This quick reference gives a concise overview of the most commonly needed features of Simple Query Syntax; see Chapter 6 of Hoffmann et al. (2008) for a comprehensive reference and tutorial. Query expressions that you can enter in BNCweb’s search box are printed in typewriter font, followed by an arrow and the matching words or word sequences in italics (e.g. st?ing → sting, stung).

**Basic word form searches**

- To search for word forms, simply type them into the query field and click [Start query]: glitterati → glitterati
- Use wildcards for unspecified letters, and prefix or suffix searches:

  ```
  ? for a single arbitrary character
  s?ng → sing, sang, song, ...
  * for zero or more characters
  *able → able, table, capable, suitable, available, ...
  + for one or more characters
  +able → table, capable, suitable, ... but not able
  ??+ for three or more characters, etc.
  ??+able → capable, ... but not able, table, unable, stable
  ```

- Combine multiple wildcards: *oo+[able,ability] → Voodoo, schoolroom, ...
- Protect wildcards and other metacharacters with backslash \ to match the literal character (called "escaping" the metacharacter):

  ```
  \? → ?
  ? → a, b, c, ..., A, B, C, ..., 1, 2, 3, ..., !, ?, ...
  ```

Simple Query Syntax uses the following metacharacters:

- List comma-separated alternatives (optionally including wildcards) in square brackets:

  ```
  ??+[able,ability] → capable, capability, availability, ...
  neighbo[u,r] → neighbour, neighbor
  ```

- Searches are case-insensitive by default: the queries bath, Bath and BATH find the same matches (viz. the three word forms bath, Bath and BATH). Set the "Query mode" drop-down menu to "Simple query (case-sensitive)" to distinguish between AIDS and aids, for example.
- Use :d modifier to ignore accents: fiancée:d → fiancée, fiancée (for details, see Hoffmann et al. 2008, Section 6.10 and Appendix 4).
Matching parts-of-speech (POS)

- Search for a word form with a specific POS tag by linking them with an underscore _. Wildcards can be used both for word form and POS tag:

  - `lights_NN` ➔ plural noun `lights`, but not the verb form `lights`
  - `*ly_AJ` ➔ adjectives ending in -ly (e.g. `daily`)
  - `super+_V` ➔ verb forms starting with `super-`

- You can also search by POS tag only: `_PNX` ➔ any reflexive pronoun
- Complete listing of POS tags used in the BNC is given on last page.
- Use simplified POS tags enclosed in curly braces: `super+_{VERB}` for verb forms starting with `super-` (no wildcards allowed in simplified tags).
- List of simplified POS tags (Table 3.8 of Hoffmann et al. (2008) shows comparison with full tagset):

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, ADJ</td>
<td>adjective</td>
</tr>
<tr>
<td>N, SUBST</td>
<td>noun</td>
</tr>
<tr>
<td>V, VERB</td>
<td>verb</td>
</tr>
<tr>
<td>ADV</td>
<td>adverb</td>
</tr>
<tr>
<td>ART</td>
<td>article</td>
</tr>
<tr>
<td>CONJ</td>
<td>conjunction</td>
</tr>
<tr>
<td>INT, INTERJ</td>
<td>interjection</td>
</tr>
<tr>
<td>PREP</td>
<td>preposition</td>
</tr>
<tr>
<td>PRON</td>
<td>pronoun</td>
</tr>
<tr>
<td>$, STOP</td>
<td>punctuation</td>
</tr>
<tr>
<td>UNC</td>
<td>other / uncertain</td>
</tr>
</tbody>
</table>

- Keep in mind that part-of-speech tags have been assigned by an automatic software tool and are not always correct (try e.g. `beer_{N} can_{N}`).

Headword and lemma queries

- Search by headword, enclosed in curly braces: `{light}` finds the forms `light`, `lights`, `lit`, `lighted`, `lighting`, `lighter` and `lightest` (but not the nouns `lighting` and `lighter`).
- In BNCweb, the lemma is a combination of headword and simplified POS tag, separated by a slash / . A lemma query distinguishes e.g. between the noun, verb and adjective reading of `LIGHT`:

  - `{light/V} ➔ light, lights, lit, lighted, lighting (tagged as verb)`
  - `{light/N} ➔ light, lights (tagged as noun)`
  - `{light/A} ➔ light, lighter, lightest (tagged as adjective)
Word sequences

- Queries can consist of multiple words, e.g. talk of the town
- All words and punctuation symbols ("tokens") are separated by blanks; possessives (Peter's) and contracted forms (they've, gonna) must be split:
  he will \, wo n't he \?  \\rightarrow he will, won't he?
- Each query item in a sequence can make full use of wildcards, part-of-speech constraints, and headword or lemma searches:
  \{number/N\} of \{A\} _NN2  \\rightarrow numbers of younger men, ...
- Use + to skip an arbitrary token, or * for an optional token. Combine + and * for larger gaps, e.g. +++* to skip between 3 and 5 tokens.
  \{eat\} * up  \\rightarrow eat up, ate up, eat it up, eaten all up, ...
  \{eat\} + up  \\rightarrow eat it up, eaten all up, ... but not eat up, ate up
  \{eat\} +++ up  \\rightarrow up at a distance of 3 or 4 tokens after eat

Advanced lexico-grammatical patterns

- Use regular expression notation (Hoffmann et al. 2008, Sections 6.8 and 12.4) for alternatives, optional elements and repetition within a sequence:
  
  \begin{align*}
  (_{A})? & \quad \text{optional adjective} \\
  (_{A})* & \quad \text{zero or more adjectives (optional)} \\
  (_{A})+ & \quad \text{one or more adjectives (non-optional)} \\
  (_{A})\{2,4\} & \quad \text{between two and four adjectives} \\
  (...|...|...) & \quad \text{matches one of the alternatives indicated by ...} \\
  (...|...|...)* & \quad \text{alternatives with repetition (optional)} \\
  (...|...|...)+ & \quad \text{alternatives with repetition (non-optional)} \\
  (...|...|...)\{2,4\} & \quad \text{between two and four repetitions of the given alternatives (may be mixed in any order)}
  \end{align*}

- Regular expression notation can be nested to match complex patterns:
  the (most _AJ0 | _AJS) {man}  \\rightarrow the biggest men, the most attractive man, ...
  the (most (_AV0)? _AJ0 | (_AV0)? _AJS) {man}  \\rightarrow plus: the very richest men, the most supremely stupid men, ...
- Complex syntactic patterns can be formed, e.g. for a prepositional phrase:
  \_\{PREP\} (_\{ART\})? ((_\{ADV\})? (_A)\}* _\{N\}
  "a preposition; followed by an optional article; followed by any number of adjectives (zero or more), each of which may optionally be preceded by an adverb; followed by a noun"
XML tags

- XML start and end tags can be inserted in query expression to match the boundaries of a region, e.g. the start (<s>) or end (<s>) of an s-unit:

  - <s> but  \( \rightarrow \) s-unit beginning with *but* (or *But*)
  - _{ART} <s> \( \rightarrow \) article at end of s-unit (mostly errors)

- To match a complete region, skip all tokens between the start and end tag:

  - <quote> (+)+ </quote>  \( \rightarrow \) list of all quotations in the BNC
  - <mw> (+)+ </mw>  \( \rightarrow \) list of all multiword units

- Some useful XML tags in the BNC:

  - <s> ... </s>  \( \rightarrow \) s-unit
  - <p> ... </p>  \( \rightarrow \) paragraph
  - <u> ... </u>  \( \rightarrow \) speaker turn
  - <head> ... </head>  \( \rightarrow \) heading or caption
  - <quote> ... </quote>  \( \rightarrow \) quotation
  - <item> ... </item>  \( \rightarrow \) list item
  - <hi> ... </hi>  \( \rightarrow \) highlighted text
  - <mw> ... </mw>  \( \rightarrow \) multiword unit

Proximity queries

- Special syntax for searching one item within a specified range of another:

  - kick <<s>> bucket  \( \rightarrow \) *kick* and *bucket* in the same sentence
  - \{kick/V\} <<s>> bucket_{NN1} (can use POS/lemma constraints)
  - day <<3>> night  \( \rightarrow \) *day* and *night* within range of 3 tokens
  - day <<5<< night  \( \rightarrow \) *night* ... *day* (within 5 tokens)
  - day >>5>> night  \( \rightarrow \) *day* ... *night* (within 5 tokens)

- Only the left element ("target") will be highlighted on the result page. The right element is considered as a "constraint" that must be satisfied.

- Multiple constraints can be chained:

  - \{day\} <<5>> \{month\} <<5>> \{year\}

  In this case, *day* must co-occur with *month* as well as *year* in a 5-token window; only *day* will be highlighted on the Query result page.

- Proximity queries can be nested with parentheses:

  - \{waste/V\} <<s>> (time <<3>> money)

  Here, the verb *waste* must co-occur with *time* as well as *money* in the same sentence; but *time* and *money* must be closer together (within a 3-token window). Again, only instances of *waste* will be highlighted.

- Proximity queries cannot be combined with lexico-grammatical patterns!
List of part-of-speech tags in the BNC (CLAWS-5 tagset)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AJ0</td>
<td>Adjective (general or positive) (e.g. good, old, beautiful)</td>
</tr>
<tr>
<td>AJC</td>
<td>Comparative adjective (e.g. better, older)</td>
</tr>
<tr>
<td>AJS</td>
<td>Superlative adjective (e.g. best, oldest)</td>
</tr>
<tr>
<td>AT0</td>
<td>Article (e.g. the, a, an, no)</td>
</tr>
<tr>
<td>AV0</td>
<td>General adverb: an adverb not subclassified as AVP or AVQ (see below) (e.g. often, well, longer (adv.), furthest)</td>
</tr>
<tr>
<td>AVP</td>
<td>Adverb particle (e.g. up, off, out)</td>
</tr>
<tr>
<td>AVQ</td>
<td>Wh-adverb (e.g. when, where, how, why, wherever)</td>
</tr>
<tr>
<td>CJC</td>
<td>Coordinating conjunction (e.g. and, or, but)</td>
</tr>
<tr>
<td>CJS</td>
<td>Subordinating conjunction (e.g. although, when)</td>
</tr>
<tr>
<td>CJT</td>
<td>The subordinating conjunction that</td>
</tr>
<tr>
<td>CRD</td>
<td>Cardinal number (e.g. one, thirty-five, thirty-six)</td>
</tr>
<tr>
<td>DPS</td>
<td>Possessive determiner-pronoun (e.g. your, their, his)</td>
</tr>
<tr>
<td>DT0</td>
<td>General determiner-pronoun: i.e. a determiner-pronoun which is not a DTQ or an AT0.</td>
</tr>
<tr>
<td>DTQ</td>
<td>Wh-determiner-pronoun (e.g. which, what, whose, whichever)</td>
</tr>
<tr>
<td>EX0</td>
<td>Existential there, i.e. there occurring in the there is... or there are... construction</td>
</tr>
<tr>
<td>ITJ</td>
<td>Interjection or other isolate (e.g. oh, yes, mhm, wow)</td>
</tr>
<tr>
<td>NN0</td>
<td>Common noun, neutral for number (e.g. aircraft, data, committee)</td>
</tr>
<tr>
<td>NN1</td>
<td>Singular common noun (e.g. pencil, goose, time, revelation)</td>
</tr>
<tr>
<td>NN2</td>
<td>Plural common noun (e.g. pencils, geese, times, revelations)</td>
</tr>
<tr>
<td>NP0</td>
<td>Proper noun (e.g. London, Michael, Mars, IBM)</td>
</tr>
<tr>
<td>ORD</td>
<td>Ordinal numeral (e.g. first, sixth, seventh, last)</td>
</tr>
<tr>
<td>PNI</td>
<td>Indefinite pronoun (e.g. none, everything, one (as pronoun), nobody)</td>
</tr>
<tr>
<td>PNP</td>
<td>Personal pronoun (e.g. I, you, them, ours)</td>
</tr>
<tr>
<td>PNQ</td>
<td>Wh-pronoun (e.g. who, whoever, whom)</td>
</tr>
<tr>
<td>PNX</td>
<td>Reflexive pronoun (e.g. myself, yourself, itself, ourselves)</td>
</tr>
<tr>
<td>POS</td>
<td>The possessive or genitive marker 's or 's</td>
</tr>
<tr>
<td>PRF</td>
<td>The preposition of</td>
</tr>
<tr>
<td>PRP</td>
<td>Preposition (except of) (e.g. about, at, in, on, with)</td>
</tr>
<tr>
<td>PUL</td>
<td>Punctuation: left bracket, i.e. ( or [</td>
</tr>
<tr>
<td>PUN</td>
<td>Punctuation: general separating mark (., !, ;, – and ?)</td>
</tr>
<tr>
<td>PUQ</td>
<td>Punctuation: quotation mark (‘ and ”)</td>
</tr>
<tr>
<td>PUR</td>
<td>Punctuation: right bracket, i.e. ) or ]</td>
</tr>
<tr>
<td>TO0</td>
<td>Infinitive marker to</td>
</tr>
<tr>
<td>UNC</td>
<td>Unclassified items which are not appropriately considered as items of the English lexicon.</td>
</tr>
</tbody>
</table>
### Verb Forms

**VBB**  The present tense forms of the verb BE (except for *is* and *'s*), i.e. *am, are, 'm, 're and be* (subjunctive or imperative)

**VBD**  The past tense forms of the verb BE: *was* and *were*

**VBG**  The *-ing* form of the verb BE: *being*

**VBI**  The infinitive form of the verb BE: *be*

**VBN**  The past participle form of the verb BE: *been*

**VBD**  The finite base form of the verb DO: *do*

**VDD**  The past tense form of the verb DO: *did*

**VDG**  The *-ing* form of the verb DO: *doing*

**VDI**  The infinitive form of the verb DO: *do*

**VDN**  The past participle form of the verb DO: *done*

**VDZ**  The *-s* form of the verb DO: *does, 's*

**VHB**  The finite base form of the verb HAVE: *have, 've*

**VHD**  The past tense form of the verb HAVE: *had, 'd*

**VHG**  The *-ing* form of the verb HAVE: *having*

**VHI**  The infinitive form of the verb HAVE: *have*

**VHN**  The past participle form of the verb HAVE: *had*

**VHZ**  The *-s* form of the verb HAVE: *has, 's*

**VM0**  Modal auxiliary verb (e.g. *will, would, can, could, 'll, 'd*)

**VVB**  The finite base form of lexical verbs, comprising the indicative, imperative and present subjunctive (e.g. *forget, send, live, return*)

**VVD**  The past tense form of lexical verbs (e.g. *forgot, sent, lived, returned*)

**VVG**  The *-ing* form of lexical verbs (e.g. *forgetting, sending, living, returning*)

**VVI**  The infinitive form of lexical verbs (e.g. *forget, send, live, return*)

**VVN**  The past participle form of lexical verbs (e.g. *forgotten, sent, lived, returned*)

**VVZ**  The *-s* form of lexical verbs (e.g. *forgets, sends, lives, returns*)

**XX0**  The negative particle *not* or *n't*

**ZZ0**  Alphabetical symbols (e.g. *A, a, B, b, c, d*)

### References