

Gaining the Perspective of a Language Family Oriented Grammar Design: Special Predicative Clitics in Slavic*

Tania Avgustinova
Department of Computational Linguistics, University of Saarland

On the abstraction level of shared grammar, a common Slavic inventory of special predicative clitics can be postulated in terms of feature specifications referring to information on TYPE, CASE and INDEX (the latter encompassing person, number and gender). This inventory is subject to parameterisation across Slavic languages. We argue that such an approach can contribute considerably to formalising clitic typology.

1 Preliminaries

A successful multilingual NLP system would eventually employ generic linguistic resources that are adaptable to specific language and application requirements. Related languages share a much wider range of linguistic information than typically assumed in standard multilingual grammar architectures. Taking grammatical relatedness seriously, we are particularly interested in designing linguistically motivated grammatical resources for Slavic languages to be used in computational linguistics. Pursuing this goal, however, one is inevitably confronted with primary problems stemming from a situation in which different linguistic theories cut up grammars in quite different ways, and grammar formalisms differ in their degree of granularity. It cannot be expected, therefore, that the minimal differences between two languages or their shared elements form easily identifiable units in the available language-specific grammars.

An ontology that will conceptualise morphosyntactic "building blocks" can certainly serve as a basis for a shared grammar of Slavic. Our use of the term ontology is fairly pragmatic – as representing a formal shared conceptualisation of a particular domain of interest. It describes concepts relevant for the domain, their relationships, as well as "axioms" about these concepts and relationships. Such a pragmatic approach does not presuppose any general all-encompassing ontology of language but rather "mini-ontologies" conceptualising the selected domain from various perspectives in a consistent way. In this study, the domain of interest is the grammatical knowledge on Slavic morphosyntax contained in linguistic theories and linguistic descriptions. This knowledge (cf. Figure 1) can be highly structured as in formal linguistic theories, semi-structured as in conventional linguistic descriptions, or even unstructured, based on linguistic intuitions.

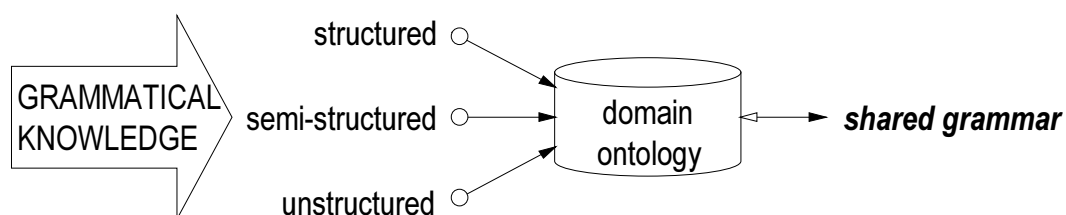


Figure 1: Grammatical knowledge, domain ontology, shared grammar

In a constraint-based theory like HPSG, the grammatical properties of linguistic entities are typically revealed in complex, multidimensional taxonomies. So, every linguistic entity can be considered from a variety of perspectives which are modelled as distinct dimensions of hierarchical classification. The type subsumption in such a multiple inheritance hierarchy can be interpreted as modelling a continuum from general – and presumably universal – types to more and still more specific instances of these types resulting from

* In: *Proceedings of The First Conference on Generative Linguistics in Poland (GLiP-1)*, Syntactic meeting, Warsaw 1999

admissible cross-classifications. Various constraints can be associated with the types in such a hierarchy, being thus formulated only once at the appropriate level of abstraction. Through the mechanism of multiple inheritance, the constraints apply to all subtypes of the type on which they have actually been formulated.¹

In the context of determining and formally describing typological similarities – the *shared grammar* – of Slavic languages, we discuss a distinguished set of elements which, based on their morphosyntactic behaviour, can be regarded as **special clitics**. The traditional distinction of simple and special clitics is introduced by Zwicky (1977). A simple clitic is an element of some basic word class, occupying the normal syntactic position for a non-clitic word of its category. In contrast, a special clitic occupies an "unusual" position as a result of additional restrictions posed on it. Special clitics have special syntactic properties which require them to cluster, therefore, special principles and mechanisms must be invoked for obtaining the correct placement, whereas simple clitics only have a phonological deficiency.

2 Predicative cliticisation

Among Slavic languages, a basic distinction has to be made between those in which pronominal, reflexive and auxiliary clitics exist (e.g., Bulgarian, Macedonian, Serbo-Croatian, Slovene, Czech, Slovak, Polish) and those which have no such clitics (e.g., Belorussian, Russian, Ukrainian). Within the first group, a further differentiating can be made on the basis of the observable prosodic behaviour. To a considerable extent, predicative clitics in, e.g., Bulgarian, Macedonian, Slovene, Czech, Slovak, or Polish are characterised by what we may generally call a *non-constant direction of prosodic attachment*, which results in certain 'prosodic autonomy' or 'prosodic neutrality'. This is to be contrasted to the situation in, e.g., Serbo-Croatian, where the discussed clitic set is always prosodically enclitic.

As argued in Avgustinova 1997, the special clitics in Bulgarian are *host-category-bound clitics*. They belong to the syntactic domain of a particular (morphosyntactic) category which also hosts them prosodically. Thus, verbal and nominal clitics are distinguished. The predicative special clitics we are interested in here are verbal clitics from this perspective, since they belong to the verb complex (set bold in ex. 1, where the vowels in all accented syllables are capitalised, and the respective prosodic words constituting the verb complex are indicated by square brackets). With respect to clustering, they are *core clitics*, i.e. pivotal components of any clitic cluster.² The clustering pattern for Bulgarian verbal core clitics is: AUX-NON-3SG < DAT < ACC < AUX-3SG.

ex. 1 (Bulgarian)

- a. TogAva [si štJAJa] [da si mi gi pokAzala] vEče.
 then AUX-2sg FUT-AUX-sg.f PRT AUX-2sg DAT-1sg ACC-3pl shown-sg.f already
 'You would have already shown them to me by that time (reportedly).'
- b. [ŠtJAJa sī] [da si mi gi pokAzala] vEče.
 FUT-AUX-sg.f AUX-2sg PRT AUX-2sg DAT-1sg ACC-3pl shown-sg.f already
 'You would have already shown them to me (reportedly).'
- c. [ŠtJAJa li sī] vEče [da si mi gi pokAzala]?
 FUT-AUX-sg.f Q AUX-2sg already PRT AUX-2sg DAT-1sg ACC-3pl shown-sg.f
 'Would you have already shown them to me (reportedly).'

¹ The two types of edges connecting types in our graphical representation of hierarchies – 'square' and 'direct' – are significant. The former indicate possible conjunction of types, and thus introduce various dimensions of multiple inheritance. The latter indicate disjunction of types within the respective dimension of classification.

² For a detailed discussion we refer the reader to the cited work.

The special predicative clitics in Bulgarian exhibit what is sometimes regarded as 'endo'-clitic behaviour. Depending on the prosodic context, they can be proclitic or enclitic with respect to the verbal element hosting them. The proclitic realisation, however, tends to be the default one, which is apparently related to constraints yielding certain "second position" effects. In particular, the core verbal clitics are non-initial in the clausal domain, but nevertheless, they are oriented towards the left edge of (a morphosyntactic sub-domain within) the verb-complex domain.

To a considerable extent, Czech special predicative clitics are characterised by what we may tentatively call **prosodic autonomy**. They are associated with a 'second' position rather than with a prosodic host to their left or to their right. So, they are typical Wackernagel clitics which (are traditionally assumed to) immediately follow the first syntactic constituent. With respect to clustering, Avgustinova and Oliva (1995) also distinguish *pure clitics* from *semi-clitics*, since the latter can but need not join the Wackernagel clitic cluster. However, as these authors show, it is not always trivially decidable what the 'first' position actually is. For instance in ex. 2 the reflexive clitic follows as much as five syntactically distinguishable constituents.

ex. 2 (Czech)

Minulý týden v neděli ráno doma v koupelně se Jan při holení nebezpečně zranil.
 last week on Sunday morning at-home in bathroom REFL Jan at shaving dangerously hurt

The clustering pattern for Czech special predicative clitics is: AUX < ETH.DAT < REFL(DAT/ACC) < DAT < ACC(GEN). There are characteristic co-occurrence restrictions which do not allow a clitic cluster to contain two phonologically identical clitics (even though, as illustrated in ex. 3a, nothing prevents the co-occurrence of two datives) or two reflexive clitics (e.g., the sentence in ex. 3b is grammatically acceptable because the (phonologically different) reflexive clitics do not cluster together).

ex. 3 (Czech)

a. Poslat kurýrem se mi mu ho dnes nepodařilo ...
 send-INF courier-INSTR REFL DAT-1sg DAT-3sg.m ACC-3sg.n today NEG-succeeded
 'I did not succeed in sending it to him by a courier today ...'

b. Jan se snažil si koupit nové auto.
 Jan REFL tried REFL buy-INF new car
 'Jan tried to buy a new car (for himself).'

Another instance of Wackernagel clitics is found in Slovene. Similarly to Czech, the special predicative clitics occur after the 'first phrase'. But they can follow an intonational pause, and even occur clause-initially (ex. 4) in the case of so-called "topic-drop". In this sense, the Slovene special predicative clitics are **prosodically neutral**.

ex. 4 (Slovene)

a. Si ga videl?
 AUX-2sg ACC-3sg.m saw
 'Have you seen him?'

b. Se mi je smejal.
 REFL DAT-1sg AUX-3sg laughed
 'He was laughing at me.'

The special predicative clitics in Serbo-Croatian are 'second-position' (or Wackernagel) clitics too. Moreover, they can be placed not only after the 'first phrase' (ex. 5a), as we saw in Czech or Slovene, but also after the 'first word' (ex. 5b). The clustering pattern is: AUX-NON-3SG < DAT < ACC/GEN < REFL < AUX-3SG.

ex. 5 (Serbo-Croatian)

a. Taj čovjek *joj* *ga* *je* danas kupio.
that man DAT-3sg.f ACC-3sg.n AUX-3sg today bought
'This man has bought it to/for her today.'

b. Taj *joj* *ga* *je* čovjek danas kupio.
that DAT-3sg.f ACC-3sg.n AUX-3sg man today bought
'This man has bought it to/for her today.'

Further crucial property of Serbo-Croatian Wackernagel clitics is that they are **prosodically enclitic**. For instance, unlike the situation in Czech or Slovene, they cannot appear after an intonational boundary, i.e. immediately after, or inside, phrases set off by comma (corresponding to an intonational pause). In addition, they allow in coordinate constructions the ellipsis of the type illustrated in ex. 6.

ex. 6 (Serbo-Croatian)

a. Ona *mu* *ga* *je* dala, a i ja *sam* *mu* *ga* dala.
she DAT-3sg.m ACC-3sg.n AUX-3sg gave and also I AUX-1sg DAT-3sg.m ACC-3sg.n gave
'She has given it to him, and also I have given it to him.'

b. Ona *mu ga je* dala, a i ja *sam mu ga*.

c. Ona *mu ga je* dala, a i ja *sam mu*.

d. Ona *mu ga je* dala, a i ja *sam*.

Far from being exhaustive indeed, this brief synopsis of predicative cliticisation across Slavic languages nevertheless suggests that special principles and mechanisms needed (at least) for (i) determining the domain of distribution, (ii) clitic climbing, and (iii) clitic cluster formation. The domain of distribution is crucial for the operation of the two general alignment constraints, namely, **non-initial** (a given clitic cannot be initial within its domain of distribution) and **left-edge** (a given clitic typically tries to occur as close to the left edge of its domain of distribution as possible). For sentential clitics, both these constraints apply within the clause, while for verbal clitics, they apply within the verb complex. The phenomenon of clitic climbing is observed only with sentential clitics and is realised out of syntactically subordinated verbal domains. The tendency of avoiding sequences of phonologically identical clitics as a result of clitic climbing is called haplology. As to clustering, in all languages with predicative clitic cluster formation, a dative pronominal clitic combines with an accusative pronominal clitic if the latter is in third person singular, and namely, in the order **DAT < ACC-3sg**. Therefore, this can be isolated as a minimal common pattern to which further language-specific constraints apply, e.g., expending it as in the case of Bulgarian where the accusative clitic is just required to be in third person.

So, the modelling of special predicative clitics in Slavic has to be based on **alignment domains** but it also has to be related to particular **anchors**. The alignment domains can be defined syntactically (e.g., word order domains with (partial) compaction), prosodically (e.g., prosodic words in an intonational phrase) and with respect to information structure (e.g., in determining the Wackernagel position). As to anchors, in syntax they can be identified with the respective licensing category; in morphology (if one adopts the phrasal-inflection view), the clitics are interpretable as features attached to lexical or syntactic categories; while in phonology, the prosodic host is relevant in this respect.

Interestingly, with respect to predicative cliticisation, Polish substantially deviates from verbal-clitic languages like Bulgarian as well as from Wackernagel-clitic languages like Czech, Slovene or Serbo-Croatian. It is common to regard Polish weak auxiliaries as floating or mobile tense inflection. Typically, they are suffixed to the verbs they morpho-syntactically belong to, but can also be attached to other syntactic items within the same

clause. When hosted by other categories, they exhibit a strictly enclitic behaviour. We will refer to Polish mobile inflections as **affixing clitics**. At the same time, the prosodic constraints on Polish pronominal and reflexive clitics are very loose. What we are actually confronted with here are special clitics with a simple-clitic behaviour. So we can tentatively call them **'free' clitics**. While optional climbing as well as haplogy of *się* are observed in Polish, there is no clitic cluster formation. So, as expected, the general co-occurrence constraint on compatibility of dative and accusative clitics mentioned above does not apply. This is illustrated by ex. 7 where, on the one hand, the order of the clitics is reversed, and on the other hand, the accusative clitic is a non-3sg one.³ Even though adjacent in ex. 7a, the two clitics – ACC-2sg and DAT-3sg.m – do not actually form a clitic cluster, and as ex. 7b shows, they could be separated.

ex. 7 (Polish)

a. Pokazali cię mu wczoraj.
 showed ACC-2sg DAT-3sg.m yesterday
 'They showed you to him yesterday.'

b. (?) Oni cię pokazali mu wczoraj
 they ACC-2sg showed DAT-3sg.m yesterday
 'They showed you to him yesterday.'

Nevertheless, certain 'second-position' effects are still detectable in the fact that the special predicative clitics also in Polish cannot be initial in the clausal domain. Moreover, they generally resist non-immediate post-verbal positioning, which is a clear resemblance of a left-edge orientation.

3 Taxonomy of special predicative clitics

While Bulgarian exhibits the richest set of special predicative clitics, there are, in contrast, no auxiliary, pronominal or reflexive clitics in Russian, with reflexivity marker being actually realised as a morpheme (i.e., *-sja/-s'*) in this language. We will look more closely at the situation in Bulgarian, Czech and Polish (Table 1), since these three languages are indeed representative of the observable morphosyntactic variation in the area of predicative cliticisation across Slavic.

Clitic Auxiliary (tense-forming)	Clitic Argument / Reflexive
<u>non-3rd-person</u> Bg: <i>sâm si sme ste</i> Cz: <i>jsem jsi jsme jste</i> Pl: <i>-m -ś -śmy -ście</i>	<u>personal pronoun</u> Bg: <i>DAT: mi ti mu ò ni vi im ACC: me te go ja ni vi gi</i> Cz: <i>DAT: mi ti mu ACC: tě ho</i> Pl: <i>DAT: mi ci mu ACC: cię go je</i>
<u>3rd-person</u> Bg: <i>e sa</i> Cz: \emptyset Pl: \emptyset	<u>anaphoric pronoun / non-anaphoric reflexive</u> Bg: <i>si se</i> Cz: <i>si se</i> Pl: <i>się</i>

Table 1: Inventory of special predicative clitics in three Slavic languages

Even though to a different extent, the predicative cliticisation paradigms for Czech and Polish contain gaps where we find "weak" non-clitic (Polish) or semi-clitic (Czech) forms. In terms of morphosyntactic properties, this is summarised in Table 2. Certainly, in languages like Serbo-Croatian, also the genitive clitics have to be taken into consideration.

³ The sentence in (ex. 7a) is taken from Franks 1998, page 84. On a representative set of examples, this author makes a very important observation that pronominal clitics in Polish do not form a cluster. Taking into consideration the typologically striking independence of clitic placement, he convincingly argues that Polish pronominal clitics have actually ceased to be special clitics.

In Bulgarian, the clitic person-number auxiliary has a full paradigm. The same holds for pronominal clitics which can even serve as "agreement markers" in clitic replication (Avgustinova 1997). Both reflexive clitics are available and, importantly, do not allow haplology.

In Czech, there are clitic forms of the person-number auxiliary for the 1st and 2nd person only. A part of the pronominal forms are semi-clitic, in the terminology of Avgustinova and Oliva 1995. Both reflexive clitics are available, allowing for haplology but not, however, for multiple reflexive marking (cf. the cited work for details).

In Polish, the person-number auxiliary occurs as a floating inflection and, again, only for 1st and 2nd person.⁴ A part of the pronominal forms is non-clitic. Moreover, the clitic pronouns behave as simple rather than special clitics (in the sense of Zwicky 1977). There is only one reflexive clitic (*się*) allowing for haplology.

	Bulgarian	Czech	Polish
<i>aux-non3</i>	sâm si sme ste	jsem jsi jsme jste	-m -ś -śmy -ście
<i>aux-3</i>	e sa	je jsou	jest są
<i>dat-non-3sg</i>	mi ti	mi ti	mi ci
<i>dat-3sg-m</i>	mu	mu	mu
<i>dat-3sg-n</i>	mu	mu	mu
<i>dat-3sg-f</i>	ì	jí	jej
<i>acc-1sg</i>	me	mě	mnie
<i>acc-2sg</i>	te	tě	cię
<i>acc-3sg-m</i>	go	ho	go
<i>acc-3sg-n</i>	go	ho	je
<i>acc-3sg-f</i>	ja	ji	ją
<i>pron-pl</i>	ni vi im ni vi gi	nám vám jim nás vás je	nam wam im nas was ich
<i>rfl-si / dat-si</i>	si	si	sobie
<i>rfl-se / acc-se</i>	se	se	się

Table 2: Special predicative clitics in terms of morphosyntactic properties

For Slavic, we assume two orthogonal dimensions of classification for the type *spec(ial)-pred(icative)-cl(itic)*. In the HPSG-style lexical hierarchy sketched in Figure 2, they are formulated as CROSS-REFERENCE of core grammatical relations (e.g., subject, direct object, indirect object), and INDEX-INFO (in the sense of externalising index information like person, number and gender).

So, the clitics we are interested in here are either *cross-referencing* or *non-cross-referencing*, with the former particularly relating to *subj(ect)* or *non-subj(ect)* core grammatical relations. With respect to externalising index information, the special predicative clitics can be *indexing* or *non-indexing*. On such a view, we can easily capture the linguistic generalisation that pronominal clitics and clitic auxiliaries supply index information

⁴ As a matter of fact, this is the set of clitics that is most grammaticalised across Slavic. The correspondents of Polish weak auxiliaries have 'disappeared' in Russian past tense forms; the third-person forms of the same auxiliary are systematically dropped in Bulgarian renarrative verb forms, and – similarly to the situation in Polish – are totally absent in Czech past-tense forms.

about the respective core grammatical relations. The non-subject cross-referencing clitics are further classified according to MORPH(ological)-CASE into *m-acc(usative)*, *m-gen(itive)* or *m-dat(ive)*, and according to REFERENCE-TYPE into *pronominal* or *anaphor*. Along the FUNCTION-TYPE dimension of classification, the types *pronominal* and *anaphor* are simultaneously instances of non-indexing clitics. The type *r(e)fl(exive)-marker* is another instance of non-indexing clitics along this dimension but, in contrast, it is a non-cross-referencing clitic. Another classification of non-indexing clitics takes into consideration their S-FORM with "se" or "si" as abstract possibilities in Slavic. So, in our model we can systematically distinguish accusative / genitive / dative pronominal clitics (*acc-prn*, *gen-prn*, *dat-prn*) from accusative / dative anaphors (*acc-se*, *dat-si*) and reflexivity markers (*rfl-se*, *rfl-si*), but also express their common properties at various levels of abstraction (e.g., the last two rows of Table 2).

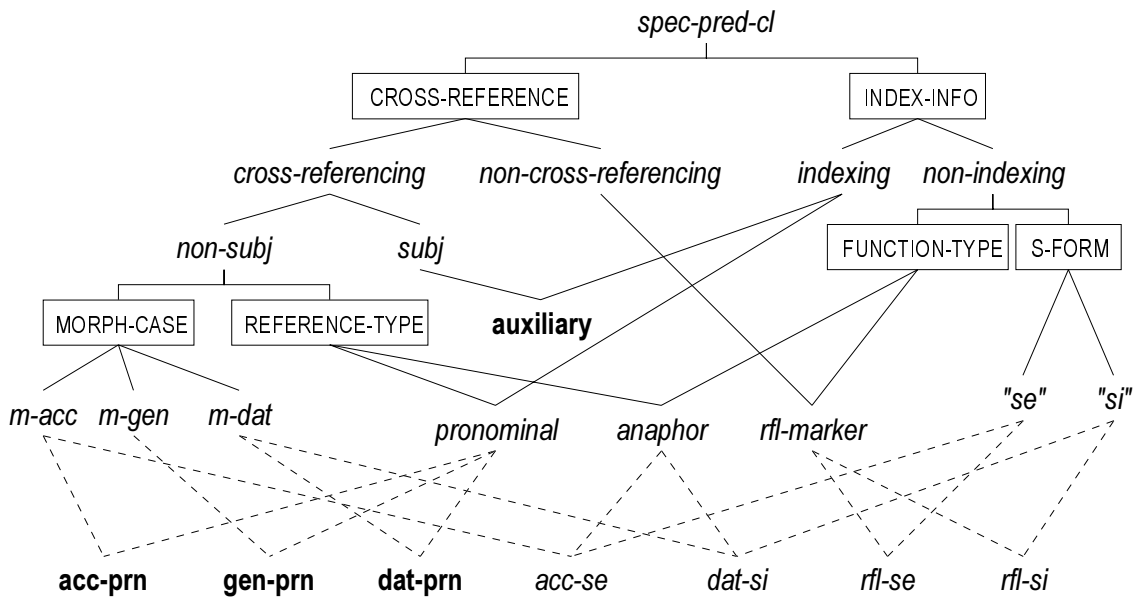


Figure 2: Taxonomy of Slavic special predicative clitics

The concrete types of clitics that are specified in Table 2 are obtained via cross-classification with relevant instances of the type *index*. This mechanism is illustrated for the auxiliary clitics in Figure 3, for the singular dative pronominal clitics in Figure 4, for the singular accusative pronominal clitics in Figure 5, and for the plural pronominal clitics in Figure 6.

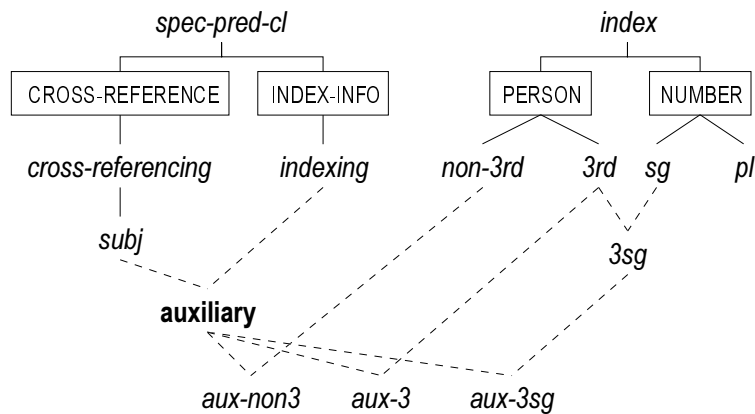


Figure 3: Auxiliary clitics

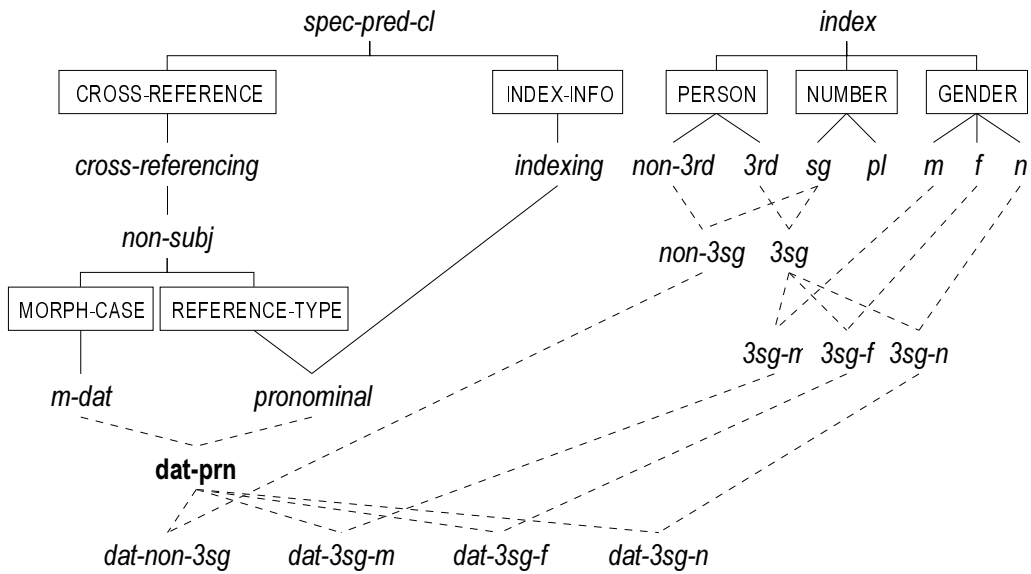


Figure 4: Singular dative clitics

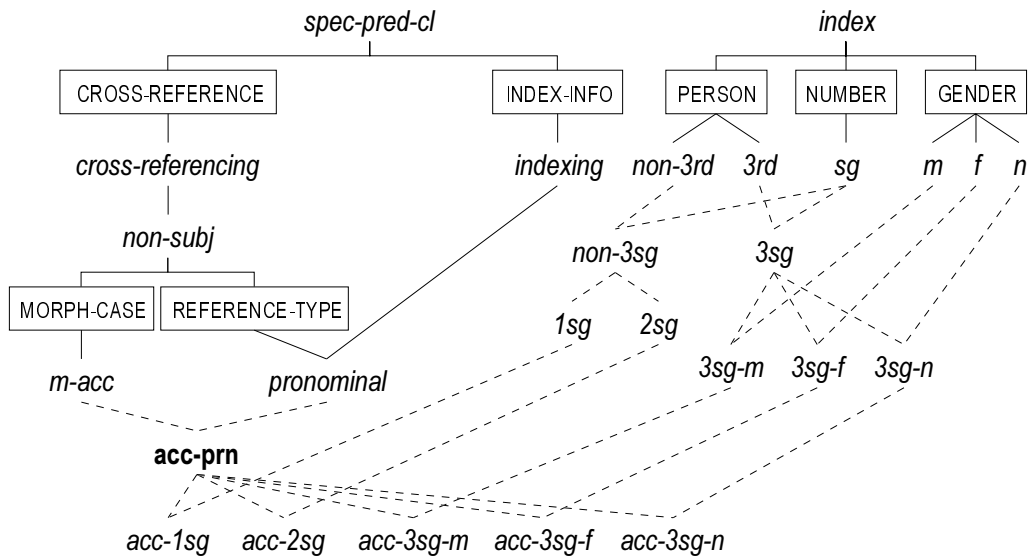


Figure 5: Singular accusative clitics

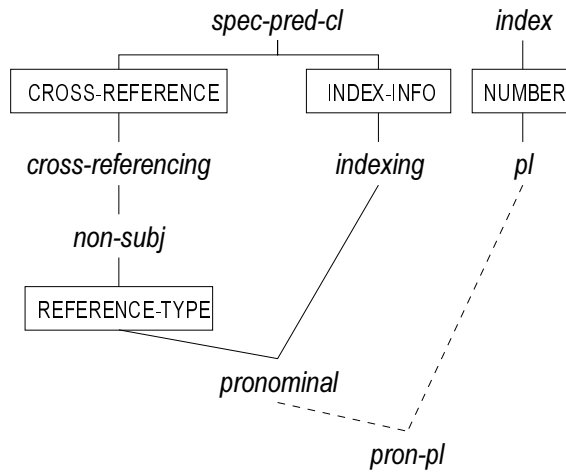


Figure 6: Plural pronominal clitics

4 Cross-Slavic parameterisation

We are now in a position to consider a possible cross-Slavic parameterisation of predicative cliticisation. First of all, there is a basic difference between Slavic languages having a distinguished set of predicative special clitics and those of them that have none (Figure 7).

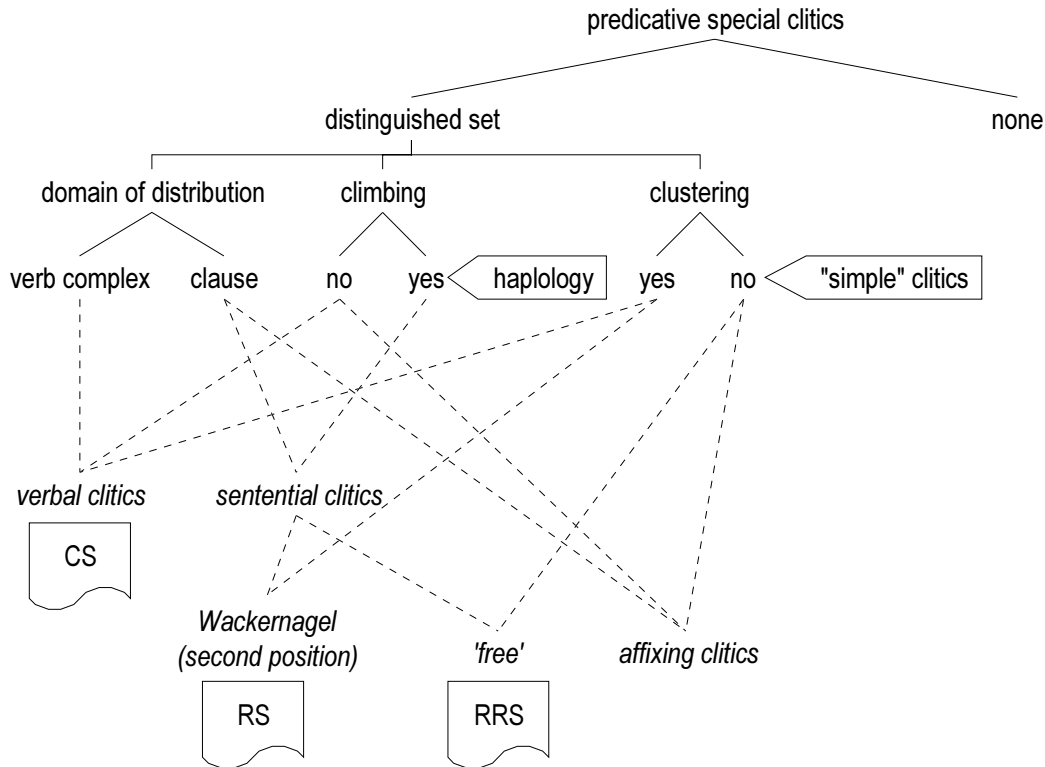


Figure 7: Cross-Slavic parameterisation of predicative cliticisation

Looking closer at Slavic languages with predicative cliticisation, we take into consideration the syntactic domain of distribution as well as phenomena like clustering and climbing. This will allow us to formally distinguish several sets of predicative clitic, with each of them being potentially associated with a bunch of constraints. In general, clitic climbing creates conditions for haplology, while the non-clustering can be explained as presupposing a simple-clitic behaviour. And indeed, clitic clusters are only formed when there is no actual possibility to simultaneously satisfy linearization constraints that are principally equal but associated with different items. The latter situation typically arises with special clitics. Simple clitics, in contrast, are just phonologically deficient, but otherwise behave like the non-clitic instances of their category.

Let us first consider the *verbal clitics*. Their domain of distribution is the verb complex, they form a cluster, but do not allow for clitic climbing, hence, no haplology is observed. Here we are confronted with the **complete set (CS)** represented by Bulgarian core verbal clitics, which includes the types *aux-non3*, *aux-3*, *dat-non-3sg*, *dat-3sg-m*, *dat-3sg-n*, *dat-3sg-f*, *acc-1sg*, *acc-2sg*, *acc-3sg-m*, *acc-3sg-n*, *acc-3sg-f*, *pron-pl*, *rfl-si*, *dat-si*, *rfl-se*, *acc-se*. For constraints on clustering cf. Avgustinova 1997, pages 47–66.

The domain of distribution of the *sentential clitics* is the clause, and in principle, they can undergo clitic climbing accompanied by haplology effects. The main difference between the two subtypes of sentential clitics – Wackernagel or 'free' – concerns clitic cluster formation. The second-position (Wackernagel) clitics form a cluster which occupies the Wackernagel position in the sentence, independently of how this position is defined for the particular language. In Czech, they correspond to a **restricted set (RS)** of special predi-

cative clitics which is obtained by removing the types *aux-3*, *dat-3sg-f*, *acc-1sg*, *acc-3sg-f* and *pron-pl* from the complete set specified above. For constraints on clustering cf. Avgustinova and Oliva 1995. The 'free' clitics, on the other hand, exhibit a free syntactic behaviour with respect to clustering, in the sense that there are almost no constraints on their positioning in Polish. This is a **reduced restricted set (RRS)** obtained by removing the types *aux-non3*, *rfl-si* and *dat-si* from the reduced set of special predicative clitics.

The domain of distribution of Polish *affixing clitics* is the clause of their predicate, they never undergo clitic climbing, and also do not cluster with other clitics. The affixing clitics are of the type *aux-non3*.

5 Concluding remarks and outlook

From a broader perspective, the observed tendency with respect to special predicative clitics actually represents a continuum from a highly articulated paradigm in South Slavic through gradual reduction in West Slavic up to a complete disappearance in East Slavic. Undoubtedly, there exists a correlation with the "reverse" tendency (i.e. from East Slavic through West Slavic to South Slavic) observable in the following areas: (i) from infinitival complements to finite complements, (ii) from elaborate case system through extended case syncretism to case remnants only with pronouns, and (iii) from copula-less predication to a morphosyntactic verb complex.

In this contribution we worked out a taxonomy of Slavic special predicative clitics, and proposed a system of linguistic parameterisation accommodating the striking variety observable across Slavic languages. This system has the potential of predicting further clitic types that are not attested in the domain of Slavic predicative cliticisation. Therefore, one possible line of research would be, for example, to consider the cliticisation in Romance languages from the perspective outlined here.

References

Tania Avgustinova: *Word order and clitics in Bulgarian*. 1997. Saarbrücken Dissertations in Computational Linguistics and Language Technology, Volume 5

Tania Avgustinova and Karel Oliva: Wackernagel Position and Related Phenomena in Czech. *Wiener Slavistisches Jahrbuch*. Band 41. Wien: Verlag der Österreichischen Akademie der Wissenschaften, 1995, pp. 21-42

Steven Franks: Clitics in Slavic. *Position paper for the workshop "Comparative Slavic Morphosyntax: The State of the Art"*, Indiana University, 1998 (<http://www.indiana.edu/~slavconf/linguistics/>)

Arnold Zwicky: *On Clitics*. Indiana University Club. Bloomington, 1977