TLG Technical Note 001: Greek Word Definition

Authored: Nick Nicholas, TLG
Maintained by tlg@uci.edu
Created: October 1999
Last Revised: 2002-11-23

The TLG maintains a word index of the Greek words occurring in its texts. Versions of the word index appear on the CD ROMs published by the TLG, and on the online TLG Search Engine. The following documents what constitutes a distinct word in the compilation of the TLG word index.

1. Definitions

Different versions of the TLG word index are referred to in the following:

- **CD ROM #D** is the CD ROM created at the Packard Humanities Institute for the TLG in 1992.
- **CD ROM #E** is the CD ROM created at the TLG in 1999. The software to create the CD ROM was developed afresh, and the resulting CD ROM differs in several details from its predecessor.
- The **Web index** is the TLG online version of the word index.

2. Word delimiters

A new word is assumed to occur whenever a blank, a dash (— ; Beta code _), a punctuation sign, or a new line occurs in the source text.

Punctuation is defined as one of the following characters:

- `.` (period: .)
- `,` (comma: ,)
- `;` (Greek question mark: ;)
- `:` (Greek raised dot: ·)
- `%4` (exclamation point: !)
- `%10` (dicolon: :)
- `%100` (Roman semicolon: ;)

However, `%1` (Roman question mark: ?), which may indicate a doubtful letter, is not considered to be punctuation.

Nor are quotation marks, which may on occasion be used as editorial brackets, or as continuing block quotation markers.

Comma is only considered punctuation when not followed immediately by a letter; otherwise, it is considered a hypodiastole.

On the Web Index, a new word is also assumed to occur whenever an apostrophe occurs in the source text; the current TLG orthographic norms do not admit word-internal apostrophe. Thus, `D'O(/S` δ'ὅς is considered two words, `D'` and `O(/S`. This also applies where the apostrophe has supplanted the coronis: `G'OUN` γ'ουν (normally `GOU)=N` γοῦν) is still analysed as the two words `G'` and `OUN`.

3. Normalization

3.1. Punctuation

Punctuation is normally stripped out of the word. (See below for exceptions.)

3.2. Hyphenation and Non-Text

Hyphenated words are joined; words in non-text brackets (`{}`) are indexed separately from words outside non-text brackets, and do not interfere with hyphenation. For example,

`A)/N- {STR.}`	ἄν-	στρ.
`QRWPOS`	θρωπος	

is indexed as `A)/NQRWPOS` ἄνθρωπος, and the marginal note `STR` στρ is indexed separately.

There is one exception: `{27 }27` (formerly occurring in work 1595.107), indicating apograph textual emendation, are treated as brackets, and are thus ignored in extracting words. For example, `ARIS{27T}27WN` αριστ*ων is indexed as `ARISTWN`, not `ARISWN`, `T`. (The escape code has been reassigned as of August 2000, so this exception is no longer relevant.)

All other non-text brackets by definition contain text extraneous to the main text (e.g. marginalia, stage directions, titles, editorial emendations); so their content is never concatenated with main text words, even if no space intervenes.

3.3. Roman Script

Words in other languages are ignored in the word index; this currently concerns only words tagged as being in Roman script, since words from other languages transliterated into Greek are not yet tagged distinctly. Thus, in a text like `*)APOMA/SAR &[1addidi]1 $E)/FH` ᾽Απομάσαρ *(addidi)* ἔφη, only the 'Greek' words `A)POMA/SAR` 'Abu-Masar' and `E)/FH` are indexed.

3.4. Beta escapes

Characters other than elision markers, the Greek alphabet, breathings, accents, iota subscripts, and commas—i.e. almost all beta escapes—are also ignored. For example, `"6$10A)$%5N?<1[H/]R>1#13` «ἀν|*[ή]ρ*※ is indexed as `A)NH/R` ἀνήρ .

Characters within a word in non-text or Roman font are also ignored (but see *Partial Words* below.)

3.5. Diacritics

Accents are regularized: grave accents become acute, only the first accent in a word is retained, and words are converted to lower case. Thus,

- `A)NH\R` ἀνὴρ is indexed as `A)NH/R` ἀνήρ,
- `A)/NQRWPO/S` ἄνθρωπός as `A)/NQRWPOS` ἄνθρωπος,
- `*STUMFALI/A` Στυμφαλία as `STUMFALI/A` στυμφαλία, and
- `*E*S*T*I*A` ΕΣΤΙΑ as `ESTIA` εστια.

3.6. Hypodiastolae

Since commas are potentially hypodiastolae, they are retained in the word, and are stripped out only if they do not correspond to a known instance of a hypodiastole; thus, `O(/,TI` ὅ,τι is distinguished from

O(/TI ὅτι. There are nine words with hypodiastolae known in our texts:

O(/,TI, O(/,TIPER, O(/,TTI, TO/,TE, O(/,TE, O(/,TOU, TA/,TE, TO/,T', H(/,TE
(ὅ,τι, ὅ,τιπερ, ὅ,ττι, τό,τε, ὅ,τε, ὅ,του, τά,τε, τό,τ', ἥ,τε)

Hypodiastolae are new to CD ROM #E; in CD ROM #D, they were stripped out of the word.

3.7. Coronides

All crasis markers in words are now retained; this is a change from CD ROM #D, in which crasis markers were dropped if they occurred after the first syllable of the word. Thus, KALOKA)/GAQOS καλοκἀγαθος is represented in the CD ROM #D index as KALOKA/GAQOS καλοκάγαθος, but KA)GW/ κἀγώ remains KA)GW/; both forms retain their crases on CD ROM #E.

On CD ROM #E the coronis is not normalized: the rough coronis is treated as distinct from the smooth coronis, although these are mere notational variants. Thus, XW(S χὡς is listed as a distinct word from XW)S χὼς. The two coronides are conflated on the online version of the index. However, a rough breathing is considered a coronis only when following an aspirated stop; otherwise, it is an internal breathing, and left as is (e.g. *)ABRAA(/M Ἀβραἁμ)

3.8. Internal breathing marks

The internal breathing marks on double rho are removed only if they are predictable; thus, R)R(is converted to RR, but R(R(is left as is. For example, A)/R)R(WSTOS ἄῤῥωστος is indexed as A)/RRWSTOS ἄρρωστος.

This is another change from CD ROM #D, which also converted R(R(, R(R) and R)R) to RR.

3.9. Character Stacking

On the web index, instances of characters superposed above other characters, as tagged with the Beta escapes <10 ... >10<11 ... >11 are deemed to be textual variants, and two versions of the word are recorded in the word index: that with the above reading, and that with the below reading. Thus, the string A)<10QH=>10<11MU/>11NAI ἀ^{μύ}_{θῆ}ναι is parsed as two words in the same spot, A)QH=NAI and A)MU/NAI.

3.10. Dittography

On the web index, if a bracket is known to indicate dittography or editorial or scribal deletion for a work, the indication is respected, and the marked portion of the word is excluded from the indexed word. If however the deletion bracket incorporates a word in its entirety, the word is indexed, as a variant reading. E.g. given the information that [3]3{ } in a given work are editorial deletions, E)/GW[3GW]3GE ἔγω{γω}γε is indexed as E)/GWGE ἔγωγε -- i.e. E)/GW~~GW~~GE ἔγω~~γε~~γε. However, in a text like E)GW\ [3DE\]3 EI)=PON ἐγὼ{δὲ} εἶπον, all three words are indexed.

4. Partial words

When a partial word is discovered in a text, the partial word is included in the current word index only if it contains two or more Greek letters (excluding diacritics) in sequence. This is at variance with CD ROM #D, which required three Greek letterals (including diacritics) in sequence. Thus, CD ROM #E includes word fragments like BA! βα. and excludes A)/! ἄ.; CD ROM #D would exclude BA! and include A)/!.

On CD ROM #D, only the missing letter code (`!`) was regarded as a partial word boundary. The repertory has been significantly expanded for CD ROM #E. A partial word boundary is discovered from the occurrence of:

- a missing letter code (`!`);
- a transition from Roman to Greek script within a word (though not vice versa):
  - `&1Iustinian$10\S` *Iustinian*ὸς;
- a final left bracket, followed by no Greek letterals between it and the word delimiter, *if* the preceding word is unaccented, oxytone (not barytone), or ends in a letter which is not legal for a Greek word (vowel, *s*, *n*, '):
  - `ABG[` *αβγ[*,
  - `$13ABG[$10_` *αβγ[*—;
- an initial right bracket preceded by a space (or equivalent `^`, though not `@` --- which acts as indentation), or the beginning of the line, and followed by text, disregarding any intervening beta escapes other than space or equivalents:
  - `]ABG` *]αβγ*,
  - `^16]1ABG` *)αβγ*,
  - `]%ABG` *]†αβγ*;
- a left bracket followed by a right bracket without any intervening characters or beta-escapes:
  - `ABG[]` *αβγ[]*,
  - `[2]2ABG` *()αβγ*,
  - `AB[]GD` *αβ()γδ*;

or:

- a line-initial instance, after a line-breaking hyphen, of `[]` (with optional digits qualifying the brackets: `[1]1` *()*, `[2]2` ⟨⟩ etc.):
  - ```
    AB- αβ-
    []  []
    ```
- or `]` preceded by:
  - nothing:
    - ```
      AB- αβ-
      ]4G ⟧ γ
      ```
  - non-spacing beta escapes (`$&<>{}`: font shifts, quasi-brackets (typographical formatting), non-text):
    - ```
      AB-   αβ-
      $10]G ]γ
      ```
  - or `[` followed by at least one space (but not an equivalent beta-escape, since these indicate *indentation* rather than lacunae):
    - ```
      AB-     αβ-
      [ AB]G [ αβ]γ
      ```
    - but not
      ```
      @@*)/AB-    ῎Α-
      [@@]DHRA [   ]βδηρα
      ```
- a beta-escape which can stand in for omitted text:
  - the asterisk `%2` (but only when followed immediately by another asterisk in the text);
  - the hyphen `%19` (but only at the beginning or end of a word or word continuation, not word-internally);

- the diacritics without letters `%30-%39, %132-%134`;
- the metrical signs `%40-%49, %140, %141, %144, %145` ;
- the partial letter sign `#7`;
- the unintelligible letter sign `#99`;
- the blot/dot sign `#459`.

Thus,

- Any one of the following has an initial partial word boundary:

  `!A` .α

  `@]A`]α

  `]A`]α

  `]$10A`]α

  `#7A` ∀α
- Any one of the following has a final partial word boundary:

  `A!` α.

  `A[@` α[

  `A[` α[

  `A%2%2` α**

  (*but not* `A%2` α*)

  `A#459` α■
- Any one of the following contains a word boundary partial on both sides:

  `A!B` α.β,

  `A[]B` α[]β ,

  `A%40%40%41B` α‿¯β.
- And in the following hyphenated words, the hyphen becomes a partial word boundary:

`A-` `[]B`	α- []β
`A-` `[ &`12 spaces$]B`	α- [12 spaces]β
`A-` `]B`	α-]β
`A-` `%19B`	α- -β

`A-`
`[%40]B-`
`]`
	α-
	[˘]β-
	]

However,

`A-` α-
`@@B]` β]

does not have the hyphen as a partial word boundary: the `@` act merely to indent, and not to indicate a lacuna. Likewise, the bracket pair in `AB[1%1]1` αβ*(?)* does not indicate a lacuna: the presence of any code between brackets is taken to indicate the bracket pair is no lacuna --- unless the code itself denotes a lacuna (e.g. `AB[1%40]1G` αβ(˘)γ).

Furthermore,

```
A-      α-
$13]B   β
```

contains a partial word boundary; the `$13` is a font shift which can be ignored as a non-spacing Beta escape, not representing any actual text intervening between the hyphen and the right bracket. However,

```
A-      α-
#13]B   ※]β
```

does not contain a partial word boundary, since the *asteriskos* is a spacing Beta escape, and is deemed to start the current text line.

By this reckoning, the following are incomplete words:

- `O(MO!` in 0059.005:

  ```
  A)NTI\ TOU= "3O(MO%19"3, "3A)%19"3
  ἀντὶ τοῦ ‘**ὁμο-**’, ‘ἀ-’
  ```

- `!NEMON` in 0006.029:

  ```
  [*U*Y.] %19NEMON A)/GAGE/ POTE/
  [ΥΨ.] **-νεμον** ἄγαγε ποτέ
  ```

- `SAP[F!` in 0009.001:

  ```
  *SAP-          Σαπ-
  [F%19 ]GXANON  φ- ]γχανον
  ```

The following are complete words, by contrast:

- `XILIONTA%19ETHRI/DAN` in 2866.001 (χιλιοντα-ετηρίδαν),
- `YALLO/%19[4MENON]4` in 2703.001 (ψαλλό- ⟦μενον⟧).

And the alternative readings `A)MAMA/CUD[1%19OS, %19ES]1` (0009.001: ἀμαμάξυδ(-ος, -ες) = `A)MAMA/CUDOS` *vel* `A)MAMA/CUDES`) are indexed as `A)MAMA/CUDOS, !ES`.

If a beta escape denoting a lacuna occurs at the beginning of a word, it is deemed part of the word unless a blank, dash, or new line follows it: the words `%40A` ˘α and `[%40]A` [˘]α contain the breve, and thus are words with incomplete beginnings; but the word `[%40] A` [˘] α is not incomplete, since the breve is considered a separate word.

Exceptionally, as of November 2002, if a lacuna is followed immediately by a vowel with a breathing mark, the word is deemed complete. This is because the breathing mark is deemed to mark the beginning of a word rather than a coronis, and the source texts are rarely consistent in delimiting words from lacunae. For instance, `!!!A)NH\R` ...ἀνήρ is treated as the complete word `A)NH/R` ἀνήρ .

5. Special hyphen rules

If a beta escape denoting a lacuna occurs on a new line after a hyphen, it is deemed part of the hyphenated word automatically; if a blank, dash or new line ensues, the word is then terminated. Thus,

```
A-    α-
%40B ˘β
```

is indexed as `A!B` = `A!`, `!B`, but

```
A-    α-
%40 B ˘β
```

is indexed as `A!`, `B`.

A hyphenated word is also terminated immediately if a space follows a left bracket in a hyphenated word continuation; what follows that space is not deemed part of the word. For instance,

```
I(STORI/-          ἱστορί-
[     A)NAGKAI/AN] [   ἀναγκαίαν]
```

is indexed as `I(STORI/!`, `A)NAGKAI/AN` — not `I(STORI/A)NAGKAI/AN` ἱστορίἀναγκαίαν .

If a hyphen follows a lower case letter within a portion of a word containing an unclosed left bracket, and the continuation of the word in the next line consists of a capital letter not followed by at least one more capital letter, then it is deemed impossible for the two fragments to be part of the same word, and the hyphen is considered a final partial word boundary. Thus, in an instance like

```
F[ILE-          φ[ιλε-
*)AS]KLHPIOU= ᾽Ασ]κληπιοῦ
```

(0020.004), the program refuses to join the two fragments, and indexes them as `FILE!`, `*)ASKLHPIOU=`. This results because the bracket in `F[ILE-` explicitly flags what follows in the word as fragmentary. The casing of `*)ASKLHPIOU=`, on the other hand, with its initial capital, indicates that it must constitute a new word. Without that casing, the algorithm would still join the two word fragments:

```
F[ILE-     φ[ιλε-
S]KLHPIOU= σ]κληπιοῦ
```

is interpreted as `FILESKLHPIOU=` φιλεσκληπιοῦ, and

```
F[ILE-            φ[ιλε-
*S]*K*L*H*P*I*O*U Σ]ΚΛΗΠΙΟΥ
```

as `FILE*S*K*L*H*P*I*O*U` φιλεΣΚΛΗΠΙΟΥ.

If a hyphen preceded by an unclosed left bracket is followed by a continuation word with a *breathing mark*, as distinct from a coronis (for which the heuristic is that the vowel with the breathing mark occur before any consonants in the line), it is likewise deemed impossible for the two words to be joined. Thus,

```
*KLEANAK[TID- Κλεανακ[τιδ-
H(            ἡ
```

(0009.003) is indexed as `KLEANAKTID!`, `H(`: by having a breathing mark, `H(` cannot be anything but the beginning of a new word.

The necessity of the bracket in such checks is shown by the instance of

PERI-	περι-
E(/LKESQAI	ἕλκεσθαι

in the non-fragmentary text 2734.013.

If a hyphen is followed in the same line by a letter or punctuation, the word is deemed to be incomplete, and is not joined with the next line. Thus, if a line ends in FILE-., that text is not joined with the next line, but is indexed as FILE! .

6. Ellipses & Abbreviations

An instance of more than one contiguous dot, either within or on the boundary of a word, is treated as a lacuna: thus, ..A, A.., and A..B are analyzed as !A, A!, and A!B = A!, !B. If a space intervenes between the multiple dots and the word, the dots are not treated as part of the word, which thus remains complete; this is how ellipses as punctuation are distinguished from lacunae. Thus, A)NH/R ... ἀνήρ ... is analyzed as A)NH/R, not A)NH/R!.

Multiple dots in the continuation of a hyphenated word are treated like the other lacunae: A- B... is indexed as AB!, and A- ...B as A!B. A single dot, on the other hand, terminates a word: A.B is indexed as A followed by B, and not as AB or A!B = A!, !B. If the dot was to represent a missing letter, it would have been encoded as ! in the Beta code instead.

Abbreviations are thus treated as separate words rather than a single word: e.g. K.O.K. κ.ο.κ. (= KAI\ OU)/TW| KAQECH=S καὶ οὕτῳ καθεξῆς 'and so forth') is indexed as K, O, K.

The one exception to this (new to CD ROM #E) is K.T.L. κ.τ.λ. ('*etc.*') which is indexed as the single word KTL κτλ (with no periods.)

The foregoing is an algorithmic approach to determining what constitutes a complete or partial word. It is by no means infallible, and several words will inevitably be indexed wrongly. Substantial further improvement of the process, however, can only be done manually.