Article

Andrew Hardie* and Sophiko Daraselia

A theory for words in Georgian: traditional constructs versus corpus annotation

https://doi.org/10.1515/cllt-2023-0107
Received November 8, 2023; accepted November 19, 2024; published online December 11, 2024

Abstract: Part-of-speech annotation, as an exercise in categorisation, necessitates a category schema, based on some model or theory of the grammar of the language. Such a model may (sometimes, *must*) deviate from traditional approaches for human understanding, as exploration of theoretical issues arising from a Georgian POS schema illustrates. Consistency on classifying by form versus function is problematised by difficult pronoun/demonstrative and adjective/noun distinctions. Adverb subcategorisation illustrates exclusion of semantic/derivational distinctions that traditional approaches readily admit. Variation in plural inflection has implications for how diachronicity is handled, as does "zero case". Postpositions make necessary a specific approach to cliticisation in which enclitic elements are handled as separate tokens bearing their own analysis. Suffixaufnahme provides a case study in inclusion versus exclusion of a rare but current phenomenon. Verb morphology illustrates how simplifying assumptions help favour abstraction of categories over descriptive exhaustiveness. Divergence between the resulting model and traditional characterisations do not invalidate either, but evidence how a model's design is inseparable from its purpose. With regard to these select issues of Georgian grammar, this discussion aims both to demonstrate the overall argument regarding theorisation/schematisation for a specific, practical purpose (POS annotation) and to justify solutions proposed to problems at hand.

Keywords: Georgian; grammar; annotation; part-of-speech; tagset

***Corresponding author: Andrew Hardie**, Lancaster University, Lancaster, UK,
E-mail: a.hardie@lancaster.ac.uk
Sophiko Daraselia, University of Leeds, Leeds, UK, E-mail: s.daraselia@leeds.ac.uk

1 Introduction

Morphosyntactic or part-of-speech (POS) tagging is a fundamental form of corpus annotation. Schemata for POS tagging, or *tagsets*, necessarily express some conceptual structure of relevant grammatical categories and features in the target language. Developing a tagset thus raises numerous issues regarding how best to characterise that language's grammar for the purpose of POS tagging. The model of a language's grammar that best serves the needs of a POS tagset (henceforth: *tagset-focused grammar model*, TFGM) may diverge from models utilised in traditional scholarship.

We aim to exemplify these issues, their complexities and solutions, with reference to a new tagset for Georgian, KATAG. We explore and justify how the solutions arrived at build on descriptive or theoretical characterisations which deviate from constructs of traditional scholarship on Georgian grammar. Thus our focus is a defence of the model of word classification that the tagset expresses (for the actual definition of KATAG see Daraselia and Hardie forthcoming). The resulting "theory" of word grammar in Georgian is not posited as definitive; rather, it is *a* suitable theory for a given purpose, and complements others.

In Section 2, we review prior work on Georgian tagging, illustrating the distinct purpose of KATAG versus earlier schemata. The main section (Section 3) considers a series of major conceptual issues in POS tagset definition. While the example at hand is always KATAG's treatment of Georgian, the conceptual problems are relevant across languages. We begin with underlying principles of schema design, most notably form versus function as target of analysis (Section 3.1); then we consider the core question of what a TFGM should *do* and how this differs from models for other purposes. Remaining issues are more Georgian-specific: clitics (Section 3.3), nominal case (Section 3.4), and massive morphological complexity, for which Georgian verbs are a case study *nonpareil* (Section 3.5).

2 Georgian grammar and automated annotation: background to KATAG

The Georgian grammatical tradition originates in the seventeenth century, when Catholic missionaries founded schools in Georgia teaching Latin and Greek. Consequently, Georgian grammars pre-1800 were strongly influenced by classical Greek and Latin grammars, e.g. Dionysius Thrax's *Tékhnē grammatikē* 'Art of Grammar', and by imported Greek terminology (Karosanidze 2017). In the nineteenth century, Georgian grammars fell instead under the influence of Russian

linguistics (Iluridze 2006). The twentieth century saw the flowering of theoretical and comparative analysis of Georgian's structures. A. Shanidze's (1953) seminal work profoundly influences grammars to this day, whether in emulation or reaction. Indeed, our source for comments on Georgian grammar where not otherwise specified is Shanidze (1980, a revision edited by M. Shanidze of her father's 1953 grammar), alongside Gogolashvili et al. (2011) and D. Melikishvili (2008a, 2014).[1]

Multiple annotation systems for POS alongside other features have been developed for Georgian. Probably the best known is Meurer's (2007) parser based on Lexical Functional Grammar (Kaplan and Bresnan 1982), used to tag the Georgian National Corpus.[2] Each token analysis is a collection of tags for specific features: major POS and morphosyntactic features like agreement, but also lexical and syntactic features, e.g. verbs' transitivity category and argument cases. Beridze et al. (2015) present a morphological analyser which they apply to Georgian dialect corpora. Lobzhanidze's (2013, 2022) similar morphological analyser, like Meurer's system, tags POS alongside features such as verb paradigm and noun animacy.

These systems' common feature is that they do not isolate morphosyntax (major POS plus inflection) from derivation, lexical features, or syntax. Given Georgian's complex morphology, this approach has many virtues. However, it has disadvantages as a substitute for straightforward POS tagging, which *categorises* word tokens rather than applying collections of feature-tags that imply categories *per feature*, not per token. Arguably, neglecting methods which rely on unitary categorisation (such as POS tag queries, or distribution analysis of tags) is suboptimal. Moreover, tagged text with feature collections per token may require specialised search interfaces (e.g. that used for Lobzhanidze's system[3]), whereas querying unitary tagging is a standard concordancer function.

Other work has developed schemata for one-analysis-per-token tagging in Georgian by extending the Slavic-oriented MULTEXT-East standard (Daraselia 2015; Lobzhanidze 2021) for tagsets of this type. However, our dissatisfaction with Daraselia's (2015) MULTEXT-East-based tagset – due to shortcomings detailed by Daraselia (2019, 2024) and Daraselia and Sharoff (2014) – inspired the creation of KATAG, a novel schema catering to (a) the structure of Georgian (rather than multilingual frameworks) and (b) anticipated needs of users of tagged text.

KATAG tags are hierarchical-decomposable, with component mnemonics for major POS, then subclassification, and then inflectional features. For example, *me* 'I'[4]

1 Quotations of works in Georgian are our own translation.

2 http://gnc.gov.ge.

3 http://iliauni.edu.ge/ge/iliauni/institutebi-451/lingvistur-kvlevata-centri-467/qartuli-jesturi-enis-korpusi.

4 Throughout, Georgian is transliterated as per Bolkvadze et al. (2019) except for using apostrophes for glottalised consonants.

receives tag *PP1SN*: a *pronoun*, a *personal* pronoun, *first* person, *singular*, and *nominative-absolutive* case. KATAG's development was informed by corpus frequency evidence from KaWaC (Daraselia and Sharoff 2014). KATAG is tagger-agnostic; Daraselia (2024) describes implementation in Schmid's (1994) TreeTagger; work on specialised software continues. Independent of such matters, however, are the theoretical problems we confronted in devising KATAG, to which we now turn.

3 Issues

3.1 Form, function, and consistency

Consistent tagset design is desirable for the sake of end-users, who must ultimately learn the tags. But consistency may conflict with other design requirements, such as the need to accurately model the language's grammar. The consideration of annotating for *form* versus annotating for *function* epitomises this tension.

In any language with inflection, a POS tagset should classify words according to the inflectional categories they exhibit. Yet for a number of reasons (the expected polysemy of linguistic forms, syncretism across paradigms, etc.) a single morphological form may correspond to multiple (morpho-)syntactic functions. In descriptive or theoretical grammar, this phenomenon needs merely to be thoroughly explained where it occurs. But for POS tagging, that will not do: annotation requires a finite set of categories and criteria for whether any word is in or out of a given category. On one hand, disambiguating forms with multiple grammatical functions is part of what POS tagging is *for* in the first place (cf. the classic English example, noun/verb ambiguity). On the other, distinct categories for variant uses of one and the same form may be impossible to apply without reference to syntactic information (e.g. English's present participle versus gerund distinction). Such multifunctionality must be distinguished from categorical *ambiguity*, the situation that pertains where it has been resolved that two categories are needed, but some forms may represent either category depending on context (Cloeren 1999: 47–48).

Simple homonymy aside, the easiest cases to decide are those where a formal distinction mapping to a functional distinction exists, but is absent for some lemmata in the class in question. For instance, some Georgian *single thema* verbs[5] exhibit no distinction in form between the nominative verbal noun and the third person

5 So-called because they exhibit the same stem (*thema*) across paradigms, whereas other verbs have distinct stems (Shanidze 1980: 387–390).

singular aorist, e.g. *dats'era*, either 'to write' (verbal noun[6]) or 's/he wrote it' (finite verb). However, most verb lemmata do exhibit a formal distinction corresponding to this functional distinction, implying a true difference in morphosyntactic category. Therefore the two forms should be tagged differently. Consistency then dictates applying that judgement also to the lemmata lacking the formal distinction. For verbs like *dats'era*, then, we *tag for function, not for form*. Consequently, annotation of *dats'era* and similarly conjugated verbs is ambiguous when considered out of context. Even English exhibits category definition issues of this kind (e.g. weak verb past tense versus past participle), but the exceptional complexity of Georgian morphology and morphotactics foregrounds such problems.

More difficult cases involve multifunctional forms where it must be determined whether or not the functional distinction should motivate distinct categories, and there *is no* comparable paradigm in which that distinction is marked morphologically. For instance, languages with an explicit copula 'to be' often use the same verb as an auxiliary, without formal distinction. Should that verb receive a different POS tag when it is an auxiliary? Consistent *tagging for function* would require that it should. A Georgian example is personal pronouns. Georgian lacks dedicated third person pronouns; demonstratives are used instead, so e.g. singular *is* translates 'that' but also 'he/she/it'. This is not unusual crosslinguistically; demonstratives so used often grammaticalise into personal pronouns (Diessel 1999), as in Germanic and Romance. To ask whether a Georgian grammar should approach these items as demonstratives or as personal pronouns is over-simplistic. A descriptive grammar can have it both ways; Shanidze (1980: 41–43) includes *is* and related forms in his account of personal pronouns, but notes first and second person pronouns as "genuine" pronouns and third person pronouns as simply a function of certain demonstratives. But this approach, tailored to support human understanding, is problematic for a POS tagset, where caveating whole categories as non-"genuine" is not possible. A Georgian tagset can have *either* a single group of tags for demonstratives (differentiated by case etc.), *or* two groups, one labelling *is* etc. demonstratives and one labelling them pronouns.[7] There is no halfway position.

These two possibilities map directly to tagging for form (one tag per demonstrative) and tagging for function (two tags for two functions). Consistency would favour tagging for function, as per ambiguous forms like *dats'era*. We can imagine end-users wishing to query tagged text for only pronominal, or only demonstrative, uses of *is* etc. A reason *not* to be consistent here is the practical consequences: adding

6 Georgian verbal nouns are often called *masdar* (< Arabic *maṣdar* 'source, gerund'), a term we avoid for clarity. Translating them as infinitives is conventional.

7 The same applies *mutatis mutandis* to other demonstratives which Shanidze (1980: 609, 616ff) additionally classes as articles and relative particles.

this distinction will impose ambiguity onto every token of *is* etc. Of course, we cannot wish away ambiguities; sometimes a form's category being underspecified is just a fact which taggers must deal with. Yet *adding* ambiguity to a set of frequent forms inevitably degrades performance (since disambiguation is never perfect). Informally: sometimes we must live with ambiguity, but we shouldn't inflict it on ourselves if we can avoid it.

Beyond practicalities, tagging according to function may imply adoption of a specific theoretical framework. Consider again copula/auxiliary status of 'to be' verbs. English grammars commonly distinguish copula and auxiliary *be*, but some scholars (e.g. Payne 2011: 266–268) argue that copula *be* is indistinguishable from an auxiliary. The same is argued for Arabic. Though *kāna* 'to be' is commonly described as able to be either copula or auxiliary, some (e.g. Holes 2004: 233) analyse it as a pure grammatical marker of anteriority, even in "copula" contexts. Assigning distinct POS tags to the 'to be' verb's two uses in English or Arabic (tagging for function) thus implies the theoretical stance that these verbs' nature *is not* as Payne or Holes argue. The aforementioned Georgian third person "pronouns" are a parallel case. By contrast, tagging for *form* in these cases does not imply that the distinction *does not* exist, since any category may exhibit functional polysemy. The (explicit or implicit) implication is rather that the demonstrative/pronoun distinction (or copula/auxiliary, etc.) is within some other domain than POS tagging (e.g. syntactic or discourse parsing), since any disambiguation of the functions must take place at one of those levels.

Given all this, need we *always* avoid tagging for function in such cases, despite the inconsistency with applying that principle elsewhere? We are loath to make that argument. A carefully structured tagset can allow functional distinctions to be neutralised in practice. If third person pronouns and demonstratives are treated as pronoun subcategories (e.g. P3P and P3D within P3, alongside P1/P2) it is easy to search for both using a wildcard (*P3**) or regular expression (*P3[PD].**). Users who want the distinction can access it; users who do not can ignore it; thereby, theoretical decisions in the tagset design are not *imposed* upon users, only *made available*. Moreover, the case for including a particular non-formalised distinction is stronger if that distinction is established in general understanding of the language, such that tagset users are likely to expect it to be present. There is then no general answer here: tagset designers must evaluate advantages and disadvantages on a case-by-case basis before deciding on formal or functional category definitions. Consistency is not the only relevant consideration.

Ultimately, we opted *not* to include distinct pronoun/demonstrative categories for *is* etc. The literature treats the distinction as marginal (not "genuine" for Shanidze; for Gogolashvili et al. 2011: 169–173 "'third person' is 'non-person'"); clearly it is not so well-established that its omission will confuse. Therefore, the hard-to-resolve

ambiguity it would cause weighs more heavily. These considerations dictate defining only one category for *is* etc. But the issue needs to be fully and publicly documented, or else users may not grasp that what they know as personal pronouns are tagged as demonstratives.

The most difficult of these problems are when a functional distinction without formal marking cuts across major categories. Consider Georgian's noun/adjective distinction. Adjectives typically premodify a head noun, but *can* postmodify, possibly with other elements intervening. They may also function predicatively with a copula, or be used independently as nominal heads. Modifying adjectives inflect for number and case of the head. All this is not unusual crosslinguistically. The question we face is *whether Georgian adjectives are a distinct POS from nouns at all*. In a grammar for human readers, this question is not crucial. Nouns and adjectives can be distinguished by prototypical semantics (entities versus properties), and readers instructed on the shared formal properties. But in a POS tagset, defining two categories with indistinguishable formal behaviour on semantic grounds is unsatisfactory. Given, say, a newly coined word (so a tagger can have no built-in information on it) that is marked by ergative *-ma*, out of context there is no way to know if it is noun or adjective, since both can take *-ma*. But there is no way to know *in* context either, since syntactic positions diagnostic of ergative nouns are equally open to ergative adjectives. (*Non*-novel forms can be classified as adjective or noun by human input, but again only via semantics.) So should there be a single noun/adjective category?

This form-versus-function issue resists simple solution. In the end, two points resolved us to retain the adjective category. First, the Georgian grammatical tradition is quite clear that it exists. Furthermore, there is a slight formal distinction. *Postmodifying* adjectives inflect for case and number like nouns. However, *premodifying* adjectives exhibit only a subset of these inflections (plural is not marked, some case suffixes are reduced); are unaffected by suffixaufnahme (see Section 3.4); and, if their root is vowel-final, appear with no suffix at all, regardless of the head's inflection. These factors justify the noun-adjective distinction. But for this to work in practice, noun/adjective ambiguity cannot be allowed, as it would be all but unresolvable. Therefore, a tagging lexicon must never class any form as ambiguous between noun and adjective; only unknown forms may be ambiguous. Hence the tagset cannot allow for noun-to-adjective/adjective-to-noun category conversion ("zero derivation"); an adjective very commonly used as a nominal head always must be tagged as adjective.

The same applies to other categories with nominal inflection: pronouns, participles, verbal nouns and denominal/deadjectival adverbs. In all but the last case, we arrived at the same conclusion as for adjectives, and tag for function not form; we return to adverbs below. As consideration of form versus function moves to broader distinctions, it becomes enmeshed with the wider issue of the model of the grammar that the tagset expresses.

3.2 The model of the grammar

POS tagging is an act of categorisation; as such it requires a structured schema of the phenomena under study. That schema should be motivated, that is, founded in a coherent conceptualisation of the domain. A POS schema's motivation is some model of the grammar of the language being tagged: a TFGM (see Section 1). The statistician's motto "all models are wrong but some are useful" (Box 1979: 202) applies here. A grammar is not equal to the whole of a language; it simplifies and approximates, as any finite model of an infinite domain *must*. In outlining a TFGM, then, what counts is how useful the model is *for a given purpose*. Usefulness for a POS tagset need not correspond to usefulness for other purposes. As we saw, the purpose of existing Georgian grammars is to support human understanding of the structure of Georgian, not POS tagging. In this section, we consider distinct qualities needed by a TFGM for Georgian, focusing on aspects that are more or less alien to other grammars. We seek, in other words, to develop not a theory *of* words but a theory *for* words: a conceptual structure suitable for implementation in token-level annotation.

This issue cuts across all others in tagset definition. It emerged at multiple points in the foregoing discussion of form versus function: in re demonstrative versus third person pronoun, copula versus auxiliary, and adjective versus noun. In each case, we noted reasons why a category asserted in existing grammars might be best omitted from a TFGM, and discussed the rationale for our decision in each case. We also observed that a TFGM may need to differ from other models because in POS tagging, in principle every token must be able to receive a single analysis, whereas some ambiguity is acceptable or even welcome in accounts for human readership.

In what other ways must a TFGM differ from other models?

3.2.1 The unit of the word

Word tokens are the unit of analysis in POS tagging. In Georgian as in many languages, although in most cases it is obvious what is and is not a word, there exist edge cases that are harder to pin down. One important example is cliticisation. Clitics possess some, but not all, qualities of a word, having the function of a complete word while lacking phonetic independence. For POS analysis, that in-between status is problematic: any element, clitic or otherwise, either is a token and receives an analysis, or is not and does not. How this is resolved, we address in Section 3.3. For now, the point is that a TFGM must draw firm lines on the status of clitics, when for other purposes we may tolerate fuzziness. Similar lines must be drawn for other elements with borderline word status: formulae, units of measurement, punctuation, abbreviations, and so on must be represented in the category system so that they can

receive an analysis. These items, labelled *residual* in tagset schemata, are of minimal interest in traditional models of grammar, appropriately. But in a TFGM, nothing can be left uncategorisable.

3.2.2 Semantic categorisation

Grammatical models for all languages, Georgian included, often subcategorise POS by semantics, e.g. for nouns concrete/abstract or animate/inanimate. In some cases (e.g. English nouns of time), semantic factors affect morphosyntactic behaviour. But otherwise, semantic classification should not form part of a TFGM. Definitionally a semantic distinction that does not affect the morphosyntax in any way is outside the appropriate scope of a schema for morphosyntactic analysis. Moreover, there is no reason *a priori* to expect semantic and morphosyntactic categories to be isomorphic. Incorporating semantic distinctions into a morphosyntactic schema is both conceptually untidy, and likely to produce a profusion of hard-to-implement subcategories.

However, in existing models of Georgian grammar semantic concepts *do* often form the basis of morphosyntactic categories. This was identified as problematic by Chikobava (1928), and is true of Shanidze (1980), for instance, as D. Melikishvili (2014) demonstrates and attempts to rectify. A relevant example is the boundary, and internal subcategorisation, of the adverb class. Shanidze (1980: 587–588) lists eight adverb categories: adverbs of place, time, manner, measure, cause, and purpose, plus interrogative and relative adverbs. Beyond the undesirable category profusion, this delineation makes little sense. Five categories are purely semantic, place/time/manner/measure/cause being prototypical adverbial meanings. The other three are structural. "Adverbs of purpose" are inflected demonstratives, e.g. *amad* 'as this', adverbial case of *am* 'this'. The "interrogative" class contains indeclinable interrogative elements (e.g. *sad* 'where'; *rodis* 'when') plus derived (pro)nominal stems (e.g. *sadaur-* 'from where'; *rodindel-* 'from when'), which take the usual case suffixes. "Relative" adverbs are indeclinable interrogatives marked by *=ts'(a)*, e.g. *sadats'*, *rots'a*, which Shanidze elsewhere (p. 607) describes as subordinating conjunctions. Hewitt (1995: 65–69) adds "adverbs of negation", a category combining predication-negating particles such as *ar* and *nu* with adverbial bases prefixed with these negators, e.g. *arsad* 'nowhere'. Shanidze (1980: 588–594) also describes a cross-cutting category of "derived" adverbs, those formed by suffixation or reduplication (e.g. *t'q'e-t'q'e* 'through forests' < *t'q'e* 'forest'). A final complexity of traditional descriptions is that nominals that can function adverbially are at times described as *being* adverbs; these "adverbs" are described as having limited case inflection consisting of use with postpositions, *not* the nominal suffixes.

For a practical tagset, simplifying assumptions are needed. First, we exclude the notion that a nominal used adverbially converts to adverb as the concern of syntactic parsing, not POS tagging. More radically, we reject all purely semantic distinctions. This drastically streamlines the model, collapsing e.g. Shanidze's first five groups. Removing assumed conversion to adverb deletes "purpose" adverbs, since all are case-inflected nominals; simplifies the "interrogative" category by excluding case-inflected nominals; and removes the need to model adverbs as exhibiting inflection. The "derived adverb" class is excluded as a matter of derivation. This done, adverbs are modelled as three jointly exhaustive sets: adverbs of negation, as per Hewitt's category but excluding actual negators; interrogative adverbs, excluding words so-described that are case-inflected and thus considered pronouns; and general adverbs, i.e. all others. This treatment of adverbs involves considerations beyond just semantically-based categories. But it clearly demonstrates the utility of a grammar model that excludes concerns prominent in grammars for other purposes.[8]

3.2.3 Treatment of derivation versus inflection

The adverbs additionally demonstrate how a TFGM may benefit from hard and fast distinctions on inflection/derivation that would not be appropriate in grammars for other purposes.

3.2.4 Categorical homogeneity of syntactic behaviour

In a TFGM, an important factor in whether two related groups of forms should be "lumped" together or "split" apart is whether the "lumped" category would be homogenous in syntactic behaviour. This is because POS tagging is often either generated by, or used as an input to, processes which rely on syntagmatic combinatory behaviour (e.g. taggers use short-range combinatory patterns to resolve POS ambiguity). This is unreliable if words within a single category have heterogenous syntactic patterning.

Georgian particles exemplify this. Traditionally (e.g. Shanidze 1980: 607–616) the term *nats'ilak'i* 'particle' is used for various elements that do not clearly form a coherent category. Some are independent words, some clitics, some affixes. Some mark verbs, others nominals, others clauses. Most grammarians make many fine-grained subcategory distinctions. Shanidze lists interrogative, relative, indefinite, prohibitive, quotative, response, emphatic and prosodic particles, and moreover gives a full listing of particles some of which are not assigned to any subcategory

8 Incidentally, since written Georgian lacks either capitalisation or distinct syntax to mark proper nouns, the common/proper distinction is likewise purely semantic, and thus omitted from our model.

(e.g. approximative numeral marker *-ode*). Some "particles" (like *-ode*) can be immediately ruled out as particles in any relevant sense, since they are morphological derivations; describing affixes as "particles" gains us nothing. Thus, we discard subcategories without distinct, homogenous syntax. But we also identify three *additional* subcategories as well-motivated by syntactic behaviour: nominal-emphatic, general-emphatic, and modal particles.

The nominal-emphatic particles are three enclitics (*=ts(a)*, *=gha*, *=ve*[9]) with specific distribution: attached to a noun, adjective, pronoun or numeral, after any other clitics. The more numerous general-emphatic particles (e.g. prototypical emphatic *kidets*; *aki* 'as it were') are independent words. They either precede or follow the element in their scope, which need not be nominal. The heterogenous behaviour of the "emphatic particle" category justifies its division.

Modal particles are uninflectable elements with modal-auxiliary function, e.g. *net'av*, which expresses obligation. They do not cliticise and may precede or follow the verb, or be non-adjacent, though immediate pre-verbal position is typical. Shanidze (1980) catalogues these without assigning a category, but the shared behaviour signals that one is needed. Overall these added categories, along with omissions from the usual list, result in a particle category more coherently defined, and thus useful for POS tagging, than traditional accounts.

3.2.5 Stance on diachronicity

Many grammatical phenomena are inexplicable except as brute facts without diachronic context. As Givón (2015: x–xi) puts it, "I have never seen a piece of synchronic data that didn't reek – instantly, to high heaven – of the diachrony that gave it rise". We concur. However, a TFGM may need to pass over, or suppress, diachronic information, for the sake of a usable schema. The continuity of Georgian's written tradition means that features distinctive of Old Georgian occur in older texts that Modern Georgian speakers nevertheless read readily; some are even productive in particular registers. For instance, both Old and Modern Georgian mark plurals in two ways: (a) with *-n* in nominative-absolutive and vocative and *-t(a)* in other cases; (b) with *-eb* across the board. In Old Georgian, the former was productive and the latter marginal, but in Middle Georgian, *-eb* became the more productive. Yet *-n/-t(a)* remains in use (Shanidze 1980: 47). What effect should this point, typically treated in detail in traditional Georgian grammar, have on our model? By the principle of tagging for function, the identity of the feature expressed dictates that both should be treated alike, and the diachronic information ignored. That said, the distinction might prove useful to users' research. For instance, high frequency of a tag for

9 *=gha* and *=ve* are not classified by Shanidze (1980) but clearly share *=ts(a)*'s behaviour.

"old-style" plural nouns might prove indexical of archaicised styles. Neither argument is without weight, but we consider the former more in line with POS tagging's goal of abstracting away from allomorphy. Thus one tag covers e.g. both *k'ats-eb-i* and *k'ats-n-i* 'men'.

3.2.6 Adherence to standards

Finally, one factor which, we argue, should *not* exert major influence on a TFGM is adherence to standards. Numerous crosslinguistic standards exist for POS tagsets, notably EAGLES (Leech and Wilson 1999), MULTEXT-East (Erjavec 2012), and Universal Dependencies (de Marneffe et al. 2014). Such systems provide "menus" of attributes and values, which can be selected from to define the categories of a compliant tagset. A predecessor to KATAG attempted to comply with MULTEXT-East (Daraselia and Sharoff 2014) with suboptimal results. In our view, adherence to such standards is useful for language-engineering interoperability, but less important than making a coherent and descriptively adequate tagset.

3.3 Tokenisation and treatment of clitics

Clitics present a challenge for POS tagging. They have the function of a word, but lack phonetic independence; in writing they may (but need not) appear attached to the host word without space division, hyphenated or unhyphenated. If we wish to tag clitics as per their function as words, a token division must be added to separate clitic and host. Other strategies have been tried; some English tagsets have genitive noun tags for words with *='s*, to avoid treating the enclitic separately. But two arguments support the clitic-separation approach. First, many clitics alternate with full words (e.g. English *not/=n't*). Consistency of tagging requires both be treated alike; only clitic separation makes this possible. Second, some units are written separately but demonstrably cliticised in speech (e.g. cannot carry stress). To analyse such clitics as part of host words, orthographically separate tokens would have to be *joined together*. Few analysts would support that procedure. Thus treating clitics separately requires less modification to naïve, whitespace-based tokenisation. But it introduces difficulty for users: finding e.g. *didn't* in tagged English text means searching for *did n't*, and users have to remember this. We consider this justified by the inconsistency avoided. But we acknowledge the downside.

Georgian is rich in enclitics, most written attached without hyphenation. Attempting to treat clitics as parts of hosts would thus have an additional disadvantage: the number of POS subcategories per lemma drastically increasing, because each clitic tokenised with its host doubles the host's possible inflectional forms

(one form per inflection with the clitic plus one without). Likewise the possibility of clitic *sequences* would necessitate yet more subcategories. Clearly separate treatment of enclitics is essential for a Georgian TFGM.

The categories most often cliticised are postpositions (see Section 3.4), particles (see Section 3.2), and the copula. The copula is a full, albeit irregular, verb. However, its third person singular present form *aris* alternates with enclitic *=a*. This alternation further evidences the need for clitic separation in Georgian. Being identical in grammatical function, *aris* and *=a* should be tagged identically, so two clauses differing only in use of *aris* versus *=a* can receive the same string of POS tags.

3.4 Case, suffixaufnahme, and postpositions

Traditionally (Shanidze 1980: 44–108), seven cases are identified for Georgian, though both enumeration and naming are somewhat problematic: nominative, ergative, dative, genitive, instrumental, adverbial[10] and vocative. The case suffix morphophonology is complex, but beyond our scope here. But discussing Georgian case does require the concept of *alignment*,[11] the comparative treatment of intransitive subject (S), transitive subject (agent-like argument A), and object (patient-like argument P). When arguments are case-marked, the A and P are treated distinctly: both appear in transitive clauses, so ambiguity must be avoided. The S is not subject to this confusion, so may share the marking of either A or P. The two simplest systems, then, have two cases for core arguments: (i) an S+A case (called nominative) and a P case (accusative); (ii) an S+P case (absolutive) and an A case (ergative). These alignments are named respectively nominative-accusative and ergative-absolutive. Other patterns include *split ergative*, where a clause's alignment depends on tense/aspect or argument animacy; and *active*, where the S's treatment depends on the semantic role the verb implies.

Georgian exhibits both ergativity split by tense/aspect and active alignment.[12] The traditional case names do not reflect their roles in the alignments. "Nominative" is both the S+A case and the S+P case, and thus better dubbed *nominative-absolutive*. Meanwhile, the P case is traditionally called "dative" rather than "accusative", as it has additional functions normally associated with dative case (e.g. marking

10 The term *adverbial case* is confusing: other cases can function adverbially. Its main meanings are goal/direction, time (at/up to), purpose, outcome state, and manner; we might dub such a case *allative*.
11 We cannot cite even a fraction of relevant literature here, beyond noting Dixon's (1979) seminal account.
12 Georgian's overall alignment is not settled. I. Melikishvili (2008b) describes it as active/ergative split; Amiridze (2006: 27–29) argues for tense/aspect-based active/accusative split.

recipients) and emerged as a syncretism of a separate accusative and dative (Skopeteas et al. 2012: 148); it is better dubbed *accusative-dative*.

Three case-related phenomena raise problems for our purposes: zero case, postpositions, and suffixaufnahme.

3.4.1 Zero case

The zero or null case is the nominal form without any of the suffixes for the seven cases. Various terms are used: *unmarked root* (Chikobava 1940); *nominative without marker* (Sarjveladze 1984); *unmarked nominative form* (Uturgaidze 1986). It remains moot whether this is truly a distinct case. Marr (1908, 1925) classifies the unmarked form as a zero case, a judgement followed by e.g. Zorell (1930), Shanidze (1934: 304; 1976: 31) and Imnaishvili (1956: 59; 1957: 21). Others disagree, considering it a variant of nominative-absolutive, e.g. Chikobava (1940: 13), Sarjveladze (1984: 357), Uturgaidze (1986: 17), Gogolashvili et al. (2011: 77). This is despite use of this zero-marked form for case functions other than nominative-absolutive (see Section 3.1's discussion of the limited inflection of premodifying adjectives).

In Old Georgian, the zero form was nominative-absolutive (Sarjveladze 1984); today's nominative-absolutive suffixes descend from definiteness markers. As definiteness came to be lost, suffixed and non-suffixed forms came into competition as nominative-absolutive, a competition now won by the marked form.[13] Native-speaker intuition suggests use of zero over the marked form to have non-standard-dialect and colloquial associations. But no consensus exists on whether there is also a functional distinction. A minority of the above-cited sources argue for zero-case functions distinct from nominative-absolutive (Danelia 1998; Imnaishvili 1957). The question is rarely given much attention; the literature emphasises the unmarked form's operation in *Old* Georgian and historical development.

The issue here is the application of a principle established previously (Section 3.2): functionally identical variants, one characteristic of earlier periods, one more modern, should not be treated as distinct categories, but the difference between them abstracted away, as exemplified with plural *-n/-t(a)* versus *-eb*. The zero/nominative-absolutive distinction is similar at first glance. In other cases, including *-n/-t(a)* versus *-eb*, we judged style/register indexicality insufficient reason to maintain a form-based category distinction with no functional difference. But for zero/nominative-absolutive case, we did not apply this principle, despite the slight inconsistency introduced, because as noted, the lack of functional distinction is less firmly established by relevant literature than for *-n/-t(a)* and *-eb*. A potential functional

13 This scenario is proposed by *inter alia* Uturgaidze (1986: 4–23) building on Marr (1925) and Vogt (1947).

distinction makes retaining the category distinction essential *for corpora tagged using this schema to be useful in research on that distinction*. Of course, the distinction may be dropped later in light of results of such research.

If we accept this argument, a question remains on the scope of tagging zero case. For vowel-final roots, the unsuffixed form is the normal nominative-absolutive, e.g. for *q'ru* 'deaf' a nominative-absolutive on the usual pattern *q'ru-i* does not occur, only bare *q'ru*.[14] But forms on the pattern of *q'ru-i* are known to occur in non-standard dialects. So should *q'ru* with nominative-absolutive function receive a tag for zero or nominative-absolutive case? This is partly a matter of the extent to which the grammar models specifically the standard variety. If forms akin to both *q'ru* and *q'rui* might occur in some dialect text to be tagged, the former should receive an analysis of zero case and the latter nominative-absolutive, as do consonant-final roots. On the other hand, if we assume only the standard variety is in scope, classifying *q'ru*-type words as zero case when they are the *only* form with nominative-absolutive function has the perverse consequence that frequency lists will make it seem as if these lemmata are never used in nominative-absolutive. This argues in favour of classifying *q'ru*-type words as nominative-absolutive (when in appropriate functional context). Once again, the goal of fostering research into a zero/nominative-absolutive distinction resolves the conundrum: such research cannot assume that data conforms to any particular dialect. Thus, our TFGM treats vowel-final nominals like *q'ru* consistently as zero case, regardless of the case function that the zero-marked form may have in context. The nominative-absolutive classification is reserved for *q'rui*-like forms, if and when they occur.

3.4.2 Postpositions

Most but not all (approx. 21 out of 36) Georgian postpositions are written attached to the preceding noun, and have thus sometimes been described as case inflections (Shanidze 1980: 73–77 calls postpositions "local cases"). Two features distinguish them from case suffixes and justify treatment as clitics. First, postpositions govern particular cases; e.g. *=tan* 'at' governs accusative-dative, *=tvis* 'for' governs genitive. But actual cases do not themselves govern any case. Second, the syntactically parallel unattached postpositions (e.g. *shesakheb* 'about') evidence the enclitic nature of the attached items.

Therefore we model a postposition-marked nominal as two tokens: the nominal and the postposition. A complication is that postpositions may suppress preceding

14 Thus some authorities (e.g. Shanidze 1980: 60) treat stem-final vowels as nominative suffixes, without other justification, meaning *any* vowel can be a nominative-absolutive suffix. For us, this complicates rather than elucidates matters.

Table 1: Frequency of suffixaufnahme in KaWaC.

Case combination	Frequency
Genitive + accusative-dative	190,695
Genitive + adverbial	75,658
Genitive + ergative	1,037
Genitive + vocative	114

suffixes. For instance, *=shi* 'in/at' governs accusative-dative, whose suffix is *-s*. The sequence *s+sh* is phonotactically impossible; the actual form is just *-shi*, e.g. *sakhl-i* 'house, nom-abs.'; *sakhl-s* 'house, acc-dat.'; but *sakhl-shi* 'in the house' from hypothetical *sakhl-s=shi*; compare *deda-s-tan* 'with mother' with non-absent *-s*. But how do we classify *sakhl* when *=shi* is separated from it: as accusative-dative case, or zero? At first glance, this looks like the same issue raised with regard to premodifying adjectives lacking the case of their nominal head, which was resolved by opting to tag such adjectives as zero case. This is not quite so, however, because in this case a purely phonetic operation drives *sakhl* to have an apparently zero case form. Before a postposition, then, a base-form nominal may receive an analysis other than zero case, so long as it meets the requirements for the suffix elision to apply (i.e. consonant-final); in the case of *sakhl-shi*, this implies the analysis of singular accusative-dative noun for the *sakhl* token. Practically, this can be accomplished without making base forms generally ambiguous, by implementing non-zero-case analyses as allowable solely before postpositions.

Suffixaufnahme or *case stacking* is the marking of a nominal modifier with its head's case affixes (whatever they may be) in addition to its own. In Georgian, this affects genitives. A possessor is marked as genitive but *may* in addition receive the case of the possessum, depending on whether the possessor precedes or follows the possessum, as with adjective case (see Section 3.1). For instance, the normal way to say 'mother's house' in accusative-dative, possessor first, is *ded-is sakhl-s*, with the possessor taking genitive *-is*. But when the possessor postmodifies, suffixaufnahme is expected: *sakhl-s ded-isa-s*. Conversely, premodification with suffixaufnahme (?*ded-isa-s sakhl-s*) and postmodification without (?*sakhl-s ded-is*) are borderline unacceptable if not ungrammatical.

Suffixaufnahme is considered more typical of Old than Modern Georgian. Empirically it remains current enough that POS tagging must account for it (see Table 1). Ergative and vocative suffixaufnahme might be rare enough to ignore, but accusative-dative and adverbial are not.

Thus we model the four combinations in Table 1 as categories parallel to the main seven. An alternative would be to assign double analyses: one for genitive, one for the suffixaufnahme case. However, this undermines the principle that any single token should be able to receive a single analysis. While having (say) genitive+ergative as its own category means that nominals so tagged will not be included in frequencies or search-capture behaviours for either genitive or ergative, to our mind this limitation is insignificant weighed against the benefit of making suffixaufnahme *as a phenomenon* searchable in tagged text.

Some vocative nominals, when used within clause syntax, are additionally marked with the relevant case for that context. This is not suffixaufnahme, since the extra cases do not come from a modified head, but is formally similar. Do similar arguments apply? The phenomenon is much rarer; KaWaC has only 645 examples. Moreover, there is evidence that this is derivation, not inflection. 627 of 645 instances are forms of *mamao* or *dedao*, historically vocatives of *mama* 'father' and *deda* 'mother' but lexicalised as terms for clergy. Asserting a category based on <20 examples in 230 million tokens would be unwise. Therefore the TFGM does not model vocatives with added cases.

3.5 The complex morphology of Georgian verbs

The key issue in modelling Georgian verbs is *exhaustiveness* versus *abstraction*, due to the high number of categories which various affixes may express on any verb. If a TFGM aims to capture these features exhaustively, the result will be a very fine-grained schema with very complex tags. Such a schema would, however, do little to abstract across that diverse array of inflectional categories; the importance of abstraction in POS annotation was discussed earlier. Conversely, abstracting away some detail to get a tractable category schema definitionally makes an analysis less than fully exhaustive.

Traditional Georgian grammars unquestionably favour exhaustiveness (as do systems like Lobzhanidze 2013; Meurer 2007: see Section 2), aiming to capture every feature expressed in verbal morphology and exhaustively catalogue formal variation. Our model instead favours abstraction: establishing categories for only the most important grammatical features and excluding distinctions arising from derivation (to be dealt with, if need be, at the level of morphological annotation). To illustrate this approach, we overview Georgian verbs as traditionally described, before arguing that practical simplifications are needed for a TFGM.

Georgian verbs are marked for voice, aspect, tense, mood, and agreement for person and number with two nominal arguments. But no straightforward association exists between any feature and a single morpheme or "slot" in the verb's

morphological template. The inflection/derivation boundary is blurred. One marker may indicate both a grammatical category and a derivational semantic change, either at the same time, or depending on context. For instance, voice is expressed partly by so-called *version* prefixes, but sometimes these instead code unpredictable semantic shifts; moreover, voice is also marked suffixally. Tense/aspect is coded partly by presence or omission of a *preverb* (prefix expressing direction), partly by other affixes. Agreement is marked by prefix/suffix combinations, where some markers sometimes reflect the first argument and sometimes the second. Often, the morphology encoding these categories varies by transitivity class and unpredictably by verb root.

Traditional grammars (Gogolashvili et al. 2011: 266–634; Shanidze 1980: 163–530) consider three verb features inflectional (person agreement, number agreement, screeve), and six derivational (direction, orientation, aspect, voice, version, contact). A *screeve* (*mts'k'rivi*, lit. 'row') is one of eleven enumerated "tenses", i.e. tense-aspect-mood combinations; each screeve consists of the full set of agreement forms of that tense-aspect-mood. Traditional approaches often treat voice, direction, version etc. as forming paradigms *despite* being derivational, this being one reason for verbs' exception-riddled complexity. But that traditional insight suggests an appropriate way to simplify verb classification. Only agreement and screeve are predictable, productive, and paradigmatic: a finite verb[15] expresses exactly one screeve and one agreement pattern. Other features may be present, absent, or multiply present, with limited predictability. Thus, we model verbs as classifiable by screeve and agreement alone.

The screeves are organised into four *sets*; two sets, Present and Future, are grouped as a *series*; two others are series on their own. Table 2 lists common terms for sets and screeves (translated where need be). Though terminology reflects it only in part, sets and screeves form a simple aspect/tense-mood matrix. Each series is an aspect, respectively imperfective, perfective, and perfect (or, according to some recent work e.g. D. Melikishvili 2014, resultative/evidential), with imperfective doubled as present and future. Then, each set has a (relative) present, (relative) past, and subjunctive, except that the perfective has no relative past. While this analysis appeals for simplicity and symmetry, our model does not employ it, due to its potential unfamiliarity to users. Rather, it lists the screeves using traditional terms (*Aorist Subjunctive* is, exceptionally, our own suggestion) and no further organisation.

Traditional accounts also catalogue how agreement targets differ across screeves (due to aspect affecting alignment). We believe this can be disregarded in a TFGM.

---

15 Non-finite verbs raise no additional problems.

Table 2: Terms for sets and screeves.

Series	Set	Screeve
I	Present	Present
		Imperfect
		Present Subjunctive
	Future	Future
		Conditional
		Future Subjunctive
II	Past	Aorist
		Optative/2nd Subjunctive/Aorist Subjunctive
III	Perfect	Perfect/1st Evidential
		Pluperfect/2nd Evidential
		Perfect Subjunctive/3rd Subjunctive/3rd Evidential

Two sets of agreement markers exist,[16] traditionally described as agreeing with subject and object. Because this actually depends upon the clause's alignment, we suggest the terms *v-agreement* and *m-agreement* for the two sets, based on the first person prefixes. V-agreement is normally but not always subject agreement, m-agreement normally object agreement; Table 3 illustrates both. However, when an agreement pattern logically implies use of two prefixes together, e.g. *v-* + *g-* for first singular subject + second singular object, in fact only one occurs (*g-* in this case). Similarly, *-t* can pre-empt other suffixes. Prefixes mostly express person and suffixes mostly number, but not always. Rather than try to disentangle all this, we focus on distinguishing the functional agreement categories, not the actual affixes.

Three persons and two numbers for two nominals means $(3 \times 2) \times (3 \times 2) = 36$ logical categories (verbs without objects get third person m-agreement, and thus motivate no additional categories). But actual morphology does not exist for all these. The eight combinations of agreement where v- and m-agreement are either both first or both second person cannot occur (this would require, e.g., prefixing both *m-* and *v-* at once, which is impossible). Because *-t* marks plural for both first and second person in both v- and m-agreement the pair *g-*…*-t* can express any of first singular + second plural, first plural + second singular, or first plural + second plural. *-t* further marks third person plural in m-agreement, so that *v-*…*-t* may be first singular + third plural, first plural + third singular, or first plural + third plural. Likewise third plural + second singular and third plural + second plural are indistinct.

16 Verbs are said to be able to agree with three (Shanidze 1980: 171) or even four arguments, counting an applicativised object as third argument. This is inaccurate. A promoted object *takes over* the demoted object's agreement. Verbs agree with at most two nominals at a time.

Table 3: Agreement markers.

Person	v-Agreement		m-Agreement	
	Singular	**Plural**	**Singular**	**Plural**
1	v-	v-…-t	m-	gv-
2	Ø-, h-, s-	Ø-, h-, s-…-t	g-	g-…-t
3	-s, -a, -o	-en, -an, -nen, -n, -es	Ø-, h-, s-	Ø-, h-, s-…t

Functionally, third singular m-agreement may be used for plural entities, depending on their semantic role (e.g. number may be neutralised for non-applicativised patientive objects but not applicativised recipient/beneficiary objects). There are, in sum, 19 patterns, 8 of which represent multiple paradigm positions. Finally, some screeves of some verbs exhibit additional syncretisms, e.g. *gzrdit* 'raise' is marked by *g-…-t* and thus has the aforementioned three-way ambiguity, but the same form is also *third* singular + second plural 's/he raises you (p)'. However, not all verbs, or all screeves of this verb, exhibit this syncretism.

Three major questions arise regarding agreement. First, should a POS schema distinguish subject versus object agreement, or v- versus m-agreement? The former implies that a verb marked by, say, *v-* should be tagged according to what it actually agrees with: a first person *subject* in most cases, a first person *object* in clauses whose alignment so dictates. The latter implies unambiguous tagging of such a verb as having first person v-agreement. This is a form-versus-function question, easily answered: the ambiguity of subject versus object agreement could not be resolved automatically, since identifying subjects and objects[17] belongs to a different annotation level (parsing), plus clauses can *lack* explicit subject/objects. Thus, the TFGM classifies each verb according to the combination of v- and m-agreement it expresses. Two verbs with first singular v-agreement and third person m-agreement receive the same analysis *even if* one has a first singular subject and the other a first singular object. This implies that voice marking that changes a verb's agreement *target* does not affect its agreement *categorisation*. The second question is whether all 36 logical categories should be present in the schema, or just the distinct 19. For consistency with issues considered earlier, we take the latter approach. Even then, the number of categories is high (19 agreements × 11 screeves = 209), though many are rare (only 77 occur at all in a manually-tagged 100,000-token sample). Third, should the neutralisation of third person object number be modelled by the schema? This last is easily answered: a third

17 This assumes theoretical consensus as to what actually *is* the subject and object in various clause types. But such a consensus may not exist for languages with complex alignment.

Table 4: Agreement patterns, exemplified by *-zrd-* 'grow, raise' (with extra example from a derived voice when *-zrd-* shows extra syncretism).

Pattern label	Example	Also covers…
S_1	*vzrdi* 'I raise'	S_1O_3
S_2	*zrdi* 'You (s) raise'	S_2O_3
S_3	*zrdis* 'S/he/it raises'	S_3O_3, $S_3O_3^P$
S_1^P	*vzrdit* 'We raise'	$S_1^PO_3$, $S_1O_3^P$, $S_1^PO_3^P$
S_2^P	*zrdit* 'You (p) raise'	$S_2^PO_3$, $S_2O_3^P$, $S_2^PO_3^P$
S_3^P	*zrdian* 'They raise'	$S_3^PO_3$, $S_3^PO_3^P$
S_1O_2	*gzrdi* 'I raise you (s)'	
$S_1O_2^P$	*gzrdit* 'I raise you (p)'; *gezrdebit* 'I raise myself for you (p)'	$S_1^PO_2$, $S_1^PO_2^P$
S_2O_1	*mzrdi* 'You (s) raise me'	
$S_2O_1^P$	*gvzrdi* 'You (s) raise us'	
S_3O_1	*mzrdis* 'S/he/it raises me'	
$S_3O_1^P$	*gvzrdis* 'S/he/it raises us'	
S_3O_2	*gzrdis* 'S/he/it raises you (s)'	
$S_3O_2^P$	*gzrdit* 'S/he/it raises you (p)'; *gezrdebat* 'S/he/it/they raise/s themself/ves for you (p)'	
$S_2^PO_1$	*mzrdit* 'You (p) raise me'	
$S_2^PO_1^P$	*gvzrdit* 'You (p) raise us'	
$S_3^PO_1$	*mzrdian* 'They raise me'	
$S_3^PO_1^P$	*gvzrdian* 'They raise us'	
$S_3^PO_2$	*gzrdian* 'They raise you (s)'	$S_3^PO_2^P$

person singular verb inflection is still a third singular form, and should be tagged as such, even if the target is plural, since the mismatch is purely functional-semantic.

We found the terms *v-/m-agreement* useful above, but users will not know them. Tag definitions therefore refer to subject/object agreement, but mean v-/m-agreement. Table 4 lists the categories, with labels and examples; for ambiguous forms, one reading's label is given, and others listed separately. (Tags use shorter forms of the labels.)

4 Conclusions

This paper has demonstrated the extensive theoretical difficulties involved in a POS tagset-focused model of grammar. Our deliberations centered on points where our Georgian model diverges from tradition; we:

- exclude derivational morphology
- exclude diachronic factors, e.g. allow no separate categories for archaic inflections
- model Georgian as having demonstratives, but not third person pronouns
- do not model adverbial nouns as converting to adverb
- exclude semantic subcategories of adverbs
- add three particle subcategories not previously identified
- model clitics (postpositions etc.) as separate tokens from the host
- model additional categories for zero case and suffixaufnahme
- limit verbal categories to screeve and agreement
- classify agreement without reference to voice/alignment-driven target variation.

These only illustrate the broader point, and contribution to theory and methodology, that we make. We have demonstrated at length how a POS tagset must embody a theory of the target language's word-level grammar. Developing that theoretical model always involves, or should involve, tagset designers delving into considerations akin to those we discussed here. This is true even if, as often the case, tagsets are published in a manner or venue that does not allow that theoretical work to be disseminated. Discussed or not, the model must exist.

Unless we believe in a Platonic ideal model of grammar out there to be discovered (and for something as complex and intertwined into cognition as language, we do not), the success of a theory cannot be measured except by reference to its purpose. As Section 3.2 noted, models can be useful, but are not the full reality. We must not mistake the map for the territory. A model for morphosyntactic tagging of Georgian (or any language) is not a model for framing human reflection on the structure of Georgian (or any language).

In sum, beyond proposed novel analyses of certain issues in Georgian grammar, this paper's contribution has been to demonstrate not only the crucial role of grammatical theorisation in defining tagging schemata, but also how simplifying assumptions and analyses that run against traditional approaches are nevertheless defensible if appropriate for the purpose at hand.

Research funding: This work was supported by Economic and Social Research Council (http://dx.doi.org/10.13039/501100000269, ES/R008906/1).

References

Amiridze, Nino. 2006. *Reflexivization Strategies in Georgian*. Utrecht: LOT.

Beridze, Marina, Liana Lortkipanidze & David Nadaraia. 2015. Dialect dictionaries in the Georgian dialect corpus. In Martin Aher, Daniel Hole, Emil Jeřábek & Clemens Kupke (eds.), *Logic, language, and computation. TbiLLC 2013*, 82–96. Berlin: Springer.

Bolkvadze, Tinatin, Maka Tetradze & Nino Kelbakiani. 2019. *Guidelines for Latin transliteration of the sound system of the Georgian language*. Tbilisi: State Language Department.

Box, G. E. P. 1979. Robustness in the strategy of scientific model building. In Robert L. Launer & Graham N. Wilkinson (eds.), *Robustness in statistics*, 201–236. New York: Academic Press.

Chikobava, Arnold. 1928. *mart'iv t'inadadebis p'roblema kartulshi: kvemdebare-damat'ebis sak'itxi dzvel kartulshi: masalebi metodologiur imanentizmisatvis* [A problem of the simple sentence in Georgian: Issues of subject-object in Old Georgian; materials for methodological immanentism]. Tbilisi: Tbilisi University Press.

Chikobava, Arnold. 1940. mesame p'iris udzvelesi nishani kartvelur enebshi [Earliest markers of third person in the Kartvelian languages]. *Bulletin of the Institute of Language, History and Material Culture* [enimk'e moambe] 4–5. 13–46.

Cloeren, Jan. 1999. Tagsets. In Hans van Halteren (ed.), *Syntactic wordclass tagging*, 37–54. Dordrecht: Kluwer.

Danelia, Korneli. 1998. *narkvevebi kartuli samtserlobo enis ist'oriidan* [Studies in the history of literary Georgian], vol. 1. Tbilisi: Tbilisi University Press.

Daraselia, Sophiko. 2015. *kartul-evrop'uli t'ip'is akhali sast'savlo leksik'onebis shedgenis sak'itkhebi k'orp'usuli metodologiis gamoq'enebit* [Issues of Compilation of New Georgian-European Learner's Dictionaries Using the Corpus Methodologies]. Tbilisi: Tbilisi University Press.

Daraselia, Sophiko. 2019. *Computational analysis of morphosyntactic categories in Georgian*. Unpublished PhD thesis. University of Leeds.

Daraselia, Sophiko. 2024. Issues in training the TreeTagger for Georgian. *Corpora* 19(3). 317–332.

Daraselia, Sophiko & Andrew Hardie. Forthcoming. The KATAG morphosyntactic annotation schema for Georgian.

Daraselia, Sophiko & Serge Sharoff. 2014. Morphosyntactic specifications for KaWaC, a web corpus for Georgian. In proceedings of *Humanities in the Information Society II, Batumi, Georgia*, 326–329.

Diessel, Holger. 1999. *Demonstratives: Form, function, and grammaticalization*. Amsterdam: John Benjamins.

Dixon, Robert M. W. 1979. Ergativity. *Language* 55(1). 59–138.

Erjavec, Tomaž. 2012. MULTEXT-East: Morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation* 46(1). 131–142.

Givón, Talmy. 2015. *The diachrony of grammar*. Amsterdam: John Benjamins.

Gogolashvili, Giorgi, Avtandil Arabuli, Murman Sukhishvili, Mariam Manjgaladze, Nino Tchumburidze & Nino Jorbenadze. 2011. *tanamedrove kartuli enis morpologia, salit'erat'uro ena* [Morphology of written Modern Georgian]. Tbilisi: Meridiani.

Hewitt, George. 1995. *Georgian: A structural reference grammar*. Amsterdam: John Benjamins.

Holes, Clive. 2004. *Modern Arabic: Structures, functions and varieties, Revised edn*. Washington, D.C: Georgetown University Press.

Iluridze, Ketevan. 2006. *sakhelta bruneba XIX sauk'unis I nakhevris kartuli enis gramat'ik'ebshi* [Nominal case inflection in Georgian grammars of the early nineteenth century]. Ph.D. thesis. Tbilisi: Arnold Chikobava Institute of Linguistics.

Imnaishvili, Ivane. 1956. *ts'rpelobiti brunvis sak'itkhi sak'utar sakhelebshi: sakhelta brunebis ist'oriisatvis kartvelur enebshi* [Issues of bare case in proper nominals: history of nominal case inflection in Kartvelian]. Tbilisi: Tbilisi University Press.

Imnaishvili, Ivane. 1957. *sakhelta bruneba da brunvata punktsiebi dzvel kartulshi* [Nominal case inflection and functions in Old Georgian]. Tbilisi: Tbilisi University Press.

Kaplan, Ronald M. & Joan Bresnan. 1982. Lexical-functional grammar: A formal system for grammatical representation. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 173–281. Cambridge, MA: MIT Press.

Karosanidze, Lia. 2017. *ant'kur-bizant'iuri teoriebi enis shesakheb da kartuli gramat'ik'uli azrovneba* [Classical and Byzantine theories on language and Georgian grammatical thought]. Tbilisi: Tbilisi University Press.

Leech, Geoffrey & Andrew Wilson. 1999. Standards for tagsets. In Hans van Halteren (ed.), *Syntactic wordclass tagging*, 55–80. Dordrecht: Kluwer.

Lobzhanidze, Irina. 2013. Morphological analyzer and generator of Modern Georgian language. In *Proceedings of Georgian Language and Modern Technologies IV*, Tbilisi, 82–83.

Lobzhanidze, Irina. 2021. MULTEXT-East morphosyntactic specifications: Georgian specifications. In Tomaž Erjavec (ed.), *MULTEXT-East: Multilingual text tools and corpora for Central and Eastern European languages, Version 6*. Available at: http://nl.ijs.si/ME/V6/msd/html/msd-ka.html.

Lobzhanidze, Irina. 2022. *Finite-state computational morphology: An analyzer and generator for Georgian*. Berlin: Springer.

de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre & Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA. Available at: https://aclanthology.org/L14-1045/.

Marr, Nikolai. 1908. *osnovnye tablicy k grammatike dreviegruzinskogo âzyka* [Basic paradigms of Old Georgian grammar]. St. Petersburg: Imperial Academy of Sciences.

Marr, Nikolai. 1925. *grammatika drevneliteraturnogo gruzinskogo âzyka* [Written Old Georgian Grammar]. Leningrad: Russian Academy of Sciences.

Melikishvili, Damana. 2008a. *The Georgian verb: A morphosyntactic analysis*. Hyattsville, MD: Dunwoody Press.

Melikishvili, Irina. 2008b. Georgian as an active/ergative split language. *Bulletin of the Georgian National Academy of Sciences* 2(2). 138–147.

Melikishvili, Damana. 2014. *kartuli zmnis sist'emuri morposint'aksuri analizi* [Systematic morphosyntactic analysis of the Georgian verb]. Tbilisi: Tbilisi University Press.

Meurer, Paul. 2007. A computational grammar for Georgian. In Peter Bosch, David Gabelaia & Jérôme Lang (eds.), *Logic, language, and computation. TbiLLC 2007*, 1–15. Berlin: Springer.

Payne, Thomas E. 2011. *Understanding English grammar*. Cambridge: Cambridge University Press.

Sarjveladze, Zurab. 1984. *kartuli salit'erat'uro enis ist'oriis shesavali* [Introduction to the history of literary Georgian]. Tbilisi: ganatleba.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK. Available at: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

Shanidze, Akaki. 1934. *dzveli kartuli ena: dzveli kartuli ena da lit'erat'ura* [Old Georgian: Old Georgian language and literature]. Tbilisi: sakhelgami.

Shanidze, Akaki. 1953. *kartuli enis gramat'ik'is sapudzvlebi I: morpologia* [Fundamentals of Georgian grammar I: morphology]. Tbilisi: Tbilisi University Press.

Shanidze, Akaki. 1976. *dzveli kartuli enis gramat'ik'a* [Grammar of Old Georgian]. Tbilisi: Tbilisi University Press.

Shanidze, Akaki (ed. Mzekala Shanidze). 1980. *kartuli enis gramat'ik'is sapudzvlebi III* [Fundamentals of Georgian grammar III]. Tbilisi: Tbilisi University Press.

Skopeteas, Stavros, Gisbert Fanselow & Rusudan Asatiani. 2012. Case inversion in Georgian: Syntactic properties and sentence processing. In Monique Lamers & Peter de Swart (eds.), *Case, word order, and prominence*, 145–171. Berlin: Springer.

Uturgaidze, Tedo. 1986. *kartuli enis sakhelis morponologiuri analizi* [Morphophonological analysis of Georgian nominals]. Tbilisi: mectsniereba.

Vogt, Hans. 1947. Le système des cas en géorgien ancien. *Norsk tidsskrift for sprogvidenskap* 14. 98–140.

Zorell, Franz. 1930. *Grammatik zur altgeorgischen Bibelübersetzung mit Testproben und Wörtvereichnis*. Rome: Pontifical Biblical Institute.