Chapter 1
Linguistic Introduction: The Orthography, Morphology and Syntax of Semitic Languages

Ray Fabri, Michael Gasser, Nizar Habash, George Kiraz, and Shuly Wintner

1.1 Introduction

We present in this chapter some basic linguistic facts about Semitic languages, covering orthography, morphology, and syntax. We focus on Arabic (both standard and dialectal), Ethiopian languages (specifically, Amharic), Hebrew, Maltese and Syriac. We conclude the chapter with a contrastive analysis of some of these phenomena across the various languages.

The Semitic family of languages [46, 57, 61] is spoken in the Middle East and North Africa, from Iraq and the Arabian Peninsula in the east to Morocco in the west, by over 300 million native speakers.[1] The most widely spoken Semitic languages today are Arabic, Amharic, Tigrinya and Hebrew. The situation of Arabic

[1] Parts of the following discussion are based on Wintner [74].

R. Fabri (✉)
University of Malta, Msida, Malta
e-mail: ray.fabri@um.edu.mt

M. Gasser (✉)
Indiana University, Bloomington, IN, USA
e-mail: gasser@cs.indiana.edu

N. Habash (✉)
Columbia University, New York, NY, USA
e-mail: habash@ccls.columbia.edu

G. Kiraz (✉)
Beth Mardutho: The Syriac Institute, Piscataway, NJ, USA
e-mail: gkiraz@gorgiaspress.com

S. Wintner
University of Haifa, Haifa, Israel
e-mail: shuly@cs.haifa.ac.il

I. Zitouni (ed.), *Natural Language Processing of Semitic Languages*,
Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-45358-8_1, © Springer-Verlag Berlin Heidelberg 2014

(and Syriac) is particularly interesting from a sociolinguistic point of view, as it represents an extreme case of *diglossia*: Modern Standard Arabic (MSA) is used in written texts and formal speech across the Arab world, but is not spoken natively. Rather, colloquial Arabic dialects (Levantine, Yemenite, etc.) are used for everyday conversation, but lack an agreed-upon script [53, p. 267].

The most prominent phenomenon of Semitic languages is the reliance on *root-and-pattern* paradigms for word formation. The standard account of word-formation processes in Semitic languages [59] describes words as combinations of two morphemes: a *root* and a *pattern*.[2] The root consists of consonants only, by default three, called *radicals*. The pattern is a combination of vowels and, possibly, consonants too, with 'slots' into which the root consonants can be inserted. Words are created by *interdigitating* roots into patterns: the consonants of the root fill the slots of the pattern, by default in linear order (see Shimron [67] for a survey).

Other morphological processes in Semitic languages involve affixation and cliticization. The various languages discussed in this chapter exhibit prefixes, suffixes, infixes and circumfixes; some of these affixes are sometimes referred to as (pro- and en-) clitics; we blur the distinction between affixes and clitics [76, 77] in the sequel.

Root-and-pattern morphology, a non-concatenative mechanism unique to Semitic languages, is one of the main challenges for computational processing. Since the vast majority of the morphological processes known in other languages *are* concatenative, existing computational solutions often fail in the face of Semitic interdigitation. The rich morphology of Semitic languages, coupled with deficiencies of the orthography of many of these languages, result in a high level of *ambiguity*, another hurdle that computational processing must overcome.

This chapter aims to describe basic linguistic facets of Semitic languages, addressing issues of orthography, morphology and syntax. We focus on five languages: Amharic (Sect. 1.2), Arabic (both standard and dialectal, Sect. 1.3), Hebrew (Sect. 1.4), Maltese (Sect. 1.5) and Syriac (Sect. 1.6). We conclude with a contrastive analysis of the similarities and differences among those languages (Sect. 1.7), a discussion that provides a backdrop for future chapters, which focus on computational processing addressing these issues. Throughout this chapter, phonetic forms are given between [*square brackets*], phonemic forms are between /*slashes*/ and glosses are listed between "double quotes".

[2] In fact, McCarthy [59] abstracts the pattern further by assuming an additional morpheme, *vocalization* (or *vocalism*), but we do not need this level of abstraction here. Many terms are used to refer to the concept "pattern". In addition to *pattern* and *template*, researchers may encounter *wazn* (from Arabic grammar), *binyan* (from Hebrew grammar), *form* and *measure*. The term *pattern* is used ambiguously to include or exclude vocalisms, i.e., *vocalism-specified pattern* and *vocalism-free pattern* [41].

1.2 Amharic

Amharic is the official language of Ethiopia, the first language of approximately 30 % of the population of the country, 21,631,370 people, according to the 2007 census. As the working language of the federal government and the language of education in many regions of Ethiopia outside the Amhara region, Amharic is also spoken by many Ethiopians as a second language. It is also the native language of perhaps several million Ethiopian immigrants, especially in North America and Israel.

Amharic belongs to the well-established Ethiopian Semitic (or Ethio-Semitic, occasionally also African Semitic) branch of South Semitic. Like the 11[3] other Ethiopian Semitic languages, Amharic exhibits typical Semitic behavior, in particular the pattern of inflectional and derivational morphology, along with some characteristic Ethiopian Semitic features, such as SOV word order, which are generally thought to have resulted from long contact with Cushitic languages.

Among the other Ethiopian Semitic languages, the most important is Tigrinya, with approximately 6.5 million speakers. It is one of the official languages of Eritrea, where it is spoken by more than half of the population, and the official language of Ethiopia's northernmost province of Tigray. Most of what appears in this section applies to Tigrinya as well as Amharic. The most important differences are as follows.

- Like other Semitic languages, but unlike Amharic, Tigrinya has productive broken plurals, as well as external suffixed plurals.
- Alongside the prepositional possessive construction, as in Amharic, Tigrinya has a possessive construction similar to the Arabic Idafa, although neither noun is marked for "state".
- Tigrinya has compound prepositions corresponding to the preposition-postposition compounds found in Amharic.
- The negative circumfix may mark nouns, adjectives, and pronouns, as well as verbs.
- The definite article is an independent word, similar to a demonstrative adjective, rather than a noun suffix.
- Accusative case is marked with a preposition rather than a suffix.
- Yes-no questions are marked with a suffix on the word being questioned.
- There is an unusually complex tense-aspect-mood system, with many nuances achieved using combinations of the three basic aspectual forms (perfect, imperfect, gerund) and various auxiliary verbs.

[3]Because the language boundaries are not well-defined within the Gurage dialects, there is no general agreement on how many Ethiopian Semitic languages there are.

1.2.1 Orthography

Unlike Arabic, Hebrew, and Syriac, Amharic is written using a syllabic writing system, one originally developed for the extinct Ethiopian Semitic language Ge'ez and later extended for Amharic and other Ethiopian Semitic languages [18, 65]. As in other *abugida*[4] systems, each character of the Ge'ez (or Ethiopic) writing system gets its basic shape from the consonant of the syllable, and the vowel is represented through more or less systematic modifications of these basic shapes.[5]

Amharic has 26 consonant phonemes, 27 if we count the /v/ that is used in foreign loan-words, and 7 vowels. For four of the consonants (/'/, /h/, /s/, and /s'/), the writing system makes distinctions that existed in Ge'ez but have been lost in Amharic. There is no single accepted way to transliterate Amharic. We use the following symbols to represent consonants (in the traditional order): *h, l, m, s, r, š, q, b, t, č, n, ň, ', k, w, z, ž, y, d, j, g, t', č', p', s', f, p.* And we use these symbols for the vowels, again in the traditional order: *ə, u, i, a, e, ɨ, o.* Thus the basic character set consists of one character for each combination of 33 consonants and 7 vowels. There are also 37 further characters representing labialized variants of the consonants followed by particular vowels. The complete system has 268 characters. There is also a set of Ge'ez numerals, but nowadays these tend to be replaced by the Hindu-Arabic numerals used in European languages.

Although the Amharic writing system is far less ambiguous than the Arabic, Hebrew, and Syriac alphabets when these are used without vocalization diacritics, the system does embody two sorts of ambiguity. First, as in other abugida systems, one of the characters in each consonant set has a dual function: it represents that consonant followed by a particular vowel, and it represents a bare consonant, that is, the consonant as the coda of a syllable. In the Ge'ez writing system, the vowel for these characters is the high central vowel /ɨ/, traditionally the sixth vowel in the set of characters for a given consonant. This ambiguity presents no serious problems because the vowel /ɨ/ is usually epenthetic so that its presence or absence is normally predictable from the phonetic context.

A more serious sort of ambiguity results from the failure of the writing system to indicate consonant gemination. Gemination is a characteristic feature of Amharic, possible with all but two of the consonants and in all positions except initially. Gemination has both lexical and grammatical functions and is quite common; most words contain at least one geminated consonant, and spoken Amharic lacking gemination sounds quite unnatural [4]. In practice, there are relatively few minimal pairs because of redundancy but the existing minimal pairs include some very common words, for example, *ələ* "he said", *allə* "there is". Syntax must be relied on

[4] In fact the name *abugida*, representing a category of writing system with scores of exemplars, comes from one name of the Ge'ez script as well as the pronunciations of the first four characters of the script in one traditional ordering.

[5] For the actual character set, see http://en.wikipedia.org/wiki/Amharic_language#Writing_system.

to disambiguate these words. The fact that there are few minimal pairs differing only in gemination means that gemination can usually be restored from the orthography, permitting text-to-speech systems to incorporate this crucial feature [4].

1.2.2 Derivational Morphology

Lexicon

As in other Semitic languages, verb roots are the basis of much of the lexicon of Ethiopian Semitic languages.

Alongside verbs proper, there is the category of so-called composite verbs, a defining feature of Ethiopian Semitic languages [3]. These consist of a word in an invariant form, the composite verb "lexeme", followed immediately by an inflected form of one or another of the verbs *alə* "say", *adərrəgə* "do, make", or *assəɲɲə* "cause to feel". A small number of the composite verb lexemes are monomorphemic, underived forms, for example, *quč'č'* "sit", *zɨmm* "be quiet", but most are derived from verb roots (see below), for example, *kɨffɨtt* from the root *k.f.t* "open". In combination with *alə*, the result is an intransitive verb; in combination with *adərrəgə* or *assəɲɲə* a transitive verb.

Nouns, adjectives, and numerals have similar morphosyntactic properties. Nouns are specified for definiteness, but there is no counterpart to the "status" (construct vs. absolute) known in Arabic or Hebrew. Many nouns are derived from verb roots, for example, *nəji* "driver" from *n.d.'* "drive", *t'ɨqs* "quotation" from *t'.q:.s* "quote", *mələt't'əfiya* "glue" from *l.t':.f* "stick (together)". Many others are not, for example, *bet* "house", *sar* "grass", *hod* "stomach".

The great majority of adjectives are derived from verb roots, for example, *kɨfu* "bad, evil" from *k.f.'* "become bad, evil", *ləflafi* "talkative" from *l.f.l.f* "talk too much", *bɨrtu* "strong" from *b.r.t.'* "become strong".

Amharic has a small number of monomorphemic, underived adverbs and conjunctions, for example, *məce* "when", *gɨn* "but". However, most expressions with adverbial or conjunctive function consist of nouns, usually with prepositional prefixes and possibly postpositions as well, or subordinate clauses of one sort or another, for example, *məjəmmərɨya* "first", lit. "beginning (n.)"; *bəgɨmmɨt* "approximately", lit. "by guess"; *sɨləzzih* "therefore", lit. "because of this". Especially common are clauses with a verb in the gerund form, including composite verbs including a gerund form of the verbs *alə* or *adərrəgə*: *qəss bɨlo* "slowly (3 pers.sing.masc.)", *zɨmm bɨyye* "silently, without doing anything (1 pers.sing.)". Note that the gerund, like all Amharic verbs, agrees with its subject, so these adverbials vary with the main clause subject.

Amharic has approximately nine prepositions. The one-syllable prepositions are treated as prefixes; the two-syllable prepositions may be written as prefixes or as separate words. Prepositions occur with and without a much larger set of postpositions. Examples: *betu* "the house", *ɨbetu* "at/in the house", *wədə betu* "to

the house", *ɨbetu wɨst'* "inside the house", *kəbetu wɨst'* "from inside the house", *ɨbetu at'əgəb* "next to the house".

Root and Pattern Processes

Amharic morphology is very complex (for details, see Leslau [56] or Teferra and Hudson [69]). As in other Semitic languages, word formation relies primarily on root-and-pattern morphology. As an example, consider the root *s.b.r*, which is the basis of many words having to do with the notion of breaking. Within the verb system, each root appears in four different tense-aspect-mood (TAM) forms and up to ten different forms representing various derivational categories such as causative and reciprocal. These derivational categories correspond to the *forms* of Arabic and the *binyanim* of Hebrew. As in the other languages, each has a rough meaning that may or may not hold for particular roots. For example, the "passive-reflexive" pattern denotes the passive for a root like *s.b.r* "break" but the transitive active for a root like *r.k.b* "receive".

Each combination of TAM and derivational category defines a particular pattern template that combines with roots to form a verb stem. For example, the pattern *C1:əC2:əC3* represents the imperfect TAM and passive-reflexive derivational category (the C's indicate the consonant slots, and ":" indicates gemination). In combination with the root *s.b.r*, this pattern yields the stem *ssəbbər* "is broken".

Nouns and adjectives are formed from a verb root through the application of other patterns. For example, the pattern *aC1:əC2aC2əC3* forms a manner noun, *assəbabər* "way of breaking", and the pattern *məC1C2əC3* forms the infinitive, *məsbər* "to break".

Most of the members of the composite verb lexeme class [3], discussed above, are also derived from verb roots through root-pattern morphology. For simple three-consonant roots, the main patterns are *C1ɨC2:ɨC3:*, *C1əC2əC3:*, and *C1ɨC2ɨC3C2ɨC3:*. Thus from the root *s.b.r*, we have *sɨbbɨrr*, *səbərr*, and *sɨbɨrbɨrr*. The meanings of these are described as "intensive", "attenuated", and "distributive" respectively [3].

As in other Semitic languages, further complexity results from the possibility of four- and five-consonant roots and from roots which may behave as though they consist of one or two consonants because of the special treatment of certain consonants and because of historical changes. For example, the root meaning "kiss" was originally *s.ʕ.m*, but the pharyngeal consonants have been lost in the modern language, resulting in a defective class with only two explicit consonants in each root. For this class, the imperfect passive-reflexive pattern is *C1:aC2*, for example, *ssam* "is/are kissed".

Within the three-consonant roots, Ethiopian Semitic languages have three lexical subclasses, traditionally referred to as the A, B, and C classes, which correspond to different derivational patterns in other Semitic languages. The B class is characterized by the gemination of the second root consonant in nearly all patterns, the C class by the presence of the vowel *a* between the first and second consonant in

all patterns. Membership of a root in one or another of these classes is lexical. For example, the A root *t'.b.q* means "be tight", while the B root *t'.b:.q* means "wait". In all, there are roughly 25 classes of roots, including subclasses of the A, B, and C classes and multiple classes for four- and five-consonant roots, each with its own set of patterns.

Other Derivational Processes

Other derivational processes are based on suffixation. For example, many adjectives are formed from nouns through the addition of the suffixes *-mma* and *-awi*: *wɨha* "water" ⇒ *wɨhamma* "watery", *abɨyot* "revolution" ⇒ *abɨyotawi* "revolutionary".

1.2.3 Inflectional Morphology

Inflectional morphology consists mostly of suffixes, but sometimes of prefixes or circumfixes, and sometimes of pattern changes.

Verbs

As noted in Sect. 1.2.2, verbs inflect for TAM through the various patterns that also convey derivational categories such as passive and reciprocal. The four TAM categories are traditionally referred to as perfect(ive), imperfect(ive), jussive-imperative, and gerund(ive). The perfect and imperfect forms, alone or in combination with various auxiliary verbs, function as indicative main clause verbs and, in combination with various prefixes, as subordinate clause verbs as well. The jussive corresponds to one sort of subjunctive in other languages: *yɨmt'u* "let them come". The gerund, also called converb or continuative, is a subordinate form used in combination with a main clause verb (see Sect. 1.2.4).

Verbs in all four TAM categories also inflect for subject person, number, and (in second and third person singular) gender.[6] Each of the TAM categories has its own set of suffixes and/or prefixes for person–number–gender. The affixes for imperfect and jussive-imperative, a combination of prefixes and suffixes, are almost identical. For perfect and gerund, the inflections are all suffixes. As an example, for the simple derivational category, the third person plural forms for the root *s.b.r* are as follows (the subject agreement affixes are in bold): perfect *səbbər**u***, imperfect ***yɨ**səbr**u***, jussive ***yɨ**sbər**u***, gerund *səbr**ew***.

[6]In some other Ethiopian Semitic languages, such as Tigrinya, a distinction is made between masculine and feminine in the second and third person plural as well.

Verbs in all four TAM categories can also take object suffixes. There is a set of direct object suffixes, and there are two sets of indirect object suffixes, distinguished by the two prepositional suffixes *-bb-* and *-ll-* that they begin with. Examples: *səbbərə**w*** "he broke it", *səbbərə**llat*** "he broke (sth.) for her". Unlike in some other Semitic languages, only one object suffix is possible on a given verb. For more on the function of these suffixes, see Sect. 1.2.4 below.

Verbs in the imperfect, perfect, and jussive-imperative (but not gerund) may also take the negative circumfix (main clause indicative verbs) or prefix (subordinate or jussive-imperative verbs): ***al****səbbərə**m*** "he didn't break", ***att****isbər* "don't break (sing.masc.)!" Imperfect or perfect verbs may also take the relative prefix *yə(mm**i**)*, used for verbs in relative clauses and some other subordinate clause types (Sect. 1.2.4): ***yə****səbbərəw* "(he) who broke it", ***yəmmi****tsəbriw* "which you (sing.fem.) break". Finally, a small set of prepositional and conjunctive prefixes and conjunctive/adverbial suffixes is possible on imperfect and perfect verbs: ***ind****atsebrut* "**so that** you (pl.) don't break it" *bisəbrew**m*** "**even** if he breaks it".

Counting the verb stem as two morphemes (root and pattern), an Amharic verb may consist of as many as ten morphemes: *ləmmayanəbbull**i**n**i**m* "also for they who do not read to us" *lə-mm-a-y-{n.b.b+aC1əC2C3}-u-ll-n-m*.

Nominals

Because of their similar morphosyntax, nouns, pronouns, adjectives, and numerals are grouped together in this section.

Nouns inflect for number, case, and definiteness. There are two number categories, singular and plural. Plurals are usually formed by the suffix *-očč*; broken plurals are rare and in all cases seem to be loanwords from Ge'ez. The plural suffix is not obligatory when plurality is clear from the context: *sost astəmari*/*sost astəmariwočč* "three teachers". When it is used, the plural suffix indicates specificity as well as plurality [52].

The definite article takes the form of a suffix; it has the masculine form *-u* and the feminine form *-wa* or *-itu*. The definite suffix follows the plural suffix if there is one. Amharic has no indefinite article.

Amharic subjects are unmarked. Definite, and some indefinite, direct objects take the accusative suffix -n. Oblique cases are signaled by prefixed prepositions, sometimes in combination with postpositions. Example: *məskot* "window", *d**i**ngay* "stone", *yohann**i**s məskotu**n** b**ə**d**i**ngay səbbərəw* "Yohannis broke the window with a stone". Amharic nouns have no genitive form; instead genitive and possession are indicated by the prepositional prefix *yə*: ***yə****betu wələl* "the floor of the house".

Noun gender can best be described as flexible [56]. While there are strong tendencies, for example, vehicles, countries, and female animals are normally feminine, it is possible to override these tendencies for pragmatic effect. Even male humans may be referred to with feminine pronouns, feminine demonstratives, and feminine verb agreement affixes, when the speaker wishes to convey affection toward the referent or attribute skillfulness or smallness to the referent.

Nouns may also take one or another of the same set of conjunctive/adverbial suffixes as verbs: *betu**m*** "**also** the house".

Demonstrative pronouns and adjectives have singular masculine, singular feminine, and plural forms: *ya bet* "that house", *yaččі kətəma* "that city". Other adjectives do not inflect for gender, but they may have plural forms, especially in headless noun phrases. Adjective plurals, unlike nouns, are often indicated by reduplication, sometimes in combination with the plural suffix *-očč*: *tɨlliq* "big", *tɨlilliq betočč* "big houses". It is also possible to use plural adjectives with the singular form of nouns: *addis* "new", *nəgər* "thing", *adaddis nəgər* "new things".

Adjectives and numerals may also take the accusative suffix, the definite suffix, and prepositional prefixes: ***kətɨlliqu*** "from the big (one)".

As in other Semitic languages, possessive adjectives take the form of suffixes on the modified noun. These follow the plural suffix and precede the accusative suffix. Examples: *bet**e*** "my house", *betočč**ačč**ɨn*** "our houses", *betočč**ačč**əw*** "their houses".

1.2.4 Basic Syntactic Structure

Noun Phrases

An Amharic noun phrase has explicit number, case, and definiteness. The accusative suffix appears obligatorily on definite direct objects and optionally on indefinite direct objects. An unusual feature of the language is the placement of the morphemes marking case (either the accusative suffix or one or another prepositional prefix) and definiteness [37,55]. These are affixed to the noun itself only when it has no modifiers. If the noun has adjective or relative clause modifiers, the morphemes are normally affixed to the first of these. Examples: *bet**un*** "the house (acc.)", *tɨlliq**un** bet* "the big house (acc.)", *yəgəzza**wn** tɨlliq bet* "the big house (acc.) that he bought",

Headless noun phrases are common. These consist of one or more relative clauses and adjectives. Examples: *tɨlliq**un*** "the big one (acc.)", *yəgəzza**wn*** "the one (acc.) that he bought",

Clauses

Unlike in other Semitic languages, all Amharic clauses are headed by verbs. The copula, *nəw*, is a defective verb with only main clause present forms. Its past is filled by the defective verb *nəbbər*, which also serves as the past of the defective verb of existence *allə*. In other cases, the copula is replaced by the perfect, imperfect, jussive-imperative, or gerund of either the verb *norə* "live" or the verb *honə* "become".

The basic word order of all Ethiopian Semitic languages is subject-object-verb (SOV), a feature that probably results from contact with Cushitic languages. As is common in SOV languages, the order of subject, object, and oblique arguments of the verb is somewhat flexible. In particular, for pragmatic reasons the subject can follow another argument: *yohannis məskotun səbbərəw*, *məskotun yohannis səbbərəw*, "Yohannis broke the window".

As in other Semitic languages, verbs agree with their subjects in person, number, and (in second and third person singular) gender.[7] Verbs also agree with definite direct or indirect objects, but not both. As discussed in Sect. 1.2.3, there are three sets of object agreement suffixes. This three-way split within the verbal morphology system maps in complex ways onto explicit syntactic arguments of the verb [45]. That is, direct object verb suffixes correspond usually but not always to accusative syntactic arguments, and indirect object verb suffixes correspond usually but not always to explicit prepositional phrases (noun phrases with prepositional prefixes). The *-ll-* indirect object suffix agrees with dative and benefactive arguments, while the *-bb-* suffix agrees with locative, instrumental, and adversative arguments. Examples: ***lə**hirut sət'ť**at*** "he gave (it) to Hirut" (direct object suffix (3 pers.sing.fem.) agrees with explicit syntactic dative), ***lə**hirut sərra**llat*** "he made (it) for Hirut" (indirect object suffix agrees with explicit syntactic benefactive).

Impersonal verbs [56] complicate the picture further. The morphological subject of these verbs is always third person singular masculine, and they take an obligatory direct object agreement suffix that refers to the experiencer: *hirut rab**at*** "Hirut is hungry". The verb of possession behaves similarly; it makes use of the verb of existence *allə* with direct object suffixes referring to the possessor. The morphological subject of the verb is the possessed object or objects. *hirut sost wəndɨmmočč alluw**at*** "Hirut has a brother" (subject agreement 3 pers.plur., object agreement 3 pers.sing.fem.)

As in other Semitic languages, pronoun subjects and pronoun objects are omitted unless they are emphasized. This fact, in combination with the elaborate derivational and inflectional verb morphology, means that sentences consisting of a verb alone or a main verb and an auxiliary verb are not uncommon: *alt'əyyəqnatɨm* "we didn't visit her", *laflallačɨhu* "shall I boil (sth.) for you (pl.)?", *awwaddədun* "they made us like each other".

Main clause verbs are either in the perfect or a compound imperfect formed from the simple imperfect plus conjugated suffix forms of the defective verb of existence *allə*. Subordinate clause verbs are in the perfect, simple imperfect, or gerund. *tɨfəlliɨgiyalləš* "you (fem.sing.) want", *bɨttɨfəlliɨgi* "if you (fem.sing.) want".

Cleft constructions are very common in Amharic [51]. Indeed for questions, cleft constructions are probably more common that non-cleft constructions. In a cleft sentence, the focused argument is placed first, followed by the conjugated copula,

[7] Some other Ethiopian Semitic languages distinguish masculine and feminine in the second and third person plural as well.

followed by other arguments of the original verb, followed by the verb in relativized form: *mɨndɨn nəw yəsəbbərəw* "what did he break?" lit. "what is it that he broke it".

Relative clauses make use of the relative imperfect or perfect form of the verb: ***yə***səbbərəw "that he broke", ***yəmm***isəbrəw "that he breaks".

Adverbial clauses are usually indicated through the use of prefix conjunctions on the relative form of the verb (in which case the initial *yə* is dropped) or the bare imperfect: ***sɨlə***mmisəbrəw "because he breaks it", ***bɨ***səbrəw "if he breaks it".

Nominalizations in Amharic take two forms. Either the verb of the nominalized clause appears in the (conjugated) relative form with the prefix conjunction ɨndə "that", or the verb appears in the (unconjugated) infinitive form, with the verb's original subject in the form of a possessive suffix. Examples: *mɨn* ***ɨndə***mmigəza *alawqɨm* "I don't know what he will buy"; *zɨgɨjju* ***məhonačč əwn*** *gəlləs'u* "they explained that they were ready", lit. "they explained **their being (acc.)** ready".

As is common in SOV languages, Amharic permits the chaining of a number of clauses together in a single sentence without explicit conjunctions indicating the relationship between the clauses. The usual interpretation is sequentiality. All verbs but the final one appear in the gerund form. The final verb may be perfect, compound imperfect, jussive, or imperative. All of the gerund forms agree with the subject of the final verb. Example: *ɨbet təməlliso rat bəlto təɲɲa* "He returned home, ate dinner, and went to bed" lit. "Returning (3 pers.sing.masc.) home, eating (3 pers.sing.masc.) dinner, he went to bed".

1.3 Arabic

The term "Arabic language" is often used to refer collectively to Modern Standard Arabic (MSA) and its dialects. MSA is the official language of the Arab World, a region of 22 countries. Arabic is also an official language of Chad, Eritrea and Israel. With a collective population of almost 300 million people, Arabic is the most commonly spoken Semitic language.

While MSA is the primary written language of the media and education, the dialects, which are historically related to MSA, are the truly native informal spoken media of communication for daily life. The dialects, which vary geographically and socially, are not taught in schools or standardized. They are the primary linguistic form used to express modern Arab culture, songs, movies and TV shows. Although the language-dialect situation of Arabic seems similar to many across the world, two aspects set it apart: (a) the high degree of difference between standard Arabic and its dialects, often compared to Latin and the Romance languages [39]; and (b) the fact that standard Arabic is not any Arab's native language [41]. The two varieties of Arabic coexist along the lines of formality in what is termed a state of *diglossia* [35]: formal written (MSA) versus informal spoken (dialect); there is a large mixing area in between [5, 6]. For more information on MSA, see Holes [47].

1.3.1 Orthography

Arabic Script

Arabic is written using the Arabic script, a right-to-left alphabet also used to write many languages around the world which are not related to Arabic such as Persian, Kurdish, Urdu and Pashto. Arabic dialects are by default written in Arabic script although there are no standard dialectal spelling systems [43].

To facilitate reading of Arabic examples, we employ a romanized "transliteration" scheme. Unlike the term "transcription," which denotes an orthography that characterizes the phonology or morpho-phonology of a language, the term "transliteration" denotes an orthography that uses carefully substituted orthographical symbols in a one-to-one, fully reversible mapping with that language's customary orthography (in our case, Arabic script). This specific definition of transliteration is sometimes called a "strict transliteration" or "orthographical transliteration" [7]. While the International Phonetic Alphabet (IPA) or some variant of it may be the best choice for phonological transcription, it is not sufficient for strict orthographic transliteration purposes. The Buckwalter Arabic transliteration scheme [14] is perhaps one of the most commonly used strict transliteration schemes for Arabic. The scheme we use is the Habash-Soudi-Buckwalter (HSB) scheme [42], which extends Buckwalter's scheme to increase its readability while maintaining the one-to-one correspondence with Arabic orthography as represented in standard encodings of Arabic, such as Unicode. The HSB scheme consists of the following symbols: (in alphabetical order) ا *A*, ب b, ت *t*, ث *θ*, ج *j*, ح *H*, خ *x*, د *d*, ذ *ð*, ر *r*, ز *z*, س *s*, ش *š*, ص *S*, ض *D*, ط *T*, ظ *Ď*, ع *ʕ*, غ *γ*, ف *f*, ق *q*, ك *k*, ل *l*, م *m*, ن *n*, ه *h*, و *w*, ي *y*; (the Hamza and Hamzated forms) ' ء, Â أ, Ă إ, Ā آ, ŵ ؤ, ŷ ئ ; (the Ta-Marbuta) *ħ* ة[8]; (the Alif-Maqsura) ý ى; and (the diacritics) a ـَ, u ـُ, i ـِ, ~ ـّ, ã ـً, ũ ـٌ, ĩ ـٍ.

Most of the Arabic script letters are written connected to neighboring letters with few exceptional non-connective or right-connective letters. When connected, an Arabic letter may change its shape, sometime radically: e.g., the letter ه *h*: ه ههه (initial-medial-final and standalone, respectively). Furthermore, some combinations of Arabic letters join to form special obligatory ligatures: e.g., the letters ا+ل *l+A* are written as لا not لا. Computer encodings abstract away from the various letter shapes and ligatures to simply encode the letter itself. The Arabic letter consists of a basic letter form and obligatory dots that together define the letter. These dots should not be confused with the optional diacritics that are used for specifying short vowels, consonantal gemination and *nunation* (a nominal word-final morpheme specifying indefiniteness and written only as a diacritic).

[8]In IPA and Maltese, the symbol *ħ* is used for *voiceless pharyngeal fricative* or (ح *H*). In HSB, it reflects the two letters that compose the Ta-Marbuta symbol (ة *ħ*): ه *h* and ت *t*.

Arabic Spelling

Arabic spelling rules utilize its script as an impure *abjad*: although diacritics are omitted most of the time, long vowels are always written in Arabic using a combination of an omittable diacritic short vowel and non-omittable compatible consonant, e.g., [u:] is written as *uw*. Additionally, Arabic uses two morphophonemic letters: ة *ħ* is the Ta-Marbuta, feminine ending morpheme, and ى is the Alif-Maqsura, derivational marker of the word's root radical ي y realizing as the vowel [a]. Some letters in Arabic are often spelled inconsistently which leads to an increase in both sparsity (multiple forms of the same word) and ambiguity (same form corresponding to multiple words), e.g., variants of Hamzated Alif, أ Â or إ Ă, are often written without their Hamza (ء '): ا *A*; and the Alif-Maqsura (or dotless Ya) ى ý and the regular dotted Ya ي *y* are often used interchangeably in word final position [22].

1.3.2 Morphology

In discussing Arabic morphology, it is useful to distinguish between operations related to the type or form of the morphological process as opposed to its function [40, 68].

In terms of morphological form, Arabic, in a typical Semitic manner, uses both *templatic* and *concatenative* morphemes. Concatenative morphemes form words through sequential concatenation, whereas templatic morphemes are interdigitated. Functionally, like most other languages of the world, we can distinguish between derivational morphology and inflectional morphology. In Arabic morphological form and function are independent although most templatic processes are derivational and most concatenative processes are inflectional. There are some important exceptions which we allude to below.

Templatic Morphology

There are two types of templatic morphemes: roots and templates.[9] The *root* morpheme is a sequence of mostly three consonantal radicals which together signify some abstract meaning. For example, the words كتب *katab* "to write", كاتب *kAtib* "writer/author/scribe", مكتوب *maktuwb* "written/letter", مكتب *maktab* "office" and مكتبة *maktabaħ* "library" all share the root morpheme *k.t.b* "writing-related". Root semantics is often, however, idiosyncratic. For example, the semantically

[9] Templates are sometimes further split into patterns and vocalisms [59].

divergent words لحم *laHm* "meat" and لحم *laHam* "to solder" also have the same root *l.H.m* [41]. The template is an abstract pattern in which roots are inserted. The template like the root provides some meaning component. For example, the template *maC1C2aC3*, often indicating *location*, combines with the roots ك.ت.ب *k.t.b* "writing-related" and ل.م.ع *ς.m.l* "work-related" to create the nouns مكتب *maktab* "office" and معمل *maς mal* "laboratory/factory".

Concatenative Morphology

The templatic morphemes form the stem onto which concatenative affixes (prefixes, suffixes and circumfixes) and clitics (proclitics and enclitics) attach. Clitics are morphemes that have the syntactic characteristics of words, but show evidence of being phonologically bound to another word [58], e.g., the proclitic conjunction و+ *w+* "and" or the object pronoun enclitic هم+ *+hm* "them". In this respect, a clitic is distinctly different from an affix, which is phonologically and syntactically part of the word and often represents inflectional features, e.g., the suffixes ة+ *+ħ* and ات+ *+At* represent the feminine singular and feminine plural inflections, respectively. A word may include multiple affixes and clitics, e.g., the word وسيكتبونها *wasayaktubuwnahA* has two proclitics, one circumfix and one enclitic: *wa+sa+y+aktub+uwna+hA* (and + will + 3person + write + masculine – plural + it) "and they will write it".

The combination of both templatic and concatenative morphemes may involve a number of phonological, morphological and orthographic adjustments that modify the form of the created word beyond simple interleaving and concatenation. For example, the feminine singular morpheme, ة+ *+ħ*, is turned into ت+ *+t* when followed by a possessive clitic: أميرة+هم *Âamiyraħu+hum* "princess+their" is realized as أميرتهم *Âamiyratuhum* "their princess".

Derivational Morphology

Derivational morphology creates new words from other words typically through a template switch and often resulting in a change in part-of-speech (POS). The root remains constant in the process. For example, the Arabic كاتب *kAtib (k.t.b+C1AC2iC3)* "writer" can be seen as derived from the verb كتب *katab (k.t.b+C1aC2aC3)* "to write" in the same way the English *writer* can be seen as a derivation from *write*. Although compositional aspects of derivations do exist, the derived meaning is often idiosyncratic. For example, the *masculine* noun مكتب *maktab* "office/bureau/agency" and the *feminine* noun مكتبة *maktabaħ* "library/bookstore" are derived from the root ك.ت.ب *k.t.b* "writing-related" with the location template *maC1C2aC3* [41].

Inflectional Morphology

In inflectional morphology, the core meaning and POS of the word remain intact and the extensions are always predictable and limited to a set of possible features. Inflectional features are all obligatory and must have a specific (non-nil) value for every word. In Arabic, there are eight inflectional features. *Aspect* (perfective, imperfective, imperative), *mood* (indicative, subjunctive, jussive), *person* (1st, 2nd, 3rd) and *voice* (active, passive) only apply to verbs, while *case* (nominative, accusative, genitive) and *state* (definite, indefinite, construct) only apply to nouns/adjectives. *Gender* (feminine, masculine) and *number* (singular, dual, plural) apply to both verbs and nouns/adjectives.

Clitics are similar to inflectional features in that they do not change the core meaning of the word; however, they are all optional. Arabic clitics include conjunctions, verb particles, nominal prepositions, the definite article and pronominal enclitics that can serve as possessives of nouns and objects of verbs and prepositions.

Form-Function Independence

Arabic morphological form and function are independent. Both templatic and concatenative morphemes can function derivationally or inflectionally, with the exception of the roots, which are always derivational. The majority of derivational morphology is templatic and the majority of inflectional morphology is concatenative. The most important exception is the templatic plural, often called "broken plural", which is formed through change of templates as opposed to the sound plurals formed through affixation. For example, compare the following two plurals of the noun كاتب *kAtib* "writer": كتاب *kut∼Ab* (broken) "writers" and كاتبات *kAtib+At* (sound) "writers [fem]". More than half of all plurals in Arabic are broken [2]. Another example of form-function independence is the derivational suffix *Ya of Nisba* ي+ *+iy∼*, which maps nouns to adjectives related to them, e.g., كتبي *kutub+iy∼* "book-related" is derived from the noun كتب *kutub* "books", a broken plural itself of the noun كتاب *kitAb* "book". For other aspects of form-function independence in Arabic, see Habash [41].

Dialectal Arabic Morphology

Arabic dialect morphology is simpler in some respects and more complex in others compared to MSA. Dialects overall lost the case and mood inflections and merged feminine and masculine plurals and duals in verbs among other changes. However, Arabic dialects introduce additional clitics, some with new functionality. For example, the non-MSA circum-clitic ش+ +ما *mA* + + *š* which is used to mark

negation appears in several dialects. Another example of change is the MSA future proclitic +س *sa*+, which is replaced by +ح *Ha*+ in Levantine and +غ *γa*+ in Moroccan. For more information on Arabic dialects, see Cowell [17], Erwin [23], Abdel-Massih et al. [1], and Harrell [44].

Morphological Ambiguity

Arabic's optional diacritics, inconsistent spelling of some letters, together with the language's complex morphology lead to a high degree of ambiguity: the Buckwalter Arabic Morphological Analyzer (BAMA), for instance, produces an average of 12 analyses per word [15] corresponding to almost 7 diacritizations and almost 3 lemmas.

1.3.3 Basic Syntactic Structure

Morphology and Syntax

In morphologically rich languages, such as Arabic, many syntactic phenomena are expressed not only in terms of word order but also morphology. For example, Arabic subjects of verbs have a nominative case and adjectival modifiers of nouns generally agree with the noun they modify in case, gender, number and definiteness.

Sentence Structure

Arabic has two types of sentences: verbal sentences (VS) and nominal (sometimes called copular, or equational) sentences (NS). The prototypical VS form is Verb-Subject-Object(s). Arabic is a pro-drop language: in the case of a pronominal subject, the verb expresses the person, gender and number of the subject. However, with non-pronominal subjects, the verb agrees in person (3rd) and gender only, while number defaults to singular. This is referred to as partial agreement.

The prototypical NS has the form of Subject-Predicate or Topic-Complement. The subject is typically a *definite* nominative noun, proper noun or pronoun and the predicate is an *indefinite* nominative noun, proper noun or adjective that agrees with the subject in number and gender. The predicate can be a prepositional phrase (PP), in which case no agreement is shown.

A complex sentence structure is formed from an NS with a VS predicate producing a Subject-Verb-Object order. Here, the subject and verb agree in full (gender, number and person).

Nominal Phrase Structure

Arabic adjectives follow the nouns they modify in a nominal phrase (NP). Adjectives and nouns agree in gender, number, definiteness and case, with the exception of irrational (non-human) plural nouns whose adjectives are feminine singular.

Arabic also has a possessive/genitive construction, called *Idafa*, which relates two nouns: the first is the possessed and the second is the possessor. The first noun is in the construct state, and the second takes a genitive case and is typically definite. An Idafa chain can be formed by adding an additional element to the beginning of the NP. For example, مفاتيح سيارة الرجل *mfAtyH syArħ Alrjl* (keys-car-the+man) translates as "The man's car keys". Adjectives modifying the head of an Idafa construction agree with it in case; but they agree with its dependent in definiteness.

Relative Clauses

Relative clauses modify the noun they follow. If the modified noun is definite, the relative clause is introduced by a relative pronoun which agrees with the noun it modifies in gender and number (irrationality gets exceptional agreement). Relative clauses of indefinite nouns are not introduced with a relative pronoun.

Arabic Dialect Syntax

Arabic dialects are not very different syntactically from MSA. For example, both SVO and VSO orders exist in the dialects although the SVO order is more dominant. Some of MSA's complex syntactic phenomena such as irrational plural agreement are maintained in the dialects, while case agreement is gone since dialects do not inflect for case. For more information on dialectal Arabic syntax, see Brustad [13].

1.4 Hebrew

Hebrew is one of the two official languages of the State of Israel,[10] spoken natively by half of the population and fluently by virtually all the (over seven million) residents of the country. Hebrew exhibits clear Semitic behavior[11]; in particular, its lexicon, word formation and inflectional morphology are predominantly Semitic.[12]

[10] The other is Arabic.

[11] In spite of some recent claims to the contrary [48, 75].

[12] Parts of the discussion in this section are based on Itai and Wintner [49].

Hebrew is unique among the Semitic languages (indeed, among the world's languages) in that it has been 'dormant' for several centuries, used primarily for religious and academic purposes, and little, if at all, as a daily spoken language. Modern Hebrew was 'revived' in the end of the nineteenth century; while it is undoubtedly based on the infrastructure of older layers of Hebrew, it was heavily influenced by Yiddish, Slavic languages, and other languages spoken by Jews during nearly two millennia. In this chapter, the term 'Hebrew' refers to Modern Hebrew only.

1.4.1 Orthography

Hebrew is written in the Hebrew alphabet, a 22-letter *abjad* [18, 65]. To facilitate readability we use a straightforward transliteration of Hebrew in this chapter, where the characters אבגדהוזחטיכלמנסעפצקרשת (in Hebrew alphabetic order) are transliterated thus: *abgdhwzxTiklmnsypcqršt*. There are two main standards for the Hebrew script: one in which vocalization diacritics, known as *niqqud* "dots", decorate the words, and one in which the dots are missing, and other characters represent some, but not all of the vowels.[13] Most of the texts in Hebrew are of the latter kind. While the Academy for the Hebrew Language publishes guidelines for transcribing undotted texts [36], they are only partially observed. Thus, the same word can be written in more than one way, sometimes even within the same document. For example, *chriim* "noon" can be spelled *chrim*.

The script dictates that many particles, including four of the most frequent prepositions, the definite article *h* "the", the coordinating conjunction *w* "and" and some subordinating conjunctions, all attach to the words that immediately follow them. When the definite article *h* is prefixed by one of the prepositions *b* "in", *k* "as", or *l* "to", it is assimilated with the preposition and the resulting form becomes ambiguous as to whether or not it is definite. For example, *bth* can be read either as *b+th* "in tea" or as *b+h+th* "in the tea". Consequently, the form *šbth* can be read as an inflected stem (the verb "capture", third person singular feminine past), as *š+bth* "that+field", *š+b+th* "that+in+tea", *š+b+h+th* "that in the tea", *šbt+h* "her sitting" or even as *š+bt+h* "that her daughter". See Table 1.1.

These features of the writing system imply that Hebrew texts tend to be highly ambiguous. First, the first and last few characters of each token may be either part of the stem or bound morphemes (prefixes or suffixes). Second, the lack of explicitly marked vowels yields many homographs. See a detailed discussion in Sect. 1.4.4.

[13] The undotted script is sometimes referred to as *ktiv male* "full script", whereas the dotted script, but with the diacritics removed, is called *ktiv xaser* "lacking script". These terms are misleading, as any representation that does not depict the diacritics would lack many of the vowels.

Table 1.1 The various morphological analyses of the form *šbth*

Lemma	POS	Number	Gender	Person	Tense	State	Def	Prefix	Suffix	Gloss
šbt	Noun	Sing	Fem	n/a	n/a	abs	No		h	"her Saturday"
bt	Noun	Sing	Fem	n/a	n/a	abs	No	*š*	h	"that her daughter"
bth	Noun	Sing	Fem	n/a	n/a	abs	No	*š*		"that a field"
th	Noun	Sing	Masc	n/a	n/a	abs	Yes	*š+b+h*		"that in the tea"
th	Noun	Sing	Masc	n/a	n/a	abs	No	*š+b*		"that in tea"
th	Noun	Sing	Masc	n/a	n/a	cons	No	*š+b*		"that in tea-of"
šbh	Verb	Sing	Fem	3	Past	n/a	n/a			"(she) captured"
šbt	Verb	Sing	Fem	3	Past	n/a	n/a			"(she) went on strike"

1.4.2 Derivational Morphology

Root and Pattern Processes

Hebrew morphology is rich and complex. The major word formation machinery is root-and-pattern. As an example of root-and-pattern morphology, consider the root *k.t.b*, which denotes a notion of writing. Hebrew has seven verbal patterns, which contribute to the meaning of the stem in productive but not fully predictable ways. Thus, construed in the *pa'al* pattern *CaCaC* (where the 'C's indicate the slots), the root *k.t.b* yields *ktb [katav]* "write"; whereas in the *hif'il* pattern *hiCCiC*, which is typically a causative, it yields *hktib [hiktiv]* "dictate". Similarly, the (nominal) pattern *haCCaCa* usually denotes nominalization; hence *hktbh [haktava]* "dictation". The pattern *maCCeCa* often denotes instruments; construed in this pattern, the root *k.t.b* yields *mktbh [maktheva]* "writing desk".

After the root combines with the pattern, some morpho-phonological alternations take place, which may be non-trivial: for example, the *hitCaCCut* pattern triggers assimilation when the first consonant of the root is *t* or *d*: thus, *d.r.š+hitCaCCut* yields *[hidaršut]*. The same pattern triggers metathesis when the first radical is *s* or *š*: *s.d.r+hitCaCCut* yields *[histadrut]* rather than the expected **[hitsadrut]*. Semi-vowels such as *w* or *y* in the root are frequently combined with the vowels of the pattern, so that *q.w.m+haCCaCa* yields *[haqama]*, etc. Frequently, root consonants such as *w* or *y* are altogether missing from the resulting form.

Other Derivational Processes

While root-and-pattern is the main word formation process in Hebrew, other processes are also operational [63]. These include several regular pattern-relating processes, such as nominalization: each verbal pattern in Hebrew is related to a pattern whose meaning is typically the nominalization of the meaning of the corresponding verb. For example, *hitCaCeC* is related to *hitCaCaCut*, so that from *htlhb [hitlahev]* "be enthusiastic about" one obtains *htlhbwt [hitlahavut]* "enthusiasm". Similarly, *CiCeC* is related to *CiCuC*, so that from *šipr [šiper]* "improve" one obtains *šipwr [šipur]* "improvement".

Other processes are more concatenative in nature, based primarily on suffixation. Thus, the suffix *wt [ut]* is frequently used to convert adjectives to nouns, as in *bria [bari] healthy* ⇒ *briawt [bri'ut]* "health". The suffix *wn [on]* can be used to construct the diminutive, as in *db [dov]* "bear" ⇒ *dwbwn [dubon]* "teddy bear". Interestingly, Hebrew also has a non-concatenative diminutive operator that is based on reduplication: *klb* "dog" ⇒ *klblb* "doggie"; *xtwl* "cat" ⇒ *xtltwl* "kittie"; *irwq* "green" ⇒ *irqrq* "light green"; etc.

1.4.3 Inflectional Morphology

Inflectional morphology is highly productive and consists mostly of suffixes, but sometimes of prefixes or circumfixes, and sometimes of pattern changes.

Verbs

Verbs inflect for number, gender and person (first, second and third) and also for a combination of tense and aspect/mood, referred to simply as 'tense' below. Some of these variations are obtained by simple concatenation rules, whereas others are better explained in terms of stem changes. The citation form of Hebrew verbs is the third person, masculine, singular past form. Consider *sgr [sagar]* "close", for example. Its past tense forms include *sgrti [sagarti]*, *sgrt [sagarta]*, *sgrnw [sagarnu]*, etc., all of which can be obtained from the citation form by concatenation. Other past tense forms, such as *sgrh [sagra]* or *sgrw [sagru]*, can still be explained in terms of concatenation and simple morpheme-boundary alternations (in this case, reduction of the last vowel of the stem). However, in the present some inflected forms require a change in the stem, as in *swgr [soger]*, *swgrim [sogrim]*, etc. The same applies to the future, some of whose forms are *tsgwr [tisgor]*, *isgrw [yisgeru]*, etc. Consequently, a good way of accounting for entire verb paradigms is by specifying, for each verb, not only the citation form but also secondary stems for some of the inflections.

Fortunately, this is not difficult because the secondary stems are determined by the root and the pattern of the verb. As noted above, verbs can belong to one of seven patterns, and each pattern is inflected in exactly the same way. Root consonants may trigger some variations, but these, too, are systematic and regular.

Verb patterns ('*binyanim*') are associated with (vague) meanings. Traditionally, the *nif'al* pattern is considered to be the passive of the *pa'al* pattern, whereas *pi'el* denotes reinforcement or intensification, and *hif'il* denotes causativization. However, these correspondences are not always clear [63]. They are highly productive in two cases: *pu'al* is almost with no exception the passive of *pi'el*, and *huf'al* is the passive of *hif'il*.

Verbs can take pronominal suffixes, which are interpreted as direct objects. Such processes, however, become less frequent in contemporary Hebrew, and are usually associated with archaic language or a high register. In some cases, verbs can also take nominative pronominal suffixes, but this is now limited to participle forms, which may be interpreted as nouns.

Participle forms indeed deserve special attention, as they can be used as present tense verbs, but also as nouns or adjectives. For example, *mTps [metapes]* "climbing" can be used as a verb (*hwa mTps yl hqir* "he is climbing on the wall"), a noun (*hmTps hgiy al hpsgh* "the climber reached the summit"), or an adjective (*qniti cmx mTps lginh* "I bought a vine [= climbing plant] for the garden").

Nominals

Several morphological properties are common to nouns, adjectives, numerals and even prepositions; they are therefore grouped together in this section. Nouns inflect for number (singular, plural and dual); the dual form is not productive and is only preserved on a few nouns, such as those denoting body parts (*id [yad]* "hand" ⇒ *idim [yada'yim]* "hands"). Even in these cases, the dual is semantically plural, except for time expressions (e.g., *šntim [shnata'yim]* "two years"). Nouns that denote animate entities inflect for gender (masculine or feminine), although some exceptions are known [62].

Adjectives inflect for number (only singular and plural) and gender; the feminine suffix is either *h [a]*, *t [et]*, or *it [it]*. Numerals also inflect for gender, and ordinals also inflect for number; several idiosyncrasies occur, especially with the low numerals.

In addition, all these three types of nominals have two phonologically distinct forms, known as the *absolute* and *construct* states; the latter are used in compounds [9, 10]. For example, *šmlh [simla]* "dress" vs. *šmlt [simlat]*, as in *šmlt klh* "bridal gown"; or *qcr [qacar]* "short" vs. *qcr [qcar]* in *[qcar ru'ax]* "impatient [=short tempered]". In the standard Hebrew orthography approximately half of the nominals appear to have identical forms in both states, a fact which substantially adds to the ambiguity.

The proto-Semitic three-case system, with explicit indication of nominative, accusative and genitive cases, is not preserved in Hebrew. Personal pronouns, however, reflect traces of cases. Distinct pronominal forms exist for the nominative, accusative, dative and genitive. All these pronouns inflect for number, gender, and person. In addition, prepositions inflect for exactly these features, in a way that fully resembles the inflectional paradigm of pronouns (and, to some extent, that of nouns and adjectives), so that the distinction between inflected pronouns and prepositions is blurred. For example, the second person, feminine, singular, dative pronoun *lk [laK̄]* can be viewed as an inflected form of the preposition *l* "to", with a cliticized pronominal suffix indicating the number, gender and person.

A related phenomenon allows nouns to take possessive pronominal suffixes that inflect for number, gender and person. The base form for such inflections is the construct state. Thus, *šmlt* can combine with *k [eK̄]* to yield *šmltk* "your dress". In sum, then, the morphological characterization of many nominal forms includes the specification of their number and gender, as well as the person, number and gender of potential pronominal suffixes.

Other Closed-Class Items

The main remaining part-of-speech category is that of adverbs. Generally, adverbs do not inflect, but few exceptions exist (e.g., *laT* "slowly" and *lbd* "alone", which inflect for person, number and gender). Many adverbs are created from nouns by adding the preposition *b* "in", e.g., *b+mhirwt* "in+speed=quickly".

1.4.4 Morphological Ambiguity

The deficient orthography, including the lack of vowels and the attachment of frequent particles to the words that follow them, and the morphological complexity outlined above, greatly contribute to the ambiguity of Hebrew word forms. Itai and Wintner [49] report that their morphological analyzer produces 2.64 analyses per word token, on average, on a corpus of newspaper articles, with many tokens associated with more than a dozen analyses. More recent results show an average ambiguity level of 3.4 (excluding punctuation), with some tokens associated with over 20 analyses (Alon Itai, p. c.). These analyses can differ at the level of segmentation; or they can reflect different morphological features (e.g., one can be a construct state noun, while the other is absolute); or, in few extreme cases, they can be identical up to the identity of the stem.

As an example, consider the output produced by the morphological analyzer of Itai and Wintner [49] on the form *šbth*, depicted in Table 1.1 (adapted from Itai and Wintner [49]). Note in particular the last two analyses, which only differ in the lemma.

1.4.5 Basic Syntactic Structure

The dominant, unmarked word order in Hebrew is subject–verb–object (SVO), although many variations are possible [8, 38]. In particular, a very common construction (especially in journalistic jargon) is "verb-second", whereby the verb follows some constituent (typically, an adverbial, but sometimes an object) and precedes the subject and the rest of its complements.

Verbs agree with their subjects in number, gender and person. This facilitates some flexibility in constituent order, even without explicit case marking. When the subject is a pronoun, it may be omitted in certain cases, especially in the past and future tenses.

While clauses may be headed by verbs, this is not mandatory, and "verbless" predicates, whose heads are adjectives, prepositional phrases or nouns, abound [20]. Typically, in such cases the predicate is indefinite but the subject *is* definite. Clauses can also be headed by modals such as *crik* "it is necessary" or *aswr* "it is forbidden", typically followed by infinitival verb phrases or by finite clauses introduced by *š* "that". A unique construction involves the existentials *iš* "there is" and *ain* "there isn't", that behave like verbs in some respects, but are highly idiosyncratic.

Yes/no questions are either formed with the explicit interrogative *ham* "is it true that", or by changing the intonation pattern of the declarative counterpart, with no change in word order. Wh-questions are formed by fronting an interrogative pronoun (e.g., *mi* "who", *lmh* "why") but with no auxiliaries and no change in word order.

Within the noun phrase, nouns typically agree with their adjuncts (adjectives, demonstrative pronouns, and numerals) in gender and number, but also on definiteness. Hebrew only has a definite article, which is marked (as a prefix, *h*) on nouns,

adjectives, numerals and demonstrative pronouns, and has to be distributed over most of the components of a definite noun phrase [73].

Hebrew provides three different constructions for specifying genitive (possessive) constructions. First, using the genitive preposition *šl* "of": *hsprim šl hmšwrr* "the-books of the-poet". Second, by using the *construct* state of the head noun: *spri hmšwrr* "books-of the-poet". In such constructions, the definite article is only marked on the complement, and is "inherited" by the head noun [73]. As noted above, when the complement is a pronoun, it is realized as a cliticized suffix on the head noun: *spriw* "books-his". Finally, a double-genitive construction combines both the pronominal suffix *and* an explicit genitive complement, as in *spriw šl hmšrr* "books-his of the-poet".

Relative clauses are introduced by an explicit relativizer *š* or *ašr* "that". A third relativizer, *h*, is used in relative clauses that begin with a present-tense verb; such constructions can also be viewed as adjectival phrases, as present-tense verbs can be viewed as adjectives, and the relativizer *h* is a homograph of the definite article. Resumptive pronouns in the relative clause must agree with the head noun, but may be omitted in certain cases (and *must* be omitted in others). Resumptive pronouns are obligatory as complements of prepositions and as possessors. When the resumptive pronoun is preceded by a preposition, and opens the relative clause, the relativizer may be omitted.

1.5 Maltese

Maltese belongs to the South Arabic branch of Central Semitic [46, 47]. It has an Arabic stratum, a Romance (Sicilian, Italian) superstratum and an English adstratum [12,30]. The influence of the non-Semitic element is most obvious in the lexis, while the most salient basic grammatical structures are of Arabic origin. Maltese is the native language of approximately 400,000 people who live in Malta and Gozo, but it is also spoken abroad in communities in Australia, Canada, the USA and the UK.

1.5.1 Orthography

Maltese is the only Semitic language that uses a Latin-based alphabet. There are 31 letters in the Maltese alphabet. Two of these are digraphs, namely, *ie [ɪː]* and *għ*, which is generally silent but can also be pronounced */ħ/* under certain conditions (see below). Three symbols require diacritic marks, namely, *ċ /č/*, *ż /z/*, *ħ /ħ/* (including *għ*). The digraph *għ* and the unbarred *h* are generally silent, but they are sounded in certain contexts, as, for example, when they occur together as *għh*, as in *tagħha* "hers" *[ʾtɐħħɐ]*. Some speakers also sound the unbarred *h* in certain contexts in which the *h* is a part of the object pronominal clitic *-ha* "her" or *hom* "them", as in *raha [ʾrɐħħɐ]* "he saw her" instead of *[rɐ ːː]*. The latter is considered to be the standard variety, but the former is also very widespread.

Some graphemes are ambiguous in terms of pronunciation. Thus, double *z*, i.e., *zz*, can either be geminate *[ts:]*, as in *razza* "race" *['rɐ ts:ɐ]*, or geminate *[dz:]*, as in *gazzetta* "newspaper" *[gɐ dz:'ɜttɐ]*. Similarly, *x* can be *[ʃ]*, as in *rixa* "feather" *['ri:ʃɐ]*, or *[ʒ]*, as in *televixin* "television" *[tɜlɜ'vɪʒɪn]*. The grapheme *i* also has the values *[ɪ]* and *[i:]*, as in *fitt* "nuisance" *[fɪt]* and *[fti:t] ftit* "a little", but it can also be pronounced as *[ɪ:]* in certain contexts, e.g., when followed by the sound corresponding to graphemic *q*, *għ*, *h* or *ħ*, as in *riħ* "wind" *[rɪ:ħ]*. The letters generally retain their sound values but there are some surface phonetic/phonological rules that bring about changes in pronunciation that are not reflected in the spelling. For example, the rule of word final obstruent devoicing is not reflected in the orthography, which retains the underlying root sound. For example, *[bɪ:p]* is rendered orthographically as *bieb* "door", not **biep*, because of an alternation with *['bɪ:ba] bieba* "a door leaf".

1.5.2 Derivational Morphology

Mixed Root-Based and Stem-Based Morphology

At different phases of its history, Maltese borrowed profusely both from Romance (Sicilian, Tuscan, and Modern Italian), as well as from English, especially in recent times (see, in particular, Brincat [12] for a historical perspective; for a recent descriptive grammar of Maltese, see Borg and Azzopardi-Alexander [11]). As a result, Maltese displays a great deal of mixture, especially at the lexical level. While some of these borrowings have been integrated into largely Semitic/Arabic morphological and syntactic patterns, others have, in turn, had innovative effects on all levels, namely, phonological, grammatical (morpho-syntactic) and semantic. The result is that Maltese morphology appears to be both root-based and stem-based, at least on a surface analysis (see Fabri [28]; Twist [70]; Ussishkin and Twist [71] for various analyses and discussion).

The Arabic vocabulary of Maltese still retains the root-and-pattern characteristic typical of Semitic languages, whereby derivational and inflectional word forms are characterized by both internal changes to the basic consonant and vowel structure, i.e., non-concatenatively, as well as through affixation, i.e., concatenatively. For example, in derivation, a number of forms are related to the triliteral (triconsonantal) radical *q.s.m*, including *qasam* "split" (1st verbal form: *CVCVC*), *qassam* "share out" (2nd verbal form: *CVCiCiVC*), *tqassam* "get shared out" (5th verbal form: *t+ CVCiCiVC*), *nqasam* "broke" (7th verbal form: *n+CVCVC*), *qasma* "a split" (feminine singular nominal form: *CVCC-a*). In contrast to words of Semitic/Arabic origin, non-Semitic borrowings, especially recent ones, often retain their stem-based, purely concatenative properties. For example, the suffix *-ata*, borrowed from Italian, can combine with stems of Arabic origin to form new lexemes, as *żiblata* "a booze up", which is made up of *żibel* "thrash" and *-ata*.

In one period of its history, borrowings, especially from Sicilian and, later, Tuscan, were often integrated into the Semitic root-and-pattern system. To take an example, the word *pejjep* "to smoke" is of Romance origin from *pipa* "pipe", but was turned into a weak verb (with a middle weak consonant *j*) and made to undergo gemination of the second consonant to be turned into a verb of the second form, which generally characterizes the causative form of a basic verbal (form one) or nominal form. The productivity of these forms in Modern Maltese is a matter of discussion and research. There is some evidence that these forms might not be actively productive anymore.

One important way in which Maltese differs from Arabic and other Semitic languages is in the morphemic status of the vowel melody. While in Modern Standard Arabic (MSA), the vowel melody itself carries morphological information, e.g., *a* for perfective active and *u-i* for perfective passive, this is not the case in Maltese. In Maltese, the vowel melody either stays the same throughout, or changes for phonological reasons or unpredictably. Thus, e.g., the imperfective active and passive, and the perfective active and passive for a verb like *kiser* "break" are *jikser* "he breaks", *jinkiser* "he is broken", *kiser* "he broke", *nkiser* "he was broken" with the melody *i-e* throughout. In other words, Maltese has lost the typical morphologically motivated vowel 'insertion'; instead, it can be argued that vowel 'insertion' is purely phonological, allowing root syllabification [28, 64].

In Modern Maltese, new verbs are generally 'derived' through loan words which take on a very specific verbal form associated with so-called 'lacking' verbs (*verbi neqsin*), i.e., verbs that traditionally are considered to have a missing final weak consonant *j* in their citation form, e.g., the verb *tefa* "extinguish" (see, in particular, Mifsud 1995 on loan verbs). Thus, an English loan verb like *to monitor* becomes *immoniterja* "to monitor" and takes on the number, person, gender inflectional affixes, as, e.g., in *jien nimmoniterja* "I monitor", *int immoniterjajt* "you monitored".

Nominal derivations (nouns and adjectives) are also based on originally Arabic patterns as well as non-Arabic ones. To give an example, one way of forming nouns in Arabic is through the prefixation of *m-*. Similarly, in Maltese one finds *miżbla* "landfill", which is derivationally related to *żibel* "rubbish". Another, different example is *ħad(d)em* "to (make to) work" and *ħaddiem* "worker". With non-Semitic-based derivation, an intriguing question in Maltese is whether what appears to be an affix actually has affix status in the language. Since most of the forms that carry derivational affixes are loan words, one could argue that the affixes do not have independent status but have simply been imported with their stems as words. Of course, if one were to use the standard contrastiveness criterion to isolate morphemes, one would be able to isolate affixes in most cases, as in the case of *-(z)zjoni* in the following examples: *ammira* "admire", *ammirazzjoni* "admiration", *applika* "apply", *applikazzjoni* "application". However, one can only unequivocally assume that an affix is identified as such when it produces new and, therefore, productive local formations, which cannot have been imported as wholes. A number of such formations exist. For example, the Romance suffix *-ata*, as in *spagettata* "a spaghetti meal", as mentioned above, has recently been used to form the new word *żiblata* from *żibel* "rubbish" meaning "a booze up".

1.5.3 Inflectional Morphology

Verbs

As in derivation, in inflection, there are two main types: one that is (or at least behaves as if it were) root-based, and one that is stem-based. Thus, for example, from *tefa* "extinguish" (*CVCV*), we get *titfi* "she extinguishes" (*t-VCCV*), *tfejt* "I extinguished" (*CCV-jt*) and *tfiet* "she extinguished" (*CCV-Vt*), with varying stem patterns. In contrast, a verb like *immoniterja* "to monitor", has an invariant stem, thus *n-immoniterja* "I monitor", *immoniterja-jt* "I monitored" and *immoniterja-t* "she monitored" [28, 60]. The two basic paradigms are the imperfective, which, except for the plural suffix *-u*, is mainly prefixing: *n-* "1 person", *t-* "2 person", *j-* "3 person masculine", *t-* "3 person feminine", and the perfective, which is wholly suffixing: *-t* "1/2 person singular", *-Vt* "3 person feminine singular", *-na* "1 person plural", *-tu* "2 person plural", *-u* "3 person plural". Apart from the affixes which trigger subject-verb agreement, Maltese also has a set of personal pronoun enclitics that agree with a direct or indirect object topic: *-ni* "me", *-k* "you", *-u* "him", *-ha* "her", *-na* "us", *-kom* "you/pl", *-hom* "them". Thus: *Dik il-mara qra-t-ha l-ittra* "That woman read the letter"; literally "That woman, she read it/her the letter". The pronominal clitics also attach to nouns in the construct (possessive) and to prepositions: *ras-ha* "her head", *ħdej-ha* "near-her" [16].

A number of contexts trigger different kinds of allomorphy both in subject agreement affixes and topic object clitics, and also in root-based stems. The default meanings of the basic imperfective and perfective forms are the present habitual and the past tense, respectively. Other tense and aspect forms are formed through the use of particles and the tense marker *kien*; thus, *nisraq* "I steal", *qed nisraq* "I am stealing", *sa nisraq* "I am going to steal", *kont qed nisraq* "I was stealing", *kont sa nisraq* "I was going to steal", *sraqt* "I stole", *kont sraqt* "I had stolen" [21, 25].

Verbs are negated by means of a preposed *ma* and either a suffix *-x* or some other negator such as *ħadd* "nobody", *xejn* "nothing", and *mkien* "nowhere". The following are examples: *ma mortx* "I/you didn't go", *ma ġie ħadd* "nobody came". Interestingly, negative imperative forms do not have a *ma* preceding them, e.g., *tiftaħx* "do not open", and positive imperatives lack person marking, thus *iftaħ* "open" and not **tiftaħ*.

Nominals

Nouns and adjectives are generally marked for gender (masculine/feminine) and number (singular, plural for adjectives; singular, plural, collective, dual for nouns). Plural forms are undefined for gender, and some forms are unspecified for number and gender, such as *martri* "martyr/s" and *interessanti* "interesting", or for gender only, e.g., *tiċer* "male/female teacher" with the plural *tiċers*. With respect to number, singular masculine is generally unmarked, feminine is generally marked by a final

-a (though not exclusively), while the plural can be either marked by means of a number of different suffixes (strong plural: e.g., *karozz-a* "car", *karozz-i* "cars", *baħri* "sailor", *baħri-n* "sailors"), or non-concatenatively, through stem change (broken plural: e.g., *barmil* "bucket", *bramel* "buckets") (see Farrugia [33] on gender, and Schembri [66] and Farrugia [32] for a discussion of the broken plural). A number of nouns also display a collective form, which is morpho-syntactically masculine singular and denotes a mass or a collective set of objects; for example, *basal* "onion", as opposed to *basla* "one onion" and *basliet* "a number of onions". An even smaller set of nouns has a dual form with the suffix *-ajn/ejn*. In Modern Maltese this is restricted to words referring to objects that typically occur in pairs (*għajnejn* "two eyes", *idejn* "two hands"), as well as to a mixed bag of objects such as time words, e.g., *xahrejn* "two months", *ġimagħtejn* "two weeks", and measure terms, e.g., *uqitejn* "two ounces", but also *ħbiżtejn* "two loaves of bread" (see Fenech [34] for a full list and discussion). The dual has not only become restricted in Modern Maltese, but is also losing its meaning and being used as a plural form. Thus, *erba' għajn-ejn* "four eyes" is acceptable. Finally, there are also a few remnants of the so-called plural-of-the-plural forms, such as *tarf* "edge", *truf* "edges", *truf-ijiet* "several edges". The difference in meaning between the last two is not so easy to characterize. To conclude, one should also mention subtractive forms like *Lhudi* "Jew", *Lhud* "Jews", and suppletive forms like *mara* "woman", *nisa* "women", *tifel* "boy", *subien* "boys" and *tifla* "girl", *bniet* "girls".

Other Closed Class Items

There is no specific morphologically marked class of adverbs in Maltese. Thus, e.g., *tajjeb* can be used to modify a noun (i.e., functions as an adjectival modifier), as in *kejk tajjeb* "a good cake" or *pasta tajba* "a good pastry", in which case it triggers agreement, or it can be used to modify a verb, as in *ikanta tajjeb* "he sings well", in which case it has the default masculine singular form. Very often a fused form with the prepositional *bi* "with" is used as an adverb, as in *ġie bil-għaġġla* "he came hurriedly/in a hurry", *mar bil-mod* "he went slowly".

1.5.4 Basic Syntactic Structure

Generally, given specific intonational melodies, all S, V, and O orders, except for VSO, are possible in Maltese, with SVO being arguably the unmarked case (though not SV). Maltese is a topic-oriented language, especially in the spoken form. This means that any complement phrases, including the subject noun phrase and object phrases (direct, indirect, prepositional, locational, etc.) can be placed in different positions within the sentence. Topicality for direct and indirect objects is obligatorily marked by means of the pronominal enclitics mentioned in Sect. 1.5.3 (see also Fabri and Borg [31]). The following is an example: *Il-ktieb il-bieraħ*

Marija xtra-t-u "Maria bought the book yesterday"; literally "the book, yesterday, Maria she bought it/him". The same can be expressed as: *il-bieraħ Marija xtra-t-u l-ktieb* and *Marija l-bieraħ il-ktieb xtra-t-u* (among others). As a result, Maltese has a structurally free constituent order in terms of S(ubject), V(erb) and O(bject) on sentence level.

It is clear from the above that Maltese displays subject verb agreement and verb (topic) object agreement [24, 29]. Moreover, agreement in Maltese is strong enough to allow both subject and object pro-drop, i.e., it allows sentences without an explicit pronominal subject or (topic) object NP. Its rich agreement morphology allows the language to reconstruct the subject/(topic) object in every case from the agreement affixes and clitics on the verb. The same applies to prepositional and nominal possessor complement phrases when a pronominal enclitic occurs on the head preposition or noun. The following are examples: *bgħat-t-hu-lha* "I sent him to her", *fuq-ha* "on her", *xagħar-ha* "her hair". Like other pro-drop languages, Maltese also lacks expletive pronouns, i.e., it does not have anything that corresponds to *it* in *it seems that John is tired* in English. It also allows subject object inversion and extraction of the subject from a subordinate clause, two other properties associated with subject pro-drop. Maltese lacks morphological case, but has a case marker, *lil*, which marks specific, human direct objects: *qrajt il-ktieb* "I read the book" but *rajt lil Pawlu* "I saw Paul", and indirect objects: *bgħatt il-ktieb lil Pawlu* "I sent the book to Paul".

Within the noun phrase there is agreement between the demonstrative and the adjective, on the one hand, and the noun, on the other. The definite article can also occur on the adjective as well as on the noun. However, this is not a case of agreement in terms of definiteness, because the definite article on the adjective is triggered under specific pragmatically driven conditions [27]. A form that is homophonous with the numeral *wieħed* "one" can sometimes occur pre-nominally as a specificity marker (*a certain X*). Another 'typical' noun internal feature is the *construct state* construction, by which two juxtaposed nouns stand in a possessive relation. However, unlike Modern Standard Arabic, the possessor noun is not marked for genitive case in Maltese, since Maltese lacks morphological case altogether. Moreover, unlike Arabic, in Maltese this construction is limited, i.e., not any noun can enter into the construct relation with any other noun. In particular, typically *inalienable* relations, i.e., body parts and family relations are found in this construction. The construct is not usually possible with alienable relations, in which case a periphrastic construction with the possessive preposition *ta'* "of" must be used (however, see Fabri [26] and Koptjevskaja-Tamm [54] for a more detailed discussion): *xagħar it-tifel* "the boy's hair", *il-ktieb ta' Pawlu* "Paul's book".

Yes/no questions are generally distinguished from declaratives through specific intonational patterns. Wh-questions are formed by fronting the question word: *lil min rajt?* "who(m) did you see?" Relative clauses are introduced by the complementiser *li*, equivalent to "that" in English: *il-ktieb li qrajt* "the book that I read". Just like "that", *li* can also introduce a subordinate clause: *naf li rebaħ* "I know he won". Resumptive pronouns can occur under specific conditions in the form of the pronominal enclitics mentioned above.

1.6 Syriac

Syriac is the official liturgical language of a number of Eastern Churches in the Middle East and the Malabar Coast of India. It is the ethnic language of the Assyrians/Chaldeans/Syriacs which may number over one million users, none of whom (apart from a few known family experiments) speak it natively (that is not to say that some do not speak Neo-Aramaic dialects natively). First attested in A.D. 6, Syriac literature continues to be produced to the present day. Syriac has two dialects: Eastern and Western. The dialects differ mostly in orthography and phonology, but are quite similar in morphology and syntax.

1.6.1 Orthography

Syriac employs the usual Semitic consonantary which consists of 22 consonantal letters. Three scripts exist: *Estrangela* 'rounded' is the oldest, and is used today in most scholarly text editions. *Serto* or West Syriac has its roots in an early informal cursive hand, but becomes more dominant after the seventh century. *East Syriac* became a distinct script around the thirteenth century. Like Arabic and Hebrew, Syriac texts are mostly consonantal with three letters (*Alaph*, *Waw*, and *Yudh*) playing two roles: consonantal but also representing vowels. Since the earliest dated manuscript from A.D. 411, there is evidence of the use of a point to distinguish homographs, which is the origin of later pointing systems. The consonantal string *mn*, for example, can be either */man/* "who" or */men/* "from": a supralinear point on *m*, or between *m* and *n*, indicates */man/*, while a sublinear point indicates */men/*. By the seventh century, this pointing system developed into a more comprehensive one where each vocalic phoneme had its own set of points. Around the tenth century, and entirely restricted to West Syriac, a new vocalization system was developed where letters from Greek were borrowed and placed as supralinear or sublinear symbols to indicate vowels. None of the vocalization systems superseded previous ones. Today, one finds phrases that employ the single diacritic point, along with full pointing, along with the Greek symbols. Having said that, most texts are unmarked and appear in consonantal form only. Two morphological diacritics, however, are mandatory: a supralinear two-point grapheme, similar to the German umlaut, marks plural words, and a supralinear point grapheme on the suffix *-h* indicates gender (it is absent in the masculine, and present in the feminine).

The four letters in the string *bdwl* act as prefixes and in such cases are attached to the word; e.g., *byt? /baytā/* "house", *lbyt? /lbaytā/* "to the house", *wlbyt? /walbaytā/* "and to the house". Possessive pronouns and object pronominal suffixes are also attached to words; e.g., *ktb /ktab/* "he wrote", *ktbh /katbeh/* "he wrote it". Spacing in Syriac has its own complexities. Two and sometimes three words are attached without space, especially in familiar liturgical phrases. Syriac does not have special graphemes for numerals; instead, the alphabet is used to denote numerals.

A sublinear, sometimes supralinear, line has a number of functions, but primarily marks silent letters.

1.6.2 Derivational Morphology

Like the rest of the Semitic languages, Syriac morphology is that of root-and-pattern morphology. In the verbal system, the main patterns are:

1. *CCaC* (*Pʕal*)
2. *CaCCeC* in East Syriac and *CaCeC* in West Syriac (*Paʕʕel*)
3. *ʔaCCeC* (*AphʕelI*).

Each of these patterns has a corresponding reflexive pattern which is marked with the prefix *ʔet-*; hence, *ʔetCCeC*, *ʔetCaCCaC* (or *ʔetCaCaC* in West Syriac), and *ʔettaCCaC* (here the */ʔ/* in *ʔaCCeC* assimilates into */t/*).

The vocalization of the first pattern differs from one verb to another and is usually lexically marked. The vocalization of the remaining patterns is more or less invariant. Having said that, phonological processes may affect the vocalism; e.g., */e/* turns into */a/* when the third consonant is */r/*. The derivational mechanism here follows the same processes described in other sections of this chapter.

1.6.3 Inflectional Morphology

Conjugation patterns marking aspect (or tense), number, person, and gender are almost invariant within a pattern. However, the patterns may differ from one root class to the next. Root classes are defined by the values of the root consonants. Typically, the consonants *ʔ*, *w*, and *y*, when they fall in any position within the root, cause idiosyncratic conjugation patterns.

Number, person, and gender are marked with suffixation in the perfect (almost equal to past tense), or circumfixation in the case of the imperfect mood (almost equal to the future tense). Hence, this part of the process is entirely concatenative, though it may cause phonological processes to be triggered within the root-and-pattern part of the stem.

Roots in which the first radical is */n/* have a different derivation in the imperfect, considering that the imperfect prefix also begins with */n/*; hence, while the imperfect of the root morpheme *k.t.b* is */nektub/*, that of the root morpheme *n.s.b* is */nessab/* < */*nensab/*. Two phonological processes take place here, the */n/* of the root is deleted, and the second radical is duplicated. A similar process takes place with roots whose second and third radical are the same, such as *b.z.z*. Here, the imperfect is */nebbaz/*.

In addition to the inflectional morphology, prefixation and suffixation takes place. Prefixation is limited to the *bdwl* letters mentioned above. Suffixation is limited to object pronominal suffixes (in the case of nouns) and possessive pronouns in the case

of nouns and prepositions. There are two sets of possessive pronoun suffixes: one attaches to singular nouns and the other to plural nouns. In the case of prepositions, each preposition attaches to one of the sets only and the choice is idiosyncratic and lexically marked.

1.6.4 Syntax

The field of Syriac syntax has not been fully explored, and most of the available studies pertain mostly to the genre of biblical texts [50, 72]. In particular, we are unaware of any computational work on Syriac syntax. We therefore do not address Syriac syntax in this chapter.

1.7 Contrastive Analysis

Much research has been done in the area of comparative Semitic linguistics [46, 57, 61]. We summarize in this section the similarities and differences among the various Semitic languages across the dimensions outlined in previous sections, with an eye to the impact of those similarities and differences on computational processing.

Overall, the dimension in which Semitic languages vary the most is perhaps their writing systems, followed by phonology, morphology and then syntax. The similarities that most uniquely define the Semitic family are their morphology and to a lesser degree syntax and some of their orthographic choices. The variations among languages in the Semitic family are generally comparable to variations among members of other families.

1.7.1 Orthography

In terms of their scripts, the Semitic languages show much variation: each of the languages we discussed above has its own script: Arabic, Hebrew, Syriac (has three by itself), Amharic and Maltese (Latin script). In the case of Arabic, Hebrew, Syriac and Maltese, the scripts of these four languages are historically related to the ancient Phoenician alphabet. Though these related languages use different scripts primarily, their scripts have been used successfully and for extended periods for a variety of other languages from other families: Arabic script used for Persian, Urdu, Pashto, Ottoman Turkish, among others and Hebrew used for Yiddish and Ladino. Semitic scripts have also been known to be used for writing other Semitic languages/dialects: e.g., Judeo-Arabic is Arabic written in Hebrew, Garshuni is Arabic written in Syriac script, and Aramaic (a variant of Syriac) is also written in Hebrew script.

Although Arabic, Hebrew and Syriac have distinctly different scripts, they share very similar orthographic conventions. Most prominent is the use of optional diacritic marks for short vowels and consonantal gemination, or what Daniels and Bright [19] call an *Abjad*. Maltese uses the Latin script with alphabetical spelling conventions that attempt one-to-one mapping of grapheme to phoneme. The Amharic writing system is quite distinct from Arabic, Hebrew and Syriac on one hand and from Maltese on the other hand. Amharic is a syllabary (*Abugida* writing system), that can be argued to be less ambiguous than Arabic/Hebrew/Syriac, but more so than Maltese.

Diacritic optionality is a big part of the challenge of handling Semitic languages of the Abjad family. All three Semitic sisters are quite morphologically rich and cliticizing only adds to the ambiguity space. Arabic, in contrast with Hebrew, has a morphophonemic spelling system where some affixes and clitics are spelled in a morphemic from that abstracts away from various allomorphs: e.g., the definite article, the feminine singular suffix and the masculine plural suffix. Syriac has two morpho-phonemic symbols for marking plurals and feminine singulars. Hebrew in contrast is phonemic in its spelling (modulo the missing diacritics). Hebrew spelling however has some unique challenges: first is the various overlapping allophones for some phonemes (*b:b/v, k:k/x, p:p/f, v:v/u, s:s/š*) and several graphemes with the same pronunciations (*q/k, T/t, x/k*), which have no parallel in Arabic or Maltese, although a similar phenomenon occurs in Syriac. Second, Modern Hebrew uses two spelling standards, with or without diacritic symbols that mark the vowels. Arabic dialects have no official standard orthographies and as such add a higher degree of complexity to computational processing. These orthographic issues (optional diacritics, clitics, complex phonology-orthography mapping and inconsistent spellings) contribute to why Semitic languages are challenging in the context of morphological analysis and disambiguation, speech recognition and text-to-speech, not to mention language modeling.

1.7.2 Phonology

Semitic phonology, like Semitic orthography, appears quite diverse. There are, however, many shared features. Emphatic, uvular and pharyngeal sounds are an important marker for Arabic and its dialects. For emphatics, Arabic has four, Syriac has two only, while Hebrew and Maltese have lost them completely (although they are retained in the script). Hebrew retains a couple in the script but not in pronunciation. The Arabic emphatics appear as ejectives in Amharic. Hebrew and Syriac, unlike Arabic and Maltese, have a few phonemes with distinct allophones: the so called *beged-kefet* (*bgdkpt*) phonemes. Syriac still makes the distinctions fully, but Modern Hebrew only does so for three (*bkp*). Arabic dialects give an interesting living example of phonological language change as the old and new pronunciations of Arabic coexist as part of the diglossic situation in the Arab world.

In terms of the Semitic vowel inventory, there is much diversity. Classical Arabic has three short and three long vowels. Arabic dialects have a wide range: Moroccan Arabic has three vowels (no length distinction) and Levantine has eight (three short and five long). Hebrew used to have long/short distinctions, some of which are retained in the script, but Modern Hebrew has no length distinction any more. Amharic and Classical Syriac have seven vowels; modern dialects of Syriac seem to have fewer (between three and five).

Shared phonemes, such as /*m*/ and /*n*/, and regularly mappable phonemes, such as /*s*/ and /*š*/, are important in establishing etymological connection across different Semitic languages. Scripts that preserve some of the historical distinction are quite important as well, especially in the case of Hebrew, establishing the relation of Arabic /*q*/ to Hebrew /*k*/ (written *q*).

1.7.3 Morphology

Morphology is the core of the "Semitic" linguistic classification. In contrast to the variable orthographies and phonologies of the Semitic languages, Semitic morphology has unique unifying aspects that define the Semitic language family and distinguishes it from other language families. The landmark of Semitic morphology is the use of templatic morphemes in addition to concatenative ones.

Although Arabic is the language famous for its templatic "broken" plurals, similar phenomena exist in Hebrew also to a limited extent. Templatic morphology is highly productive in Semitic languages and it depends on the concept of the *root* morpheme, a typically triliteral abstraction that captures some general meaning. Many of these roots seem to have shared meaning across multiple Semitic languages, e.g., the famous *k.t.b* (writing-related) root. Others have different or polysemous meanings, some of which are shared (more on this in Sect. 1.7.5 below). Different root types typically involving gemination of second and third radical, or weak radicals (*w/y/h/'*), seem to create similar challenges for the different Semitic sisters.

Concatenative morphology in Semitic languages has some shared features: verbs have typically two basic forms: the prefixing-stem (Arabic imperfective, Hebrew future) and the suffixing-stem (Arabic perfective, Hebrew past). The map from the morphemes to their meaning or function will vary however: the perfective/past is similar across Semitic languages, but the imperfective has a wider range, including present, future, subjunctive, etc.

The set of inflectional features (as opposed to morphemes which realize these features) are generally similar. Arabic has a larger set of inflectional features since it includes nominal case and verbal mood. Arabic has been described as a "conservative" language as opposed to other "progressive" languages, which would include Hebrew and other Arabic dialects which have lost case/mood in a similar fashion to what happened in the transition from Latin to Romance languages. It is often thought that Arabic (classical, modern standard) may reflect

some earlier forms of the proto-Semitic morphology that gave birth to the Semitic family. Of course, Arabic dialects and Maltese provide an interesting insight into how language evolves, since unlike Hebrew, which was "academic/religious" for centuries before being revived, these languages evolved naturally so to speak. The dialects seem to add to the complexities of Arabic morphology, but at the same time they remove some: case/mood are gone, but negative suffixes and indirect object clitics are introduced. This suggests that the Semitic family is continuing to change and evolve in new directions.

1.7.4 Syntax

In general, the syntactic properties of Semitic languages do not introduce specific difficulties for computational processing. There is much variety in the syntax of the Semitic family. In the extreme, Amharic is distinct from the rest of the languages discussed in this chapter. Amharic is a head-final language: its dominant sentential order is S-O-V and nominal phrases are Adjective-Noun. In contrast, Arabic is strongly head-initial: the dominant sentential order is V-S-O and nominal phrases are Noun-Adj. Arabic also allows S-V-O order; and Arabic dialects and Hebrew tend to be more prominently S-V-O with some V-S-O cases. Maltese and Syriac are also primarily S-V-O languages. All of the Semitic languages in this chapter, except Amharic, have post-nominal modifiers.

In Arabic dialects, Hebrew, and Maltese, the possessive construction (*idafa*, *smikhut* or *construct*) co-exists with an alternative prepositional possessive construction. Standard Arabic does not allow the prepositional construction. Both Hebrew and some Arabic dialects allow different forms of the double possession construction (e.g., *his house of the man*). Syriac primarily uses the prepositional possessive construction.

In Arabic, Hebrew, and Maltese, adjectives follow and agree with their head nouns in definiteness, gender and number. Arabic adds case agreement and odd rules for irrational plurality. Different combinations of definite/indefinite nominals are important in forming and parsing different types of syntactic constructions.

1.7.5 Lexicon

Semitic languages share numerous lexical items with cognate roots that are readily identifiable or easily identifiable once some of the phonological correspondences are adjusted for: *ktb* has to do with "writing", *qwl* with voice/speech, etc. Other roots are faux amis: *rqd* is "dance" in Hebrew or "lie down" in Arabic. Other roots have multiple senses, some of which match and some do not: the root *lHm/lxm* means "bread" in Hebrew or "meat" in Arabic in its first sense, but it means "solder" in both languages in its second sense.

The spread of the Semitic languages and colonization by other language groups has led to extensive borrowing. Arabic dialects have several colonizing English, French, and Turkish influences as well as colonized Berber, Coptic and Syriac influences. Maltese can be seen as a hybrid of Arabic and Italian, with much English influence. Hebrew has numerous Russian borrowings, as well as Arabic borrowings (by design as part of its revival and by ongoing interactions with Arabic speakers). Hebrew borrowing in standard Arabic is rather ancient and restricted to religious concepts, but it is prevalent in contemporary Palestinian Arabic.

1.8 Conclusion

We provided in this chapter a necessarily brief overview of some of the most prominent linguistic features of Semitic languages, covering the living languages in this family for which some computational work has been done. Our aim was not to give a thorough, extensive account of any language or any particular phenomenon (many books do precisely this). Rather, we tried to provide sufficient detail that would emphasize the challenges involved in computational processing of Semitic languages, the similarities and also the differences among them. We hope that this will prove useful for readers of subsequent chapters of this book.

References

1. Abdel-Massih ET, Abdel-Malek ZN, Badawi ESM (1979) A reference grammar of Egyptian Arabic. Georgetown University Press, Washington, DC
2. Alkuhlani S, Habash N (2011) A corpus for modeling morpho-syntactic agreement in Arabic: gender, number and rationality. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics (ACL'11), Portland
3. Amberber M (2002) Verb classes and transitivity in Amharic. Lincom, Munich
4. Anberbir T, Takara T, Gasser M, Yoon KD (2011) Grapheme-to-phoneme conversion for Amharic text-to-speech system. In: Proceedings of conference on human language technology for development, Bibliotheca Alexandrina, Alexandria
5. Badawi ESM (1973) Mustawayat al-'Arabiyya al-mu'asira fi Misr (the levels of modern Arabic in Egypt). Dar al-Ma'arif, Cairo
6. Bassiouney R (2009) Arabic sociolinguistics: topics in diglossia, gender, identity, and politics. Georgetown University Press, Washington, DC
7. Beesley KR (1997) Romanization, transcription and transliteration. http://www.xrce.xerox.com/Research-Development/Historical-projects/Linguistic-Demos/Arabic-Morphological-Analysis-and-Generation/Romanization-Transcription-and-Transliteration
8. Berman RA (1978) Modern Hebrew structure. University Publishing Projects, Tel Aviv
9. Borer H (1988) On the morphological parallelism between compounds and constructs. In: Booij G, van Marle J (eds) Yearbook of morphology 1. Foris, Dordrecht, pp 45–65
10. Borer H (1996) The construct in review. In: Lecarme J, Lowenstamm J, Shlonsky U (eds) Studies in Afroasiatic grammar. Holland Academic Graphics, The Hague, pp 30–61
11. Borg A, Azzopardi-Alexander M (1997) Maltese. Routledge, London

12. Brincat JM (2011) Maltese and other languages: a linguistic history of Malta. Midsea, Malta
13. Brustad K (2000) The syntax of spoken Arabic: a comparative study of Moroccan, Egyptian, Syrian, and Kuwaiti dialects. Georgetown University Press, Washington, DC
14. Buckwalter T (2002) Buckwalter Arabic morphological analyzer. Linguistic Data Consortium (LDC), Philadelphia. Catalog number LDC2002L49 and ISBN 1-58563-257-0
15. Buckwalter T (2004) Buckwalter Arabic morphological analyzer version 2.0. LDC, Philadelphia. Catalog number LDC2004L02, ISBN 1-58563-324-0
16. Camilleri M (2009) Clitics in Maltese. BA thesis, University of Malta
17. Cowell MW (1964) A reference grammar of Syrian Arabic. Georgetown University Press, Washington, DC
18. Daniels PT (1997) Scripts of Semitic languages. In: Hetzron R (ed) The Semitic languages. Routledge, London/New York, chap 2, pp 16–45
19. Daniels PT, Bright W (eds) (1996) The World's writing systems. Oxford University Press, New York
20. Doron E (1983) Verbless predicates in Hebrew. PhD thesis, University of Texas at Austin
21. Ebert K (2000) Aspect in Maltese. In: Dahl O (ed) Tense and aspect in the languages of Europe. Mouton de Gruyter, Berlin
22. El Kholy A, Habash N (2010) Techniques for Arabic morphological detokenization and orthographic denormalization. In: Proceedings of LREC-2010, Malta
23. Erwin W (1963) A short reference grammar of Iraqi Arabic. Georgetown University Press, Washington, DC
24. Fabri R (1993) Kongruenz und die Grammatik des Maltesischen. Niemeyer, Tübingen
25. Fabri R (1995) The tense and aspect system of Maltese. In: Thieroff R (ed) Tempussysteme in Europaeischen Sprachen II. Niemeyer, Tübingen, pp 327–343
26. Fabri R (1996) The construct state and the pseudo-construct state in Maltese. Rivista di Linguistica 8(1):229–244
27. Fabri R (2001) Definiteness marking and the structure of the NP in Maltese. Verbum 23(2):153–172
28. Fabri R (2009) Stem allomorphy in the Maltese verb. Ilsienna – Our Language 1/2009:1–20
29. Fabri R (2009) To agree or not to agree: suspension of formal agreement in Maltese. In: Fabri R (ed) Maltese linguistics: a snapshot; in memory of Joseph A. Cremona (1922–2003). Il-Lingwa Taghna – Our Language. Brockmeyer, Bochum, pp 35–61
30. Fabri R (2010) Maltese. Revue Belge de Philologie et d'Histoire 88(3):791–816
31. Fabri R, Borg A (2002) Topic, focus and word order in Maltese. In: Abderrahim Y, Benjelloun F, Dahbi M, Iraqui-Sinaceur Z (eds) Aspects of the dialects of Arabic today. Proceedings of the 4th conference of the international Arabic Dialectology Association (AIDA), Amapatril, Rabat, pp 354–363
32. Farrugia A (2008) Maltimorph: a computational analysis of the Maltese broken plural. BSc thesis, University of Malta
33. Farrugia G (2010) Il-Ġens grammatikali fil-Malti. PhD thesis, University of Malta
34. Fenech E (1996) Functions of the dual suffix in Maltese. Rivista di Linguistica 8:89–100
35. Ferguson CF (1959) Diglossia. Word 15(2):325–340
36. Gadish R (ed) (2001) Klalei ha-Ktiv Hasar ha-Niqqud, 4th edn. Academy for the Hebrew Language, Brooklyn (in Hebrew)
37. Gasser M (2010) A dependency grammar for Amharic. In: Proceedings of the workshop on language resources and human language technologies for Semitic languages, Valletta
38. Glinert L (1989) The grammar of modern Hebrew. Cambridge University Press, Cambridge
39. Habash N (2006) On Arabic and its dialects. Multiling Magazi 17(81). Getting Started Guide: Middle East Insert, pp 12–15
40. Habash N (2007) Arabic morphological representations for machine translation. In: van den Bosch A, Soudi A (eds) Arabic computational morphology: knowledge-based and empirical methods. Springer, Dordrecht
41. Habash N (2010) Introduction to Arabic natural language processing. In: Synthesis lectures on human language technologies. Morgan & Claypool, San Rafael. doi:http://dx.doi.org/10.2200/S00277ED1V01Y201008HLT010

42. Habash N, Soudi A, Buckwalter T (2007) On Arabic transliteration. In: Soudi A, Neumann G, van den Bosch A (eds) Arabic computational morphology, text, speech and language technology, vol 38. Springer, Dordrecht, chap 2, pp 15–22. http://dx.doi.org/10.1007/978-1-4020-6046-5_2
43. Habash N, Diab M, Rabmow O (2012) Conventional orthography for dialectal Arabic. In: Proceedings of the language resources and evaluation conference (LREC), Istanbul
44. Harrell R (1962) A short reference grammar of Moroccan Arabic. Georgetown University Press, Washington, DC
45. Hetzron R (1970) Towards an Amharic case grammar. Stud Afr Linguist 1:301–354
46. Hetzron R (ed) (1997) The Semitic languages. Routledge, London/New York
47. Holes C (2004) Modern Arabic: structures, functions, and varieties. Georgetown University Press, Washington, DC
48. Horvath J, Wexler P (1997) Relexification in Creole and Non-Creole languages – with special attention to Haitian Creole, Modern Hebrew, Romani, and Rumanian, Mediterranean Language and culture monograph series, vol xiii. Harrassowitz, Wiesbaden
49. Itai A, Wintner S (2008) Language resources for Hebrew. Lang Resour Eval 42(1):75–98
50. Joosten J (1996) The Syriac language of the Peshitta and Old Syriac versions of Matthew: syntactic structure, inner-Syriac developments and translation technique. In: Studies in Semitic languages and linguistics. Brill, Leiden
51. Kapeliuk O (1988) Nominalization in Amharic. Franz Steiner, Stuttgart
52. Kapeliuk O (1994) Syntax of the noun in Amharic. O. Harrassowitz, Wiesbaden
53. Kaye AS, Rosenhouse J (1997) Arabic dialects and Maltese. In: Hetzron R (ed) The Semitic languages. Routledge, London/New York, chap 14, pp 263–311
54. Koptjevskaja-Tamm M (1996) Possessive NPs in Maltese: alienability, iconicity and grammaticalization. Riv di Linguist 8(1):245–274
55. Kramer R (2009) Definite markers, Phi features, and agreement: a morphosyntactic investigation of the Amharic DP. PhD thesis, University of California
56. Leslau W (1995) Reference grammar of Amharic. Harrassowitz, Wiesbaden
57. Lipiński E (2001) Semitic languages, outline of a comparative grammar. Peeters, Leuven
58. Loos EE, Anderson S, Dwight HJ Day, Jordan PC, Wingate JD (2004) Glossary of linguistic terms. http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/
59. McCarthy JJ (1981) A prosodic theory of nonconcatenative morphology. Linguist Inq 12(3):373–418
60. Mifsud M (1995) Loan verbs in Maltese: a descriptive and comparative study. Brill, New York
61. Moscati S (1969) An introduction to the comparative grammar of the Semitic langauges, phonology and morphology. Otto Harrassowitz, Wiesbaden
62. Ordan N, Wintner S (2005) Representing natural gender in multilingual lexical databases. Int J Lexicogr 18(3):357–370
63. Ornan U (2003) The final word. University of Haifa Press, Haifa (in Hebrew)
64. Puech G (Forthcoming) Syllabic structure and stress in Maltese. In: Caruana S, Fabri R, Stolz T (eds) Variation and change: the dynamics of Maltese in space, time, and society. Akademie Verlag, Berlin
65. Rogers H (2005) Writing systems: a linguistic approach. Blackwell Publishing, Malden
66. Schembri T (2006) The broken plural in Maltese – an analysis. B.A. thesis, University of Malta
67. Shimron J (ed) (2003) Language processing and acquisition in languages of semitic, root-based, morphology. In: Language acquisition and language disorders, vol 28. John Benjamins, Amsterdam/Philadelphia
68. Smrž O (2007) Functional Arabic morphology. Formal system and implementation. PhD thesis, Charles University in Prague
69. Teferra A, Hudson G (2007) Essentials of Amharic. Rüdiger Köppe Verlag, Köln
70. Twist AE (2006) A psycholinguistic investigation of the verbal morphology of Maltese. PhD thesis, University of Arizona
71. Ussishkin A, Twist A (2009) Auditory and visual lexical decision in Maltese. In: Comrie B, Fabri R, Mifsud M, Hume E, Mifsud M, Stolz T, Vanhove M (eds) Introducing Maltese linguistics. Benjamins, Amsterdam, pp 207–231

72. Williams PJ (2001) Studies in the Syntax of the Peshitta of 1 kings. In: Monographs of the Peshita institute. Brill, Leiden
73. Wintner S (2000) Definiteness in the Hebrew noun phrase. J Linguist 36:319–363
74. Wintner S (2009) Language resources for Semitic languages: challenges and solutions. In: Nirenburg S (ed) Language engineering for lesser-studied languages. IOS, Amsterdam, pp 277–290
75. Zuckermann G (2003) Language contact and lexical enrichment in Israeli Hebrew. Palgrave Macmillan, London/Ney York
76. Zwicky AM (1985) Clitics and particles. Language 61(2):283–305
77. Zwicky AM, Pullum GK (1983) Cliticization vs. inflection: English n't. Language 59(3): 502–513