Development of Maximally Reusable Grammars: Parallel Development of Hebrew and Arabic Grammars

Tali Arad Greshler[1], Livnat Herzig Sheinfux[1], Nurit Melnik[2] and Shuly Wintner[1]

[1]Department of Computer Science, University of Haifa
[2]Department of Literature, Language and the Arts, The Open University of Israel

1 Introduction

Our goal in this paper is to develop deep linguistic grammars of two different yet related languages. We show that such grammars can be developed and implemented in parallel, with language-independent fragments serving as shared resources, and language-specific ones defined separately for each language. The desirability of reusable grammars is twofold. From an engineering perspective, reuse of code is clearly parsimonious. From a theoretical perspective, aiming to maximize the common core of different grammars enables better identification and investigation of language-specific and cross-linguistic phenomena.

The two grammars in the focus of this paper are of Modern Hebrew (MH) and Modern Standard Arabic (MSA). The basic infrastructure, or core, of the grammars is "standard" HPSG (Pollard and Sag, 1994; Sag et al., 2003). As these two languages are related, they exhibit a number of shared phenomena which can be attributed to their Semitic roots. Nevertheless, since the languages diverged several millennia ago the end grammars are quite different and do require language-specific accounts.

More generally, we identify four types of relations that exist between the grammars of two languages: (i) The two languages share some construction or syntactic phenomenon. (ii) Some phenomenon is present in one language but is absent from the other. (iii) The two languages share some construction, but impose different constraints on its realization. (iv) Some phenomenon seems similar in the two languages, but is in fact a realization of different constructions. While the challenge is to maximize the common parts of the grammars, it is important to be cautious with seemingly similar phenomena across the two languages. In some cases, as we will show, the solution is to define a shared construction with different language-specific constraints. Conversely, other cases are best accounted for by the definition of distinct constructions.

This paper demonstrates how the different types of relations can be implemented in parallel grammars with maximally shared resources. The examples pertain to the MH and MSA grammars, yet similar issues and considerations are applicable to other pairs of languages that have some degree of similarity.

2 Reusable grammars of Modern Hebrew and Modern Standard Arabic

Our starting point is HeGram, a deep linguistic processing grammar of Modern Hebrew (Herzig Sheinfux et al., 2015). HeGram is grounded in the theoretical framework of HPSG and is implemented in the LKB (Copestake, 2002) and ACE systems. AraGram, the MSA grammar, utilizes the types defined in HeGram, as long as they are relevant for Arabic. In cases where the two languages diverge with respect to particular phenomena, language-specific types are defined in separate language-specific modules. More technically, the two grammars make extensive use of the ":+" operator provided by the

LKB in order to define a type in a shared file, and to separately add language-specific constraints to its definition (see (9) and (10) below).

The parallel development of the two grammars with their shared resources requires a careful examination of the common and distinct properties of the two languages. Types, features, values and constraints can only be added or modified in a way that does not negatively affect the grammar of the other language (see the discussion of grammar testing with [incr tsdb()] (Oepen, 2001) in section 3).

In the following sections we focus on a number of phenomena which illustrate different types of relations between the two languages and their implementation. We begin with a discussion of the way subcategorization is handled by the two grammars. We show that while semantic selection is found to be language-independent, the syntactic realization of arguments may be subject to language-specific constraints. Next, we describe the way the nominals of the two languages are represented in the lexical type hierarchy. In this case, the MH hierarchy is found to be a sub-hierarchy of the MSA one. Finally, we move on to clause structure. We discuss one case where two seemingly similar constructions are found to be licensed by distinct mechanisms, and another where the two languages share the same basic construction, yet impose different constraints on its realization.

2.1 Maximally shared resources: subcategorization

The architecture of HeGram embodies significant changes to the way argument structure is standardly viewed in HPSG. The main one is that it distinguishes between semantic selection and syntactic selection, and provides a way of stating constraints regarding each level separately. Moreover, one lexical entry can account for multiple subcategorization frames, including argument optionality and the realization of arguments with different syntactic phrase types (e.g., *want food* vs. *want to eat*). This involves the distribution of valence features across ten categories. Each valence category is characterized in terms of its semantic role, as well as the types of syntactic phrases which can realize it (referred to as *syntactic realization classes*). Consequently, the semantic relations denoted by predicates consist of coherent argument roles, which are consistent across all predicates in the language.

The non-standard argument structure representation of HeGram is instrumental for distinguishing between general and language-specific properties of the grammar. The 10-VALENCE-feature architecture of HeGram is intended to be as language-independent as possible; indeed, corpus investigations of MSA verbs showed that they share the semantic frames identified for their MH counterparts, and consequently no changes were required in the overall argument representation scheme. Nevertheless, the languages differ with respect to the syntactic realization of semantic arguments. As examples, consider the following types which correspond to the variants of the verb 'come' in MH (1) and MSA (2). Both the MSA verb ʒaːʔa ('*came*') and the MH verb *higiʕa* ('*came*') have the same semantic frame, selecting an *Actor*, a *Source*, and a *Goal*. However, while *Goal* arguments can be realized only as PPs in MH, in MSA they can also appear as NPs. Thus, the (semantic) R(EALIZATION)-FRAME of the two verbs is identical (i.e., *arg-1-15-16-156*), but the syntactic realization of the *Goal* argument (i.e, Arg6) differs between the two languages as shown by the different types: *arg1-15-156_p_p* and *arg1-15-16-156_p_np*, which inherit from *arg6_p* and *arg6_np*, respectively.

(1) MH *higiʕa* ('*came*')

```
arg1-15-16-156_p_p := arg1_n & arg5_p & arg6_p
      [ SYNSEM.LOCAL.CAT.VAL.R-FRAMES arg1-15-16-156 ].
```

(2) MSA ʒaːʔa ('*came*')

```
arg1-15-16-156_p_np := arg1_n & arg5_p & arg6_np
      [ SYNSEM.LOCAL.CAT.VAL.R-FRAMES arg1-15-16-156 ].
```

In sum, the realization classes associated with different semantic roles are found to vary to some extent between languages while the semantic roles themselves appear to be more general.

2.2 Similarities between the languages: nominals in the lexical type hierarchy

MH and MSA are languages with rich, productive morphologies. Nouns in the two languages have natural or grammatical gender, and are marked for number. Adjectives decline according to a number-gender inflectional paradigm. Both categories are also morphologically marked for definiteness. Consequently, the grammars of the two languages require an elaborate nominal type hierarchy, where types are cross-classified according to the three dimensions: NUMBER, GENDER and DEFINITENESS.[1]

The nominal type hierarchy described above is sufficient for MH, while MSA requires an extension of the hierarchy in order to account for two additional properties: *dual number* and *Case*. A sketch of the basic shared hierarchy, along with the MSA extensions (in the boxes) is given in Figure 1. All MH nominals (i.e., nouns and adjectives) are instances of types which realize all the cross-classification combinations of the three MH-relevant dimensions (e.g., *sm-def-nom-lex*).

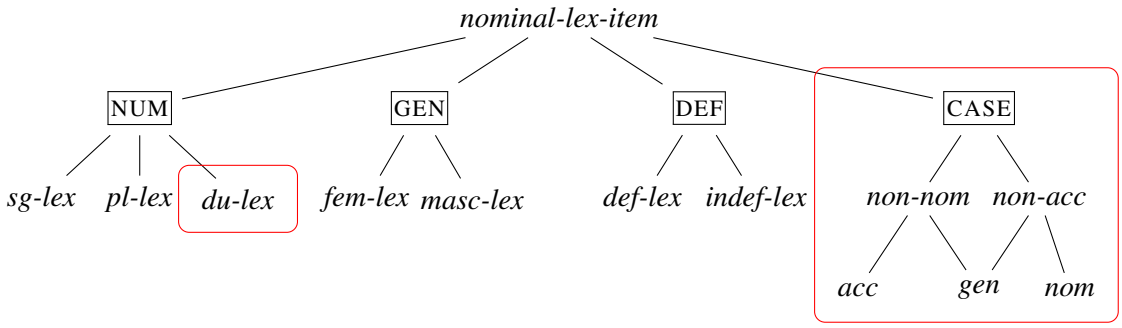


Figure 1: The nominal type hierarchy

Case in MSA is morphologically marked on all nominals by word-final vowels. Thus, in principle, all lexemes are cross-classified according to four dimensions: NUMBER, GENDER, DEFINITENESS, and CASE. The MH lexical entry for 'boy' (3) is an instance of a lexical type cross-classified according to three dimensions, whereas its MSA counterpart (4) is an instance of a lexical type which is additionally classified as accusative (marked in a box).[2]

(3) MH *yeled* ('*boy*')

```
ild := indef-cmn-3sm-noun-lex &
       [ STEM < "ild" >,
       SYNSEM.LKEYS.KEYREL.PRED _boy_n_rel ].
```

(4) MSA *walad-an* ('*boy*')

```
wlda := indef-cmn-acc-3sm-lex &
       [ STEM < "wlda" >,
       SYNSEM.LKEYS.KEYREL.PRED _boy_n_rel ].
```

[1] Since the PERSON dimension is only relevant to nouns, not to adjectives, it is not presented here as part of the nominal type hierarchy.

[2] In our grammars we use 1:1 transliteration schemes for both MH and MSA. These schemes lack vowel representations as vowels are not represented in MH and MSA scripts. In glossed examples, however, we use phonemic transcription that includes vowels.

Note that the hierarchy below Case is structured to represent two different disjunctive groupings: non-nominative and non-accusative. As some MSA nominals are orthographically underspecified for Case, this intermediate level of the hierarchy was added as an engineering choice, in order to avoid repetition in the lexicon.

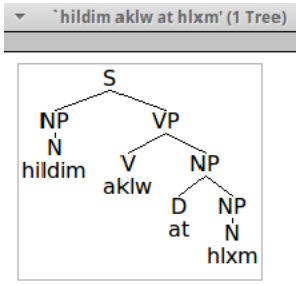
2.3 Deep and superficial similarities: clause structure

MH and MSA have different unmarked clause structures. In MH, SVO is the canonical word order, while in MSA it is VSO. Nevertheless, the unmarked clause order of MH is a marked structure in MSA, and vice versa. In addition, a notable property of MSA clauses is that subject-verb agreement depends on the subject position; verbs in SVO clauses exhibit full person-number-gender agreement with the subject, while in VSO clauses number agreement is suppressed and the verb is invariably singular. This is not the case in MH, where the verb fully agrees with the subject regardless of its position.[3]

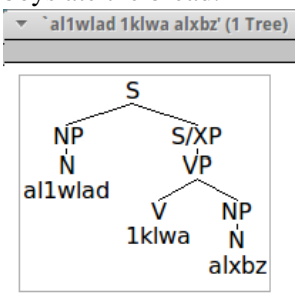
2.3.1 Superficial similarities, different constructions: SVO

The SVO clauses of the two languages are remarkably similar; the finite verb exhibits full person-number-gender agreement with the subject which precedes it. As examples, consider the following SVO clauses in MH (5) and MSA (6).

(5) *ha-yeladim axlu et ha-leħem*
*the-boys ate.*3PM ACC *the-bread*
'The boys ate the bread.'



(6) *ʔl-ʔawlaːd-u ʔakaluː l-xubz-a*
the-boys-NOM *ate.*3PM *the-bread*-ACC
'The boys ate the bread.'



While superficially almost identical, the SVO clauses of the two languages are given distinct analyses in our grammars. The unmarked MH SVO clause is licensed by a *subject-head-phrase* phrase type. This is not the case for MSA: Arabic clauses in which the subject precedes the verb have been thoroughly discussed in the literature (Fassi Fehri, 1993; Mohammad, 2000; Aoun et al., 2010; Alotaibi and Borsley, 2013, among others). We adopt the analysis discussed in Aoun et al. (2010) and elaborated and cast in HPSG by Alotaibi and Borsley (2013), according to which Arabic has only one genuine subject position, which is postverbal, whereas the allegedly preverbal subject position is really a topic position. According to this analysis, there is no real SVO order in Arabic. Instead, what looks on the surface like SVO is really a VSO construction in which the subject has been topicalized. Thus, the syntactic phenomenon of a "true" SVO clause is present in MH but absent from MSA. The two distinct analyses are shown in the syntactic trees pertaining to each of the sentences.

Consequently, *subject-head-phrase* is defined only in the MH grammar, while marked preverbal subjects in MSA are accounted for as instances of long-distance dependency. The types dedicated to long-distance dependency constructions are shared by the two languages. Nevertheless, the MH grammar is more restrictive with regard to topicalization; it confines the phenomenon only to non-subjects in order to avoid vacuous structural ambiguity with SVO clauses. MSA, on the other hand, allows all de-

[3]Exceptions to this generalization are colloquial verb-initial constructions.

pendents to be topicalized. This disparity is implemented by using *extracted-subject-phrase* as a shared resource, and adding a language-specific constraint just for the MH grammar.

The *extracted-subject-phrase* type is thus used in both grammars; in MH it is used only to license subject *Wh*-questions, while in MSA it is used for both subject *Wh*-questions and subject topicalizations. The use of a shared type reflects the generalization that both languages have subject *Wh*-questions and allows maximal reusability of the type hierarchy below the shared *extracted-subject-phrase* type.

2.3.2 Different constraints on the same construction: VSO

VSO constructions in both MH (7) and MSA (8) have a *head-subj-comp-phrase* phrase type, and thus its type definition is shared.[4]

(7) *et ha-leħem axlu ha-yeladim*
ACC *the-bread ate.*3PM *the-boys*
'The bread, the boys ate it.'

(8) *ʔakala l-ʔawlaːd-u l-xubz-a*
*ate.*3SM *the-boys*-NOM *the-bread*-ACC
'The boys ate the bread.'

There are, however, additional language-specific constraints which further restrict this clause type. In Hebrew, VSO constructions are only licensed in a V2 configuration, where some clause-initial material precedes the verb, e.g., *et ha-leħem* ('*ACC the-bread*') in (7). An additional Hebrew-specific constraint restricts this phrase type only to cases where the verb has undergone extraction (9). The MSA grammar, on the other hand, imposes its own language-specific constraint: the verb is invariably singular (10).

(9) MH Head Subject Complement constraint

```
VS-basic-head-subj-phrase :+
 [  HEAD-DTR basic-extracted-arg-phrase-lex-head-dtr  ].
```

(10) MSA Head Subject Complement constraint

```
VS-basic-head-subj-phrase :+
 [  HEAD-DTR.SYNSEM.LOCAL.CAT.HEAD.CNCRD png-s  ].
```

This mechanism, where two languages share a construction and each language adds a different constraint to it without damaging the rest of the hierarchy, is an excellent utilization of HPSG type hierarchies, allowing maximal reusability in developing and implementing two grammars with a common core.

3 Current status and future prospects

We have adapted HeGram (Herzig Sheinfux et al., 2015) to Arabic along the lines discussed above. AraGram currently covers a plethora of syntactic phenomena, including Case marking, subject-verb and noun-adjective agreement, SVO and VSO word order, relatively free complement order, multiple subcategorization frames, selectional restrictions of verbs on their PP complements, topicalization, passive and unaccusative verbs. Many of these phenomena required only minor adaptations to the Hebrew grammar. Therefore, the development of AraGram took only several weeks (excluding corpus investigation and literature review). For comparison, the development of HeGram to its stage when we started developing AraGram took about a year. AraGram currently shares 95.5% of its types with HeGram, while HeGram currently shares 99.2% of its types with AraGram.

In order to guarantee that the changes introduced by the grammar of one language do not damage the grammar of the other we developed test suites of grammatical and ungrammatical sentences for both

[4]Since only unary and binary branches are employed in the grammar, the *head-subj-comp-phrase* phrase type is implemented with two types: *head-subject* and *head-comp* (with a realized subject).

Arabic (160 sentences, 41 ungrammatical) and Hebrew (432 sentences, 106 ungrammatical) and test the grammar rigorously with [incr tsdb()] (Oepen, 2001).

The development of AraGram is ongoing. In the near future, we will focus on additional constructions, including wh-questions, control, raising, the copular construction, and multi-word expressions. We also intend to work on automatic translation between the languages using generation by MRSs (semantic representations).

References

Mansour Alotaibi and Robert D. Borsley. Gaps and resumptive pronouns in Modern Standard Arabic. In Stefan Müller, editor, *Proceedings of the 20th International Conference on HPSG*, pages 6–26, Stanford, 2013. CSLI Publications.

Joseph E. Aoun, Elabbas Benmamoun, and Lina Choueiri. *The syntax of Arabic*. Cambridge University Press, 2010.

Ann Copestake. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford, 2002.

Abdelkader Fassi Fehri. *Issues in the structure of Arabic clauses and words*. Kluwer, Dordrecht, 1993.

Livnat Herzig Sheinfux, Nurit Melnik, and Shuly Wintner. Representing argument structure in computational grammars. Submitted, 2015.

Mohammad A. Mohammad. *Word order, agreement, and pronominalization in Standard and Palestinian Arabic*. John Benjamins, Amsterdam, 2000.

Stephan Oepen. [incr tsdb()] — competence and performance laboratory. User manual. Technical report, Computational Linguistics, Saarland University, Saarbrücken, Germany, 2001.

Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications, 1994.

Ivan A. Sag, Thomas Wasow, and Emily M. Bender. *Syntactic Theory: A Formal Introduction*. CSLI Publications, Stanford, California, 2 edition, 2003.