Principles of Syriac Lemmatisation

Summary version

Nicolas Atas (Freie Universität Berlin) – August 2021

Lemmatisation is the operation identifying the lemma (headword) of each word-form attested in a given text. Each lemma is also accompanied by its part-of-speech tag (POS-tag) indicating its morphosyntactic category (noun, verb, adjective, pronoun, etc.), and, for Semitic languages as Arabic and Syriac, its root, and its origin. This paper describes these elements in a summarizing way[1]. For more information, see the website of the GREgORI Project (including complete bibliography)[2].

1. Word-forms

Syriac words are divided into two categories, monolexical word-forms and polylexical word-forms. Our system of analysis takes into account each of the constitutive lexical elements of these forms, whether they are monolexical or polylexical.

1. Monolexical word-forms which are constituted by a single word and can be directly related to a single lemma:

   a. ܟܬܒܐ (<ܟܬܒ̈ܐ “letter” Sokoloff, p. 9)
   b. ܩܕܡ (< ܩܕܡ “before” Sokoloff, p. 1318)
   c. ܐܠܗܐ (< ܐܠܗ̈ܐ “God” Sokoloff, p. 47)
   d. ܐܢܐ (< ܐܢ̈ܐ “I” Sokoloff, p. 58)
   e. ܡܢܐ (< ܡܢ̈ܐ “what” Sokoloff, p. 778)
   f. ܐܫܟܚ (< ܐܫ̈ܟܚ “to find” Sokoloff, p. 1556)
   g. ܨܒܝ̈ܐ (< ܨܒ̈ܝ̈ܐ “to desire” Sokoloff, p. 1435)
   h. ܐܬܦܗܡ (< 2 ܐܬ̈ܦ̈ܗ̈ܡ “to understand” Sokoloff, p. 1010)
   i. ܝܫܘܥ (< ܝܫ̈ܘܥ “Jesus” Payne Smith, col. 1638)

2. Polylexical word-forms which first need to be divided into multiple *word-forms* before they can be related to a lemma[3]:

   a. ܢܦܫܗ: ܢ@ܦܫ̈ܐ (< ܢܦ̈ܫܐ “soul” Sokoloff, p. 938 and ܗܘ “he” Sokoloff, p. 333)
   b. ܘܐܡܪ: ܘܐܡܪ@ܐ (<ـܐ “and” Sokoloff, p. 357 and ܐܡ̈ܪ “henceforth” Sokoloff, p. 758)
   c. ܠܗ ܡܨܐ: ܠܗ ܡܨ̈ܐ@ܐ (<ـܐ “and” Sokoloff, p. 357 and ܨ̈ܠ “to pray” Sokoloff, p. 1288)
   d. ܕܢܫܟܚ: ܕܢܫ̈ܟܚ@ܕ (<ـܕ “in order that” Sokoloff, p. 268 and ܐܫ̈ܟܚ “to find” Sokoloff, p. 1556)
   e. ܠܡܚܠ: ܠܡܚ̈ܠ@ܠ (<ـܠ “to, for” Sokoloff, p. 665 and ܡ̈ܚܠܐ “labor” Sokoloff, p. 1112)
   f. ܘܚܒܐ: ܘ@ܚܒ̈ܐ@ܐ (<ـܐ “and” Sokoloff, p. 357; ܚ̈ܒܐ “love” Sokoloff, p. 419 and ܗܘ “he” Sokoloff, p. 333)
   g. ܒܫܒܬܐ: ـ̈ܒ@ܫ̈ܒ@ܬ@ܐ (<ـܐ “and” Sokoloff, p. 357; ـܬ “in, through” Sokoloff, p. 114; ܫ̈ܒܩ “suffering” Sokoloff, p. 497 and ܚ̈ܢ “we” Sokoloff, p. 472)
   h. ܘܠܠܥܠܐ: ܘ@ܠ̈ܥ̈ܠ@ܠ@ܐ (<ـܐ “and” Sokoloff, p. 357; ـܠ “to, for” Sokoloff, p. 665; ܥ̈ܠܝܬܗ “glory” Sokoloff, p. 1518 and ܗܘ “he” Sokoloff, p. 333)
   i. ـ̈ܒܫܠܘܚܬܐ: ـ̈ܒ@ܫ̈ܠ@ܘ@ܚ@ܬ@ܐ (<ـܐ “and” Sokoloff, p. 357; ـܬ “of” Sokoloff, p. 268; ܫ̈ܒܩܠܘܚ “negligence” Sokoloff, p. 720 and ܚ̈ܢ “we” Sokoloff, p. 472)

---

[1] Thanks are due to all our collaborators in the field of Syriac lemmatisation for the GREgORI Project (alphabetical order): Naima Afif, Yury Arzhanov, Philip Michael Forness, Jean-Claude Haelewyck, Bastien Kindt, Manhal Makhoul, David Phillips, Marcel Pirard, Andrea Barbara Schmidt, Antonio Stefano Sembiante and Guido Venturini. A special gratitude is also due to Andrea Barbara Schmidt and Philip Michael Forness, who reviewed these lines and their English translation.

[2] See https://uclouvain.be/fr/instituts-recherche/incal/ciol/gregori-project.html.

[3] Polylexical word-forms are indicated in the corresponding column by the letter K (from the Greek “crasis”, contracted forms, appellation also used for prefixed and suffixed forms) and lexical elements included in polylexical word-forms are divided by means of the sign “@”.

2. Lemmas

Lemmas represent words as they appear in dictionaries. Our system of Syriac lemmatisation is based on Sokoloff's *Syriac lexicon*[4] used as "reference dictionary":

1. Lemmas correspond to the headwords found in Sokoloff's dictionary with eastern vocalization, all diacritical marks (including spirantization and linea occultans), in the emphatic state[5]:
    a. Form ܪܚܡ: lemma ܪܚܡܐ ("love" Sokoloff, p. 419)
    b. Form ܚܕܬ: lemma ܚܕܬܐ ("new" Sokoloff, p. 418)

2. Lemmas of proper names (anthroponyms and toponyms) follow the spelling found in Payne Smith's *Thesaurus*[6], without vocalization:
    a. Form ܝܫܘܥ: lemma ܝܫܘܥ ("Jesus" Payne Smith, col. 1638)
    b. Form ܕܝܕܝܫܘ: lemma ܕܝܕܝܫܘ ("Dadišoʿ" Payne Smith, col. 824)

3. Homographic lemmas are distinguished by a number conforming to the usage of Sokoloff's dictionary:
    a. Form ܣܒܪ: lemma "1 ܣܒܪ" ("to hope" Sokoloff, p. 964)
    b. Form ܣܒܪ: lemma "2 ܣܒܪ" ("to announce" Sokoloff, p. 965)

4. Word unrecorded neither in the reference dictionary nor in the database of the GREgORI project is recorded with lemma as intuitively expected by linguists:
    a. Form ܚܫܒܬܦܢܝܬܐ: lemma ܚܫܒܬܦܢܝܬܐ "being recompensed"
    b. Form ܟܚܝܠܢܝܐ: lemma ܟܚܝܠܢܝܐ "prone to disease"

3. POS-tags

Each lemma receives a tag that indicates the part of speech, the morphosyntactic category, to which it belongs. This tag characterizes the lemma at the lexical level in general. It does not describe in any way the particular use of the lemma in a specific context. POS-tags used in our system of analysis are as follows:

ADJ: adjective (ܚܣܪܐ "lacking" Sokoloff, p. 1024)
ADV: adverb (ܬܘܒ "again" Sokoloff, p. 1626)
CARD: cardinal number (ܚܕ "one" Sokoloff, p. 413)
INTJ: interjection (ܘܝ "woe!" Sokoloff, p. 357)
NAME: proper noun
    NAME_ant: anthroponym ܐܕܡ "Adam" (Payne Smith. col. 38)
    NAME_top: toponym ܡܨܪܝܡ "Egypt" (Payne Smith. col. 2196)
NOUN: noun (ܐܚܐ "brother" Sokoloff, p. 25)
ORD: ordinal number (ܩܕܡܝܐ "first" Sokoloff, p. 1319)
PART: particle[7] (ܘ "and" Sokoloff, p. 357)
PRO_dem: demonstrative pronoun (ܗܢܐ "this one" Sokoloff, p. 346)
PRO_ind: indefinite pronoun (ܡܕܡ "something" Sokoloff, p. 715)
PRO_int: interrogative pronoun (ܐܝܟܢܐ "how" Sokoloff, p. 34)
PRO_pers: personal pronoun (ܐܢܐ "I" Sokoloff, p. 58)
V1–V33: verbs[8]
    V1: ܫܒܩ "to confine" (Sokoloff, p. 411)
    V3: 2 ܣܒܪ "to announce" (Sokoloff, p. 965)
    V5 ܐܬܒ "to terrify" (Sokoloff, p. 374)
    …

[4] M. SOKOLOFF, *A Syriac Lexicon: A Translation from the Latin, Correction, Expansion, and Update of C. Brockelmann's Lexicon Syriacum*, Piscataway, 2009 (2d corrected print, 2012).
[5] For the sake of readability, we will only present from this point forward monolexical occurrences. Passive participles in the absolute state in Sokoloff will still receive a lemma in the emphatic state.
[6] R. PAYNE SMITH, *Thesaurus syriacus*, Oxford, 1871-1901 (reprint Hildesheim, 2006).
[7] Particles are tagged as conjunction and preposition in Sokoloff's *Syriac Lexicon*.
[8] Complete table of the verbal POS-tags can be found at the following address: https://www.gregoriproject.com/pdf/POS_SYC.pdf.

4. Roots

At the level of the root, two categories are distinguished.

1. **Base Nouns** (primary nouns)[9] (vocalized, if the vocalization is attested):
    a. Lemmas: ܟܒܝܪ, ܟܒܝܪܘܬܐ and ܟܒܝܪܘܬܢܘܬܐ: root ܟܒܝܪ (Sokoloff, p. 1)
    b. Lemmas: ܟܦܬܐ, 2 ܦܬ̈ܐ: root ܟܦܬܐ (Sokoloff, p. 965)
2. **Consonantal Roots** (not vocalized), linked to a base verb from which nouns and adjectives are derived:
    a. Lemmas: ܟܬܒ̈ܐ, ܟܬܒܢܐ, ܟܬܒܢܝܐܝܬ̈ܐ, ܟܬܒ̈ܐ, ܟܬܒܢܝܐܝܬ̈ܐ: root ܟܬܒ (Sokoloff, p. 374)
    b. Lemmas: ܡܠ̈ܐ, ܡܠܠܐ, ܡܠܝܐܝܬ̈ܐ: root ܡܠܠ (Sokoloff, p. 1352)

Particular cases:

1. Homographic roots are distinguished by a number conforming to the usage of Sokoloff's dictionary:
    a. Lemma 1 ܚܙ̈ܐ: root 1 ܚܙ
    lemma 2 ܚܙ̈ܐ: root 2 ܚܙ
    lemma 3 ܚܙ̈ܐ: root 3 ܚܙ
    lemma 4 ܚܙ̈ܐ: root 4 ܚܙ
    (as Sokoloff, pp. 1133-1134)
    b. Lemmas 1 ܫܐ̈ܪ, 1 ܫܐ̈ܪܝܐ, 1 ܫܦ̈ܐ, 1ܫܐ̈ܪܝܐ and 1 ܫܐ̈ܪܐ: root 1 ܫܐܪ;
    lemmas 2 ܫܐ̈ܪܐ and 2 ܫܐ̈ܪܝܢܐ: root 2 ܫܐܪ
    (as Sokoloff, pp. 1044-1045)
2. If the word is not recorded in reference dictionaries:
    a. If the root (consonantal or base noun) is known, this root is used
        i. Lemma ܟܒܝܫܘܬ̈ܐ "being recompensed": root 3 ܦܣ
        ii. Lemma ܟܚܝܫ̈ܐ "prone to sickness": root ܚܫ
    b. If the root remains unknown, lemma is used as root
        i. Lemma ܚܠܠܘܢܝܐ: root (base noun) ܚܠܠܘܢ
        ii. Lemma ܠܡܝܢܐ: root (base noun) ܠܡܝܢ
        iii. Lemma ܣܠܠܚ: root (base noun) ܣܠܠܚ

5. Origin

Each root receives an indication specifying if the word is borrowed or not from another language than Syriac. In our database, Syriac words are characterized by the sign "0" and non-Syriac word by the sign "-1" (e.g. ܐܝܪ, from Greek ἀήρ). In our concordances in PDF format, non-Syriac roots are therefore automatically characterized by the sign "Ⓛ" (for "loanword"). Proper names are not taken into account and are conventionally characterized by the "0" sign.

[9] In this case, it is the noun or the adjective that is original. The majority of base nouns falls into the following categories: family (father, brother, etc.), body parts (heart, foot, etc.), animals (bull, rabbit, etc.), plants (herbs, garlic, etc.), stones and metals (clay, gold, etc.), temporal concepts (year, eternity, etc.), geographical concepts (earth, rivers, etc.), architectural concepts (dike, roof, etc.), tools (basket, vase, etc.) and foreign words (palace, paradise, etc.). They can themselves be at the origin of verbs, which are qualified as *denominative verbs*, generally used in a derived *form*. Taking into account the base nouns obviates the need to create artificial verbal roots that are supposed to account for them (see B. KINDT, J.-Cl. HAELEWYCK, A.B. SCHMIDT, N. ATAS, *La concordance bilingue grecque-syriaque des* Discours *de Grégoire de Nazianze*, in *BABELAO*, 7 (2018), p. 63-64 ; document freely available here).

Appendix: Keyboard Shortcuts

	Belgian Keyboard	German Keyboard	English Keyboard
Vowels			
a ◌̣ (◌ܰ)	SHIFT + W	SHIFT + Y	SHIFT + Z
ā ◌̈ (◌ܲ)	SHIFT + X	SHIFT + X	SHIFT + X
e ◌̤ (◌ܶ)	SHIFT + C	SHIFT + C	SHIFT + C
ī/ū ◌̇ (◌ܺ/◌ܽ)	SHIFT + V	SHIFT + V	SHIFT + V
o ◌̣ (◌ܴ)	SHIFT + B	SHIFT + B	SHIFT + B
ē ◌̤ (◌ܹ)	SHIFT + N	SHIFT + N	SHIFT + N
Spirantisation			
Rukkākhā ◌̮ (◌ܼ)	SHIFT + ,	SHIFT + M	SHIFT + M
Quššāyā ◌̱ (◌ܻ)	SHIFT + U	SHIFT + U	SHIFT + U
Lengthening Line (*kashīda*)[10]: ـ (ـܿ/ـܼ)	SHIFT + J	SHIFT + J	SHIFT + J
Plural Marker (*syāmē*): ◌̈ (◌ܲ)	SHIFT + I	SHIFT + I	SHIFT + I
Linea occultans (*mbaṭṭlānā*):			
Below ◌̲ **(◌ܵ)**	SHIFT + L	SHIFT + L	SHIFT + L
Above ◌̲ **(◌ܲ)**	SHIFT + O	SHIFT + O	SHIFT + O
Syriac abbreviation mark: ◌̈	² (before the word)	^ (before the word)	` (before the word)

[10] In order to distinguish single letter particles ܕ, ܐ, ܗ, ܠ from numbers, particles are followed by the kashida ـܕ, ـܐ, ـܗ, ـܠ. In accordance with Sokoloff's dictionary (p. 295), we add also the kashida after the possessive particle -ܠܗ.