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 ―

梅谷 綾

1. はじめに

近年の情報技術の著しい発展により，インターネットを用いた情報収集は私たちの日常生活の中で非常に身近な手段となっている．さらにここ数年の間に，パソコンよりも直感的な操作ができるスマートフォンやタブレットが広く使用されるようになってきており，インターネットを使用した情報収集は今後もますます身近なものになるであろうと考えられる．そのような情報社会の中で北欧諸国がインターネットを利用して言語，社会，文学，歴史などに関する情報の収集や管理，公開をいかに行っているかについて『IDUN』20 号で様々な報告がされた．その中でもスウェーデン語学やスウェーデン語学習の分野については，清水(2013)にインターネット上で利用できる辞書サイトやスウェーデン語学習サイト，スウェーデン語コーパス，さらにはスマートフォンやタブレットで使用できる辞書アプリに至るまで詳細にまとめられている．本稿では清水（2013）でも紹介されているコーパス検索システム Korp に焦点をあて，Korp を使用した現代スウェーデン語コーパスの基本的な使用方法や，コーパス検索によって得られた頻度データの出力・保存方法についてまとめることを目的とする．なお本稿で紹介する Korp の使用方法は 2014 年 9 月時点で公開されている Korp version 3.0 を対象とするものであり，今後の更新により検索画面のレイアウトや検索機能などに変更が生じる場合があることをまず断っておきたい．本稿で紹介するコーパスの情報についても 2014 年 9 月時点のものである．

2. コーパスと Korp について

コーパスという言葉自体，言語学に触れたことのない人にとってはあまり馴染みのない言葉かもしれない．コーパスとは，現実の言語を大規模かつ体系的に収集し，データとして保存したものである（石川 2012：33）．Korp はそのコーパスの膨大なデータの中から，特定の条件に合った単語やフレーズを検索するための検索システムである．Korp は Göteborgs universitet が運営するウェブサイト Språkbanken (http://spraakbanken.gu.se/) で一般公開されており，誰でも無料で利用できる．Språkbanken には Korp 以外にもスウェーデン語のネット辞書などのコンピュータ言語学を駆使したスウェーデン語関係のツールが集約されており，

スウェーデン語を学ぶ者やスウェーデン語学を研究する者にとっては欠かせないサイトである．

Korp で検索できるコーパスは，言語データの年代やジャンルによって 14 種類のテーマに分類されており，テーマごとに検索画面が設けられている．

- Moderna：1900 年以降の様々なカテゴリー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．本稿ではこの Moderna に分類されているコーパスを総じて「現代スウェーデン語コーパス」と呼び，その使用方法を紹介する．
- Parallella：スウェーデン語と他言語のパラレルコーパス．パラレルコーパスとは，対応性をもった 2 言語のデータを同時に収録したものである（石川 2012：44）．ここには，欧州議会のスウェーデン語版ホームページのテキストとデンマーク語版，英語版，フィンランド語版，フランス語版，ギリシャ語版，イタリア語版，オランダ語版，ポルトガル語版，スペイン語版，ドイツ語版のページのテキストを収録したパラレルコーパスや，スウェーデン語テキスト（主に小説など）のオランダ語訳およびオランダ語テキストのスウェーデン語訳を集めたパラレルコーパスなどが収録されている．
- Fornsvenska：古スウェーデン語の文献を扱ったコーパス．古スウェーデン語文献のデータベース Fornsvenska textbanken (http://project2.sol.lu.se/fornsvenska/) で公開されている古スウェーデン語の資料および Riksarkivet 所蔵の中世スウェーデンの外交書簡のデータで構成されたコーパスなどがある．
- Litteraturbanken：スウェーデンの古典文学を中心に文学作品の電子書籍を無料で公開しているウェブサイト Litteraturbanken (http://litteraturbanken.se/) で公開されている文学作品を使用したコーパス．
- Kubhist：Riksarkivet や Kungliga biblioteket などによる新聞記事の電子データ化プロジェクト Digidaily (http://digidaily.blogg.kb.se/) で電子化された新聞のテキストを使用したコーパス．20 種類の新聞が対象になっており，テキストの年代は 1850 年代～1900 年代が中心である．
- Historiskt：上記の Fornsvenska，Litteraturbanken，Kubhist や後述の Runebergtidskrifter，Bibelstället に含まれるコーパスおよび 1600 年代～1800 年代の法律文など，古スウェーデン語から現代スウェーデン語まで様々な年代のテキストが集められている．
- Spf 1800–1900：1800 年，1820 年，1840 年，1860 年，1880 年，1900 年に出版されたスウェーデン語の文学作品を集めたデータベース Svensk prosafiktion 1800–1900 (http://spf1800-1900.se) のデータを使用したコーパス．
- Äldre finlandssvenska：1700 年代から 1950 年代までのフィンランド系スウェーデン語で書かれたテキストを集めたコーパス．

・Färöiska：フェロー語のテキストを集めたコーパス.
・Sibirientyska：シベリア中部のクラスノヤルスク地方で話されているドイツ語を書き起こしたコーパス.
・Kiöpings resor：Nils Matsson Kiöping により 1674 年と 1743 年に書かれた旅行記を使用したコーパス.
・Runebergtidskrifter：版権の切れた文学作品を中心に北欧語の出版物を電子データ化してインターネットで公開している Projekt Runeberg のウェブサイト (http://runeberg.org/) で公開されている新聞や雑誌（主に 1880 年代～1940 年代発行のもの）のデータを使用したコーパス.
・Bibelstället：1873 年と 1917 年に翻訳された聖書のコーパス.
・Lagrummet：古スウェーデン語から現代スウェーデン語まで，様々な年代の法律関係の資料を収集したコーパスを収録している.

このようにスウェーデンでは様々なジャンルのスウェーデン語資料を電子データ化してインターネットで公開するプロジェクトが複数存在し，Korp ではそれらの言語データを使った検索が可能である.

3. 現代スウェーデン語コーパスについて

現代スウェーデン語コーパスに収録されているコーパスは主に以下の 9 種類のカテゴリーに分類されている．カテゴリー名の後の丸括弧内の数字はそのカテゴリー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が収録されている.
・August Strindberg (2)：August Strindberg の作品と手紙が収録されている.
・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，雑誌などのデータが収録されている.
・Skyddade korpusar (7)：アクセス制限がかかっており，非公開になっているコーパス.
・Medicinska texter (12)：医学関係の新聞や雑誌を使ったコーパスが収録されている.
・Skönlitteratur (5)：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，そして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 (48)：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，各種ソーシャルメディア

のテキストが収録されている.

・Tidningstexter (37)：新聞記事を使ったコーパス．以下が収録されている.

・GP (15)：Göteborgs-posten (＝GP) の記事のコーパス（1999 年，2001 年～2013 年).

・Press (6)：Dagens Nyheter (＝DN) や Svenska Dagbladet (=SvD) などの複数の新聞の記事を集めたコーパス（1965 年，1976 年，1995 年～1998 年).

・Webbnyheter (13) ：DN や SvD などの日刊紙のインターネット上の記事を集めたコーパス（2001 年～2013 年).

・8 SIDOR：やさしいスウェーデン語で書かれた新聞 8 SIDOR の記事のコーパス（2000 年代).

・DN 1987：DN の 1987 年の記事のコーパス.

・ORDAT: Svenska dagbladets årsbok 1923-1958：Svd の年鑑（1923 年～1945 年，1948 年，1958 年）のコーパス.

・Tidskrifter (1)：通俗科学雑誌 *Forskning & Framsteg* 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 (demo)：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の電子データ化・一般公開を行っているウェブサイト Dramawebben (http://www.dramawebben.se/) で公開されているデータを使ったコーパス.

・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 で作ら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.

・Psalmboken (1937)：1937 年発行の賛美歌集のコーパス.

・SNP 78-79 (Riksdagens snabbprotokoll)：1978～79 年の国会の予備調査報告書のコーパス.

・SUC 2.0：ストックホルム・ユーメオコーパス 2. 0.

・SUC 3.0：ストックホルム・ユーメオコーパス 3. 0.

・SALT svenska-nederländska：スウェーデン語・オランダ語パラレルコーパス.

・Europarl svenska：欧州議会のスウェーデン語版ウェブサイトのテキストを使用したコーパス.

・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.

・Svensk författningssamling：スウェーデン法令集のコーパス.

・Svenskt frasnät (SweFN)：Göteborgs universitet で行われている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.

・Svenska Wikipedia (april 2014)：スウェーデン語版 Wikipedia (http://sv.wikipedia.org/) の 2014 年 4 月時点の記事を使用したコーパス.

・Talbanken：1970 年代に作られたスウェーデン語の書き言葉と話し言葉のコーパス.

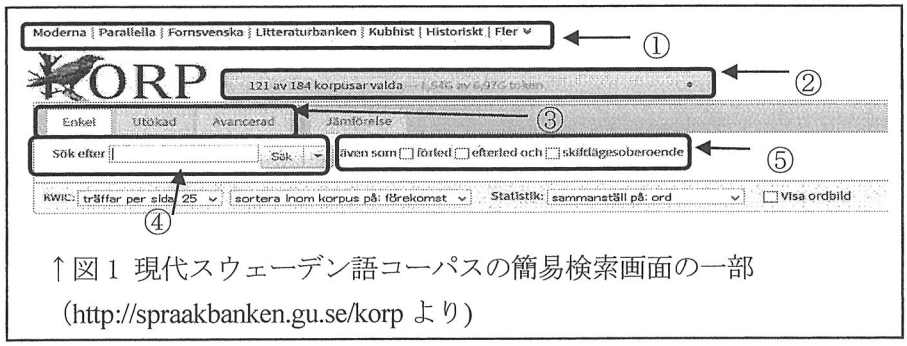
筆者が 2014 年 4 月下旬に Korp を使用した時，現代スウェーデン語コーパスに収録されているコーパスの数はアクセス制限がかかっている非公開のものを含めて合計 168 種類（一般公開されていたものは 159 種類）でその総語数は約 41 億 1000 万語だったが，同年 9 月 25 日時点では合計 184 種類（一般公開されているものは 177 種類）で総語数は約 69 億 7000 万語にまで増加していた．今後もコーパスの種類や総語数はますます増えていくと思われる.

4. Korp を使った検索方法

現代スウェーデン語コーパスでは検索の目的に合わせて［Enkel］〈簡易検索〉，［Utökad］〈拡張検索〉および［Avancerad］〈上級者用検索〉の 3 種類の検索方法を選ぶことができる．上級者検索には専門的な知識を必要とするため，今回は簡易検索と拡張検索の 2 種類に絞って紹介したい.

4.1. 簡易検索

Korp のページを開くとまず表示されるのが，現代スウェーデン語コーパスの簡易検索画面である.



↑図 1 現代スウェーデン語コーパスの簡易検索画面の一部
（http://spraakbanken.gu.se/korp より）

① コーパスのテーマ（2 章を参照）

② コーパスの選択：ここをクリックすると現代スウェーデン語コーパスの一覧が表示されるので，その中から検索目的に合ったものを選択する．一度に複数のコーパスを選択することも可能である．コーパスの名前にマウスのポインターを合わせると，そのコーパスに収録されている言語データの出典やコーパスの規模（収録語数および文数），最終更新の日付などの情報が表示

される．ちなみに［1,54G av 6,97G token］とは，現在選択中のコーパスが含む延べ語数 (token) を表しており，この場合〈184 のコーパスに収録されている約 69. 7 億語のうち，現在選択中の 121 のコーパスの延べ語数は約 15. 4 億語〉となる．

③ 検索方法を切り替えるタブ

④ 検索語入力欄

⑤ 検索オプション

［förled］：前方一致検索も同時に行う．

［efterled］：後方一致検索も同時に行う．

［skiftlägesoberoende］：大文字・小文字を区別せずに検索

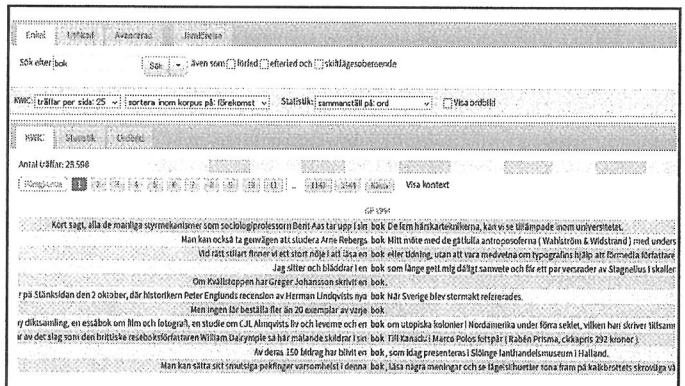
ここで一つ注意したい点は，検索オプション［skiftlägesoberoende］にチェックマークを入れないと，大文字と小文字を区別して検索が行われることである．例えば［skiftlägesoberoende］にチェックマークを入れずに全て小文字で ”språk“〈言語〉と入力して検索した場合，”Språk“ や ”SPRÅK” などの，スペルは同じだが大文字を含むものは検索語 ”språk“ とは別物とみなされてしまい検索結果に表示されない．そのため，大文字・小文字を区別せずに少しでも多くの用例を収集したい場合はこの［skiftlägesoberoende］を利用することをおすすめする．

簡易検索は主に以下のような検索を行いたい場合に使用する．

・特定の文字列（単語単体または複数の単語からなるフレーズ）を完全一致検索したい場合 → 4. 1. 1. を参照．

・特定の単語とその変化形の用例を一括で検索したい場合（lemgram 検索）→ 4. 1. 3. を参照．

・特定の単語の前後でよく使用される語句（コロケーション）を調べたい場合 → 4. 1. 4. を参照．



↑図 2 簡易検索の検索結果表示画面の一部
（http://spraakbanken.gu.se/korp より）

4.1.1. 完全一致検索

これは検索したい単語やフレーズを検索画面の検索語入力欄に入力し，[Sök] をクリックするまたはキーボードの Enter キーを押すだけの最もシンプルな検索方法である．検索結果は図2ように検索語を中心としてその前後の文脈を表示する KWIC (keyword in context) と呼ばれる形式で表示される．

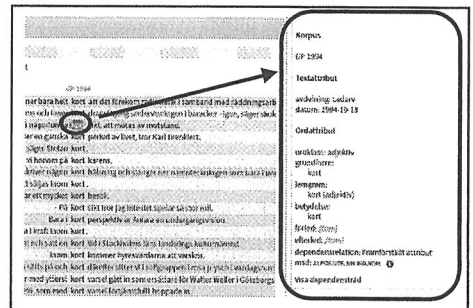
ただし，この検索で得られる検索結果はあくまで「入力された文字列に完全に一致するもの」であることに注意が必要である．例えば bok〈本〉を検索した場合，得られる検索結果はあくまで bok という文字列のみであり，ここには boken（単数・既知形）や böcker（複数・未知形）などの bok の変化形や，bokhandel〈本屋〉やbarnbok〈児童書〉などの bok を前要素または後要素とする合成語は含まれない．bok という文字列から始まる単語を検索したい場合には検索オプションの [förled] を，bok という文字列で終わる単語を検索したい場合には [efterled] を選択する．なお，さらに絞り込んで特定の単語とその変化形の用例（bok, boken, böcker, böckerna など）を一括して検索したい場合や，特定の単語を前要素・後要素とする合成語（bokhandel, barnbok など）を検索したい場合は lemgram 検索（4.1.3. を参照）を使用する．

また，例えばこの完全一致検索で kort を検索すると，名詞の kort〈カード，写真〉と形容詞の kort〈短い〉，副詞の kort〈短く，手短に〉が混在した検索結果となる．このような多義語の中から特定の品詞に絞って検索したい場合は，lemgram 検索（4.1.3. を参照）や拡張検索の AND 検索（4.2.2. を参照）を使用する．

さらに完全一致検索では単語単体の検索のみだけでなく，hör talas om〈～についての話を耳にする〉のような複数の単語からなるフレーズも検索することもできる．ただし前にも述べたとおり，この検索方法で得られる用例はあくまで入力した文字列と完全に一致するものなので，この場合に得られる用例は動詞 höra〈聞く〉の現在形 hör を使用したものに限定されてしまう．このフレーズに関して過去形 hörde や完了分詞 hört などを使用した用例もあわせて検索したい場合は，拡張検索のフレーズ検索（4.2.3. を参照）を使用する．

4.1.2. コーパスのタグについて

コーパスに収録されている膨大な言語データの中から目的にあった詳細な検索を行うためには，品詞などの言語学的な情報をあらかじめテキストに組み込んでおく必要がある．この情報付与のことをコーパス言語学では「アノテーション」と呼び，アノテーションで付与される品詞などの情報を「タグ」と呼ぶ（石川 2012：72-73）．アノテーションが行われたコーパスでは，特定の品詞を指定した検索などの詳細な検索条件の設定が可能である．Korp のコーパスもアノテーション

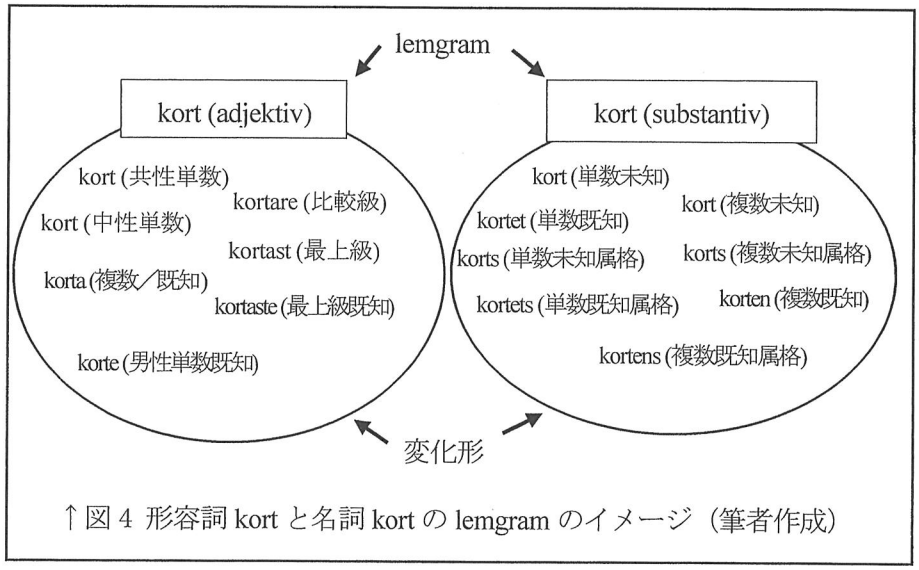


↑図3 検索結果表示画面の一部
（http://spraakbanken.gu.se/korp より）

がされており，検索結果画面で情報を確認したい単語をクリックすると，図 3 のように画面右側にその単語の情報が表示される．ここで表示されている情報には，その用例が収録されているコーパス名，そのテキストが発行された日付などの言語データに関する情報および，その単語の品詞，基本形，lemgram，その単語が合成語の場合はその前要素・後要素，そして格配列などの言語学的情報が含まれる．lemgram とはある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である．

この lemgram を使用することで，検索語が動詞であれば能動態・受動態それぞれの不定詞，現在形，過去形，完了分詞，命令形など，名詞であれば単数形・複数形それぞれの未知形，既知形や属格など，検索語のあらゆる変化形が用いられた用例を一度の検索で全て網羅できる．lemgram を使用した検索方法については 4. 1. 3. で紹介する．

ただし，Korp のコーパスのアノテーションはテキストを自動的に形態素分析する



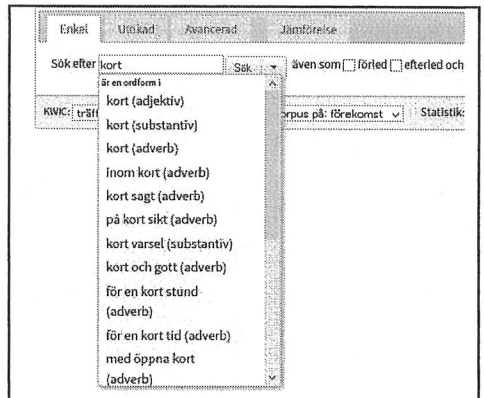
↑図 4 形容詞 kort と名詞 kort の lemgram のイメージ（筆者作成）

コンピュータプログラムを使って行われているため，タグの情報が必ずしも正しいとは限らないという問題点もある．特に［förled］〈前要素〉や［efterled］〈後要素〉の情報には，例えば final〈決勝戦〉の前要素が fin〈素敵な〉，後要素が al〈ハンノキ〉と

なってしまっていたり，enklare〈より単純な（enkel の比較級）〉の前要素が en〈一（数詞），（植物の）ネズ〉，後要素が klar〈澄んだ，明確な〉の男性既知形 klare となってしまっていたりと，明らかな異分析が目立つ．このような誤った情報も一部存在するため，タグの情報を 100 パーセント信用することは残念ながらできないが，このような誤りが混在していることを念頭に置いてうまくタグを利用すれば，各々の目的にあったピンポイントな検索ができるであろう．

4. 1. 3. lemgram 検索

検索語入力画面で検索語を入力した後少し時間をおくと，図 5 のようなリストが表示されることがある．このリストは検索語を含む lemgram の一覧である．図 5 は kort と入力した際に表示されたリストであり，ここには形容詞の kort，名詞の kort や副詞の kort だけでなく，kort を含む名詞句や副詞句などもリストアップされている．このリストから［kort (substantiv)］を選択して検索を実行すると，名詞の kort の各変化形を含む用例を一度の検索で得ることができる．



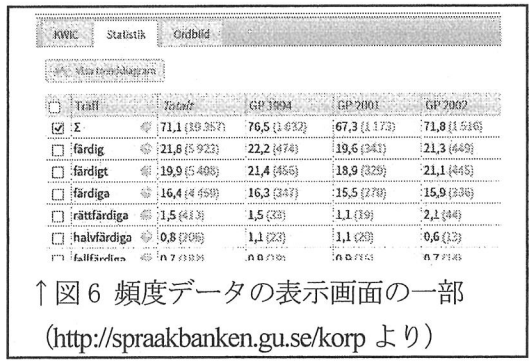
↑図 5 lemgram のリストの例
（http://spraakbanken.gu.se/korp より）

また，lemgram 検索と検索オプションの前方一致検索・後方一致検索の併用も可能である．例えば形容詞 färdig および 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，検索オプションの［efterled］にチェックを入れたうえで färdig の lemgram を検索すると，färdig とその変化形および färdig を後要素とする複合形容詞の各変化形の用例を一括して得ることができる．［efterled］にチェックマークを入れて färdig を完全一致検索しても färdig を後要素とする複合形容詞の用例は得られるが，ここに含まれるのはあくまで färdig という文字列で終わる単語であり，それらの変化形は含まれない．

4. 1. 4. 検索結果の統計とコロケーション

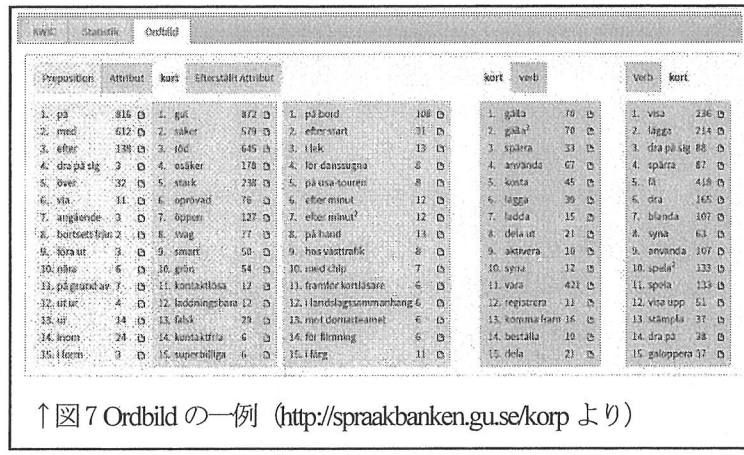
検索結果画面で［Statistik］のタブをクリックすると，検索で得られた用例の頻度データの表が図 6 のようにコーパスごとに表示される．この表の左の数字は相対頻度（100 万語あたりの頻度）で，括弧のついている右の数字は絶対頻度（実際に得られた用例数）である．この表は用例数順に並んでおり，見出し列にある

［Träff］をクリックするとアルファベット順の並び替えが，コーパス名をクリックするとそのコーパス内での用例数順の並び替えができる．［Totalt］をクリックすると再び用例数順に表示される．また，この頻度データの表は出力・保存することもできる（5 章を参照）．

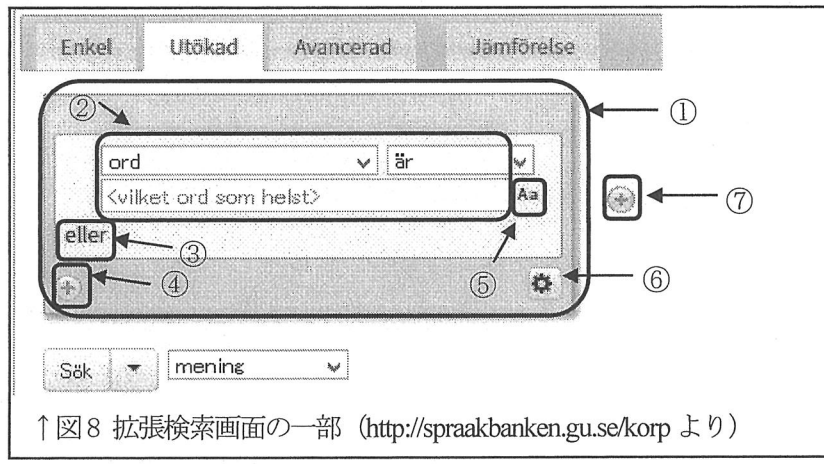


↑図 6 頻度データの表示画面の一部（http://spraakbanken.gu.se/korp より）

［Statistik］のタブの右側にある［Ordbild］のタブをクリックすると図 7 のような画面が表示される．この Ordbild は，検索語が単語一語の場合，または lemgram 検索をした場合にのみ表示される．図 7 は名詞 kort の lemgram を検索した際のものである．ここには，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．また，リストの各語句の右側にある小さなアイコンをクリックすると，その語句と検索語がセットで使われている用例が KWIC 形式で表示される．



↑図 7 Ordbild の一例（http://spraakbanken.gu.se/korp より）



↑図 8 拡張検索画面の一部（http://spraakbanken.gu.se/korp より）

4.2. 拡張検索

拡張検索では複数の検索条件を組み合わせることで，簡易検索よりも高度な検索が可能である．Korp の検索画面で［Utökad］のタブをクリックすると以下のような拡張検索画面が表示される．

①検索語枠　②検索語の条件
③検索条件の追加（OR 検索）　④検索条件の追加（AND 検索）
⑤大文字・小文字の区別の有無設定
⑥検索オプション
［Upprepa］：検索語の繰り返し設定
［Meningsbörjan］：検索語が文頭にくる用例を検索
［Meningsslut］：検索語が文末にくる用例を検索
⑦後続の検索語の追加

拡張検索で行える検索の例として，以下のようなものがある．

・複数の単語の用例を一度に検索したい場合（OR 検索）→ 4.2.1. を参照.
・特定の単語を前要素または後要素とする合成語を，品詞を指定して検索する場合（AND 検索）→ 4.2.2. を参照.
・特定の構文やフレーズの用例を集めたい場合（フレーズ検索）→ 4.2.3. を参照.

4.2.1. OR 検索

ここでは以下の検索例を用いて，複数の条件のうちいずれかに該当するものを検索する OR 検索の方法および，用例数の推移をグラフとして表示して年代ごとに比較する方法を紹介する．

【検索内容】

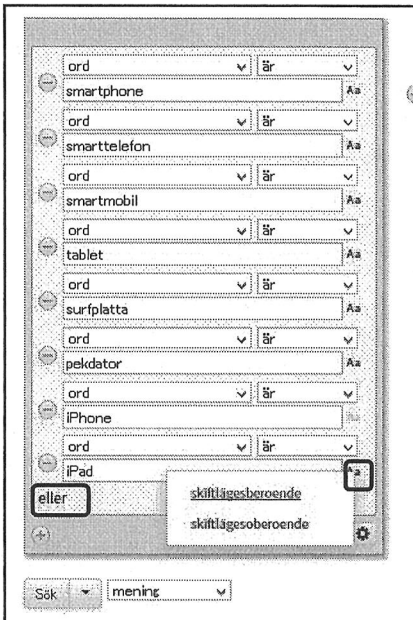
スウェーデン語で「スマートフォン」を意味する単語はスウェーデン語版 Wikipedia によると smarttelefon, smartmobil など複数あり（記事の見出し語は smarttelefon となっている），同様に「タブレット」を意味する単語にも surfplatta や pekdator など複数の単語が掲載されている（記事の見出し語は surfplatta）．これらの単語および英語表記の smartphone や tablet の用例を集め，実際にどの表記が主に使用されているのかを調べる．また，「iPhone」や「iPad」をスウェーデン語版 Wikipedia で調べると Iphone, Ipad と頭文字が大文字で書かれており，これらの表記と iPhone, iPad という表記のどちらが実際にスウェーデン語のテキストで使用されているかについても調べる．

【使用するコーパス】

検索対象となる単語がここ数年の間に使用されるようになった新語のため，2007 年以降の新聞記事が収録されている GP 2007～2013 および Webbnyheter 2007～2013 の計 7 つのコーパスを使用する．

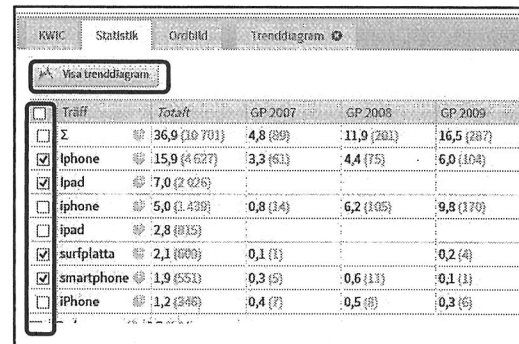
【検索手順】

まず拡張検索画面で 1 つ目の検索条件に［ord är smartphone］〈smartphone という単語である〉と設定する．検索語枠の左下にある［eller］（図 9）をクリックすると次の検索条件入力欄が表示されるので［ord är smarttelefon］〈smarttelefon という単語である〉と設定する．同様の手順で smartmobil, tablet, surfplatta, pekdator, iPhone, iPad も検索条件に加える．iPhone と iPad については，頭文字が大文字のものも同時に検索するために，検索語入力欄の右側にあるアルファベットのマーク（図 9）をクリックし，［skiftlägesoberoende］を選択しておく．



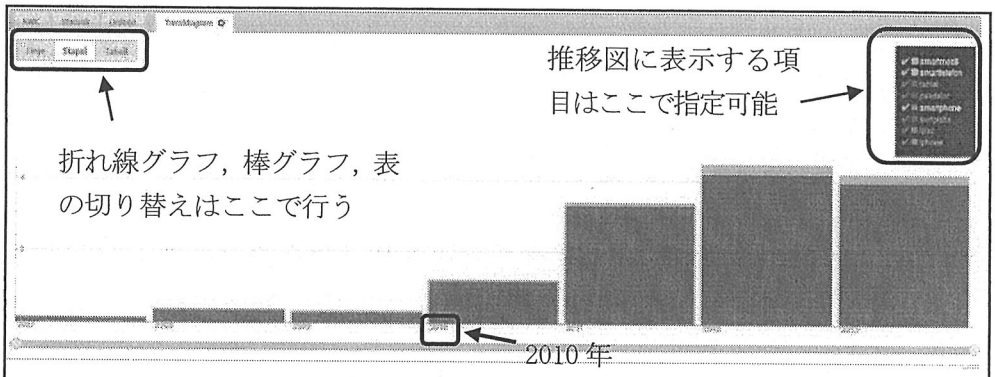
↑図 9 OR 検索の条件設定画面
（http://spraakbanken.gu.se/korp より）

以上の条件で検索を実行すると，先述の 8 つの検索語の用例が KWIC 形式で表示される．［Statistik］タブをクリックすると，コーパスごとの検索語の頻度データ表（図 10）が表示される．この頻度データを見ると，スウェーデン語の新聞では iPhone，iPad よりも頭文字を大文字にした Iphone，Ipad という表記のほうがよく使用されていることが分かる．



↑図 10 頻度データ表の一部
（http://spraakbanken.gu.se/korp より）

さらにここで得られた頻度データを推移図で表示するには，推移図に表示したい項目にチェックマークを入れて，頻度データ表の上にある［Visa trenddiagram］（図 10）をクリックする．推移図は折れ線グラフと棒グラフ，そして年代別の統計表の 3 種類がある．図 11 は smartphone，smartmobil，smarttelefon の頻度を棒グラフにしたもので



↑図 11 推移図（棒グラフ）の例（http://spraakbanken.gu.se/korp より）

あり，ここからスウェーデンの新聞において「スマートフォン」を意味する単語の中では英語の smartphone が圧倒的な割合を占めていることや，2010 年頃から smartphone という単語の使用回数が著しく増加していった様子がうかがえる．

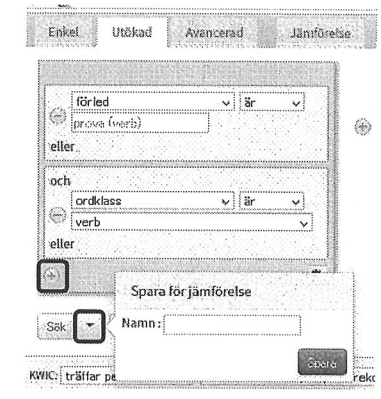
一方，同様に「タブレット」を意味する tablet，surfplatta，pekdator の 3 語の頻度データで推移図を作成すると，英語の tablet ではなく surfplatta が最もよく使用されている単語だということが明らかになる．

4. 2. 2. AND 検索と検索結果の比較

先ほどの OR 検索は複数の条件のうちいずれか 1 つでも該当するものを検索する方法であったが，次は複数の条件を全て満たすものを検索する AND 検索について検索の例を挙げる．

【検索内容】

「試す」を意味するスウェーデン語の動詞 prova と testa をそれぞれ前要素とする合成語のうち，品詞が動詞のものに絞って検索する．また，それらの検索結果で後要素に用いられる動詞を比較するために，検索結果の比較機能を使用する．



↑図 12 AND 検索の条件設定画面（http://spraakbanken.gu.se/korp より）

【使用するコーパス】

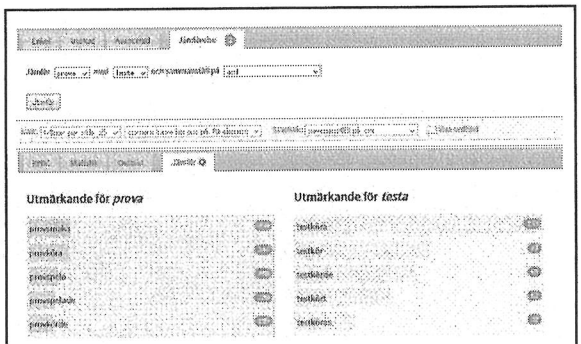
カテゴリー［Tidningstexter］に分類される 37 のコーパスを使用する．

【検索手順】

まずは前要素が prova の複合動詞を検索する．つまり「前要素が prova である」と「品詞が動詞である」という 2 つの条件を同時に満たすものを検索する．拡張検

索画面で１つ目の検索条件に［förled är prova (verb)］〈前要素が動詞 prova（の lemgram）である〉と設定する．検索語枠の左下にある＋マーク（図 12）をクリックすると次の条件の入力欄が表示されるので，そこに［ordklass är verb］〈品詞が動詞である〉と設定する．ここで［Sök］をクリックすると検索が実行されるが，今回はこの検索結果を次に行う検索結果と比較するために一時的に保存する．保存するには［Sök］の右側にある三角のボタン（図 12）をクリックし，検索結果に名前を付けて保存する．ここでは［prova］と名付ける．次に，先ほどと同じ手順で「前要素が動詞 testa（の lemgram）である」と「品詞が動詞である」の２つの条件を設定し，この検索結果に［testa］と名前を付けて保存する．

次に検索方法選択タブの右側にある［Jämförelse］のタブをクリックすると，検索結果の比較画面が表示される．先ほど保存した［prova］と［testa］が比較対象に選択されていることを確認して［Jämför］をクリックすると，それぞれの検索結果で用例数が多かったものから順に表示される（図 13）．これを比較することで，prova と testa がそれぞれどのような動詞と結びついて複合動詞を形成しているのかを比較することができる．



↑図 13 ２つの検索結果の比較画面の一部
（http://spraakbanken.gu.se/korp より）

4.2.3. フレーズ検索と繰り返し機能

次に拡張検索で複数の単語からなるフレーズを検索する例を挙げる．この方法は特定の語法や構文の用例を集めたい場合に使用できる．また，検索オプションの繰り返し機能についても紹介する．

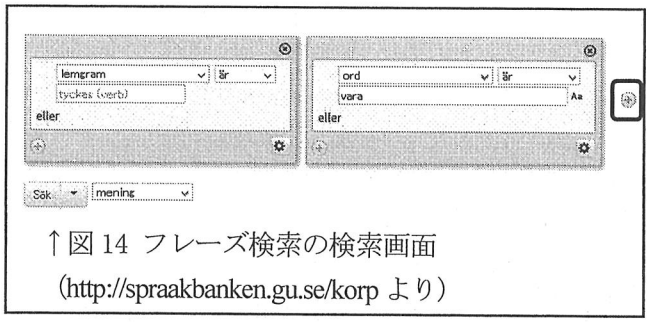
【検索内容】

［tyckas＋vara＋［形容詞］］〈［形容詞］であるように思われる〉というフレーズの用例をなるべく多く集めるため，lemgram を使用して動詞 tyckas の各変化形を使用した用例を一括検索する．また，さらに多くの用例を得るため，先ほどのフレーズに否定の inte を加えた［tyckas＋inte＋vara＋［形容詞］］も同時に検索する．

【使用するコーパス】

先ほどと同じくカテゴリー［Tidningstexter］の 37 のコーパスを使用する．

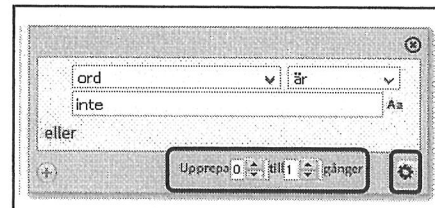
【検索手順】



↑図 14 フレーズ検索の検索画面
(http://spraakbanken.gu.se/korp より)

1つ目の検索語は動詞 tyckas とその各変化形であるので，検索条件に［lemgram är tyckas (verb)］〈lemgram が tyckas (verb) である〉と設定する．この後に続く検索語を設定するには検索語枠の右側にある＋マーク（図 14）をクリックして検索語枠を増やす．次の単語は vara なので，2つ目の検索語枠に［ord är vara］〈vara という単語である〉と条件設定をする．同様に＋マークをクリックして検索語枠を増やし，3つ目の検索語枠に［ordklass är adjektiv］〈品詞が形容詞である〉と設定し，検索を実行する．

次に，先ほどよりもさらに多くの用例を得るために［tyckas (lemgram)＋vara＋［形容詞］］および［tyckas (lemgram)＋inte＋vara＋［形容詞］］を一括で検索する．検索語枠を4つに増やし，それぞれに［lemgram är tyckas (verb)］，［ord är inte］〈inte という単語である〉，［ord är vara］，［ordklass är adjektiv］と条件設定を行う．これだけでは［tyckas (lemgram)＋inte＋vara＋［形容詞］］という否定文の用例しか得られないので，繰り返し機能を使って「inte を 0 回～1 回繰り返す」という設定を行うことで肯定文と否定文を同時に検索できるようにする．繰り返しを行いたい単語の検索語枠（inte を入力した枠）の右下にある歯車のマーク（図 15）をクリックするとオプションメニューが表示されるので，［upprepa］を選択．検索語枠下部に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］〈0～1 回繰り返す〉と設定する（図 15）．これで検索を実行すると［tyckas (lemgram) ＋（inte が 0 回繰り返される＝inte を含まない）＋vara＋［形容詞］］と［tyckas (lemgram) ＋ inte＋vara＋［形容詞］］という構造の用例を一括して検索できる．



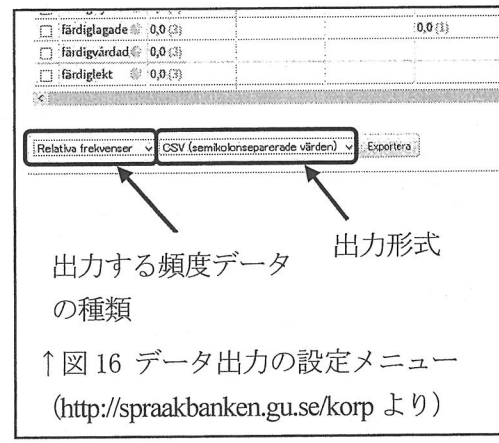
↑図 15 繰り返し設定
(http://spraakbanken.gu.se/korp より)

5. 頻度表の出力・保存方法

コーパスを使用するメリットとして，実際に使用された「生きた」言語を用例として収集できることだけでなく，検索した語やフレーズの用例数を頻度データ

として得られることも挙げられる．この頻度データを分析することにより，検索した語やフレーズが一般的に定着して使用されているものなのかを調べることができる．また，複数の類義語の頻度データを比較することにより，各語の振る舞いについて統計にもとづいた考察が可能となる．検索語が新語の場合は，年代の異なる複数のコーパスから頻度データを得ることで，その検索語が使用され始めた時期や，その後の使用頻度の変移を年代ごとに把握することも可能である．このように，Korp で得られる頻度データはスウェーデン語学における様々な研究で使用できる．本章では Korp で得られた頻度データを Excel ファイルに出力し保存する方法について説明する．なお，ここで説明する方法および説明に使用する図は Windows8 搭載のパソコンで Excel2013 を使用した場合のものであり，他の OS や他のソフトを使用する場合は手順などが多少異なる可能性がある．

まず頻度データを得たい用例を検索し，検索結果画面で［Statistik］のタブをクリックして頻度データを表示する．ここでは，形容詞 färdig を前要素とする複合形容詞を［Tidningstexter］に分類される 37 のコーパスで検索し，その用例数（絶対頻度）の表を出力する場合を例とする．



↑図 16 データ出力の設定メニュー（http://spraakbanken.gu.se/korp より）

画面の左下にデータ出力の設定メニューがあるので，ここで出力する頻度データの種類および出力形式を設定する（図 16）．ここでは出力形式は［CSV (Semicolonseparerade värden)］〈CSV 形式（セミコロン区切り）〉を選択し，出力データの種類は今回出力する値は用例数（絶対頻度）なので［Absoluta frekvenser］を選択して［Exportera］をクリックする．すると［export.csv を開く］というウィンドウが表示されるので，［ファイルを保存］を選択して［OK］をクリックする．

	A	B	C	D	E	F	G
1	Träff	Totalt	GP 2009	GP 2008	GP 2001	GP 2003	GP 2002
2	Σ	355	16	11	10	13	15
3	färdigslitna	10	0	1	1	1	0
4	färdigrustade	1	0	0	0	0	0
5	Färdigblandade	4	0	0	0	0	0
6	färdigbrutet	1	0	0	0	0	0
7	färdiglagad	16	2	0	0	1	0
8	färdigberedda	1	0	0	0	0	0
9	färdigproducerat	3	0	1	1	0	0
10	färdigutvecklad	2	0	0	0	0	0
11	färdigbeställda	1	0	1	0	0	0
12	färdigfrysta	4	0	0	0	0	0
13	färdigbruten	2	0	0	0	0	1
14	färdigkört	30	2	1	1	3	2

↑図 17 Excel に出力した頻度データ表の一部

次に Excel を起動させる．画面上部のメニューから［データ］を選択し，その下の［テキストファイル］

を選択（図 17）すると，［テキストファイルのインポート］というウィンドウが表示される．そこで先ほど保存したファイルを選択し，［インポート］をクリックする．すると［テキスト ファイルウィザード - 1/3］というウィンドウが表示されるので，［次へ］をクリックする．次のウィンドウでは［区切り文字］設定メニューで［セミコロン］を選択して［完了］をクリックする．さらに次の［データの取り込み］のウィンドウで［OK］をクリックすると表が完成する（図 17）．これに罫線を加えるなど見やすいように加工して保存しておくことで，レポート執筆や研究の際にいつでもデータの確認ができる．

6. おわりに

本稿ではコーパス検索システム Korp で検索できる様々なコーパスを紹介し，現代スウェーデン語コーパスを使った基本的な検索方法についてまとめた．Korp で検索できるコーパスに収録されている言語データは，スウェーデン語学習者にとっては辞書に掲載されている例文だけでは得ることのできないスウェーデン語語彙の用法のお手本となり，学習の助けとして活用できるであろう．また，スウェーデン語学を研究する者にとっては，研究の対象とする単語やフレーズなどについての用例を出典が明らかな言語データから効率的に集めることができるだけでなく，その検索結果をもとにした統計データを得ることのできる非常に有益なツールである．本稿で紹介した検索例は一部に過ぎず，特に拡張検索では様々な機能や検索条件を組み合わせることで，各々の目的にあった検索方法が可能になるであろう．さらに，Korp で使用できるコーパスの中には新聞記事，文学作品，政治・法律関係の文書，医学関係・学術関係のテキストやデジタル化された歴史資料など実に様々な分野のデータが含まれているため，スウェーデン語学の研究のみならず，他分野の研究においても活用が期待される．Korp はスウェーデン語学に関わる人以外にはまだあまり知られていないようであるが，本稿がスウェーデン語学に関わる人々だけに限らず，日本でスウェーデン語を学んでいる人々やスウェーデンに関する言語以外の分野を学び研究している人々が Korp の存在を知り，興味を持つきっかけとなれば幸いである．

Om korpussökningsverktyget Korp och dess användning

– Särskilt om sökning i det moderna materialet –

Aya Umetani

Sammanfattning

Denna artikel handlar framför allt om användning av Göteborgs universitets korpussökningsverktyg Korp. Syftet med artikeln är att beskriva och förklara hur man gör olika sökningar i Korp för japaner som forskar i eller studerar svenska språket och inte använt korpusar förut. Det andra och tredje kapitlet handlar om vilka sorters material man kan söka i i Korp. I fjärde kapitlet anges några sökningsexempel i enkel sökning och utökad sökning, t.ex. lemgramsökning, förled- och efterled-sökning, hur man bygger ihop olika sökuttryck, hur man jämför resultatet från två sökningar osv. Sedan handlar det femte kapitlet om hur man exporterar och sparar statistiktabellen som Excel-fil. Jag hoppas att fler personer i Japan som forskar i eller studerar svenska språket, och förhoppningsvis även de som studerar något annat ämne än språkvetenskap, kommer att bli intresserade av korpusarna och Korp genom att läsa denna artikel.

参 考 文 献

石川慎一郎. 2012.『ベーシック コーパス言語学』. 東京：ひつじ書房.
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』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

Digidaily. http://digidaily.blogg.kb.se/
Dramawebben. http://www.dramawebben.se/
Fornsvenska textbanken. http://project2.sol.lu.se/fornsvenska/
Litteraturbanken. http://litteraturbanken.se/
Projekt Runeberg. http://runeberg.org/
Riksarkivet. http://riksarkivet.se/
Svensk prosafiktion 1800–1900. http://spf1800-1900.se
Språkbanken. http://spraakbanken.gu.se/
Wikipedia. http://sv.wikipedia.org/