Lacking Integrity: HPSG as a Morphosyntactic Theory

Guy Emerson
University of Cambridge

Ann Copestake
University of Cambridge

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 75–95

http://csli-publications.stanford.edu/HPSG/2015

Emerson, Guy, & Copestake, Ann. 2015. Lacking Integrity: HPSG as a Morphosyntactic Theory. In Müller, Stefan (Ed.), *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, 75–95. Stanford, CA: CSLI Publications.

Abstract

Standard accounts of HPSG assume a distinction between morphology and syntax. However, despite decades of research, no cross-linguistically valid definition of 'word' has emerged (Haspelmath, 2011), suggesting that no sharp distinction is justified. Under such a view, the basic units are morphemes, rather than words, but it has been argued this raises problems when analysing phenomena such as zero inflection, syncretism, stem alternations, and extended exponence. We argue that with existing HPSG machinery, a morpheme-based approach can in fact deal with such issues. To illustrate this, we consider Slovene nominal declension and Georgian verb agreement, which have both been used to argue against constructive morpheme-based approaches. We overcome these concerns through use of a type hierarchy, and give a morpheme-based analysis which is simpler than the alternatives. Furthermore, we can recast notions from Word-and-Paradigm morphology, such as 'rule of referral' and 'stem space', in our framework. We conclude that using HPSG as a unified morphosyntactic theory is not only feasible, but also yields fruitful insights.

1 Word Segmentation and Lexical Integrity

The Lexical Integrity Principle holds that syntactic rules do not have access to internal parts of words (Bresnan & Mchombo, 1995; Asudeh et al., 2013). Although this principle is often not explicitly stated in HPSG, it is usually implicitly assumed that there there is some notion of 'word', with a corresponding division of labour between lexical and phrasal rules. For example, Sag et al. (2003, p.228ff.) describe the use of lexemes as abstract structures from which we can derive families of wordforms differing only by inflection, where this process is carried out using lexical rules. However, while they take it for granted that we can identify words, the difficulties in defining the term 'word' have been known for some time:

> "Many forms lie on the border-line between bound forms and words, or between words and phrases; it is impossible to make a rigid distinction" — (Bloomfield, 1933)
>
> "What we call 'words' in one language may be units of a different kind from the 'words' in another language" — (Lyons, 1968)
>
> "There may be clear criteria for wordhood in individual languages, but we have no clear-cut set of criteria that can be applied to the totality of the world's languages" — (Spencer, 2006)

More recently, Haspelmath (2011) identifies ten possible criteria for defining words: potential pauses, free occurrence, mobility, uninterruptibility, non-selectivity, non-coordinability, anaphoric islandhood, non-extractability, phonological idiosyncrasies, and non-biuniqueness. They argue that none of these criteria, nor any combinations of them, coincide with linguistic or orthographic practice. Furthermore, we cannot retreat by saying that words are simply a language-specific

concept. If we are forced to define words separately for each language, then we can quite easily define several word-like levels in any particular language, and we have no reason to give special status to one particular level. Moreover, given a language like Mandarin Chinese, where linguists cannot agree on what to call the Mandarin Word (Mair, 1990; Packard, 2000; Sun, 2006; Tang, 2010), this is not an obscure thought experiment, but a fundamental issue affecting the most widely spoken language on the planet. Haspelmath concludes:

> "Linguists have no good basis for identifying words across languages, and hence no good basis for a general distinction between syntax and morphology" — (Haspelmath, 2011)

Under this view, the distinction between morphology and syntax vanishes, leaving us with a single domain of morphosyntax, with abstract morphemes as the basic units. This pushes us towards an Item-and-Arrangement view of morphological phenomena, rather than Item-and-Process or Word-and-Paradigm (WP) views, since the latter approaches require a notion of 'word'. In Stump (2001)'s terms, we are pushed towards a lexical and incremental theory, rather than an inferential or realizational one. In Blevins (2006)'s terms, we are pushed towards a constructive theory, rather than an abstractive one. However, this is not to say that we must abandon progress made in these other frameworks. Far from it – many generalizations stated in word-based accounts can be re-expressed in morphemic terms, and we will discuss several in this paper.[1] Doing so allows us to frame them in a theory that is cross-linguistically more consistent, and where the analyses can mesh seamlessly with syntax above the 'word' level.

If we accept Haspelmath's conclusion, we are prompted to consider whether we can reformulate HPSG in terms of morphemes. In the following section, we argue not only that this is possible, but further that the use of type hierarchies makes HPSG particularly appealing as a morphosyntactic theory, as it can sidestep many problems attributed to morphemic approaches. In sections 3 and 4, we apply this framework to Slovene stem alternations and Georgian verb agreement, which have been claimed to pose problems for a morphemic approach. Along the way, we show how insights drawn from WP morphology can be recast in our framework.

2 Morphosyntactic HPSG

Recasting HPSG as a morphosyntactic theory can be done without fundamental changes to the architecture. HPSG is usually regarded as a lexicalist theory, but while the term 'lexicalism' has often been associated with lexical integrity, particularly as the term is used by transformational grammarians, we only require a relatively minor change to Sag et al.'s definition of 'strong lexicalism'. This states

[1] Roark & Sproat (2007) also demonstrate that lexical-incremental theories and inferential-realizational theories are computationally equivalent, since both can be implemented in the same model, using an FST.

firstly that the locus of grammatical and semantic information is the lexicon, and secondly that lexical entries correspond directly to the words present in a sentence.[2] We must only state instead that lexical entries correspond to morphemes, not words:

$$morpheme \rightarrow LE_1 \vee \cdots \vee LE_n$$

These lexical entries must be minimal, rather than derived by lexical rules. To formalize this idea, we propose the following (meta)principle:

> **The Morphemic Principle:** Phonological material may only be stipulated in lexical entries, not in syntactic or lexical rules.

This implies that the only way to combine phonological material is by combining lexical entries through non-unary syntactic rules, i.e. by combining morphemes. Furthermore, phonological material is not split between lexical rules and lexical entries – all morphemes are stored directly in the lexicon. This would remain true no matter what the orthographic conventions are, so adhering to such a principle would make grammars more consistent cross-linguistically.

A second reference to words lies in the Head-Complement Schema, which builds a phrase out of a word and its complements (Pollard & Sag, 1994; Sag et al., 2003). Without a notion of 'word', this instead becomes a process of building one type of phrase out of a second type of phrase and its complements. What this means is that the Head-Complement Schema must be restated in terms of pairs of types (t_1, t_2):

$$\begin{bmatrix} t_1 & \\ \text{...HEAD} & \boxed{1} \\ \text{...COMPS} & \langle\rangle \end{bmatrix} \rightarrow \begin{bmatrix} t_2 & \\ \text{...HEAD} & \boxed{1} \\ \text{...COMPS} & \langle \boxed{2} \rangle \end{bmatrix}, \boxed{2}$$

Allowing phrases to be the head daughter of a head-complement construction has in fact been motivated independently. Instead of a flat structure where the head combines with all complements at once, we can use a binary-branching structure where the head combines with one complement at a time, which allows adjuncts or subjects to intervene between the head and its complements. For example, such an approach is used in the English Resource Grammar (Flickinger et al., 2000), in the Grammar Matrix (Bender et al., 2002), to analyse the German Mittelfeld (Crysmann, 2003), and to analyse partial-VP fronting (Müller, 2015).

In conclusion, neither of the above changes are inherently problematic. However, after removing lexical rules from the theory, and assuming morphemes to be the basic units, we need to justify that it is still possible to capture phenomena traditionally regarded as morphological. In section 2.1, we clarify what we mean by 'morpheme'; in section 2.2, we review the difficulties attributed to a morphemic view; and in section 2.3, we show how the criticisms made on morphosyntactic grounds do not apply when using feature structures and a type hierarchy.

[2]The second half of this statement is also known as the Word Principle.

2.1 What is a Morpheme?

In order to shift to a morpheme-based view of morphosyntax, we have to ask whether morphemes can be more easily identified than words. However, a number of different definitions of 'morpheme' have been proposed in the literature, with some more problematic than others.

We follow Bender & Good (2005)'s notion of an 'abstract morpheme'. Under this view, we assume that a language can be split between the *morphophonology*, which establishes a correspondence between surface forms and sequences of abstract morphemes, and the *morphosyntax*, which establishes a correspondence between sequences of abstract morphemes and syntactic/semantic representations of utterances.

In this way, an abstract morpheme is a Saussurean sign, because it contains both semantic and phonological information. Furthermore, it is a minimal sign, because it is the smallest unit with both kinds of information.

While this definition may be 'weaker' than some, it is a substantive claim to say that we can analyse language in terms of abstract morphemes, and this view makes two assumptions explicit. Firstly, language is discrete,[3] in the sense that we can represent an utterance in terms of a finite number of elements from a discrete set. Secondly, morphophonology and morphosyntax are largely independent.

Where we differ from Bender and Good is to assume that the morphophonology acts not on an individual 'word', but on the whole utterance. This allows us to deal with mismatches between phonological and syntactic structure, for example Kwak'wala [kwk] definiteness and case markers, which are phonological suffixes but syntactic prefixes (Boas et al., 1947).

Assuming this overall architecture, the questions we need to ask are: can we systematically map between surface forms and abstract morpheme sequences? Can we systematically assign suitable structures to individual morphemes? And can we systematically build the semantics of the whole from the semantics of the parts? In the following sections, we discuss the challenges these questions raise, although the focus of this paper is on the second and third questions.

2.2 Challenges for Morphemes

Many objections have been raised against analysing language in terms of morphemes (Anderson, 1992; Matthews, 1991; Bochner, 1993), and they can be broadly split between considerations of phonological, semantic, and syntactic phenomena. The focus of this paper is on the latter, but we briefly discuss the first two issues now.

Various phonological phenomena resist segmentation, including metathesis, subtraction, discontinuous elements, infixation, reduplication, suprasegmental features, and apophony. However, a correspondence between surface forms and ab-

[3] An acoustic signal varies continuously in both time and amplitude, but it is nonetheless perceived categorically (Goldstone & Hendrickson, 2010)

stract morphemes does not need to explicitly involve segmentation; the correspondence is with the whole sequence of abstract morphemes, which may not be individually tied to parts of the input. This kind of analysis can be represented using a finite state transducer (FST), a simple and efficient formalism described in detail by Beesley & Karttunen (2003). Finite state techniques can express many phonological/morphological theories (such as autosegmental phonology (Kay, 1987), context-sensitive rewrite rules (Kaplan & Kay, 1994), and Paradigm Function Morphology (Karttunen, 2003), among others) and have been used to describe a variety of 'morphologically rich' languages (such as Finnish (Koskenniemi, 1983) and Turkish (Oflazer, 1994), among others). We believe that the above phenomena can be described using abstract morphemes and finite state techniques, although details are beyond the scope of this paper. What is important to note is that where we use PHON in the rest of the paper, we are not referring to the surface form, but to the representation of the abstract morpheme used by an FST.

Semantic idiosyncrasies, such as 'cranberry' morphemes and Latinate prefixes (*re-ceive, per-ceive*), have been proposed as posing difficulties for morphemic approaches. However, such phenomena are not limited to sub-word combinations, and idiosyncratic multi-word expressions are widespread (Sag et al., 2002). If the semantic objections to morphemes are valid, then we must also object to any constituent within a multiword expression. We view this conclusion as absurd, and we believe techniques used to analyse multiwords, such as those discussed by Sag et al., can also be applied to morphemes.

We now turn to syntactic objections, which can be reduced to the following:

1. Extended exponence (multiple overt morphemes expressing a feature)
2. 'Zero' inflection (no overt morphemes expressing a feature)
3. Syncretism (alternative feature values associated with the same morpheme)
4. Stem alternations (alternative morphemes associated with the same features)

Extended exponence can be dealt with using unification. Each exponent of a feature has that specified in its feature structure, and when multiple exponents occur, the features are unified, analogously to agreement.

Syncretism can be modelled using underspecified types. In some cases, this will involve a single type hierarchy for multiple featural dimensions, a technique which has been successfully used to analyse various languages, for example by Flickinger (2000) for person and number in English, and by Crysmann (2005) for number, gender, and case in German. Indeed, Krieger & Nerbonne (1993) argue that 'matrix-based' descriptions of paradigms can always be given a 'form-based' analysis, where each form is underspecified for a set of agreement values.

Although we could try to model zero inflection using morphemes without phonological material (since this is expressible using an FST), this would lead to rampant homophony between such morphemes. Instead, we first note that it only makes sense to postulate a zero element if it can be identified via overt elements

competing for the same slot (Sanders, 1988). When an overt morpheme fills a slot, the type of the mother (the whole phrase) and the type of the other daughter (the rest of the phrase) will in general be different. We can therefore replace 'zero morphemes' by unary syntactic rules, with appropriate types for the mother and daughter, and which stipulate the features associated with the 'zero'.

It has been claimed that contextually-determined stem alternations and similar kinds of allomorphy constitute a problem, because multiple morphemes are associated with a set of features, but only one morpheme is used in a given context. However, in such cases, we can associate each stem or morpheme with the contexts in which it appears. The typed feature structure corresponding to the set of contexts may be highly underspecified, but this does not present a challenge to the theory. This is also true for 'morphomic stems' (Aronoff, 1994), where many features may play a role, and where values of these features may depend on one another – we will see such an example in the Slovene data below. In more extreme cases, some elements are called 'empty' morphemes, because they are allegedly associated with no features at all. However, we reject such a view, since such morphemes will only appear in some contexts but not others, and we can therefore associate the morpheme with the relevant features for those contexts. In a sense, because we can represent morphomic stems and empty morphemes with underspecified forms, we can see this as a special case of syncretism.

In short, none of the above objections represent an obstacle to a type-driven morphemic approach. However, it should also be noted that the same cannot be said of all morphemic theories. For example, our arguments do not apply to the influential framework of Distributed Morphology (Halle & Marantz, 1993), because that theory lacks the notions of underspecification and unification. Instead, they are forced to introduce other devices, such as competition between morphemes, which we will argue against in our analysis of Georgian. Of all the mechanisms that have so far been proposed, underspecification and unification seem to us to be the only straightforward way of capturing many-to-many mappings between morphemes and features.

2.3 Modelling Morphological Paradigms

Having described the general approach, we now describe the mechanical details. We focus on inflection in this paper, but we note that our approach could be extended to include derivational morphology. Indeed, Lieber (2004) and Booij (2005) argue that derivation can be handled in an Item-and-Arrangement theory, which fits neatly with our morpheme-driven framework.

Inflectional paradigms can often be represented in terms of a root and a number of affixes, falling into discrete position classes, or slots.[4] To model the affixation, we must decide whether the root or the affixes should act as heads.

[4] As noted by Crysmann & Bonami (2015), morpheme positions can vary. While we do not deal with morphotactics in detail here, we note that variable morpheme orders can in principle be dealt with in the same way as variable constituent orders in syntax.

If the root should act as head, we can introduce an MCOMPS list, with one item for each slot in the paradigm. This list should intuitively be separate from the COMPS and SPR lists, because inflection is separate from argument structure. Affixation would then be represented using a Head-MComp Schema:

$$\begin{bmatrix} t_1 & \\ \ldots\text{HEAD} & \boxed{1} \\ \ldots\text{SUBJ} & \boxed{2} \\ \ldots\text{COMPS} & \boxed{3} \\ \ldots\text{MCOMPS} & \langle\,\rangle \\ \ldots\text{ARG-ST} & \boxed{2} \oplus \boxed{3} \end{bmatrix} \rightarrow \begin{bmatrix} t_2 & \\ \ldots\text{HEAD} & \boxed{1} \\ \ldots\text{SUBJ} & \boxed{2} \\ \ldots\text{COMPS} & \boxed{3} \\ \ldots\text{MCOMPS} & \langle \boxed{4} \rangle \\ \ldots\text{ARG-ST} & \boxed{2} \oplus \boxed{3} \end{bmatrix}, \boxed{4}$$

For an affix to share its features with the whole expression, we can introduce a re-entrancy between the head features of the root and the affix, as shown below. In the case of zero inflection, we can use a unary rule which removes an element from the MCOMPS list and unifies the appropriate head features with the root.

$$\begin{bmatrix} \textit{root} & \\ \ldots\text{HEAD} & \boxed{1} \\ \ldots\text{MCOMPS} & \left\langle \begin{bmatrix} \textit{affix} & \\ \ldots\text{HEAD} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix} \begin{bmatrix} \textit{affix} & \\ \ldots\text{HEAD} & \begin{bmatrix} \text{AGR} & \textit{agr} \end{bmatrix} \end{bmatrix}$$

If the affix should act as head, we can avoid introducing an MCOMPS list, and instead take the root or stem to be the specifier of the affix:

$$\begin{bmatrix} \textit{affix} & \\ \text{SYNSEM|LOC|CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{AGR} & \textit{agr} \end{bmatrix} \\ \text{SPR} & \begin{bmatrix} \textit{stem} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

As above, we introduce re-entrancies between the head features of the root, affix, and whole expression, which we can do in the phrasal type:

$$\begin{bmatrix} \textit{affixed-stem} & \\ \text{SYNSEM} & \boxed{3} \begin{bmatrix} \text{LOC|CAT|HEAD} & \boxed{1} \end{bmatrix} \\ \text{HEAD-DTR} & \begin{bmatrix} \textit{affix} & \\ \text{SYNSEM|LOC|CAT} & \begin{bmatrix} \text{HEAD} & \boxed{1} \\ \text{SPR} & \boxed{2} \end{bmatrix} \end{bmatrix} \\ \text{SPR-DTR} & \boxed{2} \begin{bmatrix} \textit{stem} & \\ \text{SYNSEM} & \boxed{3} \end{bmatrix} \end{bmatrix}$$

For zero inflection, we stipulate the information in a unary rule with the same pair of types as used in the above head-specifier construction:

$$\left[\begin{array}{ll} \textit{affixed-stem} & \\ \text{SYNSEM} & \boxed{3}\left[\text{LOC}|\text{CAT}|\text{HEAD} \quad \left[\text{AGR} \quad \textit{agr}\right]\right] \\ \text{HEAD-DTR} & \left[\begin{array}{ll} \textit{stem} & \\ \text{SYNSEM} & \boxed{3} \end{array}\right] \end{array}\right]$$

In the following sections, we will use the affix-as-head analysis. This creates a natural similarity between auxiliaries and affixes, which is lost in the MCOMPS analysis. However, we want to stress that the general claim of this paper is not affected by the choice of mechanism: in either case, the claimed problems with morphemes can be overcome using a type-driven approach.

3 Slovene Stem Alternations

Here we consider a situation where the choice of a noun's stem is sensitive to number and case features. This situation exhibits all four of the issues mentioned above, and we will show how the use of a type hierarchy can overcome each of them. We will further show how notions developed in WP approaches, such as 'stem space' and 'rule of referral', can not only be re-expressed in our type-driven morphemic approach, but can in fact be expressed more robustly.

Slovene nouns inflect for three numbers (singular, dual, plural) and six cases. An example of the simplest kind of declension is shown in table 1, involving a single stem, and a slot for one case/number suffix. Some suffixes are syncretic for either case or number, such as *-oma* (dative or instrumental) and *-ih* (dual or plural). This can be modelled by organizing number and case in type hierarchies, with an underspecified type for each observed syncretism, as shown in figure 1.

	SINGULAR	DUAL	PLURAL
NOMINATIVE	*mést-o*	*mést-i*	*mést-a*
ACCUSATIVE	*mést-o*	*mést-i*	*mést-a*
GENITIVE	*mést-a*	*mést*	*mést*
DATIVE	*mést-u*	*mést-oma*	*mést-om*
INSTRUMENTAL	*mést-om*	*mést-oma*	*mést-i*
LOCATIVE	*mést-u*	*mést-ih*	*mést-ih*

Table 1: Declension with a single stem

Taking the suffix to be the head, and the noun stem to be its specifier, we get phrasal types and lexical entries as shown in figure 2. Where there is a 'zero' suffix, we introduce a unary rule.

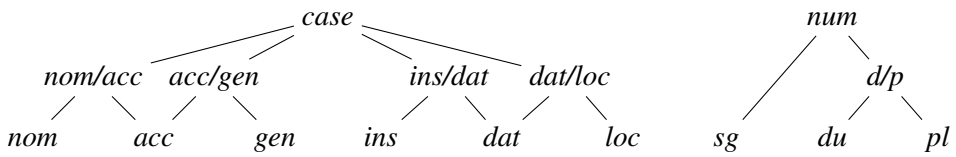

Figure 1: Case and number type hierarchies for Slovene

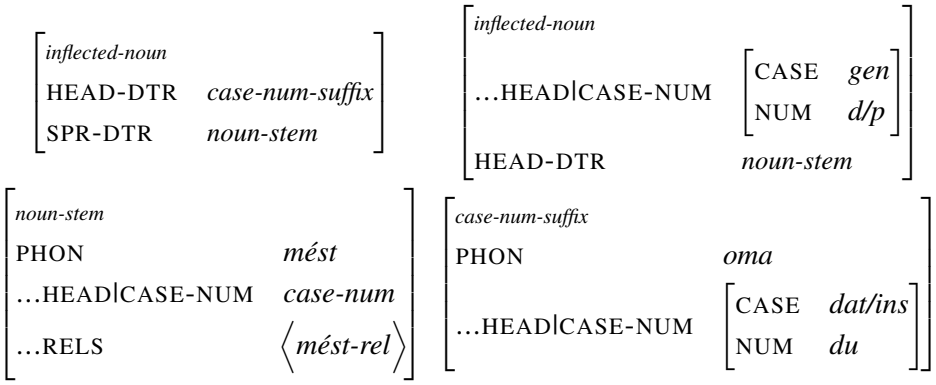

Figure 2: Phrasal types and lexical entries for case-number suffixes

A more complicated declension is shown in table 2, where an additional infixing element is present for all dual and plural forms, appearing between the noun root and the case suffix. This is an example of extended exponence, since each of the suffixes already indicates dual/plural number, and the *-ôv-* infix redundantly specifies it again. We can model this declension using the phrase structure shown in figure 3. The infix takes a noun root as its specifier, to yield a noun stem, which can then combine with a case-number suffix as before.[5] Phrasal types and lexical entries for this declension are shown in figure 4.

	SINGULAR	DUAL	PLURAL
NOMINATIVE	*grád*	*grad-ôv-a*	*grad-ôv-i*
ACCUSATIVE	*grád*	*grad-ôv-a*	*grad-ôv-e*
GENITIVE	*grad-ú*	*grad-ôv*	*grad-ôv*
DATIVE	*grád-u*	*grad-ôv-oma*	*grad-ôv-om*
INSTRUMENTAL	*grád-om*	*grad-ôv-oma*	*grad-ôv-i*
LOCATIVE	*grád-u*	*grad-ôv-ih*	*grad-ôv-ih*

Table 2: Declension with a distinct dual/plural stem

[5]There are differences in endings between the declensions for *grád* and *mést*, which exemplify two of the many declensions in Slovene. To model inflectional classes, each noun should also have a feature indicating its class, and each suffix should impose a constraint on the class of its specifier. If some suffixes appear in multiple classes (as is the case for Slovene), the classes can be organized in a hierarchy, and each suffix selects for an underspecified class.

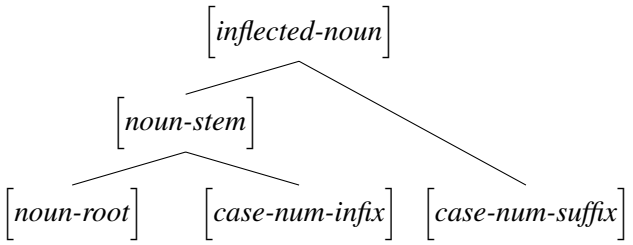

Figure 3: Phrase structure of an inflected noun

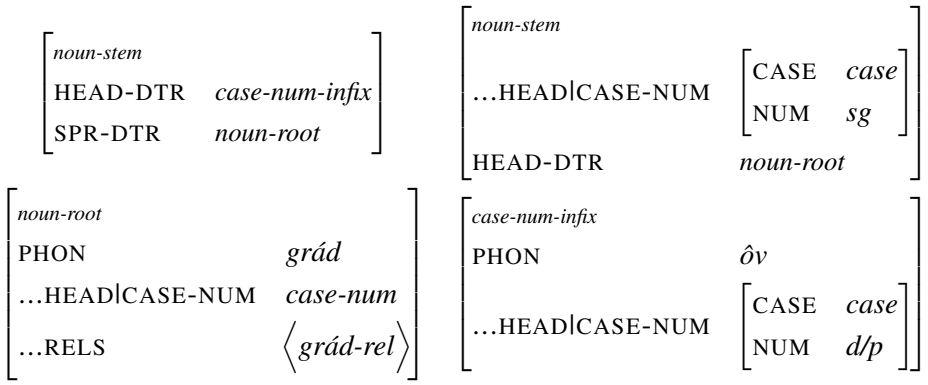

Figure 4: Phrasal types and lexical entries for case-number infixes

A number of Slovene nouns change stem, but with a pattern that involves both number and case. For example, *nágelj* ('carnation') has the stem *nágelj-n* for all forms other than nominative and accusative singular. We can deal with this in the same way as for *grád*, but unlike the *-ôv-* infix, which could be described using a pure number feature, the *-n-* infix requires a combined case-number feature.

The unique pair of stems *člôvek* and *ljud* ('man/men'), exhibits an unusual pattern of suppletion, where *člôvek* is used for the singular, *ljud* is used for the plural, and they are split in the dual, as shown in table 3 (Priestly, 1993). Furthermore, the cases where the plural stem *ljud* is used for the dual are precisely those which display syncretism in the suffixes, suggesting a deeper generalization is to be found.

	SINGULAR	DUAL	PLURAL
NOMINATIVE	*člôvek*	*človék-a*	*ljud-jé*
ACCUSATIVE	*človék-a*	*človék-a*	*ljud-í*
GENITIVE	*človék-a*	*ljud-í*	*ljud-í*
DATIVE	*človék-u*	*človék-oma*	*ljud-ém*
INSTRUMENTAL	*človék-om*	*človék-oma*	*ljud-mí*
LOCATIVE	*človék-u*	*ljud-éh*	*ljud-éh*

Table 3: Declension with suppletive stems

Corbett (2015) analyses this at the level of a paradigm, within the framework of Network Morphology, introducing 'generalized referral' rules that stipulate that the forms for the genitive and locative dual should be identical to the plural forms. Under such an analysis, however, we cannot immediately infer that using the wrong stem for the genitive dual is ungrammatical, as we need to compare it to other parts of the paradigm.

Instead, we give an analysis where the ungrammatical forms are directly ruled out by unification failure. By combining number and case into a single hierarchy, it is possible to introduce types so that each stem can only appear in the appropriate combinations of number and case. The fact that the two stems are part of the same paradigm is captured by the semantic predicate being the same for both.

$$\begin{bmatrix} \textit{gen.sg} & \\ \text{CASE} & \textit{gen} \\ \text{NUM} & \textit{sg} \end{bmatrix} \qquad \begin{bmatrix} \textit{d/p-cn} & \\ \text{CASE} & \textit{case} \\ \text{NUM} & \textit{d/p} \end{bmatrix} \qquad \begin{bmatrix} \textit{ljud-cn} & \\ \text{CASE} & \textit{case} \\ \text{NUM} & \textit{d/p} \end{bmatrix}$$

Figure 5: Combined number and case types

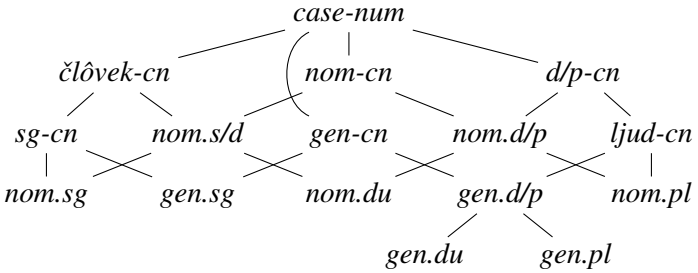

Figure 6: Case-number type hierarchy for Slovene

$$\left[\begin{array}{ll} \textit{noun-stem} & \\ \text{PHON} & \textit{člôvek} \\ \ldots\text{HEAD|CASE-NUM} & \textit{člôvek-cn} \\ \ldots\text{RELS} & \left\langle \textit{člôvek-ljud-rel} \right\rangle \end{array}\right]$$

$$\left[\begin{array}{ll} \textit{noun-stem} & \\ \text{PHON} & \textit{ljud} \\ \ldots\text{HEAD|CASE-NUM} & \textit{ljud-cn} \\ \ldots\text{RELS} & \left\langle \textit{člôvek-ljud-rel} \right\rangle \end{array}\right]$$

Figure 7: Lexical entries for *člôvek* and *ljud*

Another WP approach to modelling this would be to use the notion of a 'stem space' (Pirrelli & Battista, 2000; Bonami & Boyé, 2003). Under such an analysis, we divide the the paradigm into 'spaces' of cells, where each space uses the same stem. The underspecified types which we propose directly correspond to such spaces. However, by organizing these types in a hierarchy, we can efficiently refer to types at varying levels of granularity. In the present case, the types for *člôvek* and *ljud* are not relevant for *grád*, and vice versa; furthermore, none of these types are relevant for *mést*. For each of these nouns, we do not want to redundantly specify the same stem for multiple spaces. For a more complex paradigm, such as Italian verbal conjugation, as discussed by Montermini & Bonami (2013), this is a serious concern, as the number of spaces increases dramatically with the irregularity of the lexemes considered. By using a type hierarchy, we can simultaneously analyse a paradigm with varying numbers of stem spaces, thereby reducing redundancy in the lexicon: each lexical entry uses types at the relevant level of granularity.

In figure 6, we give a partial type hierarchy, with only nominative and genitive cases, which are sufficient to demonstrate the split in the dual for *člôvek* and *ljud*. The analysis follows similarly for the other cases.

So that we can still refer to case and number individually (which is important to get the correct semantics), each of these types has features for case and number, with examples given in figure 5. To distinguish types in the combined hierarchy from those in the separate hierarchies, we write *-cn* in the type name. For types with 'irregular' (non-rectangular) spaces in the paradigm, such as *ljud-cn*, the values for these features will the be the most specific ones that cover all relevant cells – the irregularity of the stem space is handled by the type's position in the hierarchy.

The generalization that the use of *ljud* in the dual matches the suffix syncretism is captured by *gen.d/p* being the only type immediately dominating *gen.du* and *gen.pl*. In fact, it would be impossible to maintain this property if we introduced a single underspecified type for singular and dual. Not only does this allow us to reproduce a 'rule of referral', but this is done without a need for directionality in the rule, which is known to be problematic to determine. Furthermore, the data is captured more directly, in the sense that each form can be described in terms of its parts, without referring to other cells in the paradigm.

In summary, our analysis of Slovene nominal declensions illustrates how all four of the problems discussed in section 2.2 can be overcome. Furthermore, we have seen how the WP notions of 'stem space' and 'rule of referral' can be robustly re-interpreted in a morphemic approach.

4 Georgian Verb Agreement

Georgian verbs present a situation involving multiple affixes which jointly determine the value of multiple features. The full verbal paradigm is notoriously complex, and Hewitt (1995, p.117) lists 11 different slots. We consider the two agreement affixes (one prefix and one suffix), which jointly agree with both subject and

object, as shown in examples (1)-(3). The order of the nouns does not affect the argument structure, and we will not discuss case marking here.

The full agreement paradigm in the present tense is given in table 4, adapted from Harris (1981). Note that reflexives are marked separately in Georgian, so it is not possible for the subject and object to both be first person, or both be second person. For ease of exposition, a few distracting details are suppressed for now, and will be discussed at the end.

(1) მე ვაქებ ექიმს
me *v-akeb* *ekim-s*
I praise.1SG.3SG doctor-DAT
'I praise the doctor'

(2) მე გაქებ შენ
me *g-akeb* *ʃen*
I praise.1SG.2SG you
'I praise you'

(3) მე მაქებს ექიმი
me *m-akeb-s* *ekim-i*
I praise.3SG.1SG doctor-NOM
'the doctor praises me'

	Object					
Subject	1SG	1PL	2SG	2PL	3SG	3PL
1SG	—	—	*g*—∅	*g*—*t*	*v*—∅	*v*—∅
1PL	—	—	*g*—*t*	*g*—*t*	*v*—*t*	*v*—*t*
2SG	*m*—∅	*gv*—∅	—	—	∅—∅	∅—∅
2PL	*m*—*t*	*gv*—*t*	—	—	∅—*t*	∅—*t*
3SG	*m*—*s*	*gv*—*s*	*g*—*s*	*g*—*t*	∅—*s*	∅—*s*
3PL	*m*—*en*	*gv*—*en*	*g*—*en*	*g*—*en*	∅—*en*	∅—*en*

Table 4: Agreement in Georgian present tense verbs

This data has been traditionally analysed by noting certain weak correlations between affixes and agreement features, such as *v-* denoting a first person subject, and *g-* a second person object. Morphemes based on these weak correlations would overgenerate, leading many to invoke some other mechanism to prevent overgeneration. Harris (1981) uses deletion rules, where all morphemes are generated, but, for instance, *v-* is deleted in the presence of *g-*. Several other authors, working in a variety of frameworks, impose some ordering on applying lexical rules or inserting lexical items, so that one rule or item blocks the others (Anderson, 1986; Halle & Marantz, 1993; Carmack, 1997; Stump, 2001). The deletion analysis is implausible phonologically (since Georgian allows long consonant clusters), requires

prediction of possible deleted elements when processing language, and makes it appear a coincidence that a Georgian verb can have at most one agreement suffix and one agreement prefix, since deletion rules would not guarantee this in general. Indeed, Harris neglects to state that the *-en* and *-t* suffixes cannot co-occur (the *-t* should 'delete'), although others do note this. The blocking analyses, however, hugely increase the complexity of the grammar, since we have to consider many alternative derivations in order to interpret a given form, or even determine if it is grammatical. Furthermore, as Blevins (2015) notes, competition between these rules cannot be regulated by a constraint which prioritizes more specific rules (such as 'Pāṇini's Principle' (Stump, 2001)), since we cannot say that a subject feature or an object feature is more specific than the other.

Here we present an alternative analysis, with a sign for each overt affix, and a unary rule for each 'zero', where each structure has both subject and object features. For example, *v-* indicates not only a first person subject, but also a third person object. For almost all the affixes, the paradigm cells form rectangular blocks, meaning that we can specify the subject and object features independently.

The exception is the suffix *-t*, which can appear with any subject except second singular and third plural, and with any object at all – but specifying these independently would lead to overgeneration. Instead, we can analyse this paradigm as having two homophonous *-t* suffixes, each with a rectangular shape. One specifies a first or second person plural subject, and any object. The other specifies a second person plural object, and a first or third person singular subject.[6] Indeed, traditional grammars often refer to these two distinct uses of *-t* separately.

PHON	Subj	Obj
v	*1*	*3*
g	*1/3*	*2*
m	*2/3*	*1sg*
gv	*2/3*	*1pl*
∅	*2/3*	*3*

PHON	Subj	Obj
t	*1/2pl*	*per-num*
t	*1/3sg*	*2pl*
s	*3sg*	*¬2pl*
en	*3pl*	*per-num*
∅	*1/2sg*	*¬2pl*

Table 5: Abbreviated lexical entries (left, prefixes; right, suffixes)

A summary of the agreement features of the full set of affixes and unary rules is given in table 5. The corresponding feature structures are shown in figure 10 for the unary rules, and in figure 11 for the overt affixes (just two are shown, for brevity). The corresponding person-number type hierarchy is given in figure 12. The phrasal types and the resulting phrase structure are shown in figures 8 and 9.

[6]We could also specify the subject as being anything but third plural, which would yield the same paradigm. However, doing so introduces a spurious ambiguity for *g–t*, in the case of a first plural subject and second plural object, since either homophone of *-t* could be used. For this reason, we prefer this more restrictive subject feature.

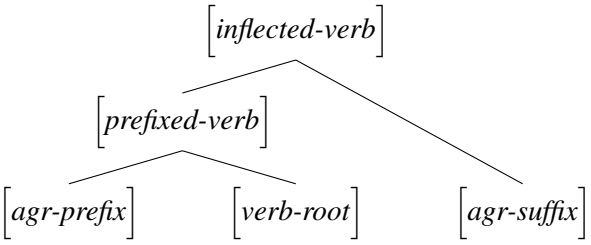

Figure 8: Phrase structure of an inflected verb

$$\begin{bmatrix} \textit{inflected-verb} \\ \text{HEAD-DTR} & \textit{agr-suffix} \\ \text{SPR-DTR} & \textit{prefixed-verb} \end{bmatrix} \qquad \begin{bmatrix} \textit{prefixed-verb} \\ \text{HEAD-DTR} & \textit{agr-prefix} \\ \text{SPR-DTR} & \textit{verb-root} \end{bmatrix}$$

Figure 9: Phrasal types

$$\begin{bmatrix} \textit{inflected-verb} \\ \text{HEAD-DTR} & \begin{bmatrix} \textit{prefixed-verb} \\ \text{...SUBJ} & \left[\text{...PER-NUM} \quad \textit{1/2s}\right] \\ \text{...COMPS} & \left\langle\left[\text{...PER-NUM} \quad \neg\textit{2p}\right]\right\rangle \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \textit{prefixed-verb} \\ \text{HEAD-DTR} & \begin{bmatrix} \textit{verb-root} \\ \text{...SUBJ} & \left[\text{...PER-NUM} \quad \textit{2/3}\right] \\ \text{...COMPS} & \left\langle\left[\text{...PER-NUM} \quad \textit{3}\right]\right\rangle \end{bmatrix} \end{bmatrix}$$

Figure 10: Unary rules

$$\begin{bmatrix} \textit{agr-prefix} \\ \text{PHON} & \textit{gv} \\ \text{...SPR} & \begin{bmatrix} \text{...SUBJ} & \left[\text{...PER-NUM} \quad \textit{2/3}\right] \\ \text{...COMPS} & \left\langle\left[\text{...PER-NUM} \quad \textit{1pl}\right]\right\rangle \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \textit{agr-suffix} \\ \text{PHON} & \textit{s} \\ \text{...SPR} & \begin{bmatrix} \text{...SUBJ} & \left[\text{...PER-NUM} \quad \textit{3sg}\right] \\ \text{...COMPS} & \left\langle\left[\text{...PER-NUM} \quad \neg\textit{2pl}\right]\right\rangle \end{bmatrix} \end{bmatrix}$$

Figure 11: Examples of expanded lexical entries

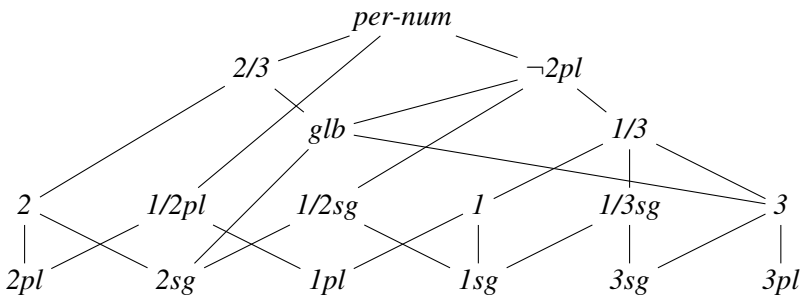

Figure 12: Person-number type hierarchy for Georgian. We introduce the type *glb* (greatest lower bound) so that the hierarchy forms a semilattice, but this type is not used in any well-formed structure.

After unification, this grammar generates all and only the forms in table 4, including leaving the gaps in the table for reflexives, without any spurious ambiguity, and without any additional ordering constraints or competition. This refutes previous claims in the literature that Georgian verb agreement cannot be modelled in a morphemic approach. Gurevich (2006) explicitly argues against the use of morphemes, but we have dealt with each of their objections (cumulative expression, zero morphs, empty morphs, and extended exponence), as explained in section 2.3. Similarly, Blevins (2015) claims that "a dynamic system of contrasts cannot be modelled by a set of static independent associations", but we have shown that this is indeed possible if the associations are with typed feature structures.

Although Blevins sets up a dichotomy between 'associative' and 'discriminative' approaches, the system of morphemes we propose can be viewed in both ways: each morpheme is associated with a feature structure, but the relevant feature values are organized in a type hierarchy so that they discriminate the appropriate meanings. For example, the *v-* prefix can be seen as being associated with a third person object, or conversely as discriminating against a second person object, since it is not unifiable with it. By organizing information using a rich type hierarchy, we can set up associations between morphemes and feature structures in a way that is perfectly compatible with a discriminative view. Indeed, the more underspecified a type is, the more it appears discriminative, rather than associative.

Some complexities of the system are evident in our analysis, such as the need for a *¬2pl* type, but this is in fact motivated twice. Moreover, a similar type is used by Flickinger (2000) to account for present tense verb agreement in English, since zero inflection indicates the subject can be anything except third person singular.

We avoid the need for blocking or competition by the use of more specific values for the person-number feature, and unlike the previously mentioned analyses, the grammaticality and interpretation of a form can be decided without reference to the rest of the paradigm.

In summary, our analysis of Georgian verb agreement illustrates how a type-driven morphemic approach can deal with many-to-many mappings between morphemes and features, contrary to previous claims.

4.1 Further Details

The agreement affixes are inverted between subject and object for a small class of verbs, and for one series of tense-aspect-mood combinations (called 'screeves' in traditional Georgian grammars). These require no change to the above analysis, and can be captured by switching how ARG-ST is linked to COMPS and SUBJ.

The third person subject suffixes are not always *-s* and *-en*, but depend on the tense-aspect-mood of the verb. To model this, we can stipulate several lexical entries as in figure 11, but each with a feature for tense-aspect-mood.

Verbs of motion require an additional agreement marker which effectively fills a separate slot – for example, *mi-v-di-var-t* 'we go', where *mi* is a directional prefix, and *var* indicates a first person subject. To model this, we can give the root *di* a distinct type from other verbs, which the affixes like *var* take as a specifier.

Agreement of intransitive verbs looks like the final column in table 4. To use the same lexical entries for agreement in both intransitives and transitives, we can define a unary rule for affixes whose mother has an empty list in ...SPR…COMPS, and whose daughter's ...SPR…COMPS…PER-NUM must be unifiable with third person. Although 'constructive', this analysis has much in common with the 'abstractive' approach to polyfunctionality described by Ackerman & Bonami (2015). In the general case, we can define a single 'abstract' lexical entry with all necessary information, and a set of unary rules modifying the morpheme for each function.

In the prestige dialect, agreement in ditransitive verbs is with the indirect object, and the direct object must be third person (first and second person objects are marked like reflexives in so-called 'object camouflage'). We can use the same lexical entries for affixes if the indirect object is the first element in the COMPS list. Some speakers have additional markers for third person indirect objects, although Harris notes that their use "is not consistent". The additional indirect object markers can be modelled by imposing an additional constraint on the verb, requiring that it is ditransitive. We neglect other dialectal variations for space reasons.

Third person plural subject agreement (with *-en*) is only triggered by animate nouns. To model this, we can extend the type hierarchy with additional subtypes of *3*, indicating both animacy and number. This does not affect the rest of the hierarchy (only the bottom right corner of figure 12), demonstrating the modular nature of our analysis.

5 Conclusion

In the light of work suggesting 'words' are not well-defined cross-linguistically, we have argued in favour of reformulating HPSG as a unified morphosyntactic theory. We have proposed the Morphemic Principle as a formalization of this approach, and shown how the use of underspecification and unification can avoid various objections to morphemic approaches. We have illustrated our framework by analysing Slovene stem alternations and Georgian verb agreement, giving simpler analyses than competing approaches, but while maintaining the same generalizations.

Acknowledgements

We would like to thank the three anonymous reviewers for their comments, Emily Bender, Dan Flickinger, and Farrell Ackerman for helpful discussion, and Khatuna Gelashvili for providing help with Georgian.

References

Ackerman, Farrell & Olivier Bonami. 2015. Systemic polyfunctionality and morphology-syntax interdependencies. In Andrew Hippisley & Nikolas Gisborne (eds.), *Defaults in morphological theory*, Oxford University Press.

Anderson, Stephen R. 1986. Disjunctive ordering in inflectional morphology. *Natural Language and Linguistic Theory* 4. 1–32.

Anderson, Stephen R. 1992. *A-morphous morphology*. Cambridge University Press.

Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes* (Linguistic Inquiry 22). MIT Press.

Asudeh, Ash, Mary Dalrymple & Ida Toivonen. 2013. Constructions with lexical integrity. *Journal of Language Modelling* 1(1). 1–54.

Beesley, Kenneth R & Lauri Karttunen. 2003. *Finite state morphology* (Studies in Computational Linguistics). CSLI Publications.

Bender, Emily & Jeff Good. 2005. Implementation for discovery: A bipartite lexicon to support morphological and syntactic analysis. In *Proceedings of the Annual Meeting of the Chicago Linguistic Society*, vol. 41 2, 1–16. Chicago Linguistic Society.

Bender, Emily M, Dan Flickinger & Stephan Oepen. 2002. The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the 2002 workshop on grammar engineering and evaluation*, vol. 15, 1–7. Association for Computational Linguistics.

Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics* 42(03). 531–573.

Blevins, James P. 2015. The minimal sign. In *The Cambridge handbook of morphology*, Cambridge University Press.

Bloomfield, Leonard. 1933. *Language*. Henry Holt.

Boas, Franz, Helene Boas Yampolsky & Zellig S Harris. 1947. Kwakiutl grammar with a glossary of the suffixes. *Transactions of the American Philosophical Society* 37(3). 203–377.

Bochner, Harry. 1993. *Simplicity in generative morphology* (Publications in Language Sciences 37). Walter de Gruyter.

Bonami, Olivier & Gilles Boyé. 2003. Supplétion et classes flexionnelles. *Langages* 102–126.

Booij, Geert. 2005. Compounding and derivation. *Morphology and its demarcations* 109–132.

Bresnan, Joan & Sam A Mchombo. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language & Linguistic Theory* 13(2). 181–254.

Carmack, Stanford. 1997. Blocking in Georgian verb morphology. *Language* 73(2). 314–338.

Corbett, Greville G. 2015. Morphosyntactic complexity: A typology of lexical splits. *Language* .

Crysmann, Berthold. 2003. On the efficient implementation of German verb placement in HPSG. In *Proceedings of RANLP 2003*, 112–116.

Crysmann, Berthold. 2005. Syncretism in German: a unified approach to underspecification, indeterminacy, and likeness of case. In *Proceedings of the 12th international conference on Head-Driven Phrase Structure Grammar*, 91–107.

Crysmann, Berthold & Olivier Bonami. 2015. Variable morphotactics in information-based morphology. *Journal of Linguistics* .

Flickinger, Dan. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering* 6(01). 15–28.

Flickinger, Dan, Ann Copestake & Ivan A Sag. 2000. HPSG analysis of English. In *Verbmobil: Foundations of speech-to-speech translation*, 254–263. Springer.

Goldstone, Robert L & Andrew T Hendrickson. 2010. Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science* 1(1). 69–78.

Gurevich, Olga I. 2006. *Constructional morphology: The Georgian version*: University of California, Berkeley dissertation.

Halle, Morris & Alec Marantz. 1993. Distributed Morphology and the pieces of inflection. In Kenneth Hale & Samuel Jay Keyser (eds.), *The view from Building 20*, 111–176. MIT Press.

Harris, Alice C. 1981. *Georgian syntax: A study in Relational Grammar* (Cambridge Studies in Linguistics). Cambridge University Press.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80.

Hewitt, Brian George. 1995. *Georgian: A structural reference grammar*, vol. 2. John Benjamins Publishing.

Kaplan, Ronald M & Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics* 20(3). 331–378.

Karttunen, Lauri. 2003. Computing with realizational morphology. In Alexander Gelbukh (ed.), *Computational linguistics and intelligent text processing* (Lecture Notes in Computer Science 2588), 203–214. Springer.

Kay, Martin. 1987. Nonconcatenative finite-state morphology. In *Proceedings of the 3rd conference of the European chapter of the Association for Computational Linguistics*, 2–10. Association for Computational Linguistics.

Koskenniemi, Kimmo. 1983. *Two-level morphology: A general computational model for word-form recognition and production*: University of Helsinki dissertation.

Krieger, Hans-Ulrich & John Nerbonne. 1993. Feature-based inheritance networks

for computational lexicons. In Ted Briscoe, Valeria de Paiva & Ann Copestake (eds.), *Inheritance, defaults, and the lexicon* (Studies in Natural Language Processing), Cambridge University Press.

Lieber, Rochelle. 2004. *Morphology and lexical semantics* (Cambridge Studies in Linguistics). Cambridge University Press.

Lyons, John. 1968. *An introduction to theoretical linguistics*. Cambridge University Press.

Mair, Victor H. 1990. Implications of the Soviet Dungan script for Chinese language reform. *Sino-Platonic Papers* (18).

Matthews, Peter H. 1991. *Morphology* (Cambridge Textbooks in Linguistics). Cambridge University Press 2nd edn.

Montermini, Fabio & Olivier Bonami. 2013. Stem spaces and predictability in verbal inflection. *Lingue e linguaggio* 12(2). 171–190.

Müller, Stefan. 2015. *German sentence structure: An analysis with special consideration of so-called multiple fronting* (Empirically Oriented Theoretical Morphology and Syntax). Language Science Press.

Oflazer, Kemal. 1994. Two-level description of Turkish morphology. *Literary and linguistic computing* 9(2). 137–148.

Packard, Jerome L. 2000. *The morphology of Chinese*. Cambridge University Press.

Pirrelli, Vito & Marco Battista. 2000. The paradigmatic dimension of stem allomorphy in Italian verb inflection. *Italian Journal of Linguistics* 12. 307–379.

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

Priestly, T M S. 1993. Slovene. In Bernard Comrie & Greville G Corbett (eds.), *The Slavonic languages*, 388–451. Routledge.

Roark, Brian & Richard William Sproat. 2007. *Computational approaches to morphology and syntax*. Oxford University Press.

Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the third international conference on computational linguistics and intelligent text processing* (Lecture Notes in Computer Science 2276), 1–15. Springer.

Sag, Ivan A, Thomas Wasow & Emily M Bender. 2003. *Syntactic theory: A formal introduction*. CSLI Publications 2nd edn.

Sanders, Gerald. 1988. Zero derivation and the overt analogue criterion. *Theoretical Morphology* 155–175.

Spencer, Andrew. 2006. Morphological universals. In Ricardo Mairal & Juana Gil (eds.), *Lingustic universals*, 101–129. Cambridge University Press.

Stump, Gregory T. 2001. *Inflectional morphology: A theory of paradigm structure*. Cambridge University Press.

Sun, Chao-Fen. 2006. *Chinese: A linguistic introduction*. Cambridge University Press.

Tang, Sze-Wing. 2010. *Formal Chinese syntax*. Shanghai Educational Publishing House.