuin_N
1 tuomio_N
3 tuomiovalta_N
3 turvallisuus_N
2 turvallisuuspolitiikka_N
1 turvallisuustarkastus_N
1 turva_N
2 tutkinta_N
1 tutkintapyyntö_N
1 työelämä_N
19 työjärjestys_N
1 työkyky_N
1 työkyvyttömyys_N
1 työllisyys_N
4 työ_N
2 työntekijä_N
2 työskentely_N
1 työttömyys_N
1 työvoima_N
7 ulkoasiainvaliokunta_N
2 ulko_N
1 ulkopolitiikka_N
1 umpilippu_N
12 unioni_N
1 upseeri_N
5 uskonto_N
6 vaalikausi_N
2 vaalikelpoisuus_N
1 vaalimääräys_N
20 vaali_N
3 vaalipiiri_N
1 vaalitapa_N
1 vaatimus_N
3 väestö_N
1 vahingonkorvaus_N
1 vahinko_N
2 vahvistus_N
1 vaihde_N
1 vaihtoehto_N
2 vaiteliaisuus_N
3 vakaumus_N
1 vakuutus_N
3 välikysymys_N
1 väli_N
1 valinta_N
44 valiokunta_N
7 valmistelu_N
1 valmiste_N
7 valtakunnanoikeus_N
4 valtakunnansyyttäjä_N
4 valtakunta_N
13 valta_N
33 valtio_N
93 valtioneuvosto_N
2 valtionhallinto_N
1 valtiontakaus_N
1 valtiontakuu_N
4 valtiontalous_N
1 valtionvelka_N
1 valtiopäiväasiakirja_N
2 valtiopäiväjärjestys_N
17 valtiopäivä_N
1 valtiopetosrikos_N
4 valtiosääntö_N
2 valtiosopimus_N
1 valtiovalta_N
3 valtiovarainvaliokunta_N
3 valtuus_N
1 valtuutettu_N
1 valtuutus_N
1 valvoa_N
1 valvontahavainto_N
2 valvonta_N
2 vammaisuus_N
1 vanhin_N
1 vanhuus_N
2 vankeus_N
3 vapaudenmenetys_N
10 vapaus_N
3 vapautus_N
1 varaedustaja_N
7 varainhoitovuosi_N
1 varajäsen_N
1 varallisuus_N
3 varapuhemies_N
1 varattomuus_N
1 varoitus_N
1 vastata_N
1 vastaus_N
3 vastuu_N
10 velvoite_N
6 velvollisuus_N
3 vero_N
1 verotusoikeus_N
2 verovelvollisuus_N
2 viesti_N
1 viikko_N
1 viittomakieli_N
3 viivytys_N
18 viranomainen_N
1 virasto_N
1 viri_N
1 virkaikä_N
6 virkamies_N
8 virka_N
1 virkasuhde_N
17 virkatoimi_N
1 virkavuosi_N
2 voimaansaattaminen_N
2 voimaantuloajankohta_N
1 voimaantulo_N
14 voima_N
10 vuosi_N
4 yhdistymisvapaus_N
3 yhdistys_N
1 yhdyskunta_N

1 yhteensovittaminen_N
4 yhteiskunta_N
1 yhteistoiminta_N
1 yhteistyö_N
2 yhtiö_N
1 yksikkö_N
6 yksilö_N
1 yksityiselämä_N
4 yleisistunto_N
1 ylijäämä_N
1 yliopisto_N
1 ylipäällikkö_N
1 ylipäällikkyys_N
2 ympäristö_N
1 yritys_N
1 äänestäjä_N
1 äänestysmenettely_N
9 äänestys_N
17 ääni_N
1 äänioikeus_N

(6)
Search string: **V**
Output:

1 aiheutua_V
4 ajaa_V
3 alkaa_V
1 ansaita_V
50 antaa_V
1 armahtaa_V
1 arvioida_V
12 asettaa_V
7 asioida_V
3 asua_V
1 avustaa_V
2 edellyttää_V
1 edeltää_V
7 edistää_V
3 edustaa_V
1 ehdottaa_V
1 elää_V
1 epäillä_V
1 erota_V
2 erottaa_V
1 esiintyä_V
5 esitellä_V
4 esittää_V
4 estää_V
1 estyä_V
1 haitata_V
1 hakea_V
6 hankkia_V
2 harjoittaa_V
2 havaita_V
22 hoitaa_V
4 huolehtia_V
1 huomauttaa_V
1 hylätä_V
23 hyväksyä_V
5 hyväksyttää_V
3 iätä_V
2 ilmaista_V
2 ilmetä_V
7 ilmoittaa_V
1 jakaantua_V
1 jakaa_V
11 järjestää_V
1 järjestäytyä_V
3 jatkaa_V
3 jatkua_V
5 jättää_V
5 johtaa_V
5 julistaa_V
8 julkaista_V
1 julkistaa_V
10 kannattaa_V
1 kantaa_V
2 karkottaa_V
1 kärsiä_V
21 käsitellä_V
8 katsoa_V
2 kattaa_V
1 kaventaa_V
2 käydä_V
10 käyttää_V
4 kehittää_V
2 keskeytyä_V
3 keskustella_V
1 kiduttaa_V
2 kieltää_V
2 kieltäytyä_V
1 koetella_V
2 kohdella_V
2 kohdistua_V
2 kokoontua_V
7 koskea_V
4 kulua_V
2 kumota_V
2 kuolla_V
2 kutsua_V
4 kuulla_V
19 kuulua_V
1 kyetä_V
1 laajentaa_V
1 laatia_V
2 lähettää_V
1 lähteä_V
1 laiminlyödä_V
1 lakata_V
4 levätä_V
1 liikkua_V
3 liittyä_V
2 lopettaa_V
2 loukata_V
5 luovuttaa_V
9 määrätä_V
1 määräytyä_V
1 mahdollistaa_V
4 menetellä_V
2 menettää_V
3 merkittää_V
1 merkityttää_V
2 muistaa_V
3 muodostaa_V
11 muuttaa_V
1 muuttua_V
7 myöntää_V
5 nauttia_V
1 neuvotella_V
17 nimittää_V
9 nostaa_V
15 noudattaa_V
1 oleskella_V
1 omata_V
18 osallistua_V
2 osoittaa_V
15 ottaa_V
32 päättää_V
2 päättyä_V
1 palauttaa_V
2 palautua_V
5 panna_V
2 perustaa_V
5 perustua_V
3 peruuttaa_V
3 pidättää_V
1 pidättäytyä_V
2 pitää_V
2 poiketa_V
1 puhua_V
1 puuttua_V
3 pyrkiä_V
1 pysyä_V

1 pyytää_V
1 rajata_V
2 rajoittaa_V
4 ratkaista_V
1 riippua_V
2 riistää_V
1 riittää_V
2 rikkoa_V
9 ryhtyä_V
54 saada_V
1 saapua_V
117 säätää_V
5 saattaa_V
4 sallia_V
1 selvittää_V
2 seurata_V
7 siirtää_V
5 sisältää_V
4 sisältyä_V
1 sitoa_V
1 sitoutua_V
2 soveltaa_V
1 suojella_V
1 suorittaa_V
1 suostua_V
1 syyllistyä_V
2 taata_V
1 taitaa_V
3 tarkastaa_V
2 tarkoittaa_V
11 tarvita_V
2 täydentää_V
6 täyttää_V
23 tehdä_V
1 todeta_V
3 toimia_V
16 toimittaa_V
1 toteutua_V
1 tukea_V
43 tulla_V
1 tunnustaa_V
6 tuomita_V
20 turvata_V
5 tutkia_V
1 työllistää_V
1 tyytyä_V
2 uhata_V
3 uusia_V
3 vaalia_V
3 vaarantaa_V
10 vaatia_V
9 vahvistaa_V
4 vaikuttaa_V
1 vakuuttaa_V
32 valita_V
1 valittaa_V
6 valmistella_V
1 välttää_V
1 valtuuttaa_V
10 valvoa_V
2 vangita_V
2 vapauttaa_V
1 varata_V
1 varmentaa_V
1 varmistaa_V
1 vastaanottaa_V
9 vastata_V
1 vastoa_V
1 viestiä_V
6 viipyä_V
1 viivästyä_V
91 voida_V
2 ylittää_V
1 ylläpitää_V
3äänestää_V

(7)
Search string: **A**
Output:

1 aiheeton_A
2 aikainen_A
2 ainoa_A
2 ajankohtainen_A
1 alainen_A
1 alemmanasteinen_A
1 allekirjoitettava_A
2 alueellinen_A
1 ammatillinen_A
1 ankara_A
2 annettava_A
21 annettu_A
2 antava_A
1 aseellinen_A
1 asiallinen_A
1 asianmukainen_A
8 asianomainen_A
1 avattu_A
2 avoin_A
2 edellyttämä_A
2 enempi_A
4 enin_A
3 ennenaikainen_A
2 eriävä_A
8 erityinen_A
1 eronnut_A
1 esitettävä_A
1 esteellinen_A
5 estynyt_A
2 etevä_A
1 evankelis-luterilainen_A
2 hankittu_A
1 hankkima_A
2 henkilökohtainen_A
2 hoitava_A
1 huomattava_A
1 hylättävä_A
2 hyvä_A
1 hyväksyttävä_A
1 ihmisarvoinen_A
1 ilmeinen_A
2 jäänyt_A
1 jakamaton_A
29 jokainen_A
1 juhlallinen_A
29 julkinen_A
4 kaikki_A
23 kansainvälinen_A
1 kansalaisuusi_A
3 kansallinen_A
1 kansanvaltaisa_A
3 käsiteltävä_A
1 käytettävä_A
2 käyttävä_A
1 keskeinen_A
4 kiinteä_A
1 kiireellinen_A
1 kirjallinen_A
2 kokoontunut_A
15 korkea_A
18 koskeva_A
1 kunnallinen_A
1 kuulumaton_A
7 kuuluva_A
1 laajakantoinen_A
6 lainvastainen_A
1 lakannut_A
1 lievä_A
6 liittyvä_A
1 lopullinen_A
1 loukkaamaton_A
1 luottamuksellinen_A

1 määräämä_A
5 määrätty_A
2 mainittu_A
1 maksuton_A
2 menettänyt_A
4 merkittävä_A
1 merkitty_A
1 monijäseninen_A
1 mukainen_A
79 muu_A
1 myöhempi_A
1 neuvoa-antava_A
4 noudatettava_A
1 oikeudellinen_A
1 oikeudenmukainen_A
1 olennainen_A
1 oleskeleva_A
8 oleva_A
7 oma_A
1 omatoiminen_A
1 osallistunut_A
2 otettava_A
2 päätösvaltainen_A
1 päättynyt_A
1 paikallinen_A
3 painava_A
1 peräkkäinen_A
3 perusteltu_A
1 perustuva_A
2 pidettävä_A
2 pysyvä_A
2 rajattu_A
1 rangaistava_A
3 rangaistu_A
1 rauennut_A
1 rehellinen_A
2 rekisteröity_A
3 riippumaton_A
1 rikkonut_A
1 rikosoikeudellinen_A
2 ruotsinkielinen_A
15 säädetty_A
3 saanut_A
2 saatu_A
1 salainen_A
2 sama_A
2 samanlainen_A
1 satunnainen_A
9 sellainen_A
5 seuraava_A
1 sisäinen_A
1 sisältävä_A
1 sivistyksellinen_A
2 sotilaallinen_A
2 soveltuva_A
2 suhteellinen_A
10 suuri_A
1 syntyperäinen_A
2 syyllinen_A
1 tahallinen_A
5 tällainen_A
1 tärkeä_A
13 tarkka_A
5 tarkoitettu_A
1 tarkoittava_A
1 tarkoituksenmukainen_A
3 tarpeellinen_A
5 tarvittava_A
1 tavallinen_A
2 täydentävä_A
1 täysi_A
1 täysilukuinen_A
1 täysivaltainen_A
1 täytäntöönpanokelpoinen_A
17 tehtävä_A
1 tehty_A
1 tekemä_A
1 tekijänoikeudellinen_A
1 terveellinen_A
2 tilapäinen_A
1 toimitettava_A
2 toimitettu_A
2 toimiva_A
1 toimivaltainen_A
3 toinen_A
1 törkeä_A
1 tuleva_A
1 tullut_A
1 tunnettu_A
1 tuomionvoipa_A
1 tuomionvoiva_A
1 uhkaava_A
4 ulkomaalainen_A
1 ulkopuolinen_A
1 ulottuva_A
1 usea_A
1 useampi_A
1 uskonnollinen_A
5 uusi_A
1 vaadittava_A
1 vaalea_A
1 vaalikelpoinen_A
1 vaarantava_A
1 vaativa_A
1 vahvistettava_A
1 vajaavaltainen_A
1 vakinainen_A
1 välinen_A
2 välitön_A
2 valitsema_A
3 valittu_A
2 valtiollinen_A
9 välttämätön_A
1 vanhempi_A
1 vanhin_A
2 vastaava_A
1 vastattava_A
1 vastuunalainen_A
4 velvollinen_A
2 verovelvollinen_A
2 viimeinen_A
2 viisijäseninen_A
1 viranomainen_A
4 vireä_A
2 vuotuinen_A
1 yhdenvertainen_A
2 yhtäläinen_A
1 yhteensopiva_A
3 yhteiskunnallinen_A
1 yksikamarinen_A
1 yksilöllinen_A
1 yksimielinen_A
1 yksittäinen_A
1 yksityinen_A
1 yksityiskohtainen_A
11 yleinen_A
7 ylin_A
8 ääninen_A
5 äänioikeutettu_A

5 Conclusion

The information retrieval system discussed here opens almost limitless possibilities for various types of searches. It allows for all traditional search methods of direct surface text search. In addition, it makes possible also accurate search, because the base form of each

word is also available. This again makes possible the production of various statistical lists from the target text.

This report describes two types of implementations for accurate information retrieval, where the base form of a word and its POS category is used for refining retrieval. It would also be possible to add more linguistic features to the system, such as cases of nominals, various verb forms, and even syntactic tags. The analysis produces this all, and what we have seen in this report, is the use of one category only, that is, the POS, and all other features were removed.

References

Hurskainen, Arvi, 2019. Accurate information retrieval using text analysis and disambiguation. *Technical Reports in Language Technology*, Report, No 36. http://www.njas.helsinki.fi/salama