Lacking Integrity: HPSG as a Morphosyntactic Theory

Guy Emerson and **Ann Copestake**
University of Cambridge
{gete2,aac10}@cam.ac.uk

Standard accounts of HPSG assume a distinction between morphology and syntax. However, despite decades of research, no consistent definition of 'word' has emerged (Haspelmath, 2011), implying that no such distinction is justified. In this paper, we argue that morphological paradigms can be straightforwardly handled using existing HPSG machinery, including cases of 'zero' morphemes. In particular, we consider two phenomena which have been used to argue for complex lexical rules, and give simpler analyses within a morphemic approach. The notion of a type hierarchy becomes crucial in overcoming problems associated with morpheme-based Item-and-Arrangement morphological analyses. We conclude that using HPSG as a unified morphosyntactic theory is both feasible and also yields fruitful insights.

1 Lexicalism in HPSG

The Lexical Integrity Principle holds that syntactic rules do not have access to internal parts of words (Bresnan and 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 a notion of 'word', with a corresponding division of labour between lexical and phrasal rules. 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, they take it for granted that we can identify words. The difficulties of defining the term 'word' have been known for some time (Bloomfield, 1933; Lyons, 1968, among others), but more recently, Haspelmath (2011) surveyed a wide variety of proposed definitions and identified problems with each of them, suggesting that wordhood is not well-defined cross-linguistically. Under this view, the distinction between morphology and syntax vanishes, leaving us with a single domain of morphosyntax, and no obvious way to define lexemes. Moreover, this pushes us towards a morphemic Item-and-Arrangement view of morphological phenomena, rather than Item-and-Process or Word-and-Paradigm views (see Stump (2001) for further comparisons).

If we accept Haspelmath's conclusion, then we have no other choice but to reformulate HPSG in terms of morphemes, rather than lexemes. In this paper, we argue not only that this is possible, but that the use of type hierarchies makes HPSG particularly appealing as a morphosyntactic theory, as it can sidestep many problems often attributed to morphemic approaches.

Furthermore, this can be done without fundamental changes to HPSG architecture. 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 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. We must only state instead that lexical entries correspond to morphemes, not words. Just as words and phrases are subsumed under the notion of 'sign', morphemes too should be considered signs. To formalize this idea, we propose the following principle:

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

This principle 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. In the following sections, we will show how lexical rules can be replaced by lexical entries for morphemes. In particular, we produce simple analyses for Slovene stem alternations and Georgian verb agreement, even though at first sight they might appear challenging to a morphemic approach.

2 Morphological Paradigms and Zero Morphemes

Inflectional paradigms can often be represented in terms of a root and a number of affixes, falling into discrete position classes, or slots.[1] In a morphosyntactic version of HPSG, each such morpheme should be assigned a feature structure. To model the affixation, we must first decide whether the root or the affixes should act as heads. If the root should act as head, then 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. If the affix should act as head, then we can avoid introducing an MCOMPS list, and have the root appear in the COMPS list of the affix. In either case, by introducing suitable re-entrancies, for instance between the HEAD features of the root and the affix, the feature structure for the phrase ('word') should look like the lexical structure given in a lexemic version of HPSG. If there are no other rules to license an affix, then it will act as a bound morpheme. Just as syntactic rules can control what information within a phrase is available later in a derivation, this affixation process could leave the internal structure unavailable, in some sense mimicking lexical integrity using syntactic principles.

In the case of zero morphemes, which we can identify via the overt elements competing for the same slot (Sanders, 1988), we do not want to postulate signs with no phonological material. Instead, for each zero morpheme, we can introduce a unary syntactic rule, which removes an element from the MCOMPS list, and unifies the rest of the structure with the features of the purported zero morpheme.

Of course, there is still the question of what to do if we *cannot* decompose morphology with affixes. However, Lieber (1992, p.165ff.) argues that autosegmental phonology and prosodic morphology always enable us to reduce apparently nonconcatenative inflection to affixation. This would guarantee us being able to apply the above principles, leaving only the question of whether such an approach is feasible. In other words, do we lose explanatory strength by abandoning lexemes? In the following sections, we show that this approach can in fact capture various challenging phenomena better than alternatives.

We focus on inflection in this paper, but we note that a morpheme-driven 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 approach.

In the following sections we consider two case studies which involve complex interactions between morphemes. In both cases, we introduce a single type hierarchy for multiple featural dimensions, a technique which has been successfully used to analyse other languages, for example by Flickinger (2000) for English, and by Crysmann (2005) for German.[2]

3 Slovene stem alternations

Here we consider a situation where the choice of a stem is sensitive to number and case features. Slovene nouns inflect for three numbers (singular, dual, plural) and six cases, with a slot for one suffix. Some nouns use the same stem for all numbers and cases, while others use one stem for the singular and another for the dual and plural. However, the pair of stems *človek* and *ljud*, meaning 'man/men', are split in the dual, as shown in table 1, from Priestly (1993). Furthermore, the cases where the plural stem *ljud* is used for the dual are precisely those which display sycretism in the affixes, suggesting a deeper generalization is to be found.

Corbett (2015) analyses this at the lexemic level, within the framework of Network Morphology, introducing 'generalised referral' rules that stipulate that the forms for the genitive and locative dual should be identical to the plural forms. Under such an analysis, 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 ungrammatical forms are directly ruled out by unification failure. We define morpheme signs for the affixes, with underspecified case-number types to match their syncretism. For a noun with a single stem, we define one morpheme for the stem, with unspecified case and number. For a noun with two stems, we define an abstract structure of type *synsem*, that includes all features shared between the two stems – but as there is no PHON feature, it is not a sign. We define one morpheme for each stem, inheriting from the same *synsem* structure, but with a PHON feature, and with an appropriate value for the case-number feature.

[1] As noted by Crysmann and Bonami (forthcoming), morpheme positions can vary. While we do not deal with morphotactics in detail here, we note that variable morpheme order can in principle be dealt with in the same way as variable (word) order in syntax.

[2] Analyses involving 'spaces' in a paradigm (Montermini and Bonami, 2013, i.a.) can also be restated using type hierarchies.

In figure 1, we give a simplified type hierarchy, with only nominative and genitive cases, which are sufficient to demonstrate the issue. Most nouns with two stems use the SG and D/P types, while *člôvek* and *ljud* use specific types defined for them. The generalization that the dual/plural stem matches the affix 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 is the data more directly captured, but we maintain the generalization about the syncretism within the type hierarchy.

	SINGULAR	DUAL	PLURAL
NOM	*člôvek*	*človék-a*	*ljud-jé*
ACC	*človék-a*	*človék-a*	*ljud-í*
GEN	*človék-a*	*ljud-í*	*ljud-í*
DAT	*človék-u*	*človék-oma*	*ljud-ém*
INS	*človék-om*	*človék-oma*	*ljud-mí*
LOC	*človék-u*	*ljud-éh*	*ljud-éh*

Table 1: Declension of Slovene *člôvek / ljud*

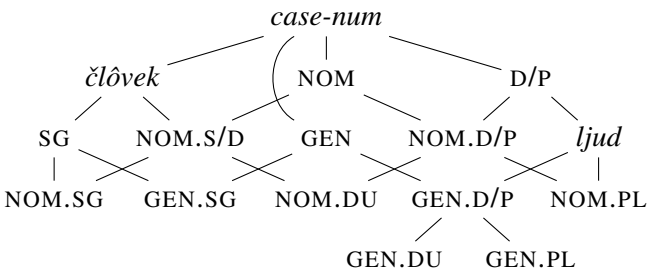

Figure 1: Case-number type hierarchy for Slovene

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. A schema for Georgian agreement in the present tense is given in table 2, following Harris (1981).[3] 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. Gurevich (2006) explicitly argues against a morphemic approach, but their objections (cumulative expression, zero morphs, empty morphs, and extended exponence) are dealt with by the approach followed here.

	Object					
Subject	1SG	2PL	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 2: 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 some to cricticize a direct morphemic approach, and instead 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, propose imposing some ordering on applying lexical rules or inserting lexical items, so that one rule or item blocks the others (Anderson, 1986; Halle and Marantz, 1993; Carmack, 1997; Stump, 2001). The deletion analysis is implausible phonologically, 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

[3]Depending on the tense-aspect-mood of the verb, the third-person singular subject forms can display suffixes other than *-s*, but we stick to present tense *-s* for ease of exposition. Furthermore, some speakers have additional markers for third person indirect objects, but Harris notes that the use of such markers "is not consistent", so for space restrictions we neglect them here.

the grammar, since we have to consider a large number of alternative derivations in order to determine if a given form is grammatical.

Here we present an alternative analysis, which assigns a feature structure to each affix (and to each unary rule for zero morphemes), as shown in figure 2. We write SUBJ and OBJ as shorthand for the paths to the person-number feature of the subject and object of the verb, respectively, and in abuse of notation, we write *unary* in the type name to mean that this structure would be unified with the verb when removing an item from its MCOMPS list, as explained in section 2. The corresponding person-number type hierarchy[4] is given in figure 3. After unification, the grammar generates all and only the forms in table 2, including leaving the gaps in the table for reflexives, and without any spurious ambiguity.

Some complexities of the system are evident in the morphemes, such as the need for a ¬2PL type, which is in fact motivated twice. However, we note that a similar type is used by Flickinger (2000) to account for present tense verb agreement in English. The need for blocking is removed by the use of more specific values for the person-number feature, and unlike the previously mentioned analyses, the grammaticality of a form can be decided without reference to the rest of the paradigm.

$$\left[\begin{array}{ll} \textit{agr-prefix} & \\ \text{PHON} & v \\ \text{SYNSEM} & \left[\begin{array}{ll} \text{SUBJ} & 1 \\ \text{OBJ} & 3 \end{array}\right] \end{array}\right] \quad \left[\begin{array}{ll} \textit{agr-suffix} & \\ \text{PHON} & t \\ \text{SYNSEM} & \left[\text{SUBJ} \quad 1/2\text{PL}\right] \end{array}\right]$$

$$\left[\begin{array}{ll} \textit{agr-prefix} & \\ \text{PHON} & g \\ \text{SYNSEM} & \left[\begin{array}{ll} \text{SUBJ} & 1/3 \\ \text{OBJ} & 2 \end{array}\right] \end{array}\right] \quad \left[\begin{array}{ll} \textit{agr-suffix} & \\ \text{PHON} & t \\ \text{SYNSEM} & \left[\begin{array}{ll} \text{SUBJ} & 1/3\text{SG} \\ \text{OBJ} & 2\text{PL} \end{array}\right] \end{array}\right]$$

$$\left[\begin{array}{ll} \textit{agr-prefix} & \\ \text{PHON} & m \\ \text{SYNSEM} & \left[\begin{array}{ll} \text{SUBJ} & 2/3 \\ \text{OBJ} & 1\text{SG} \end{array}\right] \end{array}\right] \quad \left[\begin{array}{ll} \textit{agr-suffix} & \\ \text{PHON} & s \\ \text{SYNSEM} & \left[\begin{array}{ll} \text{SUBJ} & 3\text{SG} \\ \text{OBJ} & \neg 2\text{PL} \end{array}\right] \end{array}\right]$$

$$\left[\begin{array}{ll} \textit{agr-prefix} & \\ \text{PHON} & gv \\ \text{SYNSEM} & \left[\begin{array}{ll} \text{SUBJ} & 2/3 \\ \text{OBJ} & 1\text{PL} \end{array}\right] \end{array}\right] \quad \left[\begin{array}{ll} \textit{agr-suffix} & \\ \text{PHON} & en \\ \text{SYNSEM} & \left[\text{SUBJ} \quad 3\text{PL}\right] \end{array}\right]$$

$$\left[\begin{array}{ll} \textit{agr-prefix-unary} & \\ \text{SYNSEM} & \left[\begin{array}{ll} \text{SUBJ} & 2/3 \\ \text{OBJ} & 3 \end{array}\right] \end{array}\right] \quad \left[\begin{array}{ll} \textit{agr-suffix-unary} & \\ \text{SYNSEM} & \left[\begin{array}{ll} \text{SUBJ} & 1/2\text{SG} \\ \text{OBJ} & \neg 2\text{PL} \end{array}\right] \end{array}\right]$$

Figure 2: Morphemes and rules for Georgian verb agreement

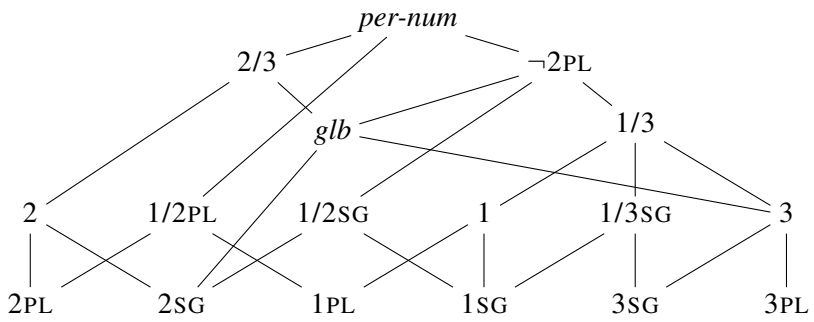

Figure 3: Person-number type hierarchy for Georgian

[4]We introduce one *glb* type so the hierarchy forms a semilattice, but this type is not used in any well-formed structure.

5 Theoretical Guarantees

Suppose we have a paradigm, defined for a lexeme and a set of features, with each feature taking values from a finite set. For any set of feature values, the paradigm provides the acceptable wordform. We can always construct morphemes and a type hierarchy to reproduce this paradigm as long as:

1. Each (overt) morpheme is associated with a unique slot.
2. We can combine all relevant features into a single hierarchy.

The first condition is empirical, but argued for by Lieber (1992). The second is methodological – in some cases it may seem perverse for certain features to be tied together. Indeed, to model Georgian verb agreement, we kept subject and object features separate, to allow re-entrancies with the nominal arguments.

To build a hierarchy, we first define an atomic type for each set of atomic feature values. For each morpheme, we find all atomic types where the wordform includes the morpheme. From the second condition, we are free to define a type which subsumes precisely these atomic types. The lexical entry for each morpheme will have an agreement feature whose value is this underspecified type. After adding greatest lower bound types, we have a well-formed type hierarchy.

If we take a sequence of morphemes and unify their values for the agreement feature, the resulting type precisely subsumes the atomic types where the wordform is composed of those morphemes. By the first condition, the order of the morphemes is the order of the slots, so the only such wordform is those morphemes in sequence. Hence, these morphemes with this hierarchy exactly reproduces the paradigm.

If either condition does not hold, the process may fail. For example (violating the first condition), if we have two morphemes *a* and *b*, such that the wordforms *ab* and *ba* take different values for some feature, then this approach will fail, since unification is commutative. As another example, suppose we have a pair of features each taking two possible values, so there are four atomic types. Suppose further that morpheme *a* is used for three of these types, but not the fourth. If (violating the second condition), we don't want to collapse both features into a single hierarchy, then no single feature structure precisely subsumes those three types.

In fact, with both types of failure, we can always postulate homonymy to make the analysis work. (In the extreme case, we can simply define one morpheme for each cell in the paradigm, although this does not generalize anything.) Limited homonymy is a common feature of natural language, but we conjecture that excessive cases of homonymy would never be required to describe any natural language.

6 Conclusion

In the light of work suggesting 'words' are not well-defined cross-linguistically, we have argued for the need to use HPSG as a unified morphosyntactic theory. We have proposed the Morphemic Principle as a formalization of this approach, and shown how inflectional paradigms can be reproduced using morphemic signs. For the cases of Slovene stem alternations and Georgian verb agreement, alleged problems with morphemes were overcome using type hierarchies, giving simpler analyses than competing approaches.

References

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

Asudeh, A., M. Dalrymple, and I. Toivonen (2013). Constructions with lexical integrity. *Journal of Language Modelling 1*(1), 1–54.

Bloomfield, L. (1933). *Language*. Henry Holt.

Booij, G. (2005). Compounding and derivation. *Morphology and its demarcations*, 109–132.

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

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

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

Crysmann, B. (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*, pp. 91–107.

Crysmann, B. and O. Bonami (forthcoming). Variable morphotactics in information-based morphology. *Journal of Linguistics*.

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

Gurevich, O. I. (2006). *Constructional Morphology: The Georgian Version*. Ph. D. thesis, University of California, Berkeley.

Halle, M. and A. Marantz (1993). Distributed Morphology and the pieces of inflection. In K. Hale and S. J. Keyser (Eds.), *The view from Building 20*, pp. 111–176. MIT Press.

Harris, A. C. (1981). *Georgian Syntax: A Study in Relational Grammar*. Cambridge Studies in Linguistics. Cambridge University Press.

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

Hewitt, B. G. (1995). *Georgian: A structural reference grammar*, Volume 2. John Benjamins Publishing.

Lieber, R. (1992). *Deconstructing morphology: Word formation in syntactic theory*. University of Chicago Press.

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

Lyons, J. (1968). *An Introduction to Theoretical Linguistics*. Cambridge University Press.

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

Priestly, T. M. S. (1993). Slovene. In B. Comrie and G. G. Corbett (Eds.), *The Slavonic languages*, pp. 388–451. Routledge.

Sag, I. A., T. Wasow, and E. M. Bender (2003). *Syntactic Theory: A Formal Introduction* (2 ed.). CSLI Publications.

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

Stump, G. T. (2001). *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge University Press.