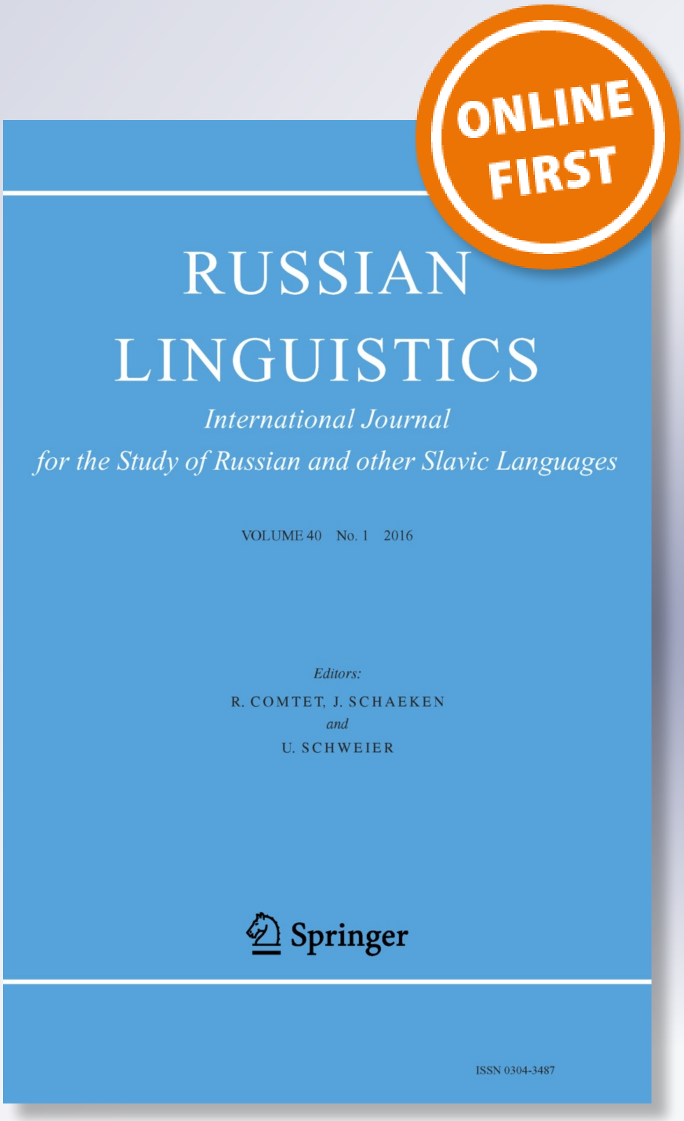
# *Who needs particles? A challenge to the classification of particles as a part of speech in Russian*

Anna Endresen, Laura A. Janda, Robert Reynolds & Francis M. Tyers

**Russian Linguistics**
International Journal for the Study of
Russian and other Slavic Languages

ISSN 0304-3487

Russ Linguist
DOI 10.1007/s11185-016-9160-2



ONLINE FIRST

RUSSIAN LINGUISTICS

*International Journal*
*for the Study of Russian and other Slavic Languages*

VOLUME 40 No. 1 2016

*Editors:*
R. COMTET, J. SCHAEKEN
*and*
U. SCHWEIER

Springer

ISSN 0304-3487

Russ Linguist
DOI 10.1007/s11185-016-9160-2

# Who needs particles? A challenge to the classification of particles as a part of speech in Russian

## Кому нужны частицы? Стоит ли определять частицы как отдельную часть речи в русском языке?

**Anna Endresen**[1] · **Laura A. Janda**[1] · **Robert Reynolds**[1] · **Francis M. Tyers**[1]

**Abstract** In 1985, Zwicky argued that 'particle' is a pretheoretical notion that should be eliminated from linguistic analysis. We propose a reclassification of Russian particles that implements Zwicky's directive. Russian particles lack a coherent conceptual basis as a category and many are ambiguous with respect to part of speech. Our corpus analysis of Russian particles addresses theoretical questions about the cognitive status of parts of speech and practical concerns about how particles should be represented in computational models. We focus on nine high-frequency words commonly classed as particles: *ešče*, *tak*, *ved'*, *slovno*, *daže*, *že*, *li*, *da*, *net*. We show that the current tagging of particles in the manually disambiguated Morphological Standard of the Russian National Corpus is not entirely consistent, and that this can create challenges for training a part-of-speech tagger. We offer an alternative tagging scheme that eliminates the category of 'particle' altogether. We show that our enriched scheme makes it possible for a part-of-speech tagger to achieve more useful results. Our analysis of particles provides a detailed account of various sub-uses that correspond to different parts of speech, their relationships, and relative distribution. In this sense, our study also contributes to the study of words that exhibit part-of-speech ambiguities.

**Аннотация** В работе 1985 года Цвикки утверждал, что 'частица'—это до-теоретическое понятие, которое нужно исключить из лингвистического анализа. Следуя установке Цвикки, мы предлагаем пересмотреть традиционный подход к русским частицам и перераспределить соответствующие слова по другим частеречным классам.

This research was supported in part by grant 222506 from the Norwegian Research Council. The authors would also like to thank their employer, UiT The Arctic University of Norway.

✉ L.A. Janda
laura.janda@uit.no

A. Endresen
anna.endresen@uit.no

R. Reynolds
robert.reynolds@uit.no

F.M. Tyers
fty000@post.uit.no

[1] UiT The Arctic University of Norway, Tromsø, Norway

Published online: 25 April 2016

Ясные содержательные основания для выделения русских частиц как отдельной категории отсутствуют, частеречная принадлежность многих частиц неоднозначна. В нашем корпусном исследовании рассмотрены теоретические вопросы о когнитивном статусе частей речи, а также практические сложности, связанные с представлением частиц в компьютерных моделях обработки данных. В центре внимания девять высокочастотных слов, традиционно определяемых как частицы: *еще*, *так*, *ведь*, *словно*, *даже*, *же*, *ли*, *да*, *нет*. В статье показано, что существующая система частеречной разметки, принятая в Морфологическом стандарте Национального корпуса русского языка (тексты со снятой омонимией), недостаточно последовательна и что это может создать проблемы при обучении частеречного анализатора. В статье предложена альтернативная система разметки, в которой категория 'частиц' как отдельной части речи полностью устранена. Благодаря этой улучшенной системе разметки частеречный анализатор может функционировать более успешно. В статье представлен подробный анализ девяти 'частиц' с разбором основных подтипов их употреблений, которые соответствуют различным частям речи, также обсуждаются взаимосвязи выделенных подтипов и их распределение в использованной выборке примеров. В этом отношении, данное исследование вносит вклад в изучение слов с неоднозначной частеречной принадлежностью.

## 1 Introduction

Who needs particles? While words commonly called 'particles' are robustly attested in Russian, accounting for approximately 4.5 % of all words in a corpus (see Table 1), we argue that their classification as a separate part of speech is not justified. The category of particle lacks a proper definition and is neither informative nor useful. We propose dispensing with the category of 'particle' altogether and reclassifying words according to an enriched scheme of conceptually motivated parts of speech. Our proposal yields an analysis that is linguistically more satisfying and descriptively more precise.

In Sect. 2 we address theoretical and practical problems associated with particles in general as a part of speech category, and detail some specific problems for particles in Russian. We close this section by selecting nine high frequency words commonly classified as particles to focus on in the remainder of the article: *ešče*, *tak*, *ved'*, *slovno*, *daže*, *že*, *li*, *da*, *net*.[1] We undertake a computational experiment in Sect. 3 that highlights a practical problem with assigning words to a 'particle' category in Russian. We find that the current practice for tagging particles in the manually disambiguated Morphological Standard of the Russian National Corpus is not sufficiently consistent to facilitate the training of a reliable automatic part-of-speech tagger. In order to rectify this situation, in Sect. 4 we offer a scheme for classifying the nine words we have identified without resorting to the label 'particle'. This scheme is based on the detailed analysis of a sample of corpus data and classes the nine words as adverb, conjunction, predicative, interjection, emphasizer, and question word. Our scheme is richer and avoids categories that lack conceptual motivation. However, this richness also comes at a cost, since added categories increase the complexity of analysis for this group of words. We implement our scheme in an experiment in Sect. 5 showing that we can achieve better accuracy without the category 'particle'. We summarize our findings and suggest further measures for and implications of eliminating the category of 'particle' from Russian

[1] Since the meanings of these particles are heavily dependent on context, in this article they are not translated when cited out of context.

grammar in Sect. 6. We conclude that without particles we can achieve better linguistic insights that improve the practical performance of natural language processing and might also improve language pedagogy.

## 2 Problems with particles

Particles present problems. These problems are both theoretical and practical, and one subsection is devoted to each domain (Sects. 2.1 and 2.2 respectively). Of course there are parallels across the two domains since a category that lacks a good theoretical description is also likely to run into trouble when we try to implement it in computational linguistics. Russian particles have their own peculiarities, which we examine in more detail in Sect. 2.3, where we also explain the selection of nine words traditionally classed as particles for further study.

### 2.1 Theoretical problems

Zwicky (1985) is a landmark article that specifically addresses the identity of particles, based on both theoretical and typological observations. Zwicky makes a compelling argument that 'particle' should be eliminated from the list of parts of speech for all languages. A major problem with particles is that they are negatively defined: "particles are the words left over when all the others have been assigned to syntactic categories" (Zwicky 1985, p. 292). As we show in Sect. 2.3.3, the situation for Russian particles is entirely in keeping with Zwicky's allegation: definitions of Russian particles consist of statements about what they lack. Instead of recognizing particles, Zwicky calls upon linguists to reassign such words to other syntactic categories, such as adverbs, interjections, and discourse markers. Our analysis in Sect. 4 implements Zwicky's recommendation for Russian particles.

'Particle' is often used as a part-of-speech category alongside other terms such as 'noun', 'pronoun', 'verb', 'adjective', 'adverb', 'numeral', 'conjunction', 'preposition', etc. Before delving further into the role of particles in particular, it is worth asking: What is a part of speech? Although most linguists would agree that the grammar of Russian, for example, contains at least nouns, verbs, adjectives, prepositions and possibly a few more parts of speech, there are several different ways to arrive at a list. One way to approach parts of speech is by examining formal characteristics such as morphological classes, observing that nouns are inflected for case, whereas verbs are inflected for tense and person. Another approach is distributional, based on facts such as that prepositions appear before nouns, pronouns substitute for nouns, and conjunctions bind phrases together. A third strategy highlights semantic differences such as that nouns signify entities and verbs signify situations. Some linguistic traditions emphasize one strategy over another: for example, post-Bloomfieldian American structuralists (Harris 1951; Fries 1952) and generativists (Chomsky 1965) focus on formal and distributional criteria, while cognitive linguists (Wierzbicka 1988, p. 488; Langacker 2013, pp. 115–117) focus on semantic grounds for part-of-speech distinctions. In practice it is likely that linguists combine strategies when identifying parts of speech, and Croft (2001, p. 92) suggests a "conceptual space for parts of speech" that does just that. Ideally, therefore, a part-of-speech category should be justified both in terms of its formal behavior and its semantic content.

While the conceptual space of parts of speech and the discovery procedure for locating items in that space might be universal at an abstract level, the details are language-specific (Croft 2001, pp. 63–107). This means that different languages will have different sets of categories, and that the 'same' categories might not coincide exactly across languages, though the focal points of certain categories, such as 'noun', 'pronoun', 'verb', are typologically more common than others. For the purposes of this article we restrict our purview to Russian.

Having categories does not necessarily entail that the categories are discrete, and indeed Croft's conceptual space makes room for categories to overlap. Langacker (2013, p. 96) cautions that while many part-of-speech categories are unavoidable because they are ubiquitous, we should not take these categories for granted: "Traditional terms lack precise definition, are inconsistent in their applications, and are generally inadequate". Like other linguistic categories, parts of speech have a network structure with prototypical members as well as less prototypical members that may overlap with other categories. A verb, for example, profiles a temporal relationship, which is prototypically an event in time conceptually dependent on its participants (Langacker 2013, pp. 108–112). In addition to this semantic characterization, a verb also appears in certain constructions; in Russian this would include the transitive construction, the intransitive construction, and various impersonal constructions. Non-prototypical instantiations of verbs can overlap with other parts of speech. In Russian we see this in the case of participles that can be interpreted both as non-finite forms of verbs and as adjectives. For example, the participle *vydajuščijsja* has largely been lexicalized as an adjective in the meaning 'remarkable' and is rarely used to literally mean 'jutting out'. *Blestjaščij* 'shining' is arguably usable both as a participle (in *blestjaščie glaza* 'shining eyes', where the eyes are literally shining) and as an adjective (in *blestjaščaja pobeda* 'shining [= remarkable] victory', where the victory is not literally shining). A number of lexemes straddle the boundary between adverbs and prepositions, such as *vokrug* 'around', which is an adverb in *osmatrivat'sja vokrug* 'look around', but a preposition in *putešestvie vokrug sveta* 'journey around the world'. More marginal examples can be found in non-inflected diminutive words like *kušan'ki* 'eat' that can be classified as both a verb (*budem kušan'ki kruasančiki* 'we will eat croissants') and a noun (*ja prigotovila kušan'ki* 'I prepared food') (Makarova 2015).

Until the end of the 19th century, Russian grammarians used the notion of particle in a broad sense: the term *časticy reči* (lit. 'particles of speech') was applied to all function words (including conjunctions and prepositions) as opposed to referential words (cf. grammars by Lomonosov, Vostokov, and Sobolevskij; see Vikul'ceva 2004, p. 8 for details). In the 20th century, Šaxmatov (1941, p. 506) was the first to define particles as a separate part of speech. Bogorodickij (1939, p. 200) explicitly distinguished particles (*časticy*) from conjunctions and prepositions, and the tradition of identifying particles as a separate part of speech becomes solidly established in the works of Vinogradov and his theory of modality. Vinogradov (1972, p. 520) defined particles as words that lack referential content and contribute additional nuances to the semantics of other words, phrases or clauses or serve to express various grammatical, logical, or evocative relationships. In his authoritative grammar of Russian however, Timberlake (2004, pp. 463–465) does use the word 'particle', but only in reference to the word *li* in questions and implied questions. In the same section Timberlake refers to *da* and *net* as 'polarity words' and briefly describes their syntax.

Some other linguistic traditions do not interpret discourse markers that correspond to Russian particles as a separate part of speech. Vikul'ceva (2004, p. 112) observes that in grammars of Italian, for example, the term 'particle' ('particella') refers to any uninflected unstressed lexeme that has a pragmatic function. When the term 'particle' is applied, one normally specifies the part of speech it belongs to: e.g. *ne* is a 'pronominal particle' ('particella pronominale', Serianni and Castelvecchi 1997, p. 353), *pure* is a 'conjunctional and adverbial particle' ('particella congiuntiva ed avverbiale', Miot 1987, p. 607). Likewise, grammars of Estonian do not consider particles to be a distinct part of speech and apply the term 'particle' ('partikkel') to all uninflected words including adverbs, prepositions, conjunctions, etc.

(Vikul'ceva 2004, p. 112; Tauli 1972, p. 91). Russian particles are functionally similar to Estonian lexemes that are classified as modal and focus adverbs, i.e. 'modaaladverbid' and 'rõhumäärsõnad' (Erelt, Erelt and Ross 2000, pp. 145–146).

Recent work by Drummen (2015, pp. 80–86) on Ancient Greek shows that it is fruitful to go beyond the assumption that particles are a discrete class of words and account for the multifunctionality of Greek particles by examining constructions in which these particles perform functions analogous to those of other parts of speech (see Drummen 2015, p. 40 for applications of construction grammar to discourse markers for a range of languages). For Drummen, particles may be problematic as a part of speech, but particles clearly share a discourse function. Similarly, the authors in Andersen and Fretheim (2000) argue that various particles in Amharic, English, German, Hausa, Hungarian, Japanese, Modern Greek, Norwegian, Occitan, and Swahili function as pragmatic markers that encode the speaker's attitude.

There are some scholars who, while they do not posit particles as a separate part of speech, do recognize them as a distinct group of discourse markers. In their introduction to a volume dedicated to particles in South Slavic languages, Dedaić and Mišković-Luković (2010) note that particles have not received enough attention in the scholarly literature and that South Slavic languages are replete with such words that serve discourse functions. Instead of claiming that particles are a separate part of speech, they find it useful to consider them as a 'class' of discourse markers that serve "as pointers to the ways the basic proposition or message should be taken by the addressee" (Dedaić and Mišković-Luković 2010, p. 2). Brinton (1996, pp. 33–34) makes similar observations for what she calls 'pragmatic markers' in English.

Wierzbicka (1976) offers a Leibnizian semantic interpretation of three items she identifies as 'particles': English *well* and Polish *no* and *że*. According to Wierzbicka, particles serve a social role, conveying a speaker's attitudes toward the hearer or speech situation; in other words, they express illocutionary forces. She reasons that since "a particle condenses in itself an entire sentence" (Wierzbicka 1976, p. 328), the best way to capture the semantics of particles is by 'reconstructing' a particle's sentence. For example, a semantic component of both Polish *no* and English *well* is reconstructed by Wierzbicka as 'I don't want more time to pass like this'.

As we argue in Sect. 2.3, the category of particle in Russian lacks coherence both in terms of its formal behavior and its conceptual content. This makes it impossible to specify a prototype that could serve to define even a fuzzy category. While overlap at the periphery is expected for marginal members of part-of-speech categories, like any linguistic categories, the extent of claimed overlap with other parts of speech is rather extreme for particles, adding to the suspicion that 'particle' is not a valid category. Like Zwicky (1985), we have not been able to find any justification for defining the particle as a part of speech.

## 2.2 Practical problems

Language users, learners, and researchers increasingly rely on electronic tools such as spelling and grammar checkers, machine translation, intelligent computer assisted language learning (ICALL) software, and linguistic corpora. These tools are usually sourced at least in part by some kind of natural language processing (NLP), and a key element in NLP is the assignment of part-of-speech tags to each token. Typically an automatic tagger is trained on a manually disambiguated 'gold standard' corpus like the Morphological Standard for the Russian National Corpus (see: http://ruscorpora.ru/corpora-usage.html: henceforth: 'RNC gold standard').

Manning (2011), in an evaluation of tagging of the Penn Treebank of English, makes the case that per-token accuracies of 97 % in automatic part-of-speech tagging give us a false sense of security. A more realistic evaluation measure is the portion of entire sentences that get tagged correctly since a single part-of-speech error can foul up the dependency parsing of a whole sentence. The same automatic tagger of English that achieves 97 % per-token accuracy has only 55–57 % sentence accuracy. This means that there is a lot of incentive to eliminate the remaining part-of-speech tagging errors since this has the potential to vastly improve NLP.

According to Manning, the major obstacle to eliminating part-of-speech tagging errors is inconsistency in the gold standard that the tagger is trained on. Other sources of trouble for taggers include situations where the local context available to the tagger is insufficient to disambiguate the part of speech, the tag itself is ambiguous, or there is a mistake in the gold standard corpus.

Of course Russian is not English, but the basic task of assigning parts of speech is roughly similar, except that the task is harder and automatic tagging is less successful in Russian. In both languages, the presence of even one part-of-speech error in a sentence can mean that it will not be parsed properly, so these errors have serious consequences (cf. Manning 2011).

Given this situation, one possible strategy is to focus on the part of speech that is associated with the most tagging errors. As we argue in Sect. 2.3, a highly error-prone part of speech in Russian is the class of words called ‘particles’, and the main culprit is the same one that Manning identified for English, namely inconsistency in the tagging of the gold standard corpus.

## 2.3 Russian particles in particular

In order to gain perspective on Russian particles, we look both at the extent to which particles are attested and at how they are identified. Identification of Russian particles is problematic in part due to ambiguity. Russian words that are classed as particles can often be classed as other parts of speech as well, similar to *blestjaščij* ‘shining’ (participle and adjective), *vokrug* ‘around’ (adverb and preposition) and *kušan’ki* ‘eat’ (verb and noun) discussed in Sect. 2.1. We have selected a subset of Russian particles that optimally represent the frequency and ambiguity of this group of words for further study.

### *2.3.1 Extent of particles in Russian*[2]

Estimates of the number of Russian particles vary. Zaliznjak (1980) designates over 100 Russian words as particles, and Russian part-of-speech tagging is typically based on this authoritative source. Nikolaeva (1985, p. 8) lists the following alternative counts: 131 particles in the 17-volume *Akademičeskij slovar’*, 110 in the *Malyj akademičeskij slovar’*, 84 in Ušakov’s *Tolkovyj slovar’*, and 75 in Ožegov’s *Tolkovyj slovar’*. Nikolaeva (citing Bartoševič 1978, p. 332) notes that only 42 particles appear in all four of these sources, while 64 of them are found in only one source each, meaning that the list of lexemes is inconsistent and unstable from one reference source to another. Starodumova (1997, pp. 8–9) claims that Russian is

[2] Abbreviations used in this paper: A—adjective, A-NUM—numeral adjective, A-PRO—pronominal adjective, ADV—adverb, ADV-PRO—pronominal adverb, CNJ—conjunction, CNJADV—adverbial conjunction, CNJCOO—coordinating conjunction, CNJSUB—subordinating conjunction, EMPH—emphasizer, INTJ—interjection, NEST—negative predicative of existence, NUM—numeral, PARENTH—parenthetical, PART—particle, PR—preposition, PRAEDIC—predicative, PRAEDIC-PRO—predicative pronoun, QST—question word, S—substantive, S-PRO—pronoun, V—verb, VMOD—modal verb.

**Table 1** Parts of speech as represented in the RNC disambiguated subcorpus (total size = 5,944,156 tokens)

| Part of speech | # of tokens in RNC disambiguated subcorpus | % tokens in RNC disambiguated subcorpus |
|---|---|---|
| S | 1,707,312 | 28.7 |
| V | 1,007,526 | 16.9 |
| A+A-PRO | 784,340 | 13.2 |
| PR | 621,857 | 10.5 |
| S-PRO | 467,440 | 7.9 |
| CNJ | 471,275 | 7.9 |
| ADV+ADV-PRO | 375,740 | 6.3 |
| PART | 268,139 | 4.5 |
| NUM+A-NUM | 126,567 | 2.1 |
| PRAEDIC+PRAEDIC-PRO | 42,998 | 0.7 |
| PARENTH | 25,891 | 0.4 |
| INTJ | 8,377 | 0.1 |
| *Sum* | | *99.2* |

among the most 'particle-rich' (*časticeobil'nyj*) languages in the world, with approximately 300 particles.

Knowledge of how and where to appropriately use discourse particles constitutes an important part of the linguistic competence characteristic of native speakers and is difficult to acquire for second language learners ("[A]ktivnoe upotreblenie častic est' odin iz pokazatelej znanija jazyka" 'Active use of particles is an indicator of language proficiency' Nikolaeva 1985, p. 7; also cf. Heinrichs 1981, p. 3 for discussion). Without particles, spoken Russian can sound impolite and 'dry', lacking engagement with the interlocutor (Nikolaeva 1985, p. 13). Wierzbicka (1992, p. 396, p. 433) claims that the rich system of particles plays a prominent role in conveying the open expression of feelings that is characteristic of Russian culture. Moreover, Russian particles are claimed to be the words most responsible for successful and effective communication (Nikolaeva 1985, p. 14).

Enough particles are of sufficiently high frequency to place 'particle' as the eighth most common part of speech in Russian. Table 1 reports the frequencies of all parts of speech tagged in the manually disambiguated subcorpus of the RNC (henceforth 'RNC disambiguated subcorpus'), which contains nearly 6 million tokens.

Particles are less common than nouns, verbs, adjectives, prepositions, pronouns, conjunctions, and adverbs. However, particles outrank numerals, predicatives, parenthetical words, and interjections. Clearly Russian particles are frequent enough to matter significantly in part-of-speech tagging.

Scholars often claim that higher use of particles is a characteristic of spontaneous spoken Russian (for example, Vasilyeva 1972). In a comparison of particles in the manually disambiguated subcorpus of the RNC, we did find that there are more words tagged as particles in the spoken subcorpus than in the written corpus. However, while this difference is statistically significant (chi-squared 4081, $df = 1$, $p < 2.2\text{e-}16$), it is also very small (Cramer's $V = 0.026$), in fact an order of magnitude below the threshold of what is standardly considered a reportable effect size (the minimum standard value for a small effect size is Cramer's $V = 0.1$).

The fact that the difference between written vs. spoken texts in terms of the use of particles is so small is surprising and counterintuitive. A possible explanation might be that the

**Table 2** Use of particles in RNC written and spoken disambiguated subcorpora

| | Total # of words | # of particles | % particles |
|---|---|---|---|
| Spoken subcorpus | 216,112 | 16,165 | 7.4 |
| Written subcorpus | 5,728,044 | 251,974 | 4.4 |

spoken subcorpus underrepresents informal types of speech communication such as dialogs that feature the use of particles. At the present time (August 2015) the spoken subcorpus is rather unbalanced: spoken public speech yields 192,275 words (= 89 %) as opposed to only 16,955 words (= 7.8 %) of spoken non-public (informal) speech. The issue of underrepresentation of informal speech in the spoken subcorpus has been raised in the literature (Grišina and Savčuk 2009, p. 147). This is particularly important because the genres of public speech that predominate in the spoken corpus include lectures, discussions, parliamentary speeches, conferences, and TV interviews. According to Voejkova (2009, p. 361), the majority of these spoken texts are monologues and lack the properties of spontaneous speech (during which people might interrupt each other and finish each other's replies, etc.).

Recall that particles have a privileged status in dialog (Starodumova 1997) and are claimed to be responsible for successful communication (Nikolaeva 1985, p. 14). Therefore, we may hypothesize that an underrepresentation of dialogic informal speech genres in the spoken subcorpus might be the reason why particles do not score higher in terms of token frequency in the overall picture in Tables 1–2. In other words, this is a feature of the spoken subcorpus in its current shape rather than a feature of Modern Russian.

However, in the data available for this study there is insufficient justification for distinguishing between the use of particles in speech vs. text, and consequently in the remainder of this article we combine results for both types of data.

### *2.3.2 How Russian particles are (not) defined*

According to Švedova (1980, §1689), Russian particles are uninflected words lacking referential content ("neizmenjaemye neznamenatel'nye [...] slova"). An absence of inflection is of course not an identifying feature since most Russian parts of speech (all except verbs, pronouns, and numerals)[3] contain some uninflected lexemes, and several parts of speech consist entirely of uninflected words, such as prepositions, conjunctions, adverbs, and interjections. Although she brands particles as lacking referential content, Švedova (1980) claims that particles do express meaning. However the meanings of particles are maximally heterogeneous ("samye raznoobraznye"), including various modal and pragmatic attitudes toward propositions. Starodumova (1997, p. 8) observes that for particles we have only a negative definition for a set of words that lack the basic formal characteristics of a category.

A further negative characteristic of particles is that they cannot be discretely distinguished from other parts of speech, a point that Švedova (1980, §1690) makes repeatedly: "vse ėti časticy imejut tesnye vnešnie i vnutrennie svjazi s drugimi klassami slov" 'all these particles have close internal and external ties to other parts of speech'; "Mnogie časticy po svoemu značeniju i po svoim sintaktičeskim funkcijam ne protivostojat rezko slovam drugix klassov" 'many particles in their meaning and syntactic functions are not strictly distinct from other parts of speech' (ibid., §1699). Among the parts of speech most often mentioned as overlapping with particles are adverbs, conjunctions, and interjections, though

[3] Even in the case of verbs and pronouns, one can find marginal examples of lexemes that are uninflected, such as the verbal interjective forms *xvat'* (< *xvatat'* 'grab'), *pryg* (< *prygat'* 'jump'), *spaten'ki* (< *spat'* 'sleep') (Klobukov 2001); adverbial pronouns *tak* 'so', *tam* 'there', *tut* 'here', *gde* 'where', *kogda* 'when'.

**Table 3** Extent of ambiguity of particles[a]

| | Part of speech in addition to PART | # of lexemes | | Illustrative examples |
|---|---|---|---|---|
| Unambiguous | – | | 54 | *vot* 'look!, here', *by* 'would' |
| 2-way ambiguity | CNJ | 28 | 65 | *ved'* 'indeed', *budto* 'as if', *i* 'and' |
| | ADV | 17 | | *ešče* 'still, in addition, more', *von* 'there, out', *ladno* 'ok' |
| | INTJ | 8 | | *nu* 'well', *aga* 'aha', *iš'* 'bah' |
| | PRAEDIC | 7 | | *net* 'no', *amin'* 'amen' |
| | PARENTH | 4 | | *požaluj* 'perhaps', *avos'* 'maybe' |
| | PR | 1 | | *vrode* 'like, sort of' |
| 3-way ambiguity | ADV; PRAEDIC | 5 | 12 | *prosto* 'simply', *klassno* 'cool' |
| | ADV; CNJ | 5 | | *poka* 'meanwhile', *kak* 'how', *tol'ko* 'only' |
| | PARENTH; PRAEDIC | 1 | | *spasibo* 'thank you' |
| | PARENTH; ADV | 1 | | *nikak* 'no way' |
| 4-way ambiguity | PARENTH; ADV; PRAEDIC | | 1 | *xorošo* 'good' |
| | PARENTH; ADV; CNJ | | 1 | *točno* 'precisely' |

[a] Note that this table is restricted to unambiguously uninflected lexemes; it does not include items that are interpreted by some as particles but have alternative interpretations as part of a paradigm such as *ėto*, which is a form of a pronoun

other options include predicatives, parenthetical expressions, and prepositions. Table 3 gives a breakdown of the 133 particles listed in grammatical dictionaries (Zaliznjak 1980; Grišina and Ljaševskaja 2008) according to their possible part-of-speech ambiguities.

While there are 54 lexemes like *vot* 'look, here' that are identified only as particles, 65 particles (49 %) are two-way ambiguous, since they can also be interpreted as other parts of speech: 28 as conjunctions, 17 as adverbs, etc. (see Table 3). If a particle has two additional possible part-of-speech interpretations, we say that it is three-way ambiguous, and there are twelve lexemes of this type. Four-way ambiguity, involving three additional part-of-speech interpretations, is found for only two lexemes: *xorošo* 'good' and *točno* 'precisely'. Our nine focus particles, presented in Sect. 2.3.3, represent three of the most common types of two-way ambiguity, namely of a particle with a conjunction, adverb, or predicative.

Švedova makes no attempt at suggesting strategies for disambiguating particles from other parts of speech. Starodumova (1997, p. 8) claims that a hallmark of particles is "gibridnost' " 'hybridity', which she defines as the combination of the function of a particle with another part-of-speech function in one and the same use.

In sum, Russian particles have no coherent profile by any measure: morphological, semantic, or syntactic. 'Particle' looks like a garbage category that is used when one feels uncertain about how to classify a word. In other words, 'particle' is not a classification, but rather a failure to classify a word. And classification as 'particle' doesn't yield any meaningful information that can be utilized in further analysis or applications. Even when the identification of a lexeme as a particle is unambiguous, this is not a very informative result because 'particle' has no positively defined semantic, formal, or behavioral profile. Worse yet, particles are systematically indistinguishable from other parts of speech.

It is no surprise that annotation guidelines for particles are problematic as well. Sičinava (2005) offers guidelines for the manual tagging of the RNC gold standard, but addresses particles only in one small paragraph devoted to one subtype of two-way ambiguity, namely when a lexeme can be identified as both a particle and a conjunction, like *ved'* in Table 3. Here we cite the entire paragraph (in our translation from Russian):

> Some words can be both particles and conjunctions. A conjunction introduces an entire clause and as a rule stands at its beginning. A conjunction bears an additional meaning (explanatory, adversative…) that is not shared by the particle. The sphere of influence for a particle is only a part of a clause. Here is a list of some words of this kind: *budto*, *ved'*, *daže*, *že*, *li*, *liš'*, *pust'*, *rovno*, *slovno*, *točno*, *xot'*, *jakoby*.
>
> (Sičinava 2005, p. 151)

In light of these instructions, which involve the function, position, and meaning of lexemes, consider the following two examples of *ved'* from the RNC gold standard. In example (1), *ved'* is tagged as a particle, as in 3,890 other examples. Example (2), by contrast, is one of 1,360 examples where *ved'* is tagged as a conjunction:

(1) *Ved'* $^{\text{part}}$ *vy ne znaete*, možet, on na vas takoe nagovoril…
‘*But you don't know*, maybe he has made up a story about you…’
(Ju. O. Dombrovskij. *Fakul'tet nenužnyx veščej*. 1978)[4]

(2) *Ved'* $^{\text{cnj}}$ *vy ne znaete* goroda…
‘*But you don't know* the city…’
(M. A. Bulgakov. *Master i Margarita*. 1929–1940)

The function, position, and meaning of the lexeme *ved'* is arguably the same in these two examples, yet they have received different tags. Examples like this are not difficult to find, nor are they limited to the particle vs. conjunction ambiguity. As shown in Table 3, *ešče* has another two-way ambiguity, namely with particle vs. adverb. Example (3) contains one of 911 attestations of *ešče* tagged as a particle in the RNC gold standard, while example (4) contains the same lexeme in the same position and collocated with the same two words, yet is among 13,871 attestations of *ešče* tagged as an adverb:

(3) *Ešče*$^{\text{part}}$ *odin primer* – trava na kartinax.
‘*One more example* – grass in the pictures.’
(Solomennye kartiny. *Narodnoe tvorčestvo*. 2004)

(4) *Ešče*$^{\text{adv}}$ *odin primer*: “Samo nazvanie Jaro-slavl', verojatno, označalo kogda-to ‘Slavnyj Jar’.”
‘*One more example*: “The very name *Jaro-slavl'* probably once meant ‘Glorious Ravine’.” ’
(A. A. Zaliznjak. Lingvistika po A. T. Fomenko. *Voprosy jazykoznanija*. 2000)[5]

These examples give us anecdotal evidence that there are inconsistencies in the tagging of lexemes often classed as particles in the RNC gold standard. We do not mean to imply that the annotators of the RNC gold standard have been careless or inadequate in any way. What we see here is that they have struggled with a truly difficult problem that has lacked a satisfactory solution. We hope that the present study will contribute to improvements in the valuable resource the RNC gold standard represents.

[4] Most of the examples we cite are taken from the database used for our experiments (The Tromsø Repository of Language and Linguistics, http://hdl.handle.net/10037.1/10291). Because our experiments were run on examples containing only one token of the nine lexemes in our study, example (48) was excluded from the database because it contained more than one token. In addition, examples (5) and (49) illustrate low frequent or obsolete uses of particles that were not represented in our database.

[5] Note that in this example, Zaliznjak is citing Fomenko's words.

**Table 4** High-frequency particles with one additional part-of-speech reading according to the RNC disambiguated subcorpus

| Lexeme | Frequency in RNC disambiguated subcorpus | ADV | CNJ | PRAEDIC | PART |
|---|---|---|---|---|---|
| *ešče* | 14,765 | ADV | | | PART |
| *tak* | 22,093 | ADV | | | PART |
| *ved'* | 5,149 | | CNJ | | PART |
| *slovno* | 1,369 | | CNJ | | PART |
| *daže* | 8,562 | | CNJ | | PART |
| *že* | 21,350 | | CNJ | | PART |
| *li* | 7,708 | | CNJ | | PART |
| *da* | 12,280 | | CNJ | | PART |
| *net* | 9,786 | | | PRAEDIC | PART |

### *2.3.3 Our nine focus particles*

Most of the objections to particles we have raised thus far have been polemical. What does this situation mean for the actual performance of an NLP model in analyzing Russian? In order to concretely assess the dimensions of the problem, it is necessary to focus on a subset of particles that can represent the group as a whole.

In order to obtain a representative subset of particles, we selected lexemes according to their frequency and degree of ambiguity with other parts of speech. Many particles found in Zaliznjak (1980) are attested very infrequently or not at all in the RNC gold standard. We focus on high-frequency particles that can give us enough data for an analysis.

Since part-of-speech ambiguity is rampant among particles, this feature should be represented in our subset. Similar ambiguity makes it possible to compare across items, so we have chosen lexemes that share the same level of ambiguity, namely those that have just two possible designations, one as particle, and one as another part of speech. Table 4 shows the particles that we have selected for further study on the basis of high frequency and two-way ambiguity.

The lexeme *ešče* appears 14,765 times in the RNC disambiguated subcorpus and is tagged as both an adverb (ADV) and a particle (PART). *Tak* has the same type of ambiguity, but is more frequent. Six of our words, *ved'*, *slovno*, *daže*, *že* and *li*, are tagged both as conjunctions (CNJ) and as particles. The last word in our group is *net*, which the RNC gold standard tags as both a predicative (PRAEDIC) and a particle. Sections 3–5 will focus on the lexemes in Table 4, subjecting them to two experiments and offering improved annotation guidelines that eliminate the class of particles altogether.

## 3 What happens if we try to tag particles: experiment 1

The first thing we want to find out is how well the current system of tagging words classed as particles works. Given the challenges of rather meagre instructions and examples like (1)–(4), we cannot expect a high degree of consistency for the RNC gold standard that an automatic tagger could be trained on. Without a consistent model, a tagger cannot produce reliable and useful results even if we assume that all the tags themselves are useful. We test the RNC gold standard tagging using the nine words in our study.

**Table 5** Distribution of original tags in RNC gold standard for our database

| Lexeme | # of ADV | # of CNJ | # of PRAEDIC | # of PART | Total # of examples |
|---|---|---|---|---|---|
| *ešče* | 83 | | | 17 | 100 |
| *tak* | 100 | | | 0 | 100 |
| *ved'* | | 33 | | 67 | 100 |
| *slovno* | | 83 | | 17 | 100 |
| *daže* | | 16 | | 84 | 100 |
| *že* | | 6 | | 94 | 100 |
| *li* | | 18 | | 82 | 100 |
| *da* | | 54 | | 46 | 100 |
| *net* | | | 58 | 42 | 100 |

Our strategy is to construct a database by extracting 100 random sentences for each of the nine focus words from the RNC gold standard. This database is used for both training and testing a Hidden Markov Model (HMM) trigram tagger (Halácsy, Kornai and Oravecz 2007), which is the standard model for training part-of-speech tagging. We then divide this database into ten chunks and perform a ten-fold cross-validation, each time using 90 sentences as the training set and 10 sentences as the test set. This means that each part of the total set is tested in the course of the ten repetitions of training and testing. Our entire database of sentences and the statistical analysis for both Experiment 1 and Experiment 2 are publicly available at http://hdl.handle.net/10037.1/10291.

Table 5 gives an overview of the database of sentences for our experiments and the distribution of tags that the two-way ambiguous lexemes were assigned in the RNC gold standard.

Our random sample did not include an example of *tak* classed as a particle, which is perhaps not surprising since only 1 % of examples of *tak* have received the PART tag in the RNC disambiguated subcorpus. As a result, getting a correct answer for *tak* is a trivial issue for Experiment 1. For the remainder of our focus words, there are two ways of determining what baseline or chance performance is. One is to say that in all cases (except *tak*) the baseline is 50 % since the tagger always has a choice between two items. An alternative that is more conservative in measuring performance gain and more appropriate for our purposes (since the tagger sometimes chooses among more options in Experiment 2) is to assume that the baseline is the frequency of the most common item. This is the best that one could achieve by simply guessing the most frequent item every time. In other words, if the tagger always chose ADV for *ešče*, it would be correct 83 % of the time (since there are 83 examples of *ešče* tagged as an adverb in our sample), and if it always chose CNJ for *da*, it would be correct 54 % of the time (since there are 54 examples of *da* tagged as an adverb in our sample). We want to know how much better our HMM tagger performs in comparison with a simple guess of the most frequent item, since this will show us how much it can learn from the existing gold standard tags. We will gauge the accuracy of the tagger in terms of its gain over baseline. Due to the necessarily limited size of our sample, it was not possible to measure sentence accuracy.

Table 6 summarizes the outcome of Experiment 1 visualized in Fig. 1. Accuracy is measured as the number of correct guesses in a testing trial over the total number of guesses.[6]

We can ignore the results for *tak* since tagging this word presented no challenge; all examples were of adverbial use according to the RNC gold standard. The overall average gain

[6] A full account of the outcomes of all ten trials for each word is presented in Table A of the Appendix.

**Table 6** Outcome of Experiment 1

| Lexeme | Baseline in % | Average accuracy in % | Gain over Baseline (percentage points) |
|---|---|---|---|
| *ešče* | 83 | 70 | −13 |
| *tak* | 100 | 100 | 0 |
| *ved'* | 67 | 78 | +11 |
| *slovno* | 83 | 89 | +6 |
| *daže* | 84 | 90 | +6 |
| *že* | 94 | 89 | −5 |
| *li* | 82 | 76 | −6 |
| *da* | 54 | 75 | +21 |
| *net* | 58 | 89 | +31 |

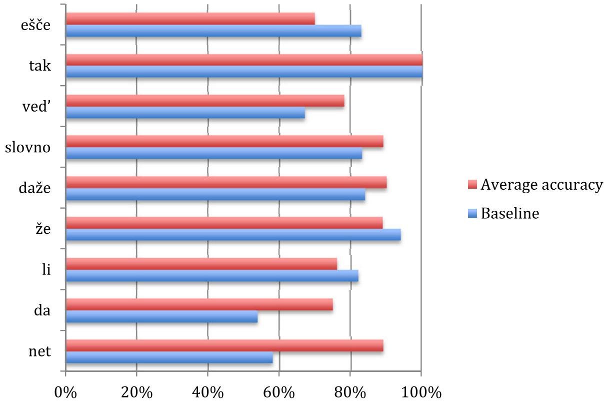


**Fig. 1** Outcome of Experiment 1

for the remaining eight lexemes is 6.4 %, but most of that gain is made up by *da* and *net*, where the HMM tagger clearly did gain an advantage by learning from the tags in the RNC gold standard. If we look at the results for *ešče*, *ved'*, *slovno*, *daže*, *že* and *li*, it is more of a mixed bag, with half of them actually showing worse results for the tagger than the baseline of a simple guess of the most frequent item.

The results of Experiment 1 confirm our suspicion that the tagging in the RNC gold standard is not very consistent in classifying the part of speech of some words (especially *ešče*, *ved'*, *slovno*, *daže*, *že* and *li*) that are considered ambiguous between particle and adverb and conjunction. Of course we must add the fact that 'particle' is not a particularly useful classification, even when it is supposedly correct.

## 4 Particle-free annotation

Can we eliminate particles from the part-of-speech classification of Russian? In this section we propose the reclassification of our nine words represented in Table 7 and detailed below. In this classification scheme, all categories are meaningful and useful for further applications. In addition, we provide a detailed explanation of how these tags should be assigned.

The numbers in Table 7 indicate the number of examples in our database of 100 randomly extracted sentences for each lexeme that received each tag in our scheme. These are the same

**Table 7** Proposed tagging scheme

| | ADV | CNJ | PRAEDIC | INTJ | EMPH | QST |
|---|---|---|---|---|---|---|
| *ešče* | ADV 100 | CNJADV 0 | | | | |
| *tak* | ADV 84 | CNJADV 7<br>CNJSUB 8 | | INTJ 1 | | |
| *ved'* | ADV 57 | CNJADV 33<br>CNJSUB 10 | | | | |
| *slovno* | ADV 49 | CNJCOO 51 | | | | |
| *daže* | ADV 85 | CNJCOO 15 | | | | |
| *že* | | CNJADV 13<br>CNJCOO 6 | | | EMPH 81 | |
| *li* | ADV 23 | CNJCOO 6<br>CNJSUB 0 | | | | QST 71 |
| *da* | ADV 19 | CNJCOO 25 | PRAEDIC 3<br>VMOD 3 | INTJ 50 | | |
| *net* | | | NEST 60<br>PRAEDIC 10 | INTJ 30 | | |

examples as those reported in Table 5, there with their original tags from the RNC gold standard. For example, *da* according to our scheme is tagged as an adverb in 19 examples, as a coordinating conjunction in 25 examples, as a predicative in 3 examples, as a modal verb in 3 examples, and as an interjection in 50 examples.

Our scheme is both more ambitious and more complex than that employed by the RNC gold standard. Shading in Table 7 indicates the tags that correspond to tags for the same words in the RNC gold standard. The adverb tag is preserved as a tag for *ešče* and *tak*, but also recognized for *ved'*, *slovno*, *daže*, *da*, and *li*. We distinguish between the types of conjunctions on the basis of: 1. the syntactic optionality vs. obligatoriness of the word, 2. the semantic contribution, and 3. replaceability of the word with semantically equivalent conjunctions. An adverbial conjunction (CNJADV; cf. example (15)) is more optional than a coordinating (CNJCOO; cf. example (29)) or subordinating conjunction (CNJSUB; cf. example (24)). An adverbial conjunction participates in the juxtaposition of two propositions or clauses (this function distinguishes it from an adverb), but at the same time this word is optional and can be excluded without causing any syntactic disruption of the sentence. By contrast, coordinating and subordinating conjunctions create an explicit contrast between syntactic constituents. The predicative classification is likewise further differentiated into three types, including special classifications for modal verbs (VMOD) and something we call 'NEST' (see Sect. 4.9 below). Interjection is an already existing part-of-speech tag in the RNC, and here we have extended its application to *tak*, *da*, and *net*. We have added two more classifications: emphasizer for *že*, and question word for *li*. The classification scheme is detailed in Sects. 4.1–4.9 and illustrated with examples from the sentences in our database used in the two experiments. Note that the classifications in the following subsections are based primarily on the 100 examples of each lexeme in our sample, supplemented where necessary with examples from grammars and dictionaries.

### 4.1 *ešče*: ADV 100, CNJADV 0

We tagged all 100 examples of *ešče* in our sample as adverbs (described below), however, the use of this word as a coordinating conjunction, which is also described as a 'concessive conjunction' in the Malyj akademičeskij slovar' (1999), is not uncommon in spoken Russian, as in (5). In examples of this type, *ešče* serves to combine and contrast two propositions (here the physical development of two individuals) where one is considerably worse than the other:

(5) Mne *ešče*$^{\text{cnjadv}}$ povezlo. Mama govorit / čto posle vojny roždalis' mladency bez nogtej i volos.
'I, *however*, was lucky. Mama says that after the war babies were born without fingernails and hair.'
(V. Basov and V. Koževnikov. *Ščit i meč*. 1968)

However, aside from this use as a coordinating conjunction, we suggest that in all other uses *ešče* should be considered an adverb and that there is strong unity to this lexeme (cf. Malyj akademičeskij slovar' 1999, which also lists *ešče* as an adverb in the following five uses; and Percov 2002 who likewise presents *ešče* as semantically coherent). The polysemous nature of *ešče* can be described in terms of a radial category with a prototype referring to the addition of items and extensions to the domains of time and qualities. The prototypical meaning is 'in addition, more' and illustrated in (6). This is also the meaning that is most frequently attested for *ešče*:

(6) Možno ja *ešče*$^{\text{adv}}$ nemnožko dobavlju k skazannomu.
'Maybe I can add something *more* to what has been said.'
(Beseda v Moskve. Fond *Obščestvennoe mnenie*. 2003)

An extension to the domain of time yields meanings equivalent to 'still' (7), 'yet' (8), and 'as early as' (9); in the domain of qualities, *ešče* presents a comparative degree, as in (10):

(7) Šrėk… vse *ešče*$^{\text{adv}}$ ostaetsja v Amerike vtorym po kasse fil'mom *prošlogo goda*…
'Shrek… *still* remains the second-highest box-office film in America from last year…'
(Detskij sad (2002). *Izvestija*, 2002.02.14)

(8) Vozmožno potomu, čto moja žena *ešče*$^{\text{adv}}$ ne razljubila menja *okončatel'no*.
'Possibly because my wife has not *yet* completely stopped loving me.'
(A. Slapovskij. *Žizn' Lagarpova*. 1999)

(9) No v uzkoj pribrežnoj polose,… rastut "prišel'cy" iz sredizemnomor'ja, pronikšie sjuda po morskomu beregu *ešče*$^{\text{adv}}$ v dalekom prošlom.
'But in the narrow coastal zone,… there are Mediterranean "intruders" growing that found their way here along the sea coast *already* in the distant past.'
(Ju. N. Karpun. *Priroda rajona Soči*. 1997)

(10) Na smenu ej vse čašče prixodit ėlektronika, kotoraja kontroliruet gidromexaničeskie mufty ili, čto *ešče*$^{\text{adv}}$ bolee progressivno, rabotaet v svjazke s sistemami ABS i ESP.
'It is more and more often being replaced by electronic devices that control the hydromechanical couplings, or work together with the ABS and ESP systems, which is something *even* more advanced.'
(N. Kačurin. *Krutjaščij moment istiny* (2002). *Avtopilot*, 2002.02.15)

This latter use, as an adverb associated with comparison, motivates the extension to the adverbial conjunction described above (which also entails comparison) as well as the intensifying use for which *ešče* has been classified as a particle in the Malyj akademičeskij slovar' (1999), as in (11); here it is natural to use an adverb of degree as an intensifier:

(11) – Ne vižu u vas svobodnogo tvorčestva, poleta mysli. Xotja by slovo ot sebja, a to vse ot djadi. – Ot kakogo *ešče*$^{\text{adv}}$ djadi? – Ot djadi Zuja.

'– I don't see in you any spontaneous creativity, active thought. If only there was just one word of your own, but instead everything comes from your uncle. – From which uncle *the heck you are talking about*? – From uncle Zuj.'
(Ju. O. Dombrovskij. *Ručka, nožka, ogurečik*. 1977)

We find that this use of *ešče* is not essentially different from its use as an adverb, as in (6) above, since it gives additional intensity to the expression. Also in the collocation *ešče by*, we tag *ešče* as an adverb since it contributes to a meaning of 'of course, even more so' and is thus associated with comparison, as in (12):

(12) No ved' sčastlivym on vas nikogda ne sdelaet, ponimaete? – *Ešče*$^{\text{adv}}$ by, – skazal Andrej, – slovo-to kakoe strašnoe.
'But after all he will never make you happy, do you understand that? – *You betcha*, said Andrej, – the word itself is frightening.' (V. Pelevin. *Željtaja strela*. 1993)

In sum, the uses in which *ešče* has previously been classified as a particle conform to this word's adverbial uses, and we argue that it is therefore reasonable to interpret and tag *ešče* as an adverb.

### 4.2 *Tak*: ADV 84, CNJADV 7, CNJSUB 8, INTJ 1

*Tak* is used most frequently as an adverb, in which case it usually bears stress and often appears in multiword constructions. *Tak* can also be used as an unstressed adverbial conjunction or as part of a subordinating conjunction. The use of *tak* as an interjection was only attested once in our dataset, though it is rather common in informal spoken Russian.

In its adverbial use, *tak* typically presents the information that could be queried with the interrogative adverb *kak?* 'how?', as in (13):

(13) *Tak*$^{\text{adv}}$ vygljadit karta Uilkinsona.
'*That* is what Wilkinson's map looks like.' (V fokuse otkrytij. *Znanie – sila*. 2003)

In an extension from this use, *tak* can appear clause-initially where it is customarily followed by a comma signifying a pause, parallel to English 'Thus, ...':

(14) *Tak*$^{\text{adv}}$, do nedavnego vremeni WU dejstvitel'no ne predostavljala vozmožnosti otpravljat' perevod posredstvom različnyx sistem udalennogo dostupa k bankovskomu sčetu.
'*For instance*, until recently the WU actually did not provide the opportunity to send transfers via various systems for remote access to bank accounts.'
(Denežnye perevody. *Voprosy statistiki*. 2004)

Adverbial use of *tak* is also observed in numerous constructions such as *tak skazat'* 'that is to say'; *tak nazyvaemyj* 'so-called'; *esli tak* 'if so'; *tak ..., čto ...* 'so..., that'; *tak ili inače* 'some way or another'; *i tak dalee* 'and so on'.

When *tak* is used as an adverbial conjunction it typically does not bear stress and functions in a similar way to *poėtomu* 'therefore', as in (15):

(15) Kurit' est'? – sprosil Poceluev. – Ja brosil, *tak*$^{\text{cnjadv}}$ s soboj ne nošu.
'Got any cigarettes? – asked Poceluev. – I quit, *so* I don't carry them with me.'
(T. Tolstaja. *Reka Okkervil'*. 1983)

*Tak* appears in the multiword subordinating conjunctions *tak kak* 'because' and *tak čto* 'that is why', where it is typically stressed as in (16) and (17); as an interjection, *tak* is roughly equivalent to English 'so' or 'well', cf. (18):

(16) Praktičeski ves' ėtot ščit s tex por nikogda ne podvergalsja razrušeniju, *tak kak*$^{\text{cnjsub}}$ soderžal očen' malo železa.
'Practically the whole shield has never undergone decay since that time *because* it contained very little iron.' (Krepkij orešek. *Znanie – sila*. 2003)

(17) *Tak čto*$^{\text{cnjsub}}$ vo vsem ėtom predatel'stve… Ritina dolja viny pobol'še moej.
'*Therefore* in all this treachery… Rita's share of the blame is greater than mine.'
(J. Trifonov. *Predvaritel'nye itogi*. 1970)

(18) *Tak*$^{\text{intj}}$ nu ja podmetu sejčas.
'*Well* then, I'll sweep up right away.' (*Domašnie razgovory*. Moskva. 1971–1977)

### 4.3 *Ved'*: ADV 57, CNJADV 33, CNJSUB 10

In all its uses, *ved'* expresses insistence based on some indisputable fact that the hearer should pay attention to (cf. Minčenkov 2001, p. 15). McCoy (2003b, p. 331) states that *ved'* is "a marker of information that is assumed by the speaker to be known to the hearer but not activated yet, it is […] a marker of encyclopedic knowledge, and is perceived as a (polite) reminder" (as opposed to *že*). This meaning derives from an obsolete aorist form of the verb **věděti* 'know'. *Ved'* invites the hearer to acknowledge that the speaker's statement is unquestionable. When *ved'* is used as an adverb, the reference is to knowledge shared by the speaker and hearer. As a subordinating conjunction *ved'* presents the reasoning behind the speaker's argument and cannot be omitted. As an adverbial conjunction, *ved'* refers to the information presented in the preceding clause and here it is typically omissible.

Example (19) illustrates *ved'* as an adverb referring to shared experiences of the speaker and hearer. In its adverbial use, *ved'* can be collocated with conjunctions *i* 'and', *no* 'but', *da* 'and', and *esli* 'if', as in (20):

(19) – Nu napugal!… Da my *ved'*$^{\text{adv}}$ tože gramotnye: syn-to *ved'*$^{\text{adv}}$ za otca čego?… Ne otvečaet?… Nu vot. – Ja *ved'*$^{\text{adv}}$ sovsem malen'kij byl…
'– He really gave (us) a scare!… *After all*, we are literate too: does a son *indeed*… answer for a father?… Now you see how it goes. – I was *after all* just a little kid…'
(B. Okudžava. *Iskusstvo krojki i žit'ja*. 1985)

(20) Položenie strannoe, no *ved'*$^{\text{adv}}$ čuvstvuetsja, čto ėto dejstvitel'no tak.
'It's a strange situation, but *after all* it feels like it really is that way.'
(S. G. Bočarov. *Iz istorii ponimanija Puškina*. 1998)

The reference of *ved'* can include encyclopedic knowledge, such as what kinds of things happen in fairy tales, as in (21):

(21) No *ved'*$^{\text{adv}}$ tak byvaet tol'ko v skazke.
'But *of course* that only happens in fairy tales.'
(V. Gubarev. *Troe na ostrove*. 1950–1960)

We also classify the use of *ved'* to confirm statements or reproach the addressee as an adverb, as in (22):

(22) Ja *ved'*$^{\text{adv}}$ prosil ostavit' moj stol v pokoe.
'*Didn't* I ask to leave my desk alone?' (Minčenkov 2001, pp. 70–71)

We interpret *ved'* as a subordinating conjunction when it joins two clauses and expresses the meaning 'considering that, because'. In this use the clause that contains *ved'* presents the motivation behind accepting the proposition in the other clause, as in (23):

(23) I vam ne stoit otstavat' – *ved'* $^{\text{adv}}$ vaš rebenok ėtogo dostoin!
'You mustn't get left behind – *because* your child deserves it!'
(*Turizm i obrazovanie*. 2000.06.15)

Minčenkov (2001, p. 57) observes that in this use the speaker affirms proposition A (here: that you mustn't get left behind) on the grounds that the speaker is certain of proposition B (that your child deserves it). Similarly, the speaker in (24) is claiming that the mouse has to be able to distinguish color (proposition A) because it needs to be able to identify food (proposition B). This is the use of *ved'* that Vikul'ceva (2004, p. 69) terms 'explanatory', and here *ved'* is necessary to mark the connection between the two propositions:

(24) Ryžaja polevka različaet želtyj i krasnyj cveta, *ved'* $^{\text{cnjsub}}$ ej nado otličat' spelye plody i zerna ot nedozrelyx.
'A red field mouse can distinguish between yellow and red colors, *because after all*, it has to distinguish between ripe fruits and grains and unripe ones.'
(Kogda svetofor budet černo-belym? *Znanie – sila*. 2003)

When *ved'* is sentence initial and refers back to previous sentences or clauses in the discourse, it is an adverbial conjunction and is optional in this role. Here *ved'* is termed 'argumentative' by Vikul'ceva (2004, p. 69) and its import is similar to *že* as an adverbial conjunction, while its impact is milder, serving as a more polite way of expressing one's insistence (see Sect. 4.8; cf. McCoy 2003a, p. 125):

(25) No čtoby stat' takim vypusknikom, neobxodimo vlit'sja v učebnuju sredu Soedinennogo Korolevstva ešče v škol'nye gody. *Ved'* $^{\text{cnjadv}}$ britanskaja sistema obrazovanija – ėto i podgotovka k postupleniju v universitet.
'But in order to graduate, it is necessary to join the ranks of the academic milieu of the United Kingdom already during one's school years. *After all*, the British educational system is also a preparation for entry into university.'
(Kak popast' v ėlitu (2000). *Turizm i obrazovanie*. 2000.06.15)

### 4.4 *Slovno*: ADV 49, CNJCOO 51

Like *ešče* and *ved'*, *slovno* can be classified either as an adverb or as a coordinating conjunction. The pragmatic import of *slovno* is the opposite of *ved'*: it expresses imprecision about a statement. Malyj akademičeskij slovar' (1999) makes the following distinction: when *slovno* is used to express comparison, it is a conjunction (*on kričit, slovno rebenok* 'he yells like a child'); when *slovno* expresses uncertainty and hesitation, it is a particle (*on slovno ne v duxe* 'he seems to be in a bad mood'). There is a close relationship between these two facets of *slovno*: comparison often introduces imprecision, and imprecision often results from a speaker's uncertainty about a statement.

In our analysis, we take both the semantic nuances and syntactic scope of *slovno* into account using the following strategy. When *slovno* appears between the subject and the predicate of a clause, modifies an adverb, or is used in constructions with gerunds and participles, we interpret it as an adverb. As an adverb, *slovno* typically expresses hesitation and uncertainty and can be omitted. In this role, *slovno* tells us about a possibility, what may be happening. When *slovno* introduces a comparison, it is analyzed as a coordinating conjunction. As an adverb, *slovno* can modify a predicate, expressing uncertainty, as in (26); *slovno* in this role can also modify another adverb, as in (27):

(26) … ėta ego naprjažennost' *slovno* $^{\text{adv}}$ sozdavala v komnate nezrimoe, no tjagostnoe silovoe pole.

'… his anxiety *seemed* to create an invisible but oppressive force field in the room.'
(Ju. O. Dombrovskij. *Ručka, nožka, ogurečik*. 1977)

(27) … mesto vozle žarkoj batarei *slovno*$^{\text{adv}}$ special'no bylo zabronirovano dlja nas…
'… the place next to the hot radiator was *seemingly* specially reserved for us… '
(B. Okudžava. *Iskusstvo krojki i žit'ja*. 1985)

*Slovno* often appears in a gerund construction where it expresses a possible simultaneous activity. In (28) the main clause tells us that the people have stopped shifting their feet, and the gerund suggests that the reason for this is that they are perhaps trying to listen for something:

(28) Oni perestupili neskol'ko raz i, *slovno*$^{\text{adv}}$ prislušivajas', ostanovilis'.
'They shifted their weight several times from one foot to the other and then, *perhaps* to hear better, they stopped.' (F. Iskander. *Moj kumir*. 1965–1990)

In all these adverbial uses, *slovno* tells us that the speaker suspects something might be the case and *slovno* can be omitted.

By contrast, in its use as a coordinating conjunction, *slovno* expresses a comparison rather than a possibility. In (29) the snail is not actually looking through a clouded glass, its vision is simply compared with what one sees through a clouded glass. Similarly in (30) the speaker is not literally dressed in his own fear, instead he is comparing the feeling of the shirt he has on to his feeling of fear:

(29) Odnako ulitka vidit vse vokrug sebja rasplyvčatym, *slovno*$^{\text{cnjcoo}}$ gljadit skvoz' matovoe steklo.
'But a snail sees everything around itself in a blur, *as if* it were looking through a clouded glass.'
(A. Zajcev. Zagadki ėvolucii: Kratkaja istorija. *Znanie – sila*. 2003)

(30) Ėta rubaška sejčas v temnote kazalas' strannoj, *slovno*$^{\text{cnjcoo}}$ ja byl odet v sobstvennyj strax.
'In the darkness now that shirt seemed strange, *as if* I were wearing my own fear.'
(F. Iskander. *Moj kumir*. 1965–1990)

In both of the examples above, *slovno* introduces an entire clause. However, a comparison can also be evoked elliptically, as in (31) and (32). Note that in all four examples cited here of *slovno* as a coordinating conjunction, it is not possible to omit *slovno* without disturbing the syntactic coherence of the sentences:

(31) Derevo zatreščalo i perelomilos', *slovno*$^{\text{cnjcoo}}$ spička.
'The tree cracked and broke in two, *like* a match.'
(V. Gubarev. *Troe na ostrove*. 1950–1960)

(32) A tut ešče ėti trevogi, tak i pokatitsja serdce, *slovno*$^{\text{cnjcoo}}$ s gory.
'And in addition, all this anxiety, my heart is tumbling, *as if* it were falling downhill.'
(I. Grekova. *Pervyj nalet*. 1960)

### 4.5 *Daže*: ADV 85, CNJCOO 15

In the majority of cases in our sample *daže* is used as an adverb and can be easily omitted without disrupting the syntactic structure of an utterance. Less frequently *daže* appears as a coordinating conjunction where it is syntactically obligatory and typically follows a comma or period.

As an adverb, *daže* often appears with a conjunction such as *no*, *a*, *esli*, *i*, *kogda*, etc., as in (33). This use includes the frequent expression *i delo daže ne v tom, čto* 'and that's not even the main point'. As an adverb, *daže* can also introduce a gerund construction, as in (34) and (35):

(33) Snačala idti legko i *daže*$^{\text{adv}}$ veselo.
'In the beginning it was easy and *even* fun to walk.' (F. Iskander. *Deduška*. 1966)

(34) N'juton ispol'zuet ego pri vyvode zakona vsemirnogo tjagotenija, *daže*$^{\text{adv}}$ ne nazyvaja Keplera…
'Newton uses it to derive the law of universal gravitation, without *even* mentioning Kepler…' (V. Ševčenko. Demon nauki: Kosmičeskij kubok. *Znanie – sila*. 2003)

(35) Anna Fedorovna s gotovnost'ju vstala iz-za stola, ne uspev *daže*$^{\text{adv}}$ posožalet' o neudavšemsja melkom prazdnike.
'Anna Fedorovna readily got up from the table, without *even* spending any time grieving over the failed attempt at a celebration.'
(L. Ulickaja. *Pikovaja dama*. 1995–2000)

As a coordinating conjunction, *daže* serves to add to and elaborate a previous item and in this function it cannot be omitted without losing the connective structure *daže* provides. This use of *daže* typically appears after a comma, as in (36), or after a period, as in (37) where it connects an utterance to the previous discourse:

(36) Kažetsja, odno neostorožnoe slovo – i skandal budet vserossijskij, *daže*$^{\text{cnjcoo}}$ meždunarodnyj.
'It's as if one careless word will bring on a scandal all across Russia, *even* on an international level.'
(K načal'stvu ne dopuskat'! (2002). *Vitrina čitajuščej Rossii*. 2002.10.25)

(37) – Odinnadcat' let, govorjat, pisali? – *Daže*$^{\text{cnjcoo}}$ s xvostikom.
'– They say you were writing for eleven years? – *Even* more than that.'
(Ju. O. Dombrovskij. *Ručka, nožka, ogurečik*. 1977)

### 4.6 *Že*: CNJADV 13, CNJCOO 6, EMPH 81

The word *že* has been traditionally assigned the status of a particle and a conjunction, as suggested by Švedova (1980, §3135) and Kasatkina (2004, p. 71). According to Švedova (1980, §3135) and Plungjan (1987, p. 36), when *že* is used as a conjunction, it is semantically equivalent either to the coordinating conjunction *a*, which expresses contrast, or to the use of *ved'* as a conjunction for presenting logical motives for reasoning. However, scholars are inconsistent on this point. Kuznecov (1998) suggests, contrary to Švedova, that when the meaning of *že* overlaps with *ved', že* is a particle. We formulate guidelines that can be operationalized and used for manual differentiation of three subuses of *že*, but we do not recognize *že* as a particle.

It is well known that in Standard Russian *že* never appears clause-initially. *Že* is a clitic[7] that forms a prosodic unit with a stressed lexeme, to which it is either preposed or postposed.

[7] Kasatkina (2004, pp. 73–74) shows that as opposed to Standard Russian, in Russian dialects *že* often bears stress and escapes vowel reduction, and this takes place in certain specific uses not found in Standard Russian: e.g. dialectal *že* can be semantically equivalent to the conjunction *i* 'and', as in this example from the Vologodskaja oblast': *V magazin prjaniki privezli, konfety že* 'Spice cookies have been brought to the store, and candy (too)'.

We suggest that the position of *že* with regard to its prosodic head is associated with different functions of *že*: postposed *že* is typically an emphasizer or coordinating conjunction, whereas preposed *že* is mostly an adverbial conjunction.

In addition to position, two factors make it possible to distinguish among uses: 1. syntactic optionality vs. obligatoriness of *že* and 2. semantic contribution / replaceability of *že* with semantically equivalent conjunctions. In the functions of emphasizer and adverbial conjunction, *že* can be syntactically optional, whereas in the function of a coordinating conjunction, *že* is obligatory for creating an explicit contrast between syntactic constituents. In the function of coordinating conjunction, *že* can be semantically replaced with the conjunction *a* (though the word order changes). In the adverbial conjunction function, *že* can be replaced with *ved'*. However, the pragmatic impact of *že* is considerably stronger and less polite than that of *ved'*; cf. McCoy (2003a, p. 125), who paraphrases *že* thus: "You are wrong! And more than that, you are capable of arriving at the correct conclusion yourself, but nevertheless you are sticking to the wrong conclusion."

The most common use of *že* is as an emphasizer. We analyze *že* as an emphasizer only when it immediately follows a phrasal stress-bearing lexical item that serves as the focus of the speaker's attention: a verb, an adjective, a pronoun, or a conjunction (e.g. *ili že*). In this use, *že* emphasizes the lexeme with which it forms a prosodic unit and is syntactically optional, since *že* can be omitted without affecting the syntax of an utterance:

(38) Seli s kraju – i tut *že*$^{\text{emph}}$ iz veščmeška Vovka izvlek butylku portvejna.
'They sat down at the edge of the table – and *right away* then Vovka pulled a bottle of portwine out of the supply bag.' (V. Makanin. *Kavkazskij plennyj*. 1995)

Plungjan (1987, pp. 36–39) also distinguishes this use of *že*, which he calls 'anaphorically bound' because it refers to two propositions that share some property (such as time or space), but their co-occurrence is somewhat unexpected. In our example (38), the location and time of the sitting down and the pulling of the bottle out of the bag are shared, but the immediacy of the second event is unexpected.

Like *že* as an emphasizer, in its use as a contrastive coordinating conjunction, *že* is postposed to a phrasal stress-bearing lexeme. This use of *že* contrasts two clauses, but also puts emphasis on the lexeme that is the focus of the contrast. However, as opposed to *že* as an emphasizer, *že* as a coordinating conjunction is not syntactically optional and cannot be removed from a sentence without losing explicit contrast, as in this example:

(39) Takže raznorečivo opredeljaetsja i otnošenie satiry k jumoru: odni ix rezko razdeljajut…, drugie *že*$^{\text{cnjcoo}}$ vidjat v jumore… smjagčennuju raznovidnost' satiry.
'Equally contradictory is the definition of the relationship between satire and humor: some people keep them strictly distinct…, others *however* see humor as a mild form of satire.' (M. M. Baxtin. *Satira*. 1945–1950)

When preposed to a stress-bearing lexeme, *že* typically functions as an adverbial conjunction that is syntactically optional and can be removed or replaced with *ved'* in the same function meaning 'because, considering that'. In this use, *že* additionally introduces a contraposition of a phrase to the previous discourse, as in (40). The adverbial conjunction function can also appear when *že* is postposed, as in (41):[8]

[8] Note that an alternative interpretation for examples (40) and (41) is that *že* appears in Wackernagel position and emphasizes the entire clause.

(40) Ved' ja *že*$^{\text{cnjadv}}$ ponimaju, čto vo mne vsego ėtogo polnym-polno, i čto?
‘After all I *do* understand that there is all too much of that in me, and so what?’
(E. Griškovec. *OdnovrEmEnno*. 2004)

(41) Konečno, sgorela – nel'zja *že*$^{\text{cnjadv}}$ v polden' ležat' na solncepeke.
‘Of course you got a sunburn – you can't lie in the hot sun in the middle of the day.’
(Maxrovaja istorija (2002). *Domovoj*. 2002.09.04)

### 4.7 *Li*: ADV 23, CNJCOO 6, CNJSUB 0, QST 71

We distinguish between four uses of *li*, three of which are attested in our dataset: *li* as an adverb, *li* as a coordinating conjunction, and *li* as a question word. The use of *li* as a subordinating conjunction is less frequent and not attested in our dataset.

In the majority of cases, *li* is a question marker and appears in interrogative clauses, as in (42):

(42) Ne soglasites' *li*$^{\text{qst}}$ vmeste použinat'?
‘*Wo*n't you join me for dinner?’ (S. Dovlatov. *Inaja žizn'*. 1984)

*Li* can be used to signal indirect questions, as in (43), embedded questions (typically marked with words like *vopros* ‘question’, *sprašivat'* ‘ask’) as in (44), or rhetorical questions as in (45):

(43) … bol'šinstvo žitelej Kvebeka jasno otvetit na vopros: xotite *li*$^{\text{qst}}$ vy otdelenija ot Kanady.
‘… the overwhelming majority of Quebec's inhabitants give a clear answer to this question: *Do* you want separation from Canada?’ (*Izvestija*. 2001.07.09)

(44) Otvet na ėtot vopros ležit v tom, priznaem *li*$^{\text{qst}}$ my Rossiju Zapadom ili Vostokom.
‘The answer to that question depends on *whether* we recognize Russia as West or East.’ (D. Lixačev. *O russkoj intelligencii*. 1993)

(45) Dvadcat' let, šutka *li*$^{\text{qst}}$! Za dvadcat' let redejut lesa, oskudevaet počva.
‘Twenty years, *is that* a joke? Twenty years is enough for the forests to become thinned out and the soil to get impoverished.’
(J. Trifonov. *Predvaritel'nye itogi*. 1970)

As a question word, *li* can also appear in affirmative sentences, where *li* contributes semantics of doubt, questioning the status of the word it modifies, as in (46). It is this use of *li* that has traditionally been classed as a particle. We argue that this is merely the same *li* signaling the modality of doubt, but applied to a different context, namely that of an affirmative rather than interrogative sentence:

(46) Malo *li*$^{\text{qst}}$ čto možet slučit'sja, i komandir staralsja byt' nagotove.
‘Anything [= *no* small number of things] could happen, and the commander tried to be ready.’ (V. Bykov. *Boloto*. 2001)

As a coordinating conjunction, *li* is used either in lists as in (47) or in expressions like *dolgo li korotko li* ‘long or short/after a while’. *Li* can also function as a coordinating conjunction in the *to li… to li…* construction, as in (48):

(47) Na gorjačuju plitu pečki stavili čajnik, a v žarkij duxovoj škaf – blincy *li*$^{\text{cnjcoo}}$, pyški, pirogi – kto čto iz doma prines.

‘A tea kettle was put on the hot stovetop, and into the warm oven went various things people brought from home, *be they* pancakes, doughnuts, or pies.’
(B. Ekimov. Fetisyč. *Novyj Mir*. 1996)

(48) To *li*$^{\text{cnjcoo}}$ zaboleval, to *li*$^{\text{cnjcoo}}$ roditeli ego za čto-to nakazyvali, to *li*$^{\text{cnjcoo}}$ tetja iz drugogo goroda v gosti priezžala, no tol’ko on, kotoryj celymi večerami slonjalsja po dvoru, kak raz v ėtot večer sidel doma.
‘*Either* he was sick, *or* his parents were punishing him for something, *or* his aunt came to visit from out of town, but it was he, who used to spend whole evenings hanging out outside, that happened to be at home that evening.’
(A. Aleksin. *Moj brat igraet na klarnete*. 1967)

*Li* is less frequent in its use as a subordinating conjunction, which is not attested in our database, but can be found in examples like (49) where *li* introduces a subordinate clause of condition. Note, however, that this use of *li* is largely obsolete, as it is being replaced by *esli* ‘if’:

(49) On očen’ nelovok: stanet *li*$^{\text{cnjsub}}$ otvorjat’ vorota ili dveri, otvorjaet odnu polovinku, drugaja zatvorjaetsja.
‘He is very clumsy: *if* he starts to open a gate or a double door, as he is opening it on one side, the other side is closing.’ (I. A. Gončarov. *Oblomov*. 1859)

Finally, *li* is classed as an adverb in the following multiword units: *vrjad li*, *edva li*, *čut’ li* ‘hardly’, as in (50):

(50) Tat’jana Vasil’evna ezdila k nemu, unižalas’, zadabrivala, *čut’ li*$^{\text{adv}}$ ne nasil’no privezla ego, p’janogo, i on zasnul na razvoročennom divane.
‘Tat’jana Vasil’evna went to see him, she groveled and coaxed, *almost* used violence to transport him when he was drunk, and then he fell asleep on the dissheveled couch.’
(I. Grekova. *Pod fonarem*. 1963)

### 4.8 *Da*: ADV 19, CNJCOO 25, PRAEDIC 3, VMOD 3, INTJ 50

The most frequently attested role of *da* in our database is interjection. *Da* as an interjection means ‘yes’ and carries stress. It can be used in affirmative sentences (51) and in tag questions (52):

(51) No Leva ne sodrognulsja. A, naoborot, soglasilsja s Alinoj: – *Da*$^{\text{intj}}$, saksofon obladaet original’nymi sredstvami muzykal’nogo vyraženija…
‘But Leva was unshaken. To the contrary, he agreed with Alina: – *Yes*, the saxophone has a unique capacity for musical expression.’
(A. Aleksin. *Moj brat igraet na klarnete*. 1967)

(52) Tut vyjasnili / vrode ėto skoree nacionalism / *da*$^{\text{intj}}$?
‘Then they discovered that it was more like nationalism, *right*?’
(Beseda v Moskve. Fond *Obščestvennoe mnenie*. 2003)

The second most frequent use of *da* is as a coordinating conjunction equivalent to the conjunctions *i* ‘and’, see (53) and (54). In some cases the semantics of *da* is equivalent to that of *no* ‘but’ (55). As a conjunction, *da* does not carry stress and can join two items:

(53) Otčim Fedor mudril poroju nad nim s mašinkoj *da*$^{\text{cnjcoo}}$ nožnicami, ostavljaja čelku na lbu i golyj zatylok.

‘His stepfather Fedor fussed for a while over him with a hair clipper *and* scissors, leaving a lock on his forehead and the back of his neck bare.’

(B. Ekimov. Fetisyč. *Novyj Mir*. 1996)

(54) A esli by ne ėtot nelepyj skvoznjak, možet byt’, žila by sebe *da*$^{\text{cnjcoo}}$ žila – xot’ by i dvaždy devjanosto devjat’?…
‘And if it hadn’t been for that absurd draft, maybe she would have lived *and* lived – maybe she would have been ninety-nine two times over?’
(M. Palej. *Pominovenie*. 1987)

(55) Možno by i smirit’sja s ėtimi ciframi, *da*$^{\text{cnjcoo}}$ na molekuljarnom urovne zakis’ azota v 300 raz ėffektivnee drugogo tepličnogo gaza – uglekislogo.
‘One could live with those numbers, *but* at the molecular level nitrous oxide is 300 times as potent as the other greenhouse gas – carbon dioxide.’
(Kto kogo. *Znanie – sila*. 2003)

In its use as a conjunction, *da* is often accompanied by the conjunction *i* or by the adverb *ešče* or the phrase *k tomu že* ‘additionally’:

(56) Kolja nalil ešče v stakašek i soprovodil zakusku vincom, posle čego oblokotilsja na ruku, *da*$^{\text{cnjcoo}}$ i zadremal umirotvorenno.
‘Kolja filled up the glass again and followed the snack with wine, and after that he leaned on his elbow *and* peacefully dozed off.’
(V. Astaf’ev. Zatesi (1999). *Novyj Mir*. 2000)

As an adverb, *da* is typically not stressed and expresses mild surprise. In this use *da* is frequently accompanied by words with compatible semantics such as *neuželi* ‘really, is it possible’ and *začem* ‘what for’:

(57) *Da*$^{\text{adv}}$ v kiteljax ėtix polgoroda xodit!
‘*Well* half the town is wearing those coats!’
(A. Azol’skij. Obldramteatr. *Novyj Mir*. 1997)

(58) – *Da*$^{\text{adv}}$ neuželi ž ty na menja obidiš’sja?
‘– *Surely* you can’t be angry with me?’ (A. Solženicyn. *Matrenin dvor*. 1960)

(59) Ja sejčas konču razgovor. Ty slušaeš’, mužik? – sprosil on trubku. – Molodec. Tak vot, ty daleko ot menja? – *Da*$^{\text{adv}}$ začem tebe ėto nužno?
‘I’m hanging up now. Do you hear me, man? – he asked the receiver. – Attaboy. So then, are you far away? – *So* why do you need this *at all*?’
(Ju. O. Dombrovskij. *Ručka, nožka, ogurečik*. 1977)

*Da* is used as an unstressed modal verb to support the predicate with imperatives (60), infinitives, and with present tense finite forms (61):

(60) – *Da*$^{\text{vmod}}$ už pogodi, Ignatič, paru dnej…
‘– Wait a few days, Ignatič…’ (A. Solženicyn. *Matrenin dvor*. 1960)

(61) *Da*$^{\text{vmod}}$ zdravstvuet nerušimaja družba narodov…
‘Long live the indestructible friendship of nations…’
(V. Makanin. *Kavkazskij plennyj*. 1995)

Predicative use of *da* is rather infrequent in our database. As a predicative *da* stands for an entire proposition and carries stress:

(62) Pomimo ėtogo, sudu neobxodimo ustanovit', … želaet li vyexavšij vozvratit'sja v žiloe pomeščenie dlja proživanija v nem; esli *da*$^{\text{praedic}}$, to čerez kakoj period vremeni.
'Furthermore, the court needs to establish whether the person who has left intends to return to the residence in order to reside there; if *so*, then after what period of time.'
(K. Aksenova. Osobennosti žiliščnyx sporov (2002). *Birža pljus svoj dom* (N. Novgorod). 2002.05.20)

### 4.9 *Net*: NEST 60, PRAEDIC 10, INTJ 30

We distinguish between *net*-interjection (INTJ), predicative *net* (PRAEDIC), and *net*-predicative of existence (NEST). *Net* is used as a negative interjection meaning 'no', where its meaning is the opposite of *da* 'yes':

(63) *Net*$^{\text{intj}}$ / ja ne skažu / čto ėto u mnogix bylo.
'*No*, I wouldn't say that many people had it.'
(*Biografija* (*beseda lingvista s informantom*), Sankt-Peterburg. Arxiv Xel'sinkskogo universiteta. 1997–1998)

Sometimes *net* as an interjection is used in combination with *da* to constitute a single interjection *da net* meaning 'not really':

(64) Potom govorit – ustala? Ja govorju – da *net*$^{\text{intj}}$.
'Then she says – are you tired? I say – *not* really.' (A. Gelasimov. *Žanna*. 2001)

Malyj akademičeskij slovar' (1999) recognizes *net* as an emphatic particle that is used to attract the hearer's attention, but we have analyzed such uses as interjection as well:

(65) *Net*$^{\text{intj}}$, ty vse-taki dura.
'*No*, you're a fool anyway.' (A. Gelasimov. *Žanna*. 2001)

Predicative *net* stands for a whole clause and is often introduced as an alternative to an option described by a preceding phrase. Predicative *net* is often preceded by the conjunction *ili* 'or':

(66) Nas nikto ne sprašivaet – skazal Andrej, – soglasny my ili *net*$^{\text{praedic}}$.
'Nobody is asking us – said Andrej, – whether or *not* we agree.'
(V. Pelevin. *Željtaja strela*. 1993)

A use similar to predicative *net* is the third classification that we recognize: the predicate of existence that refers to the absence of objects. In this use, *net* means 'there is none of' and is the opposite of *est'* 'there is…'. We tag this use as NEST, which stands for *ne est'* 'there is none of'. *Net* means that something is not available and the missing object is marked by the genitive case. In our sample this is the most frequently attested use of *net*:

(67) Problem u menja ser'eznyx *net*$^{\text{nest}}$.
'I *don't* have any serious problems.'
(Beseda v Novosibirske. Fond *Obščestvennoe mnenie*. 2003)

## 5 Life without particles: Experiment 2

Experiment 2 used the same database of 900 sentences (100 for each lexeme) used in Experiment 1, but trained the Hidden Markov Model (HMM) tagger on the part-of-speech tags assigned according to the scheme we describe in Sect. 4 (see outcomes in Table B of appendix). The task of guessing tags in Experiment 2 is considerably more difficult than in

**Table 8** Comparison of part of speech tags for Experiment 1 and Experiment 2

| Lexeme | Tags in RNC gold standard | # tags in Exp 1 | Tags in our tagging scheme | # tags in Exp 2 |
|---|---|---|---|---|
| *ešče* | ADV, PART | 2 | ADV, [~~CNJADV~~] | 1 |
| *tak* | [~~ADV~~], PART | 1 | ADV, CNJADV, CNJSUB, INTJ | 4 |
| *ved'* | CNJ, PART | 2 | ADV, CNJADV, CNJSUB | 3 |
| *slovno* | CNJ, PART | 2 | ADV, CNJCOO | 2 |
| *daže* | CNJ, PART | 2 | ADV, CNJCOO | 2 |
| *že* | CNJ, PART | 2 | CNJADV, CNJCOO, EMPH | 3 |
| *li* | CNJ, PART | 2 | ADV, CNJCOO, [~~CNJSUB~~], QST | 3 |
| *da* | CNJ, PART | 2 | ADV, CNJCOO, PRAEDIC, VMOD, INTJ | 5 |
| *net* | PRAEDIC, PART | 2 | NEST, PRAEDIC, INTJ | 3 |

Experiment 1 due to the fact that there are more tags in our scheme than in the one employed in the RNC gold standard. Table 8 compares the tags across the two tagging schemes.

The left half of Table 8 pertains to the tags in the RNC gold standard that were used to train the tagger in Experiment 1, whereas the right half of the table pertains to the tags in our tagging scheme used to train the tagger in Experiment 2. All nine lexemes had two potential tags in the RNC gold standard, but one tag (ADV) was not realized among our examples for *tak*, and this is indicated by [~~ADV~~] in Table 8. Accordingly, only one tag was assigned to *tak* in Experiment 1, while all others could be assigned two different tags.

In our tagging scheme there was also one lexeme, *ešče*, that received only one tag in Experiment 2 because there were no examples of *ešče* as an adverbial conjunction, indicated by [~~CNJADV~~] in Table 8. Two lexemes, *slovno* and *daže* received only two tags in Experiment 2. Four lexemes received three tags each (*ved'*, *že*, *li*, and *net*), *tak* received four tags, and *da* received 5 tags. Thus the HMM tagger potentially faced a bigger challenge in Experiment 2 than in Experiment 1.

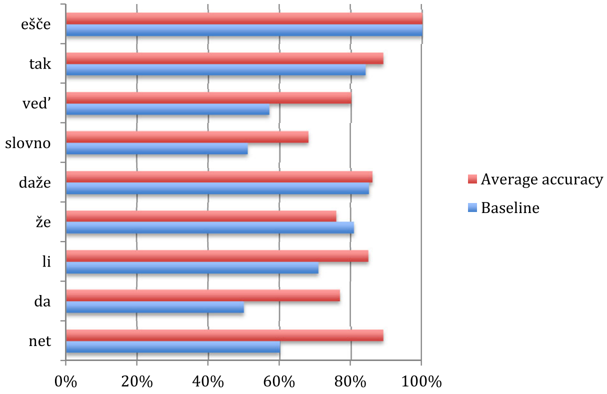
However, despite the added challenge, it is possible to argue that the HMM tagger performed better in Experiment 2 than in Experiment 1, as shown in Table 9 and Fig. 2. Note that the baseline in Experiment 2 is based on the highest frequency tag in our tagging scheme, as shown in Table 7, and this is different from the baseline in Experiment 1 (the baseline for Experiment 1 is the highest number in each row of Table 5, whereas the baseline for Experiment 2 is the highest number in each row of Table 7).

Shading in Table 9 indicates lexemes that fared better in Experiment 2 than in Experiment 1 in terms of gain over baseline. In Experiment 2, all lexemes show a positive gain over baseline except *ešče* (which had 100 % for baseline) and *že* (which lost only 5 percentage points, the same as its loss in Experiment 1). The total gain over baseline in Experiment 2 is 111 percentage points, which is more than double the total gain over baseline in Experiment 1. Analysis of *slovno* and *li* was considerably better in Experiment 2 than in Experiment 1. Only *daže* (5 percentage points lower) and *net* (2 percentage points lower) showed some modest losses. Of course the standard for measurement is also different, since when the number of interpretations was expanded, the baseline also shifted downward for seven of the nine lexemes (all but *ešče* and *net*).

**Table 9** Outcome of Experiment 2

| Lexeme | Baseline in % | Average accuracy in % | Gain over Baseline | Gain over Baseline Exp 2 – Exp 1 (percentage points) |
|---|---|---|---|---|
| *ešče* | 100 | 100 | 0 | +13 |
| *tak* | 84 | 89 | +5 | +5 |
| *ved'* | 57 | 80 | +23 | +12 |
| *slovno* | 51 | 68 | +17 | +11 |
| *daže* | 85 | 86 | +1 | −5 |
| *že* | 81 | 76 | −5 | 0 |
| *li* | 71 | 85 | +14 | +20 |
| *da* | 50 | 77 | +27 | +6 |
| *net* | 60 | 89 | +29 | −2 |

**Fig. 2** Outcome of Experiment 2



Overall, we take these results to mean that the HMM tagger reached similar or better accuracy in part-of-speech tagging of the same data after we eliminated the PART tag, even though the resulting tagging scheme was considerably more complex.

## 6 Conclusion

We find that Zwicky (1985) was justified in asking for 'particle' to be removed from the inventory of parts of speech. We argue that 'particle' is not useful as a part-of-speech classification in Russian. The lexemes traditionally classed as particles are often frequent and ambiguous across part-of-speech categories, causing significant problems for natural language processing of Russian, and therefore also negatively impacting all language technology applications that are sourced by NLP. Furthermore, our Experiment 1 shows that current practice in the manual tagging of such lexemes in the RNC does not yield data that is consistent enough to be used to train reliable automatic taggers.

We offer a revised and expanded scheme for the tagging of nine high-frequency lexemes that have been traditionally classed as ambiguous across two parts of speech: as a particle and a conjunction, adverb, or predicative. In our tagging scheme the 'particle' designation has been eliminated. The nine lexemes are classed as adverb, conjunction (adverbial, coordinating, or subordinating), predicative, modal verb, 'nest' (indicating that something is not available), interjection, emphasizer, and question word. All of the classifications in our tagging scheme are conceptually motivated, and therefore more useful in the long run. Despite the added complexity of this tagging scheme, the automatic tagger performs well, though there is still a long way to go to produce a tagging scheme that would facilitate reliably accurate part-of-speech automatic tagging.

We hope that this study can contribute to the further refinement of the Russian National Corpus, ultimately making this superb resource even more valuable. In particular, it should be possible to reclassify the remaining lexemes classed as particles and overall improve part-of-speech tagging, yielding better results for NLP and downstream applications.

## Appendix[9]

**Table A** Outcome of all ten trials for Experiment 1

| | trial 1 | trial 2 | trial 3 | trial 4 | trial 5 | trial 6 | trial 7 | trial 8 | trial 9 | trial 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *ešče* | 80 | 70 | 70 | 50 | 80 | 50 | 80 | 80 | 80 | 60 |
| *tak* | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| *ved'* | 80 | 80 | 80 | 70 | 70 | 70 | 80 | 90 | 80 | 80 |
| *slovno* | 90 | 80 | 90 | 90 | 100 | 90 | 80 | 90 | 90 | 90 |
| *daže* | 90 | 80 | 70 | 80 | 100 | 100 | 90 | 100 | 100 | 80 |
| *že* | 80 | 90 | 100 | 100 | 80 | 90 | 80 | 90 | 80 | 100 |
| *li* | 70 | 90 | 60 | 80 | 70 | 60 | 70 | 100 | 90 | 70 |
| *da* | 90 | 80 | 90 | 60 | 70 | 90 | 80 | 60 | 80 | 50 |
| *net* | 90 | 80 | 90 | 100 | 80 | 70 | 100 | 90 | 100 | 90 |

**Table B** Outcome of all ten trials for Experiment 2

| | trial 1 | trial 2 | trial 3 | trial 4 | trial 5 | trial 6 | trial 7 | trial 8 | trial 9 | trial 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *ešče* | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| *tak* | 90 | 90 | 90 | 80 | 90 | 100 | 80 | 100 | 70 | 100 |
| *ved'* | 70 | 70 | 90 | 90 | 90 | 90 | 60 | 80 | 80 | 80 |
| *slovno* | 80 | 50 | 50 | 90 | 60 | 60 | 70 | 80 | 60 | 80 |
| *daže* | 90 | 90 | 90 | 100 | 80 | 100 | 90 | 70 | 90 | 60 |
| *že* | 80 | 80 | 60 | 90 | 80 | 80 | 80 | 90 | 60 | 60 |
| *li* | 80 | 100 | 90 | 80 | 80 | 80 | 80 | 90 | 80 | 90 |
| *da* | 80 | 80 | 60 | 90 | 50 | 90 | 80 | 90 | 60 | 90 |
| *net* | 100 | 100 | 80 | 100 | 100 | 60 | 90 | 80 | 90 | 90 |

## References

Andersen, G. & Fretheim, T. (Eds.) (2000). *Pragmatic markers and propositional attitude* (Pragmatics & Beyond. New Series, *79*). Amsterdam, Philadelphia.

Bartoševič, A. (1978). Časticy i leksikografičeskaja praktika. *Slavia orientalis*, *XXVII*(3), 331–334.

[9] Please note that all of the numbers given in the tables in the Appendix are cited in percentages.

Bogorodickij, V. A. (1939). *Očerki po jazykovedeniju i russkomu jazyku*. Moskva.

Brinton, L. J. (1996). *Pragmatic markers in English. Grammaticalization and discourse functions* (Topics in English Linguistics, *19*). Berlin.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge.

Croft, W. (2001). *Radical Construction Grammar. Syntactic theory in typological perspective*. Oxford.

Dedaić, M., & Mišković-Luković, M. (2010). South Slavic discourse particles: introduction. In M. Dedaić & M. Mišković-Luković (Eds.), *South Slavic discourse particles* (Pragmatics & Beyond New Series, *197*, pp. 1–22). Amsterdam.

Drummen, A. (2015). *Dramatic pragmatics. A discourse approach to particle use in ancient Greek tragedy and comedy* (Doctoral dissertation). Ruprecht-Karls-Universität Heidelberg, Heidelberg.

Erelt, M., Erelt, T., & Ross, K. (2000). *Eesti keele käsiraamat*. Tallin.

Fries, Ch. C. (1952). *The structure of English*. New York.

Grišina, E. A., & Savčuk, S. O. (2009). Korpus ustnyx tekstov v Nacional'nom korpuse russkogo jazyka: sostav i struktura. In *Nacional'nyj korpus russkogo jazyka. Novye rezul'taty i perspektivy*. Available at http://ruscorpora.ru/corpora-biblio-2008.html.

Grišina, E. A., & Ljaševskaja, O. N. (2008). *Grammatičeskij slovar' novyx slov russkogo jazyka*. Available at http://dict.ruslang.ru/gram.php.

Halácsy, P., Kornai, A., & Oravecz, C. (2007). HunPos: an open source trigram tagger. In *ACL 2007. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Interactive Poster and Demonstration Sessions, June 25–27, 2007, Prague* (pp. 209–212). Stroudsburg.

Harris, Z. S. (1951). *Methods in structural linguistics*. Chicago.

Heinrichs, W. (1981). *Die Modalpartikeln im Deutschen und Schwedischen* (Linguistische Arbeiten, *101*). Tübingen.

Kasatkina, R. F. (2004). Častica *že* v roli tekstovogo konnektora (na materiale russkoj dialektnoj reči). In T. M. Nikolaeva (Ed.), *Verbal'naja i neverbal'naja opory prostranstva mežfrazovyx svjazej* (pp. 71–82). Moskva.

Klobukov, E. V. (2001). Analitičeskie glagoly v russkom jazyke. In S. M. Kuz'mina (Ed.), *Žizn' jazyka. Sbornik statej k 80-letiju Mixaila Viktoroviča Panova* (pp. 77–87). Moskva.

Kuznecov, S. A. (Ed.) (1998). *Bol'šoj tolkovyj slovar' russkogo jazyka*. Sankt-Peterburg.

Langacker, R. W. (2013). *Essentials of cognitive grammar*. Oxford.

Makarova, A. (2015). One type of verbal diminutives in Russian: verbs ending in *-n'kat'*. *Russian Linguistics*, *39*(1), 15–31.

Malyj akademičeskij slovar' (1999). *Slovar' russkogo jazyka v četyrex tomax*. Moskva.

Manning, C. D. (2011). Part-of-speech tagging from 97 % to 100 %: is it time for some linguistics? In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing. 12th International Conference, CICLing 2011, Tokyo, Japan, February 2011. Proceedings, Part I* (Lecture Notes in Computer Science, *6608*, pp. 171–189). Berlin.

McCoy, S. (2003a). Unifying the meaning of multifunctional particles: the case of Russian *ŽE*. *University of Pennsylvania Working Papers in Linguistics*, *9*(1), 123–135.

McCoy, S. (2003b). Connecting information structure and discourse structure through "Kontrast": the case of colloquial Russian particles – *TO*, *ŽE*, and *VED'*. *Journal of Logic, Language and Information*, *12*, 319–335.

Minčenkov, A. G. (2001). *Russkie časticy v perevode na anglijskij jazyk*. Sankt-Peterburg.

Miot, G. (1987). *Il mio primo dizionario*. Firenze.

Nikolaeva, T. M. (1985). *Funkcii častic v vyskazyvanii (na materiale slavjanskix jazykov)*. Moskva.

Percov, N. V. (2002). O vozmožnom semantičeskom invariante russkix frazovyx častic *uže* i *ešče*. In N. D. Arutjunova (Ed.), *Logičeskij analiz jazyka. Semantika načala i konca* (pp. 137–145). Moskva.

Plungjan, V. A. (1987). Ocenka verojatnosti v značenii časticy *že* (k formalizacii semantičeskogo opisanija služebnyx slov). *Naučno-texničeskaja informacija. Serija 2: Informacionnye processy i sistemy*, *8*, 36–40.

Russian National Corpus (RNC). Available at www.ruscorpora.ru.

Šaxmatov, A. A. (1941). *Sintaksis russkogo jazyka*. Leningrad.

Serianni, L., & Castelvecchi, A. (1997). *Italiano: grammatica, sintassi, dubbi*. Milano.

Sičinava, D. V. (2005). Obrabotka tekstov s grammatičeskoj razmetkoj: instrukcija razmetčika. Retrieved from http://ruscorpora.ru/sbornik2005/09sitch.pdf (1 March 2016).

Starodumova, E. A. (1997). *Russkie časticy (pis'mennaja monologičeskaja reč')* (avtoreferat dissertacii). Moskva.

Švedova, N. Ju. (Ed.) (1980). *Russkaja grammatika. Tom I: Fonetika, fonologija, udarenie, intonacija, slovoobrazovanie, morfologija*. Moskva.

Tauli, V. (1972). *Eesti grammatika* (Vol. I). Uppsala.

Timberlake, A. (2004). *A reference grammar of Russian*. Cambridge.

Vasilyeva, A. N. (1972). *Particles in colloquial Russian (manual for English-speaking students of Russian)*. Moscow.

Vikul'ceva, N. (2004). *Semantičeskie, sintaksičeskie i pragmatičeskie osobennosti leksemy* VED' (Master's thesis). University of Tartu, Tartu.

Vinogradov, V. V. (1972). *Russkij jazyk. Grammatičeskoe učenie o slove*. Moskva.

Voejkova, M. D. (2009). Problemy ispol'zovanija podkorpusa ustnoj razgovornoj reči (na primere analiza russkix diminutivov). In *Nacional'nyj korpus russkogo jazyka. Novye rezul'taty i perspektivy*. Available at http://ruscorpora.ru/corpora-biblio-2008.html.

Wierzbicka, A. (1976). Particles and linguistic relativity. *International Review of Slavic Linguistics*, *1*(2–3), 327–367.

Wierzbicka, A. (1988). *The semantics of grammar* (Studies in Language Companion Series, *18*). Amsterdam, Philadelphia.

Wierzbicka, A. (1992). *Semantics, culture, and cognition. Universal human concepts in culture-specific configurations*. New York, Oxford.

Zaliznjak, A. A. (1980). *Grammatičeskij slovar' russkogo jazyka*. Moskva.

Zwicky, A. M. (1985). Clitics and particles. *Language*, *61*(2), 283–305.