［資料］

スウェーデン語コーパス検索システム Korp ver.9 の使い方
－現代スウェーデン語コーパス Moderna を中心に－

梅谷 綾

1. はじめに

Korp[1]は Göteborg 大学が運営するウェブサイト Språkbanken[2]で公開されているスウェーデン語コーパスの検索システムである. 筆者は『IDUN －北欧研究－』21 号に Korp の ver.3 の使い方についての研究ノート[3]を寄稿したが，それ以降 Korp は更新を重ね, 2022 年 2 月時点では ver.9.1.0 が公開されている. Korp ver.3 に収録されていた現代スウェーデン語のコーパスは 184 種類, 延べ語数は約 69.7 億語であったが，ver.9.1.0 ではコーパス数は 255，延べ語数は約 144 億語に増えている. また, 更新に伴って検索画面のレイアウトなどにも一部変更がみられる. 本稿では梅谷（2015）の内容を更新し，本学スウェーデン語専攻の学生が Korp を使う際にまず把握しておきたい基本的な検索方法について紹介する.

2. 簡易検索（Enkel sökning）

Korp のサイトを開くとまず表示されるのは，現代スウェーデン語のコーパスを集めた Moderna の簡易検索画面（図 1）である.

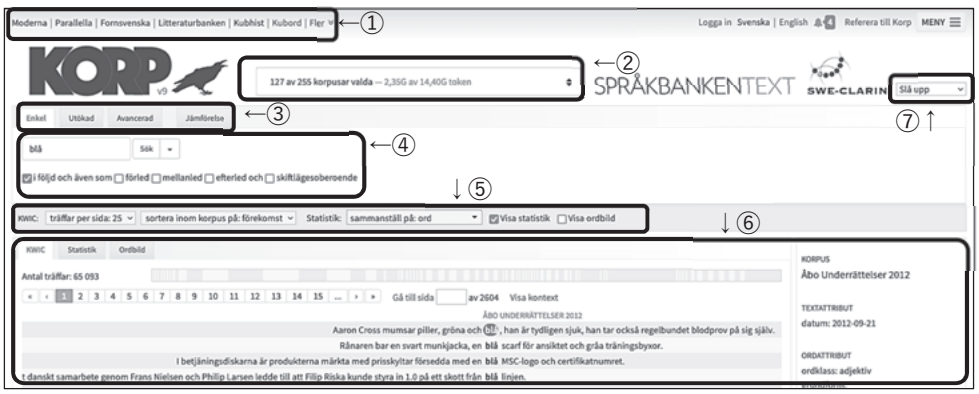


図 1　Moderna の簡易検索画面

[1] https://spraakbanken.gu.se/korp/
[2] http://spraakbanken.gu.se/
[3] 梅谷（2015）

① コーパスのジャンル[4]

Korp で検索できるコーパスは Moderna（現代スウェーデン語）や Fornsvenska（古スウェーデン語）などのジャンルに分類されており，ジャンルごとに検索インタフェースが用意されている．

② コーパス選択

現在選択中のコーパスの数とその延べ語数がここに表示される．図 1 では[127 av 225 korpusarna valda – 2,35G av 14,40G token]となっており，これは「225 のコーパス（延べ 144 億語）のうち，127 のコーパス（延べ 23.5 億語）が選択されている」ことを意味する．また，ここをクリックすると Moderna に分類されるコーパスの一覧から検索するコーパスを選択できる．各コーパスの名前にマウスポインタを合わせると，そのコーパスの説明や収録語数・文数，最終更新時期などの情報を確認できる．Moderna で検索できるコーパスについては 4 章を参照．

③ 検索方法

検索目的に合わせて簡易検索（Enkel），拡張検索（Utökad），上級者用検索（Avancerad）を選択できる．本稿では簡易検索と拡張検索に絞って紹介する．簡易検索については本章を，拡張検索は 3 章を参照．

④ 検索語入力欄・検索条件

簡易検索では以下の条件を設定することができる．詳しくは 2.1 と 2.2 を参照．

i följd：完全一致
förled：前方一致
mellanled：部分一致
efterled：後方一致
skiftlägesoberoende：大文字・小文字を区別しない

⑤ 検索結果の表示設定

1 ページに表示する検索結果の数や検索結果の並び替えなどの設定．

⑥ 検索結果

検索結果は以下の 3 種類の方法で表示される．

KWIC：keyword in context の略．検索語を中心としその前後の文脈を表示する．
Statistik：検索結果の統計データ．
Ordbild：検索語とよく一緒に使われる語句（コロケーション）．

⑦ 検索履歴

[4] 梅谷（2015）では「コーパスのテーマ」としていたが，本稿では「コーパスのジャンル」とする．また，梅谷（2015）では Korp ver.3 で検索できるコーパスのジャンルにどのようなものがあったかを紹介したが，本稿では割愛する．

2.1. 完全一致検索と大文字・小文字の区別

検索条件（図 1 の④）の[i följd]を選択すると，検索語の文字列と完全に一致するものを検索する「完全一致検索」ができる．例えば blå を完全一致検索すると，全て小文字で書かれた blå という文字列だけが検索対象となり，blå の変化形 blått, blåa などや大文字を含む Blå は検索結果に含まれない．大文字・小文字を区別せずに検索したい場合は，検索条件の[skiftlägesoberoende]を選択する．

同様に slå upp を完全一致検索した場合，slår upp や Slå upp は検索結果に含まれない．一方で[i följd]を選択解除して slå upp を検索すると「1 つの文の中に slå と upp という文字列の両方が含まれるもの」が検索対象となる．

2.2. 前方一致・後方一致・部分一致検索

例えば blå から始まる語を検索したい場合は，検索条件の[förled]を選択して「前方一致検索」を行う．この場合，形容詞 blå とその変化形の blått や blåa, blåbär などの blå を含む複合語はもちろんのこと，形容詞 blå とは関係のない動詞 blåsa とその変化形など，blå という文字列から始まる語が全て検索結果に含まれる．

一方で[efterled]を選択すると，検索語の文字列で終わる語を検索する「後方一致検索」ができる．blå を後方一致検索すると，mörkblå などの blå を含む複合語に加え，形容詞 blå とは関係のない名詞 tablå なども検索結果に含まれる．形容詞 blå を含む複合名詞に絞って検索したい場合は，次節の lemgram 検索を使用する．

また，[mellanled]を選択した場合は，検索語の文字列を含む全ての語が検索対象となる「部分一致検索」[5]ができる．

2.3. 簡易検索で lemgram[6]を使用する

簡易検索では，検索語を入力した後少し待つと，図 2 のようなリストが表示される．これは検索語を含む lemgram の一覧である．例えば kort という語には同じ綴りで形容詞・名詞・副詞の 3 つの品詞があるが，形容詞の kort の用例のみを検索したい場合はこのリストから[kort (adjektiv)]を選択する．する

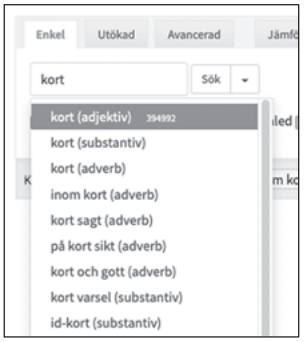


図 2 kort を含む lemgram

[5] mellan の意味を考えると「中間一致」とすべきかもしれないが，実際に[mellanled]を選択して検索した結果は「中間一致」ではなく「部分一致」だったことから，ここでは「部分一致検索」とした．

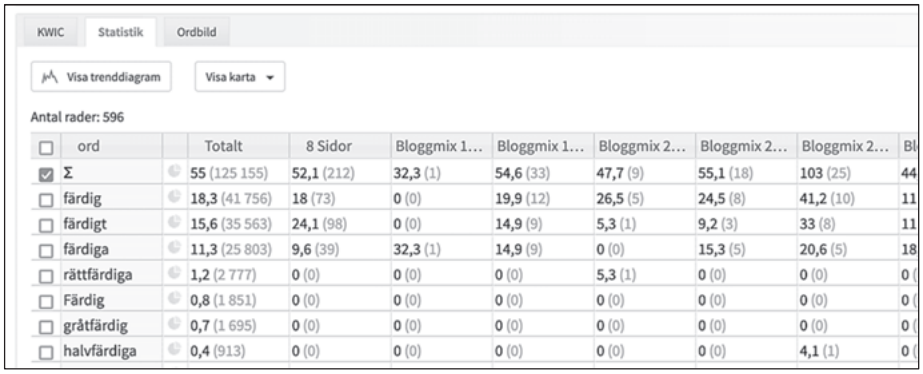
[6] lemgram は単語がもつ様々な変化形を集約して基本形を用いて表したもののことで，日本語のコーパス言語学用語ではレマ (lemma) と呼ばれる．Korp で検索できるコーパスの多くは品詞や lemgram など様々な情報がタグ付けされている．タグ付けはテキストを形態素分析する

と形容詞 kort とその変化形 korta，kortare などの用例をまとめて得ることができる．

また，この lemgram 検索は前節で紹介した前方一致検索・後方一致検索との併用も可能である．例えば，形容詞 färdig 単独での用例と färdig を後要素とする複合形容詞の用例を一括して収集したい場合は，検索条件の[efterled]を選択して färdig の lemgram を検索するとよい．

2.4. 検索結果の統計

[Statistik]というタブ（図 1 の⑥）を選択すると，検索で得られた用例の統計データの表が表示される（図 3）．この表で示される数値は，左が相対頻度（100 万語あたりの頻度），右の括弧付きの数値が絶対頻度（実際に得られた用例数）である．



ord	Totalt	8 Sidor	Bloggmix 1…	Bloggmix 1…	Bloggmix 2…	Bloggmix 2…	Bloggmix 2…	Bl
Σ	55 (125 155)	52,1 (212)	32,3 (1)	54,6 (33)	47,7 (9)	55,1 (18)	103 (25)	44
färdig	18,3 (41 756)	18 (73)	0 (0)	19,9 (12)	26,5 (5)	24,5 (8)	41,2 (10)	11
färdigt	15,6 (35 563)	24,1 (98)	0 (0)	14,9 (9)	5,3 (1)	9,2 (3)	33 (8)	11
färdiga	11,3 (25 803)	9,6 (39)	32,3 (1)	14,9 (9)	0 (0)	15,3 (5)	20,6 (5)	18
rättfärdiga	1,2 (2 777)	0 (0)	0 (0)	0 (0)	5,3 (1)	0 (0)	0 (0)	0
Färdig	0,8 (1 851)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
gråtfärdig	0,7 (1 695)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
halvfärdiga	0,4 (913)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	4,1 (1)	0

図 3 統計データ表（一部）

統計データは出力・保存することもできる．画面の左下にデータ出力の設定メニューがある．出力したいデータ（[Relativa frekvenser]〈相対頻度〉または[Absoluta frekvenser]〈絶対頻度〉）を選択し，出力形式が CSV になっていることを確認し，[Generera export]を選択して CSV ファイルを保存する．保存した CSV ファイルは Excel などの表計算ソフトで閲覧・加工することができる．

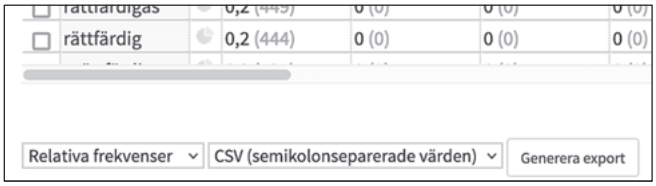


図 4 統計データの出力設定

コンピュータープログラムを使って自動的に行われているため，タグの情報が必ずしも正しいとは限らないことに注意が必要である．コーパスのタグ付けやレマについて詳しく知りたい場合は石川（2021）などコーパス言語学について書かれた文献を参照のこと．

2.5. コロケーション

簡易検索で単語や lemgram を検索する際，検索結果の表示設定（図 1 の⑤）で[Ordbild]を選択して検索を実行し，検索結果の[Ordbild]というタブ（図 1 の⑥）を選択すると，検索語のコロケーション（検索語の前後に頻繁に出現する語句）が頻度順に表示される．図 5 は検索したコーパスで動詞 åka がどのような主語・目的語・副詞句と一緒によく使われるかを表した ordbild である．

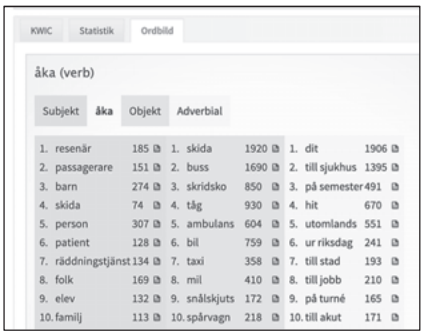


図 5 動詞 åka の ordbild

3. 拡張検索（Utökad sökning）

拡張検索ではコーパスに付けられたタグの情報を用いて簡易検索よりも詳細な条件設定ができる．Korp の検索画面で[Utökad]のタブをクリックすると図 6 の拡張検索画面が表示される．

図 6 拡張検索画面

① 検索語の条件

② OR 検索の条件追加

複数の条件のいずれかに該当するものを検索する「OR 検索」の条件を追加.

③ AND 検索の条件追加

複数の条件の全てを満たすものを検索する「AND 検索」の条件を追加.

④ 大文字・小文字の区別設定

⑤ 検索オプション

検索語の繰り返し設定や文境界の設定.

⑥ 検索語の追加と文境界の追加

3.1. AND 検索と OR 検索

本節では以下の検索例を用いて，AND 検索と OR 検索の条件の設定方法について紹介する．

【検索例】

強調の接頭辞として svin が用いられている形容詞と副詞を検索する．つまり設定する条件は「複合語の前要素が svin」で**なおかつ**「品詞が形容詞」**または**「品詞が副詞」となるため AND 検索と OR 検索を組み合わせる．

【手順】

(1) 1 つ目の条件は「複合語の前要素が svin である」なので，条件の一覧から [sammansättningar] を選択．その右に [förled] が選択されているのを確認した上で検索欄に svin と入力．

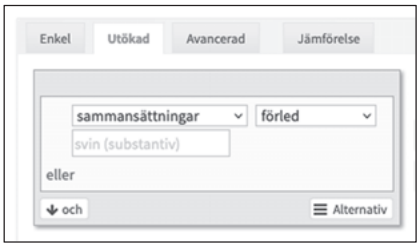


図 7 「複合語の前要素が svin」

(2) 次に AND 検索の条件を追加するために検索画面左下の [↓och] をクリック．2 つ目の条件は「品詞が形容詞である」なので ordklass är adjektiv とする（図 5）．

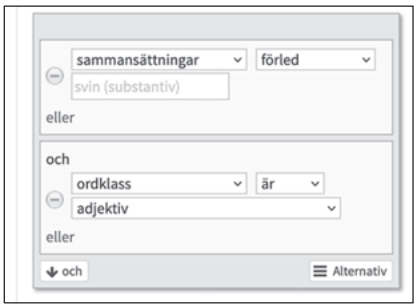


図 8 AND 検索「前要素が svin の複合語で品詞が形容詞」

(3) OR 検索の条件を追加するために検索画面左下の [eller] をクリック．3 つ目の条件「品詞が副詞である」を追加するために ordklass är adverb と設定する（図 9）．これで全ての検索条件の設定ができたので検索を実行する．なお，検索条件を削除したい場合は各条件の左にある㊀をクリックする．

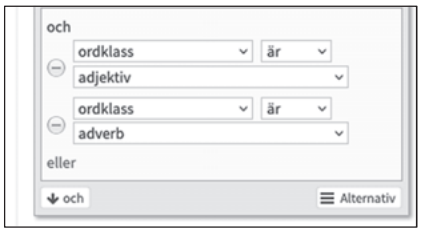


図 9 OR 検索「品詞が形容詞または副詞」

3.2. フレーズ検索と繰り返し機能

前節では 1 つの検索語に複数の条件をつける方法を扱ったが，本節では複数の検索語からなるフレーズを検索する例を挙げる．また，検索条件の繰り返し機能についても紹介する．

【検索例】

tyckas vara [形容詞] 〈[形容詞]であるように思われる〉の用例をなるべく多く集める．動詞 tyckas の各変化形を使用した用例を一括検索するために tyckas の lemgram を使用する．さらに，繰り返し機能を使って tyckas inte vara [形容詞]という否定文も同時に検索する．

【手順】

(1) 1 つ目の検索語は動詞 tyckas とその各変化形であるので，[lemgram är tyckas (verb)]と設定する．次に検索画面の右側にある[lägg till token]（図 6 の⑥）をクリックして検索語を 3 つ追加する．2〜4 つ目の検索語の条件をそれぞれ ord är vara 〈vara という語である〉, ord är inte 〈inte という語である〉, ordklass är adjektiv 〈品詞が形容詞である〉とする（図 10）．

図 10 フレーズ検索の検索画面

(2) (1) で設定した条件では tyckas (lemgram)＋inte＋vara＋[形容詞]という否定文しか検索できないため，3 つ目の検索語 inte に繰り返し設定を行う．「検索語 inte を 0 回〜1 回繰り返す」と設定することで tyckas (lemgram)＋vara＋[形容詞]と tyckas (lemgram)＋inte＋vara＋[形容詞]を同時に検索することができる．繰り返し設定を行う検索語の右下にある[Alternativ]をクリックすると検索オプション一覧が表示される．そこから[Upprepa]を選択すると繰り返し設定が表示されるので，繰り返し回数を Upprepa 0 till 1 gånger と設定する（図 11）．これで tyckas (lemgram)＋vara＋[形容詞]と tyckas (lemgram)＋inte＋vara＋[形容詞]の用例を一度の検索で得ることができる．

図 11 繰り返し設定

4. Moderna に収録されているコーパス

Moderna には 1900 年以降の様々な種類のスウェーデン語資料を扱ったコーパスが収録されている．収録されているコーパスは主に以下の 11 のカテゴリーに分類されている．カテゴリー名の後の括弧付きの数字はそこに分類されているコーパスの数を表す.

- Akademiska texter (2)：人文科学系と社会科学系の学術テキスト.
- August Strindberg (2)：August Strindberg の作品と手紙.
- Finlandssvenska texter (56)：フィンランド系スウェーデン語で書かれたブログなどのインターネット上のテキストや政治・法律関係の文書, 文学, 新聞, 雑誌などのデータ.
- L2 Korp (8)：第二言語としてのスウェーデン語の教材などを収録したコーパス．一部非公開.
- Medicinska texter (13)：医学関係の新聞や雑誌の記事のコーパス.
- Myndighetstexter (26)：政府関係の文書を使ったコーパス．Riksdagen が公開している文書やスウェーデン法令集（Svensk författningssamling）などが収録されている.
- Nyhetstexter(39)：ニュース記事のコーパス．Göteborgs-posten（＝GP）や Dagens Nyheter（＝DN）や Svenska Dagbladet（=SvD）といった日刊紙に加え，SVT nyheter や 8 SIDOR の記事も検索できる.
- Skyddade korpusar (18)：アクセス制限があり非公開になっているコーパス.
- Skönlitteratur (7)：文学作品のコーパス．Bonnier 社から 1976 年～1977 年と 1980 年～1981 年に出版された小説や Norstedts 社から 1999 年に出版された小説，スウェーデン語とオランダ語のパラレルコーパスに収録されている小説，版権の切れた古典作家の作品などが収録されている.
- Sociala medier (68)：スウェーデン語で書かれたブログやツイッター，インターネットフォーラム（電子掲示板）の書き込みなどのソーシャルメディアのテキストが収録されている.
- Tidskrifter (2)：通俗科学雑誌 *Forskning & Framsteg* とクラシック音楽の雑誌 *Podiet* の記事のコーパス.

さらに上記のカテゴリーには分類されていないコーパスもある.

- Dramawebben：スウェーデン語で書かれた 1600 年代から現代までの演劇台本を電子データ化して公開しているサイト Dramawebben[7]のデータを使ったコーパス.

[7] http://www.dramawebben.se/

・Etnologiska frågelistor：Nordiska museet が収集した民族学資料を使用したコーパス.
・IVIP demo：スウェーデンとフィンランドの大学や研究機関によるプロジェクト IVIP（Interaktion och variation i pluricentriska språk – Kommunikativa mönster i sverigesvenska och finlandssvenska）で作成されたコーパスのデモ版.
・LäSBarT – Lättläst svenska och barnbokstext：やさしいスウェーデン語で書かれたテキストと児童書のテキストを集めたコーパス.
・PAROLE：EU のプロジェクト PAROLE で作成されたスウェーデン語コーパス．小説，日刊紙，雑誌，インターネット上のテキストが含まれている.
・Poeter.se：詩の投稿サイト Poeter.se[8]に投稿された詩のテキストのコーパス.
・Psalmboken (1937)：1937 年発行の賛美歌集のコーパス.
・SUC 2.0：Stockholm-Umeå コーパス 2.0．2006 年に公開されたコーパス．1990 年代の様々なタイプのスウェーデン語のテキストが収録されている.
・SUC 3.0：Stockholm-Umeå コーパス 3.0．SUC 2.0 に修正や新たなテキストデータを加えたもの.
・sv-COVID-19：新型コロナウイルス感染症関係の記事を集めたコーパス.
・Svenska partiprogram och valmanifest 1887-2010：1887～2010 年のスウェーデンの政党の綱領と選挙のマニフェストのコーパス.
・Svenska Wikipedia (januari 2017)：スウェーデン語版 Wikipedia[9]の 2017 年 1 月時点の記事を使用したコーパス.
・Svenskt frasnät (SweFN)：Göteborg 大学のスウェーデン語の語彙情報資源構築プロジェクト SweFN のデータを使用したコーパス.
・Talbanken：1970 年代に作成されたスウェーデン語の書き言葉と話し言葉のコーパス.

上記と梅谷(2015)を比較すると Etnologiska frågelistor や L2 Korp, sv-COVID-19 など Korp ver.3 にはなかった新しいコーパスが追加されていることがわかる.

6. おわりに

本稿では本学スウェーデン語専攻の学生が Korp を使い始める際にまず把握しておきたい検索方法について紹介した．コーパスのタグ付けはコンピュータープログラムで自動的に行われているため，タグの情報に間違いが多くみられるなど

[8] https://www.poeter.se/
[9] https://sv.wikipedia.org/wiki/

問題点も確かにあるが，それでもこれほど膨大なデータ量のスウェーデン語コーパスを日本に居ながら使用できることはスウェーデン語学習者やスウェーデン関係の研究者にとって非常に有難いことである．コーパスを初めて使う際は lemgram などの用語に戸惑うことも多いと思われるが，本稿を参考にしながら Korp を使った様々な検索を試してみて頂きたい．

また今後の目標として，本稿の内容をさらに更新すると同時に本稿で紹介しきれなかった Korp の検索機能や Språkbanken で利用できる Korp 以外のツールの使い方についても日本語でまとめ，日本のスウェーデン語学習者やスウェーデン関係の研究者がこれらのツールを活用しやすい環境を整えていきたい．

参 考 文 献

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*. 474–478. Istanbul: ELRA.

石川慎一郎．2021．『ベーシック コーパス言語学』第 2 版．東京：ひつじ書房．

梅谷綾．2015．「コーパス検索システム Korp の基本使用方法 ― 現代スウェーデン語コーパスを中心に ―」，『IDUN －北欧研究－』21 号．161-178．大阪：大阪大学言語文化研究科言語文化専攻デンマーク語・スウェーデン語研究室．

清水育男. 2013.「スウェーデン語の情報が得られる電子媒体 ― 辞書を中心に ―」，『IDUN －北欧研究－』20 号．181-198．大阪：大阪大学言語文化研究科言語社会専攻デンマーク語・スウェーデン語研究室．

インターネット上の資料

Dramawebben. http://www.dramawebben.se/

Interaktion och variation i pluricentriska språk (IVIP). https://www.su.se/svefler/ivip

Konsertmagasinet Podiet. https://www.gso.se/upptack/podiet/

Litteraturbanken. https://litteraturbanken.se/

Nordiska museet. https://www.nordiskamuseet.se/

Poeter.se. https://www.poeter.se/

Språkbanken. https://spraakbanken.gu.se/