Challenges and Perspectives in Italian Clitic Tagging: A Case Study on the SerbItaCor3 Corpus

Scientific paper DOI: 10.18485/judig.2025.1.ch15

Saša Moderc[1] *0009-0007-5059-2007*

Abstract

Italian clitics are functionally heterogeneous: some function as pronouns, others as various types of adverbial adjuncts, or as markers of passive and impersonal constructions. Clitics are also used pleonastically and are integrated into pro-complement verbs (e.g., *farcela* 'to manage'). They occupy three positions relative to the verb and can form clusters of two or three clitics; in such cases, the entire cluster usually occupies one of the three aforementioned positions.

In the SerbItaCor3_it corpus, the tagging of Italian texts was performed using TreeTagger (by Achim Stein). During analysis, inaccuracies were observed in the tagging of the clitics *ci* and *si*, as well as inconsistencies in the processing of homographs. The clitic *si* was tagged either as a reflexive pronoun (PRO:refl) or as a personal pronoun (PRO:pers), while its passive and impersonal uses were not marked. These tagging inaccuracies compromise the reliability of the corpus's statistical data and limit its usefulness for linguistic analysis and language teaching. However, thanks to the accompanying Serbian translations, in most cases it was possible to determine the exact function of the clitics. This enabled the proposal of improvements to the Italian tagger, contributing to more accurate tagging of clitics and homographs. The paper presents examples of incorrect tagging and translation-based solutions, on the basis of which suggestions for improved tagging of Italian texts can be formulated.

Key words: Italian, Serbian, SerbItaCor3_it, homographs, clitics, tagging, improving tagger performances

[1] University of Belgrade – Faculty of Philology, moderc.sasa@gmail.com

1. Introduction: Italian Clitics

Italian clitics pronominalize grammatical persons, sentence arguments, or constituents. These include *mi*, *ti*, *ci*, *vi*, *lo*, *la*, *li*, *le*, *gli*, *ne*, and *si*. They are unstressed and bound to the verb or verb phrase, appearing in proclitic, enclitic, or mesoclitic position. Clitics form a prosodic unit with the verb and may be written either attached to it or separated from it, depending on their position. In terms of syntax, Italian clitics can occur in sequences of two, and occasionally even three, most commonly when *si* is used in impersonal or passive constructions (on these uses, a detailed discussion is given in Bentley 2006). The addition of a third clitic other than *si* is possible, though rare. Each clitic must have a unique referent (which may consist of one or more lexical items); when no referent is present, the clitic functions pleonastically or as part of a phraseme.

With certain frequent verbs, clitics may become lexicalized and significantly modify the verb's meaning. Such verbs are referred to as *verbi procomplementari* in Italian linguistics (see Russi 2008 for a detailed discussion). The syntax and functions of clitics are addressed in all Italian grammars, with a more extensive treatment in Renzi (1988), Serianni (1989), Salvi & Vanelli (2004) and Moderc (2021a, 2021b); most common uses of Italian clitics are discussed in a Serbian-Italian contrastive perspective in Moderc (2015).

2. The Polyfunctionality of Clitics

In standard Italian, clitics perform multiple functions, summarized in the following table (translated into English from Moderc 2021a: 21):

Table 1 Multiple functions of Italian clitics

FUNCTION	MI	TI	CI	VI	LO	LA	LI	LE	GLI	NE	SI
DIRECT OBJECT	●	●	●	●	●	●	●	●			●
INDIRECT OBJECT	●	●	●	●				●	●		●
REFLEXIVE FUNC.	●	●	●	●							●
PARTITIVE FUNC.										●	
SPATIAL FUNC.			●							●	
SOCIATIVE FUNC.			●								
INSTRUMENTAL FUNC.			●								
POSSESSIVE FUNC.										●	
IMPERSONAL FUNC.			●								●
PASSIVE FUNC.											●
PROFORM					●						
IDIOMATIC FUNC.			●		●	●		●		●	●

In addition to the functions mentioned above, other clitic functions emerge under the influence of colloquial language and dialects, or as a result of shifts in communicative strategy during spontaneous oral production. Despite these variations, the linguistic competence of speakers facilitates the correct association of clitics with their referents. Resolving clitic referentiality is a key component of language acquisition, as it requires an integrated understanding of syntax, verb valency, frequent collocations, and world knowledge, especially when context does not provide direct information about the referent and it must instead be inferred from extra-linguistic cues. Italian grammars and their accompanying exercises typically address only the most frequent clitic functions and the most common combinations of two clitics. However, language teaching demands a more comprehensive approach, aimed at enabling learners to identify each clitic's function and to substitute it with the appropriate referent or sentence constituent, typically expressed lexically (primarily with nouns).

In this context, annotated language corpora can facilitate the acquisition of clitic functions, as they are subject to automatic tagging. A considerable body of research has been conducted in this area (Schmid et al. 2007; Tamburini 2000, 2009; Dell'Orletta 2009; Schmid 2013), and the resulting findings are largely satisfactory, although they still show inaccuracies in the specific cases discussed in this paper. The tagging results can serve as a tool for testing and, if necessary, correcting learners' hypotheses about clitic functions in texts from the corpus. On this basis, we analyzed the extent to which the bilingual Serbian–Italian parallel corpus SerbItaCor3_it2 is reliable and accurate in identifying and distinguishing the functions of Italian clitics.

3. SerbItaCor3_it Corpus and Homographs: an Instrument for Successful Disambiguation

We begin our analysis by examining how homograph pairs are tagged in the corpus mentioned above. In the following examples, Italian nouns and verbs share the same form. Nouns are preceded by a definite article (la or lo), while verbs are preceded by a clitic (la or lo as unstressed personal pronouns). Since articles and clitics are themselves homographs, this results in what we may call "double homography" or "homographic syntagms". The homographic syntagms in the following list represent an illustrative sample.3 In the English translations, nouns appear first, followed by verbs:

[2] More information in Moderc S.; Stanković R.; Tomašević A.; Škorić M. (2023).

[3] In English translations an indefinite article was preferred instead of the equivalent determinative *the*. In Italian there are no neutral nouns, therefore in some cases we had to use the pronoun *it*. Where needed, lexemes are added in order to stress the meaning or the English verb.

la caccia (a hunt; he/she chases her)
la cava (a quarry; he/she takes something out)
la guida (a guide; he/she guides her)
la leva (a lever; he/she removes her)
la manovra (a maneuver; he/she maneuvers it [e.g. *a car*]/he manipulates her [fig.])
la mostra (an exhibition; he/she shows her/it)
la piega (a fold; he/she folds it)
la posta (a mail; he/she posts it)
la sega (a saw; he/she saws it)
la sposa (a bride; he marries her)
la sveglia (a clock; he/she wakes her)
la veste (a dress; he/she dresses her)
le serve (the maids; she needs something)
il/lo perdono (a forgiveness; they lose it; I forgive him)[4]
lo sbaglio (a mistake; I get it wrong)
lo sconto (a discount; I discount [e.g. *this product by 10%*])
lo sfondo (a background; I break through [e.g. *the police checkpoint*])
lo sporco (dirt; I make/get it dirty)
lo sposo (a bride; I marry him/I'm marrying him)

Although some homograph pairs are correctly tagged exclusively as nouns (e.g., *sega* 'a saw' or *posta* 'mail', since the corresponding verb forms are not present in the corpus), or as verbs (since the corresponding nouns are not used), in several cases the tagging of homographs proves inaccurate. In a number of instances, *lo* is tagged as an article rather than as a clitic (i.e., a personal pronoun), even when it precedes a verb, a context in which determiners cannot appear. For example, the string "la cava" appears 14 times in the corpus SerbItaCor3_it: four times as a noun ('a quarry'), once as a verb ('to take something out'), and nine times as part of the idiomatic expression *cavarsela* (from *cavare* + reflexive *si* + idiomatic *la*, meaning 'to manage, to get by'). Yet, in all 14 cases, *cava* is tagged as a noun (NOM). Correspondingly, the word *la* that precedes *cava* is always marked as an article (DET:def), which is only correct in the four cases where *cava* is a noun. In the remaining instances, *la* functions as a clitic, specifically, an unstressed personal pronoun without a definite referent, causing the verb to adopt an idiomatic meaning detached from any specific feminine singular noun as direct object. Similarly, the string "la conta" occurs seven times in the corpus SerbItaCor3_it. In one case, it is incorrectly tagged as a noun when it is actually a verb, as in *Chi non la*

[4] In this case, two verbs are used: *pErdere* 'to lose' and *perdonAre* 'to forgive'.

conta giusta a noi? ('Who's not being straight with who?', from *contare*, colloquial for 'to tell' + idiomatic *la*). Here again, *conta* is preceded by *la*, but both words are misclassified, *conta* as a noun [NOM] and *la* as an article [DET:def]. In the same corpus (SerbItaCor_it) the Italian string "le serve" is recorded 59 times; *serve* is dominantly tagged (56 times) as a noun ('the maids') preceded by the article *le*, and three times only as a verb ('she needs something'), preceded by the personal clitic *le* (in the dative case, 'to her'). In reality, the string "le serve" in 37 occurrences contains a noun and in 19 a verb, so that a revision of the POS tagging would be needed in this case also.

From these examples, it can be concluded that TreeTagger lacks the necessary instructions to distinguish between homographic nouns and verb forms. To enhance its performance, particular attention should be devoted to homographs, and specific rules or guidelines should be developed and implemented to facilitate their disambiguation. A possible control mechanism for this task could be derived from the data of the bilingual corpus SerbItaCor3_it, allowing an following a technical enhancement of the corpus itself. Such an enhancement would involve establishing links between semantically equivalent nouns, verbs, adjectives, and adverbs in the Italian and Serbian texts. After these adaptations, the SerbItaCor3_it bilingual corpus could be exploited to achieve a more accurate tagging of homographs, using the translations as reference points for refining the Italian tagger. For example, if the noun *serva* corresponds to *maid* or a similar term in Serbian ('sluškinja', or a synonym), the tagging is appropriate; otherwise, it is reasonable to assume that *serve* represents a verb form of *servire*, as confirmed by the use of a verb in the Serbian translation (for instance, 'služiti'), which reflects the actual function of this word in the original text. Although the list of homographs discussed here is not exhaustive, it nevertheless underscores that the issue has not been adequately addressed in the current tagging system, reducing the quality of linguistic annotation and, ultimately, distorting statistical data related to word classes. We assume that in the future, the integration of taggers with bilingual corpora and Large Language Models will allow for more accurate tagging of Italian words, while parallel bilingual or multilingual corpora will serve as valuable resources for verifying tagging accuracy. That said, we realize that the interpretation of ambiguous cases, such as *La porta la porta dal falegname*[5] ('He/She carries the door to the carpenter'), will

[5] Depending on the interpretation and – in spoken language – on the intonation, "porta" can be interpreted in the first case as a noun and in the second as a verb, or vice versa, in the first case as a verb and in the second as a noun. The structure of the Serbian language in this case does not allow two different focalizations (left dislocation and right dislocation).

likely continue to require human intervention. Finally, we briefly mention the related issue of grammatical congruence in tagging. Specifically, the tagger should be instructed that a clitic cannot precede a noun, just as a determiner cannot precede a verb, except in the case of substantivized verbs, as in *Il bere fa male* ("Drinking is harmful to health"). Better results are obtained, for instance, in the tagging of the Paisà corpus (https://www.corpusitaliano.it/), where lemmatization and part-of-speech (POS) annotation, along with the indication of syntactic dependencies, have been applied. The TreeTagger used for the SerbItaCor3_it corpus, by contrast, does not appear to include syntactic dependencies.

This observation is prompted by the word *perdono*, which can mean either 'forgiveness' (in which case it is a masculine noun) or a verb form: 'they lose', from *pErdere*, or 'I forgive', from *perdonAre* (Italian accented vowels are represented by capital letters). In the corpus, we searched for the string "la perdono/i", and in the six examples found, the tagger marked *la* as an article and *perdono/i* as a noun. This contradicts standard grammar rules, which require agreement in gender and number between the article and the noun (the correct forms being *il perdono*, *i perdoni*). Since in all six examples *perdono/i* is actually used as a verb, *la* preceding it can only be interpreted as a clitic pronoun (e.g., 'to forgive her', 'to lose it'). In each corresponding Serbian translation, a verb is used to denote the action of forgiving or losing, rather than a noun. With the development of linguistic tools and the aforementioned corpus enhancements, translations into other languages (in this case, into Serbian) could serve as a valuable auxiliary resource for achieving a more accurate tagging of Italian parts of speech (POS).

4. SerbItaCor3_it Corpus and the Tagging of Clitics *si* and *ci*.

A search for the clitic *si* in the proclitic position in the "Ammaniti" subcorpus (part of the SerbItaCor3_it corpus), specifically within the novel *Io e te* (*Me and You*), yielded 217 results out of a total of 23,133 words. The clitic *si* is tagged either as a personal pronoun (*PRO:pers*, in 100 cases) or as a reflexive pronoun (*PRO:refl*, in 117 cases). In a sample consisting of the first 20 examples from the list of 217 results, the tagging was incorrect in six cases. In four instances (below, examples 3, 8, 13, 14), the label *PRO:refl* should have been used instead of the generic *PRO:pers* label; in two instances, *si* is an impersonal clitic, yet the tagger lacks a specific tag for this function. As expected, the dominant use of *si* is reflexive, corresponding to the *PRO:refl* label. However, the tagger also employs *PRO:pers* for the same reflexive function. To improve accuracy, two new

labels should be introduced: *PRO:imp* for impersonal *si* and *PRO:pass* for passive *si*. Undeniably, distinguishing between these two functions is not always straightforward, but successfully doing so would be a significant contribution to language teaching, especially by enabling the targeted extraction of sentences containing impersonal *si* and passive *si* constructions from the corpus. In the sample of 20 examples (fitting a single screen on the corpus interface), in two cases (4, 15), *si* was incorrectly tagged; the correct label for each is added in parentheses:

1. Dalla rabbia avevo preso un pietrone e l'avevo scagliato contro un albero, mentre quel ritardato **siREFL** rotolava a terra dalle risate. ✓
2. Mia madre e mio padre non lo sopportavano perché dicevano che **siREFL** prendeva troppe confidenze. ✓
3. Alla fine ha mollato la scopa e **siPERS (=REFL)** è avviato verso la guardiola con il suo passo dondolante e l'ho visto sparire sulle scale che portavano al suo. ✖
4. … e al prato all'inglese con le panchine di marmo dove non ci **siREFL (=IMP**) poteva sedere. ✖
5. Due lunghi neon scarichi **siREFL** sono accesi illuminando un corridoio stretto e senza finestre…✓
6. La porta **siREFL** è spalancata su una grande stanza rettangolare…✓
7. … un fluido rosso mi saliva per le gambe, mi inondava lo stomaco e mi **siREFL** irradiava fino alla punta delle mani…✓
8. Ma qui ci **siPERS (=REFL)** mettono tutti quelli che hanno problemi? ✖
9. mi avrebbe trasmesso, come un corpo caldo che trasmette calore a un corpo freddo, i pensieri dei bambini che **siREFL** erano sdraiati prima di me. ✓
10. Un Lorenzo che **siREFL** vergognava a parlare con gli altri ma che voleva essere come gli altri. ✓
11. Ho scoperto di avere un serbatoio nello stomaco, e quando **siREFL** riempiva lo svuotavo attraverso i piedi…✓
12. … penetrava nelle viscere del mondo e siREFL consumava nel fuoco eterno. ✓
13. … manager americani e italiani facoltosi che **siPERS (=REFL)** potevano permettere la retta. ✖
14. Uno **siPERS (=REFL)** è arrampicato sopra un albero e ha appeso lo zaino di una ragazza su un ramo e quella gli tirava le pietre. ✖
15. Chi aveva deciso che quello era il modo giusto ? Non **siPERS (=IMP)** poteva vivere diversamente?✖

16. Io ho il sé grandioso, - ho sussurrato, mentre tre bestioni che **siREFL** tenevano a braccetto mi spingevano via come fossi un birillo…✓
17. I predatori in quella scuola erano molto più evoluti e aggressivi e **siREFL** muovevano in branco. ✓
18. Mi sono messo le stesse cose che **siREFL** mettevano gli altri. ✓
19. Il solco che mi divideva dagli altri **siREFL** faceva più profondo.✓
20. E sotto la giacca dura come un esoscheletro **siREFL** agitavano cento zampette da insetto. ✓

In [21] another example of incorrect tagging of *si* is given. The appropriate annotation is provided in parentheses:

21. Gli unici rumori che **siREFL (=PASS)** sentivano erano la pioggia che batteva contro la finestra. ✖

With regard to the clitic *si*, we tested if TINT (The Italian NLP Tool, https://dh.fbk.eu/research/tint/, which allows users to test its functionalities in demo mode) would produce more accurate tagging results. We entered the Italian sentence: *Si dice che si sia convertito e a casa sua adesso si adorino gli idoli sumeri* ('They say that he converted and, in his home, now Sumerian idols are worshipped'). In all three instances, the clitic *si* was tagged identically (*Clitic=Yes, Person=3, PronType=Prs*), despite the fact that each *si* has a different function: impersonal (*Si dice*), reflexive (*si sia convertito*), and passive (*si adorino*), respectively. This uniform tagging implies that the user must manually determine the specific function of *si* in each context. To address this limitation, we recommend introducing distinct labels for the different uses of *si*, namely *PRON:Imp* (impersonal), *PRON:Refl* (reflexive), and *PRON:Pass* (passive), and enhancing the linguistic instructions required for a more advanced identification of each of these three functions.

As for the clitic *ci*, it appears 86 times in the aforementioned "Ammaniti" subcorpus. For the purposes of this study, we analyzed the first 20 occurrences in the list; inaccurate tagging is marked with the symbol ✖. We argue that tagging should distinguish among various functions of the clitic *ci*: locative (LOC), reflexive (REFL), phrasal or idiomatic (FRAS), pronominal (PERS), and, possibly, sociative (SOC) and instrumental (INSTR) uses. It can be assumed that proper differentiation of *ci* functions would necessitate tagging instructions accounting for syntax, semantics,

and textual coherence, an undertaking that is undoubtedly demanding and complex. In the following examples, we provide the correct label (or the most plausible one, in cases where the function of *ci* is ambiguous) in parentheses:

22. Così mi **ciREFL (=INSTR)** lavo e ti ho addosso. (clitic *mi* is reflexive; *ci* has an instrumental function, since it refers to a bar of soap, mentioned in the previous sentence) ✖
23. **CiREFL (=LOC)** hai messo dentro il termometro? ✖
24. Fortuna c'era un camion della spazzatura che **ciREFL (=PERS)** rallentava. ✖
25. Non avevo calcolato che mia madre **ciPERS (=FRAS)** tenesse tanto ad accompagnarmi. ✖
26. Non **ciREFL (=LOC)** vado. ✖
27. **CiREFL (=FRAS)** hanno messo un sacco a prepararsi…✖
28. Allora **ciREFL** sentiamo stasera così la ringrazio. ✓
29. Il Cercopiteco **ciREFL (=FRAS)** ha messo parecchio a sentirlo. ✖
30. ….e al prato all'inglese con le panchine di marmo dove non **ciLOC** si poteva sedere. ✓
31. Ma quanto **ciREFL (=LOC)** devo stare? ✖
32. Ma qui **ciPRO:demo[nstrative] (=REFL)si** mettono tutti quelli che hanno problemi? ✖
33. Non **ciREFL (=FRAS)** voleva molto a fregarlo. ✖
34. Questo **ciREFL (=PERS)** sta dicendo il professore? ✖
35. Mi spiegava che gli amici **ciREFL (=FRAS)** mettono un attimo a dimenticarsi di te…✖
36. Se **ciREFL (=PERS)** parla mia madre, - ha risposto Alessia Roncato. ✖
37. Io a Cortina **ciREFL (=LOC)** andavo da quando ero nato. ✖
38. Vedi che non **ciREFL** dobbiamo preoccupare. ✓
39. **CiREFL (=FRAS)** pensavo un po' e rispondevo tranquillo: «Va bene vengo».✖
40. Mamma, ho deciso di non andare a sciare perché nonna sta male e se muore quando io non **ciREFL (=FRAS/LOC)** sono? ✖
41. Quanta neve **ciPERS (=LOC)** poteva essere? ✖

As shown by the examples provided, the clitic *ci* is tagged inaccurately, even more so than *si*. It is particularly surprising that the label for the reflexive function (REFL) is applied to verb forms that do not refer to the first-person plural, as *ci* can only function as a reflexive pronoun in this combination (e.g., *Ci troviamo bene a Pisa* 'We feel comfortable in Pisa-). However, in this clitic-verb combination (followed by an object), a locative interpretation is also possible (e.g., *Ci troviamo un bel ristorante* 'We find a nice restaurant there'). The inherent interdependence of context and the potential arguments of verbs further complicate the interpretation of *ci* and, as a consequence, its automatic tagging. Consequently, the development of more precise instructions for taggers is necessitated, and ultimately, human supervision appears to be needed.

5. Conclusion: why and how to Improve Programs for Tagging Italian Clitics

Improving taggers with more precise instructions regarding parts of speech and their functions may not be a primary focus in contemporary corpus linguistics. However, there is undoubtedly room for improvement in existing taggers and for applying a more advanced, detailed tagger in the next revision and expansion of the SerbItaCior3_it corpus, assuming such a tagger becomes accessible at that time. Greater accuracy in tagging and processing homographs and multifunctional words like clitics would enable linguists to conduct more focused and precise research within the corpus, allowing them to locate relevant examples more efficiently and test their hypotheses more effectively. In the field of foreign language teaching, the application of an improved tagger would provide corpus users with an efficient tool for extracting examples of specific clitic functions, as well as of other linguistic phenomena. This would make the corpus a more reliable learning tool, allowing students to observe a wide range of linguistic phenomena in Italian. Additionally, it would help clarify the usage and functions of clitics, particularly given the tendency in language teaching to overlook the complexities of referentiality and clitic functions (except for the most frequent uses and combinations), focusing instead on comprehension and production.

On the other hand, despite the mentioned flaws, we deem that the fact the parallel texts we have compiled over the years were integrated into the SerbItaCor3_it corpus is a great achievement for Serbian and Italian studies. This accomplishment is largely due to the efforts of Prof. Dr. Ranka Stanković, her associates, and their extensive experience in computational linguistics. Our comments on the tagging of the Italian portion of the corpus are intended to highlight areas where the corpus could be further improved; this task can be achieved by identifying and handling homographs,

especially in cases where the same form is used for different parts of speech (such as nouns and verbs, as demonstrated in this paper). Translations into Serbian can be especially useful in many cases, as shown with the homograph *serve*, which can correspond to either the Serbian noun *sluškinja* 'maid' or one of the verbs expressing necessity (e.g. *služiti* 'to serve'). For clitics, the Serbian language can serve as a control parameter in cases when clitics in both languages are used similarly (such as unstressed personal pronouns and reflexive pronouns).

The multifunctionality of Italian clitics and their dependence on context through referential relationships complicate the development of tagging instructions that can achieve a more reliable degree of accuracy. The congruence of the clitic *ci* with the first-person plural is a common example where *ci* functions as a reflexive pronoun. However, for some verbs, even this criterion is not entirely reliable. For instance, the phrase *ci troviamo* may mean 'we are located', or 'we gather [in a place]', or 'we find ourselves [there]' – in the last case a direct object is necessary, and this instruction should also be embedded in the tagger. Similarly, the reflexive pronoun *si* (which has distinct forms for each person) could be tagged according to its agreement with the corresponding person of the verb. However, even in this case, the third-person *si* can be mis-tagged when taken out of context. For example, in the phrase *Si dice bravo*, the verb can be either reflexive ('He says of himself that he is good') or impersonal ('One says bravo'). The mismatch between reflexive verbs in Italian and Serbian represents an unreliable criterion for extracting tagging parameters, so this aspect of the tagger improvement would likely need to rely solely on elements of the Italian language, possibly with human revision to fine-tune the parameters.

Finally, given the complexity of tagging issues, a possible aid could lie in the integration of AI into the tagging process. In our experience, AI has proved to be enough accurate in identifying the functions of the clitic *ci*, as illustrated in the following response, obtained after a query asking ChatGPT to analyze the functions of the four occurrences of *ci* in the sentence below. The interpretations provided by ChatGPT are linguistically correct and are given in brackets:

42. **Ci** (IMPERSONAL, IDIOMATIC) vuole tanta fatica per riuscir**ci** (PREPOSITIONAL OBJECT), ma se **ci** (LOCATIVE) vai e **ci** (PREPOSITIONAL OBJECT) provi, avrai successo. ('**It takes** a lot of effort to succeed [**in it**], but if you go for it [litt. "**go there**"] and give **it** a try, you'll succeed.')

Further integration of AI with the SerbItaCor3_it corpus and its Serbian translations could prove useful in the tagging process.

References

[1] Bentley, D. (2006). *Split intransitivity in Italian*. Berlin-New York. Mouton de Gruyter.

[2] Dell'Orletta, F. (2009). Ensemble system for part-of-speech tagging. In P. Basile, F. Cutugno, M. N. Malvina, V. Patti, R. Sprugnoli (Eds.), *Proceedings of EVALITA 2009 – Evaluation of NLP and speech tools for Italian* (pp. 1-7). Reggio Emilia, Italy: EVALITA.

[3] Moderc, S., Stanković, R., Tomašević, A., Škorić, M. (2023). An Italian-Serbian sentence-aligned parallel literary corpus. *Review of the National Center for Digitization*, *43*, 1-20.

[4] Moderc, S. (2015). *Gramatika italijanskog jezika: Morfologija s elementima sintakse*. Beograd. Luna crescens.

[5] Moderc, S. (2021a). *I clitici italiani: Usi, ambiguità, interpretazioni. Volume primo: Il sistema dei clitici*. Beogrd. Filološki fakultet.

[6] Moderc, S. (2021b). *I clitici italiani: Usi, ambiguità, interpretazioni. Volume secondo: I nessi di clitici*. Beograd. Filološki fakultet.

[7] Renzi, L. (1988). Renzi, L., Salvi, G., Cardinaletti, A. (eds.). *Grande grammatica italiana di consultazione* (Vol. 1). Bologna. Il Mulino.

[8] Russi, C. (2008). *Italian clitics: An empirical study*. Berlin-New York. Mouton de Gruyter.

[9] Salvi, G. & Vanelli, L. (2004). *Nuova grammatica italiana*. Bologna. Il Mulino.

[10] Schmid, H., Baroni, M., Zanchetta, E., Stein, A. (2007). Il sistema "TreeTagger arricchito" – The enriched TreeTagger system. In B. Magnini & A. Cappelli (Eds.), *EVALITA 2007: Evaluation of NLP tools for Italian* (pp. 22-23). Retrieved from http://www.evalita.it/2007/proceedings [accessed October 24, 2024]

[11] Schmid, H. (2013). Probabilistic part-of-speech tagging using decision trees. D. B. Jones, H. Somers (Eds.). *New methods in language processing*, No. 5. London. Routledge.

[12] Serianni, L. (1989). *Grammatica italiana: Italiano comune e lingua letteraria*. Torino. Utet.

[13] Tamburini, F. (2000). Annotazione grammaticale e lemmatizzazione di corpora in italiano. In R. Rossini (Ed.), *Linguistica e informatica: Multimedialità, corpora e percorsi di apprendimento* (pp. 157-171). Roma. Bulzoni.

[14] Tamburini, F. (2009). PoS-tagging Italian texts with CORISTagger. In *EVALITA 2009: Workshop on evaluation of NLP and speech tools for Italian* (Vol. 1, pp. 1-7). Reggio Emilia, Italy: Accademia University Press.

Izazovi i perspektive tagovanja italijanskih klitika. Studija slučaja na materijalu korpusa SerbItaCor3

Saša Moderc

Sažetak

Italijanski klitici predstavljaju grupu reči koju karakteriše homogeno sintaksičko ponašanje i raznovrsnost funkcija. Njihova referencijalna polivalentnost, s jedne strane, i nekompletne instrukcije u jezičkim tagerima s druge, doprinose nedovoljno preciznom ili čak i pogrešnom označavanju klitika u jezičkim korpusima. Pored klitika, i homografi čine klasu reči koju tageri ne označavaju dovoljno precizno. Na primer, leksema *perdono* može imati značenje imenice (‘oproštaj’), ali predstavlja i oblik glagola *perdonare* (‘opraštati’: *io perdono* ‘ja opraštam’), odnosno glagola *perdere* (‘gubiti’: *loro perdono* ‘oni/one gube’). Program TreeTagger, korišćen za morfološko označavanje reči u dvojezičnom korpusu paralelnih tekstova SerbItaCor3, ne raspolaže dovoljno preciznim instrukcijama za dodeljivanje ispravne oznake homografima. U ovom radu iznosimo pretpostavku da se dvojezični korpus može iskoristiti za podizanje preciznosti u tagiranju homografa, imajući u vidu da je polisemija iz italijanskog teksta leksički razrešena u prevodu na srpski jezik, te se odgovarajuće instrukcije za tagere mogu dedukovati iz korpusa SerbItaCor3. Označavanje polifunkcionalnih klitika *si* i *ci* takođe predstavlja izazov jer pomenuti tager ne sadrži instrukcije za preciznu obradu klitika. Usled strukturnih razlika između italijanskog i srpskog jezika, dvojezični korpus može samo u ograničenoj meri da pruži preciznije instrukcije za prepoznavanje specifičnih funkcija klitika. Stoga je realnija pretpostavka da se za italijanski jezik razviju posebni moduli za postojeći tager, sa posebnim instrukcijama za označavanje klitika *ci* i *si*. S obzirom na sintaksičku i semantičku složenost njihove upotrebe, neophodno je razmotriti i primenu manuelne provere i korekcije ispravnosti tagova koji se dodeljuju kliticima *ci* i *si*. Precizno razrešavanje homografa i ispravno tagovanje klitika, uz integraciju tagera sa potencijalima VI značajno bi povećalo pouzdanost lingvističkih podataka kojim je opremljen korpus SerbItaCor3 i doprinelo bi njegovoj većoj upotrebljivosti u lingvističkim istraživanjima i u didaktici italijanskog jezika.

Ključne reči: italijanski jezik, srpski jezik, SerbItaCor3_it, homografi, klitici, tagovanje, unapređenje tagera