Representation and Classification of Polyfunctional Synsemantic Words in Monolingual Dictionaries and Language Corpora: The Case of the Croatian Lexeme *Dakle*

Virna Karlić[1], Petra Bago[2]

[1] Department of South Slavic Languages and Literatures
[2] Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3, HR-10000
Email: {vkarlic, pbago}@ffzg.hr

Abstract

The paper will discuss the central issues concerning lexicographic descriptions of synsemantic words, with special regard to those with multiple syntactic and pragmatic functions. This topic will be exemplified through a description of a representative example, the Croatian lexeme *dakle* (Eng. *well*, *now*; *consequently*; *accordingly*, *so*, *then*, *therefore*, *thus*). We will focus on the shortcomings of lexicographic descriptions of such words in four contemporary monolingual dictionaries of the Croatian (standard) language. We pay particular attention to the inconsistent part of speech classification in these dictionaries, as well as to the type and content of their definitions, which generally do not take into account multiple syntactic and pragmatic functions of the word. This paper will analyse the functions and the use of lexeme *dakle*, an analysis based on language material extracted from the Croatian web corpus hrWaC, and processed by two independent annotators. We have attained fair agreement between annotators for the first task of determining the (supra)syntactic function (Cohen's κ is 0.4332), and poor agreement for the second task of determining the semantic-pragmatic function (Cohen's κ is 0.2908). Ultimately, the data collected, when compared to dictionary content, can serve as a starting point for a general discussion of an adequate methodology for lexicographic description of polyfunctional synsemantic words.

Keywords: monolingual lexicography; language corpora; pragmatics; synsemantic words; polyfunctionality; Croatian language; lexeme *dakle*

1. Introduction

Lexicographic descriptions of polyfunctional synsemantic (functional / grammatical / closed class) words are often problematic, particularly since they have numerous syntactic and pragmatic functions. Contemporary (but theoretically and methodologically traditional) monolingual dictionaries of the Croatian language reduce the description of this kind of lexeme to its main syntactic-semantic function. However, these lexemes have important and frequently employed pragmatic roles in written and spoken discourse, roles that are generally left out of dictionary definitions. The shortcomings of such descriptions are especially salient in the annotation process of language corpora, resulting in an overly generic categorization of polyfunctional synsemantic words in these annotations. This problem becomes exacerbated as new dictionaries are compiled based on inaccurately annotated language corpora. We believe this vicious cycle can only be broken by the application of a pragmatic approach to dictionary descriptions of such words.

These issues become clear when specific lexemes are examined. This analysis will focus on the use and syntactic/pragmatic functions of the lexeme *dakle* (Eng. conj. *well, now; consequently* [Bujas, 1999] / *accordingly, so, then, therefore, thus* [Bujas, 2005]). In Croatian monolingual dictionaries[1] the lexeme is categorized as a conjunction or an adverb, while in the Croatian web corpus hrWaC (Ljubešić & Klubička, 2014) over 99% of the occurrences of the word *dakle* are annotated as conjunctions, which is inconsistent with previous linguistic research, as well as our analysis of hrWaC. According to Dedaić (2010), the lexeme has developed four predominant functions in discourse: conclusional, reformulational, argumentative/rhetorical, and attitudinal. In spoken language, especially in scientific discourse, *dakle* is also frequently used as a filler word (Pintarić, 2002; Silić & Pranjković, 2005).

Our research was conducted on a random sample of 400 KWIC examples of the word *dakle* extracted from hrWaC. Every example has been annotated by two annotators on two levels. The first level contains five distinct labels: sentence connective (conjunction), textual (discourse) connective, modifier (particle/adverb), filler word, or “other”. The second, discourse function level also contains five distinct labels, as identified by Dedaić (2010): conclusional, reformulational, argumentative/rhetorical, attitudinal, or “other”. We analysed and compared the distribution of the labels with the descriptions and categorizations of the word in Croatian monolingual dictionaries and web corpora.

The paper is structured as follows: Section 2 discusses general issues observed in lexicographic descriptions of synsemantic words, with emphasis on the contemporary Croatian monolingual (standard) language dictionaries. Section 3 analyses dictionary entries (the types and content of definitions, and part of speech classification) of the lexeme *dakle* as an example of a synsemantic polyfunctional word. Lexicographic descriptions are also compared to the features described in contemporary linguistic studies and grammar textbooks, as well as with the classification applied in the Croatian web corpus hrWaC. Section 4 focuses on the experimental methodology and the annotation results of labelling grammatical/discourse and pragmatic functions on corpus examples of the lexeme *dakle*, followed by Section 5 with a discussion and conclusion.

[1] *Hrvatski jezični portal / Croatian Language Portal* [HJP] (1991–2004), *Rječnik hrvatskoga jezika / Croatian Language Dictionary* [RHJ] (1998), *Školski rječnik hrvatskoga jezika / School Dictionary of Croatian Language* [ŠRHJ] (2012); *Veliki rječnik hrvatskoga standardnog jezika / Comprehensive Dictionary of Croatian Standard Language* [VRH] (2015).

2. Synsemantic words in (Croatian) dictionaries

On the semantic level, words are classified into two major classes: autosemantic (content / lexical / open-class) and synsemantic (empty / grammatical / functional / closed-class) word-forms. While autosemantic words have lexical meaning and refer to the extralinguistic world independent of their use, synsemantic words serve as functional units with grammatical (operational) meaning; they are used to mark the relations between the language units at a syntactic, semantic, and pragmatic level (Kunzmann-Müller, 1998: 239). In some cases, it is difficult to determine the border between autosemantic and synsemantic words, which is why Kordić (2002) introduces the intermediate category of *words on the border of lexicon and grammar*. The description of words in this intermediate category is a difficult task due to their oscillation between lexical and grammatical status, an alternation which can be observed in dictionaries and grammars of the Croatian language.

Based on an analysis of the descriptions of synsemantic words in Croatian dictionaries, Kunzmann-Müller (1998) concludes that the Croatian lexicography of synsemantic words is just beginning to develop. These language units have so far received fairly little attention as a result of the absence of an adequate theoretical and methodological apparatus, although they have always been included in Croatian dictionaries (ibid. 241-242). For this reason, Hoekstra (2010: 1009) points to the importance of implementing contemporary linguistic insights into lexicographic practice:

> To sum, it is important that lexicography stays in touch with the advances that are made in the disciplines of phonology, morphology, syntax and semantics as these disciplines may provide tools for structuring the encyclopedic information about words and collocations that is presented to the laymen who are the primary target group of dictionaries.

The example of the lexeme *dakle* allows us to present the problem of determining how part of speech makes lexicographic analysis and corpus annotation more difficult, and to identify the possible causes for problems with further classification.

While Croatian lexicography currently does not give much attention to synsemantic words, dictionaries specialized for particular synsemantic word classes do exist for some languages.[2] These approach the subject differently – while some merely list synsemantic words, others describe them in detail, across all language levels. The level of analysis here is, in large part, determined by dictionary type (e.g. a language learning dictionary vs. a monolingual dictionary). For example, Kobozeva and Zakharov (2004) note that a dictionary of discourse markers should include graphic, phonetic, syntactic, semantic, communicative, pragmatic, paralinguistic and derivational information in order to serve

[2] As an example we list only a few particle dictionaries: *Lexikon deutscher Partikeln* by Helbig (1988); *Dictionary of Slovenian Particles* by Žele (2015); Shimchuk & Shchur: *Slovar' russkix chastic* (1999); *A Dictionary of Japanese Particles* by Kawashima (2000); *A Dictionary of the Chinese Particles* by Dobson (1974) etc. It is worth emphasizing that the lexicographic analysis of individual types of synsemantic words varies greatly, according to their specific grammatical, semantic and functional features.

as a source of study for Russian language learners, but also as a source of further linguistic study. In a discussion about the definition of a lexeme, Hoekstra (2010)—calling upon the work of Bergenholtz (1985) and Coffey (2006)—states that an intentional definition (a paraphrased meaning) is not a suitable solution for synsemantic words, and calls for detailed syntactic descriptions followed by relevant examples of the word's use.

Osswald (2015) emphasizes that the lexicographic analysis of synsemantic words in monolingual dictionaries is especially problematic, because the definition cannot rely on a denotative meaning. He also explains that such dictionaries usually do not include the syntactic features (or functions) of synsemantic words because "the user is expected to have some basic knowledge of the respective language, and mastering the use function words is considered part of general grammatical competence" (ibid. 7). However, the author points to the "duty of documentation" in monolingual reference dictionaries and calling upon the work of Lang (1989), he concludes that lexicographic descriptions of synsemantic words should "[…] follow grammatical insights; syntactic constructions and their constraints should be part of the entry; and building the entry should consist of two stages, first, recording the relevant facts and, second, designing the final entry presentation" (Osswald, 2015: 7).

A lexicographic entry, thus, needs to mark the non-denotative meaning of the word; that is, according to Adamska-Sałaciak (2012), it needs to define the word "without describing the thing behind the word". She claims such metalinguistic definitions that describe usage and function have been in use for a long time:

> Thus, instead of defining an expression by describing its referent (i.e. the thing or situation named), a metalinguistic definition focuses on how the expression is used. It starts with a phrase such as: "(is) used to/for…", "when you/people say…", "you call sb a…", and proceeds to specify the function(s) which the expression serves in communication.

An analysis of synsemantic words in Croatian monolingual dictionaries (HJP, RHJ, ŠRHJ, VRH) reveals that metalinguistic definitions are, in most cases, absent. Observed definitions do not contain detailed information on the words' syntactic features, language use, and pragmatic functions. Grammatical descriptions are, in large part, reduced to part of speech classification, and this classification is inconsistent among observed dictionaries.

Synsemantic lexemes with multiple syntactic and pragmatic functions introduce additional problems. Descriptions of such words in Croatian dictionaries generally only partly describe their polyfunctionality. Thus, we will demonstrate this tendency in the following sections using the lexeme *dakle* as a case study.

3. An example of the polyfunctional synsemantic lexeme *dakle*

Entry 'dakle'[3]	**HJP/RHJ**	**ŠRHJ**	**VRH**
Part of speech categorization	conjunction	adverb	adverb
Lexicographic definitions' content			
Syntactic function	-	connective function in a compound sentence	connective function in a compound sentence
Semantic-pragmatic function	conclusional function	conclusional function	conclusional function
Synonym(s)	+	-	+

Table 1: The description of dictionary entries for the lexeme *dakle* within contemporary monolingual dictionaries of the Croatian (standard) language

The analysis of dictionary entries for the lexeme *dakle* (Eng. conj. *well, now; consequently* [Bujas, 1999] / *accordingly, so, then, therefore, thus* [Bujas, 2005]) within contemporary monolingual dictionaries of Croatian (standard) language (see Table 1) lead us to the following conclusions:

(1) Definitions of this lexeme in the analysed dictionaries are metalinguistic (followed by examples, and, in some cases, synonyms), but point to just one or two semantic-pragmatic functions: introducing a conclusion and/or a consequence. The function of introducing a conclusion is featured in relevant examples in all of the analysed dictionaries. An exception can be found in VRH, which lists *Što, dakle, ja tu mogu!?* (Eng. *So what can I do!?*), as an example for introducing a conclusion, an example we deem inappropriate, as it primarily represents the rhetorical and/or expressive function of the word. VRH is also the only dictionary to feature an example for introducing a consequence, although such a decision is questionable as well, as the function it serves better illustrates the function of introducing a conclusion (*Uzeo je stvari, dakle odlazi na put* / Eng. *He took his stuff; therefore, he is going on a trip*).

[3] (1) **dȁklē** *vezn.* – označuje zaključak ili posljedicu [*dakle, to smo se dogovorili*; *dakle, stigao si*]; prema tome, onda, i zato, pa zato [HJP, RHJ]; (2) **dȁklē** *pril.* 1. uvodi zaključak [*Ti, dakle, odlaziš.*] 2. ima vezničku funkciju u nezavisnosloženoj zaključnoj rečenici [*Uzeo je stvari, dakle odlazi na put.*] [ŠRHJ]; (3) **dàkle** *pril* 1. uvodi zaključak [*Ti, dakle, odlaziš.*; *Alkohol šteti, dakle valja ga izbjegavati.*; *Što, dakle, ja tu mogu?!*]; 2. <*u vezn. službi*> u nezavisnosloženoj zaključnoj rečenici označuje posljedicu [*Uzeo je stvari, dakle odlazi na put.*]; *Sin.* elem, ergo, prema tome [VRH].

According to a pragmatic study by Mirjana N. Dedaić (2010)[4], the lexeme *dakle*, when observed as a discourse particle, accomplishes multiple functions: "*Dakle* seems to have developed four principal functions in discourse: (a) conclusional; (b) reformulational; (c) argumentative/rhetorical; and (d) attitudinal" (ibid. 129). Considering the first two functions, the author states:

> *Dakle* occurs by and large in two environments roughly defined as environment (1), in which dakle marks a **causative-resultative relationship** [sic] between S1 and S2, and (2) in which it marks S2 to be a reformulation of S1, with consequential inferences. (ibid.)

The author additionally states that these two functions (conclusional and reformulational) are not necessarily mutually exclusive. The other two functions of the lexeme *dakle* (argumentative/rhetorical and attitudinal) are listed as secondary and originate from its conclusional function, "[...] which allows for occasional manipulation in recipient's reasoning. It also incites attitudes towards unfulfilled expectations, allowing for attitude-revealing explicatures" (ibid. 110).

Considering that the analysed entries of the lexeme *dakle* capture only one of its four listed functions (*cf.* Dedaić, 2010), namely the conclusional function, the representation of other functions (reformulational, argumentative/rhetorical, and attitudinal) is a matter requiring further investigation and inclusion in the lexicographic descriptions of the word.

(2) Part of speech classification of the lexeme *dakle* is inconsistent among the analysed dictionaries. While in two dictionaries (HJP, RHJ) it is categorized as a conjunction, the other two (ŠRHJ, VRH) categorize it as an adverb, wherein the lexicographic definition contains information of its connective function. Such inconsistencies are likewise consistent in Croatian language grammar textbooks and linguistic studies, in which the lexeme is listed as a conjunction, a textual connector, a particle, a modal word, a modifier, a discourse marker/particle, an adverb, or a filler word. This can be seen as a reflection of the polyfunctionality of the lexeme, but also a consequence of applying different approaches to uninflected words. The origin of the observed methodological problems include the following: (1) difficulties with differentiating the traditional part of speech categories—in this case, conjunctions, particles, and adverbs; (2) limitations of traditional grammar focused only on the sentence level; and (3) more recent application of contemporary linguistic (text/discourse oriented) approaches, an application which opens new issues (notably terminological inconsistencies and diverse interpretations of "new" terms)[5].

While the lexeme *dakle* is classified as either an adverb or a conjunction in the four

[4] The study is based on an analysis conducted on the examples "collected from conversation events, media talk shows and reports, various written material (Internet, newspapers, and books), and the Croatian National Corpus, which includes journalistic texts, essays, and fiction—more than three thousand occurrences in total)" (Dedaić, 2010: 210-112).

[5] More discussion on this topic is available in works of Badurina (2009) and Glušac (2012).

analysed dictionaries, in the online language corpus hrWaC it is labelled as a conjunction in over 99% of instantiations.

For these reasons, we conducted a corpus-based study to investigate the (supra)syntactic and pragmatic polyfunctionality of the lexeme *dakle* in order to identify the correlation between the existing linguistic/lexicographic descriptions and its (written) language use.

4. Polyfunctionality of lexeme *dakle*: a corpus-based experiment

4.1 Methodology

We conducted a corpus-based experiment on two different annotation tasks to investigate polyfunctionality of lexeme *dakle*. We calculated the sample size needed for the experiment taking into account the total size of the population (the size of hrWaC containing over 1.3 billion tokens), a margin of error of 5%, and a confidence level of 95%. The number of 385 was rounded up to 400 random KWIC examples from hrWaC[6].

Step 1

Conjunction[7]	Conjunctions are uninflected words which connect words, word groups, or clauses within complex sentences.
Textual connector[8]	Connectors organize and signal relations between the text/discourse components.
Modifier **(particle or adverb)**[9]	Syntactically independent words that modify the sentence meaning.
Filler words[10]	Syntactically independent words used unconsciously/automatically, without any connection to their meaning.
Other	

Table 2: Annotation scheme for determining the (supra)syntactic function

[6] We believe that hrWaC is an adequate Croatian corpus for pragmatic research, as it contains documents from varied sources, and not only documents written in standard language like newspaper articles and literary texts (e.g. Croatian National Corpus and Croatian Language Corpus).

[7] An example from instructions for annotators: *Danas ne mogu doći na košarku, dakle igrat ćete bez mene.* (Eng. *Today I cannot come to a basketball practice so you'll play without me.*)

[8] An example from instructions for annotators: *Dakle, na temelju svega što je u članku izneseno proizlaze sljedeći zaključci ...* (Eng. *Therefore, based on everything in the article, the following conclusions are ...*)

[9] An example from instructions for annotators: *To, dakle, stvarno nije bilo lijepo od tebe.* (Eng. *Well, that was really not nice of you.*)

[10] From instructions for annotators: *Although the filler words are a feature primarily of oral language production, they are listed here as a possible category. If, in the examples presented, the annotators notice an unnecessary accumulation of the lexeme* dakle*, it is possible to categorize it as a filler word.*

Step 2

Conclusion[11]	Introducing the conclusion which logically stems from the previous discourse, but is not explicitly stated.
Reformulation[12]	Reformulating a statement which has previously been explicitly stated in the discourse. The reformulation can include: (a) expansion of the previous statement (b) summary of the previous statement
Argumentative / rhetorical function[13]	(a) discourse organization (initiating the act of communication, changing the subject, returning to the subject etc.) (b) rhetorical questions (c) enticing the collocutor (d) persuading the collocutor
Attitudinal function[14]	Expressing the locutor's emotions, attitudes, or states in reference to the collocutor or the contents of the utterance.
Other	

Table 3: Annotation scheme for determining the semantic-pragmatic function

Annotation of the examples from the corpus was undertaken in two steps: (1) determining the (supra)syntactic function; (2) determining the semantic-pragmatic function of the word in discourse. In both steps, the annotators were required to choose one of the five possible categories (see Tables 2 and 3). In determining the (supra)syntactic function, annotators had the option of labelling the lexeme *dakle* as a conjunction, a textual connector, a modifier or a filler word. The fifth category was the option "other" if the annotators could not decide on one of the offered possibilities. In the second step, to determine the semantic-pragmatic function of the word in discourse, we followed Dedaić's classification (2010). The annotators had the option of choosing if the word functioned as a conclusion, a reformulation, had an argumentative/rhetorical or an attitudinal function. As in the first step, the final category was the option "other" if the annotators could not decide on one of the offered possibilities.

The two annotators had a high level of education in linguistics. They were remotely trained and given precise instructions containing definitions and illustrative examples for each of the categories offered. They had no prior experience in corpus annotation, worked separately during the annotation tasks, and had no restriction on time.

In order to evaluate the annotated examples we used accuracy as well as Cohen's κ

[11] An example from instructions for annotators: *A: Spremi se, doći ćemo po tebe u sedam. B: Dakle, na večeru idemo poslije predstave.* (Eng. *A: Get ready, we'll pick you up at seven. B: So, we're going to dinner after the show.*)

[12] An example from instructions for annotators*: Takvu ružnu stvar si rekla mom najboljem prijatelju, dakle, Ivanu.* (Eng. *You said this ugly thing to my best friend, [dakle] to John.*)

[13] An example from instructions for annotators: *Dakle, zovem se Andrej i imam 16 godina.* (Eng. *So, my name is Andrej and I am 16 years old.*)

[14] An example from instructions for annotators: *Mislim, dakle, stvarno si neodgovoran.* (Eng. *I mean, [dakle], you are really irresponsible.*)

(Cohen 1960), as it is the predominant reliability measure of corpus annotation used in NLP due to the work of Carletta (1996). Cohen's κ was developed for two annotators and nominal data, as is the case with our experiment. We considered using Krippendorff's α, but Antoine et al. (2014) concluded that there is no benefit in using this measure on nominal data. Additionally, we would like to point out that we are aware the annotation process in the domain of pragmatics is highly affected by the annotators' subjectivity. In the next section we present the results of our research.

4.2 Results

4.2.1 Distribution of the annotation categories

Table 4 presents the distribution of the annotation categories for determining the (supra)syntactic function of the lexeme *dakle*.

	Annotator A	**Annotator B**	**Total**
Conjunction	150 (37.5%)	69 (17.25%)	219 (27.38%)
Textual connector	246 (61.5%)	211 (52.75%)	457 (57.13%)
Modifier (particle or adverb)	3 (0.75%)	117 (29.25%)	120 (15%)
Filler words	1 (0.25%)	3 (0.75%)	4 (0.5%)
Other	0 (0%)	0 (0%)	0 (0%)
Total	400	400	800

Table 4: Distribution of annotation categories for determining the (supra)syntactic function

It is obvious that the categories are not balanced, as the textual connector accounts for more than half (57.13%) of all labels. The next two categories vary between annotators. While annotator A's second most frequent choice was conjunction (37.5%), for annotator B it was the third most frequent choice (17.25%). Modifier is a category with the most drastic difference between annotators: while annotator A chose it in only 0.75% of the cases, annotator B chose it in 29.25% of the cases. Both annotators agreed that the lexeme *dakle* was rarely a filler word (0.5%), and none of them selected the option "other".

Table 5 presents the distribution of the annotation categories for determining the semantic-pragmatic function of the word in discourse. From the data we can conclude that the distribution for the second step is overall more balanced between three categories (the argumentative/rhetorical function 40.5%, reformulation 32.38%, conclusion 26%). However, when examining each annotator separately, we can observe that each annotator has a different category prevailing. Annotator A chose the rhetorical and interactional function 56% of the time, while annotator B chose reformulation 43.5% of the time. As with the first step, both annotators agree that the lexeme *dakle* rarely serves as an attitudinal marker (1.13%), and none of them selected the option "other".

	Annotator A	Annotator B	Total
Conclusion	88 (22%)	120 (30%)	208 (26%)
Reformulation	85 (21.25%)	174 (43.5%)	259 (32.38%)
Argumentative / rhetorical function	224 (56%)	100 (25%)	324 (40.5%)
Attitudinal function	3 (0.75%)	6 (1.5%)	9 (1.13%)
Other	0 (0%)	0 (0%)	0 (0%)
Total	400	400	800

Table 5: Distribution of annotation categories for determining the semantic-pragmatic function

4.2.2 Data reliability

We used accuracy as well as Cohen's κ to measure data reliability for both steps, since it considers the possibility of the agreement occurring by chance. The results are shown in Table 6. The accuracy for determining the (supra)syntactic function is 0.655, while for the semantic-pragmatic function it is 0.5025. Before interpreting the results, we calculated Cohen's κ for both annotation tasks. For the first task of determining the (supra)syntactic function, the result is 0.4332, while for the second task of determining the semantic-pragmatic function it is 0.2908. It is still not agreed upon as to what constitutes a good agreement, i.e. how to interpret Cohen's κ. According to Landis and Koch (1977)[15], for the (supra)syntactic function we have a moderate agreement, while for the semantic-pragmatic function we have a fair agreement. Altman (1990) proposed a slightly modified interpretation[16], but the interpretation of our results stays the same (moderate and fair agreement, respectively). On the other hand, Fleiss et al. (2013) proposed another interpretation[17]. According to them, for the (supra)syntactic function we have fair to good agreement, but for the semantic-pragmatic function we have poor agreement.

We tend to agree with Fleiss et al.'s (2013) interpretation of Cohen's κ. We believe that we have attained a fair agreement between annotators for the first task of determining the (supra)syntactic function of the lexeme *dakle*. However, we are aware of the disproportionate distribution of categories for this task, which might skew the results in our favour. For the second task of determining the semantic-pragmatic

[15] Landis and Koch (1977) proposed the following interpretation of Cohen's κ: < 0.0 poor agreement; 0.00 – 0.20 slight agreement; 0.21 – 0.40 fair agreement; 0.41 – 0.60 moderate agreement; 0.61 – 0.80 substantial agreement; 0.81 – 1.00 almost perfect agreement.

[16] Altman (1990) proposed the following interpretation of Cohen's κ: 0.00 – 0.20 poor agreement; 0.21 – 0.40 fair agreement; 0.41 – 0.60 moderate agreement; 0.61 – 0.80 good agreement; 0.81 – 1.00 very good agreement.

[17] Fleiss et al. (2013) proposed the following interpretation of Cohen's κ: < 0.40 poor agreement; 0.40 – 0.75 fair to good agreement; > 0.75 excellent agreement.

function of the word in discourse, we attained poor agreement between annotators. We believe the reason for this is that the categories in this task are not mutually exclusive, as Dedaić (2010) pointed out. In order to investigate this matter further, in following sections we analyse in more detail: (1) agreements and disagreements between annotators for each task, and (2) the combination of categories between annotation tasks.

	Accuracy	**Cohen's κ**
(Supra)syntactic function	0.655	0.4332
Semantic-pragmatic function	0.5025	0.2908

Table 6: Reliability measures

4.2.3 Analysis of agreements and disagreements between annotators

The next step was to analyse how many times the annotators agreed and on what categories, as well as how many times they disagreed and what were the categories that could be interpreted as "interchangeable". Table 7 presents the frequency distribution of the agreements and disagreements for the first task of determining the (supra)syntactic function.

Agreements		**Disagreements**	
Categories	**Frequency**	**Categories**	**Frequency**
Textual connector	204 (51%)	Conjunction (for annotator A) and Modifier (for annotator B)	87 (21.75%)
Conjunction	56 (14%)	Textual connector and Modifier	29 (7.25%)
Modifier	3 (0.75%)	Conjunction and Textual connector	18 (4.5%)
		Conjunction and Filler words	2 (0.5%)
		Textual connector and Filler words	2 (0.5%)

Table 7: Distribution of agreements and disagreements for determining the (supra)syntactic function

We will first focus on agreements, and then on disagreements. The annotators agreed the most on when the lexeme *dakle* had the function of a textual connector (51%), which is expected since over half of the labels for this task were annotated with this category. The annotators agreed 14% of the time the lexeme had the function of a conjunction and only 0.75% of the time that it had a modifier function. It is worth mentioning that none of the annotators chose the option "other", which is also considered an agreement.

Analysing disagreements, we found an anomaly in that annotator A labelled an example as a conjunction, while annotator B labelled the same example as a modifier. However, there is not one instance of a vice versa case (annotator A labelling an example as a modifier and annotator B labelling it as a conjunction). We find this result very peculiar and one that needs to be investigated further, possibly by increasing the number of annotators. Other cases of disagreements had instances of a vice versa case (e.g. annotator A choosing X and annotator B choosing Y, as well as annotator A choosing Y and annotator B choosing X). In 7.25% of the instances, the annotators interchanged the labels of a textual connector with a modifier, and in 4.5% of the instances interchanged a conjunction and a textual connector.

Table 8 presents the frequency distribution of the agreements and disagreements for the second task of determining the semantic-pragmatic function of the word in discourse.

Agreements		**Disagreements**	
Categories	**Frequency**	**Categories**	**Frequency**
Argumentative/rhetorical function	87 (21.75%)	Conclusion and Argumentative/rhetorical function	89 (22.25%)
Reformulation	79 (19.75%)	Reformulation and Argumentative/rhetorical function	53 (13.25%)
Conclusion	35 (8.75%)	Conclusion and Reformulation	48 (12%)
		Argumentative/rhetorical function and Attitudinal function	8 (2%)
		Conclusion and Attitudinal function	1 (0.25%)

Table 8: Distribution of agreements and disagreements for determining the semantic-pragmatic function

As with the previous task, we will first focus on agreements, and then on disagreements. The annotators agreed the most on when the lexeme *dakle* had the argumentative/rhetorical function (21.75%). Similarly, the annotators agreed 19.75% of the time the lexeme was used for reformulation. Only 8.75% of the agreements were on the conclusional function. Analogous to the first task, none of the annotators chose the option “other”, which we also consider an agreement. When analysing disagreements, the annotators mostly disagreed between the conclusional function and the argumentative/rhetorical function (22.25%). In 13.25% of instances the annotators interchanged the reformulational and the argumentative/rhetorical functions, while disagreement between the conclusional and the reformulational functions occurred 12% of the time.

4.2.4 Analysis of combination of categories between annotation tasks

In this section, for each annotator we analyse combinations of categories between the two annotation tasks, i.e. what category they selected for the first task and what category they selected for the second. The detailed results for annotator A and annotator B are presented in Table 9.

From the data it is evident that annotator A has more stable combinations of categories than annotator B. For example, annotator A covers 97% of all annotations with the top 5 combinations or 98.5% with the top 6. On the other side, annotator B has more combinations. With the top 5 combinations they cover 84% of all annotations, while with the top 6 they cover 88.5%. It takes the top 9 combinations for annotator B to cover 98% of all annotations. This data shows that every time annotator A selects a certain category in the first task, they are more likely to consistently select the same category in the second. On the other hand, every time annotator B selects a certain category in the first task, they are more likely to change categories for the second task.

5. Discussion and conclusion

The analysis of the functions and the use of the lexeme *dakle*, based on language material extracted from the Croatian web corpus hrWaC, has shown discrepancies between corpus data and dictionary descriptions. In Croatian monolingual dictionaries the lexeme is categorized as either a conjunction or an adverb, while in the corpus over 99% of occurrences are labelled as a conjunction. Our experiment has shown that in most cases (57.13%) the annotators have labelled the lexeme as a textual connector, while in considerably fewer cases they labelled it as a conjunction (27.38%) or a modifier (particle or adverb) (15%). However, we are aware of the great imbalance between annotators regarding the modifier category: while annotator A selected this category in only 0.75% of cases, annotator B selected it in 29.25%.

The disagreement between annotators regarding the first task is expected, due to the already mentioned issues with part of speech categorizations and grammatical descriptions of synsemantic (poly)functional words (presented in Section 3). It is also expected that the function of the filler word is confirmed in only 0.5% of the cases, due to hrWaC not containing spoken language material.

Annotator A		**Annotator B**	
Category of task 1 and Category of task 2	**Frequency**	**Category of task 1 and Category of task 2**	**Frequency**
Textual connector and Argumentative/rhetorical function	192 (48%)	Textual connector and Conclusion	92 (23%)
Conjunction and Reformulation	79 (19.74%)	Modifier and Reformulation	87 (21.75%)
Textual connector and Conclusion	46 (11.5%)	Textual connector and Argumentative/rhetorical function	71 (17.75%)
Conjunction and Conclusion	42 (10.5%)	Textual connector and Reformulation	44 (11%)
Conjunction and Argumentative/rhetorical function	29 (7.25%)	Conjunction and Reformulation	42 (10.5%)
Textual connector and Reformulation	6 (1.5%)	Modifier and Argumentative/ rhetorical function	18 (4.5%)
Modifier and Argumentative/rhetorical function	2 (0.5%)	Conjunction and Conclusion	17 (4.25%)
Textual connector and Attitudinal function	2 (0.5%)	Modifier and Conclusion	11 (2.75%)
Modifier and Attitudinal function	1 (0.25%)	Conjunction and Argumentative/rhetorical function	10 (2.5%)
Filler words and Argumentative/rhetorical function	1 (0.25%)	Textual connector and Attitudinal function	4 (1%)
		Modifier and Attitudinal function	1 (0.25%)
		Filler words and Argumentative/rhetorical function	1 (0.25%)
		Filler words and Attitudinal function	1 (0.25%)
		Filler words and Reformulation	1 (0.25%)

Table 9: Distribution of combination of categories between annotation tasks for annotator A and annotator B

We would like to point out one unexpected result regarding a disagreement between a textual connector and a modifier. The traditional grammar focused only on the sentence level includes the textual connectors within adverbs. Therefore, we expected the disagreement between these two categories to be larger than our data confirmed (only 7.25%). The analysis of semantic-pragmatic function of the lexeme *dakle* confirmed its polyfunctionality. In Croatian monolingual dictionaries, the definitions point to just one or two semantic-pragmatic functions: introducing a conclusion and/or a consequence. Our experiment has shown that in most cases (40.5%) the annotators labelled the lexeme with the argumentative/rhetorical function. Unexpectedly, even the reformulation is more frequent (30.38%) than the conclusional function (26%). Since Dedaić (2010) stated that the conclusional and the reformulational functions are not necessarily mutually exclusive, we expected these two categories to be interchangeable among annotators. However, our data demonstrates that the annotators disagree on these two categories in only 12% of the cases. A larger disagreement is confirmed between the argumentative/rhetorical function and the conclusional function (22.25%), while a similar disagreement is confirmed between the argumentative/rhetorical function and the reformulational function (13.25%). According to Dedaić (2010), the argumentative/rhetorical function originates from the conclusional function, which explains the aforementioned disagreement. We deduce that the lexeme *dakle* simultaneously performs more than one of these three functions proposed by Dedaić (2010). Our experiment hardly found the fourth attitudinal function (1.13%).

We find the combination of categories between annotation tasks very intriguing, as we are not certain if the (in)consistency of an annotator is indicative of their quality (due to the highly subjective annotation task in the field of corpus pragmatics research). As both annotation tasks are performed simultaneously, we cannot be sure of how one task influenced the other. In future work it would be beneficial to perform the annotation tasks separately.

We believe the experiment proves: (1) the polyfunctionality of the lexeme *dakle*, (2) the simultaneous multiple functionality of the lexeme, and (3) vague boundaries between (supra)syntactic and the semantic-pragmatic categories. It is our opinion that monolingual dictionaries for native speakers, like the ones analysed in our study, should contain lexicographic descriptions of all (or at least most frequent) functions of synsemantic words. Our pilot study has indicated that the functions of the lexeme *dakle* are not equally distributed. However, to identify a more precise frequency distribution of its functions, it is necessary to conduct a more extensive study that would include more annotators and, possibly, more corpus examples. With such information lexicographers can define and apply the criteria for structuring dictionary entries (e.g. the order or selection of functions defined). Dictionary entries of polyfunctional synsemantic words should contain metalinguistic definitions and usage descriptions, supported by illustrative examples based on language corpora. The analysis of language corpus data can improve linguistic (and thereby lexicographic) descriptions of such words, which will become a much-needed form of reciprocal feedback for adequate processing of language corpora. Since Croatian monolingual dictionaries do not offer a methodical, exhaustive, and thorough lexicographic descriptions of polyfunctional synsemantic words, our pilot study offers an insight into developing an accepted procedure of their corpus-based processing and presentation.

6. Acknowledgements

We would like to thank the two annotators Sanja Šakić and Mateja Zirdum.

7. References

Adamska-Sałaciak, A. (2012). Dictionary definitions: problems and solutions. *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, *2012*(4), 323-339.

Altman, D. G. (1990). Practical statistics for medical research. CRC press.

Anić, V. (1998). *Rječnik hrvatskoga jezika.* Novi liber. [RHJ]

Antoine, J. Y., Villaneau, J., & Lefeuvre, A. (2014, April). Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *EACL 2014* (10 p).

Badurina, L. (2008). *Između redaka: studije o tekstu i diskursu.* Hrvatska sveučilišna naklada, Zagreb.

Bergenholtz, H. (1985). Vom wissenschaftlichen Wörterbuch zum Lernerwörterbuch. In *Lexikographie Und Grammatik. Akten Des Essener Kolloquiums Zur Grammatik Im Wörterbuch 28.-30.6. 1984.* Max Niemeyer Verlag.

Birtić, M., Blagus Bartolec, G., Hudeček, L., Jojić, L., Kovačević, B., Lewis, K., & Vidović, D. (2012). *Školski rječnik hrvatskoga jezika.* Školska knjiga: Institut za hrvatski jezik i jezikoslovlje, Zagreb. [ŠRHJ]

Bujas, Ž. (1999). *Veliki hrvatsko-engleski rječnik.* Globus, Zagreb.

Bujas, Ž. (2005). *Veliki englesko-hrvatski rječnik.* Globus, Zagreb.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, *22*(2), pp. 249-254.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), pp. 37-46.

Coffey, S. (2006). High-frequency grammatical lexis in advanced-level English learners' dictionaries: From language description to pedagogical usefulness. *International Journal of Lexicography*, *19*(2), pp. 157-173.

Dedaić, M. (2010). Reformulating and concluding: The pragmatics of the Croatian discourse marker *dakle.* In M. Dedaić & M. Mišković-Luković (eds.) *South Slavic Discourse Particles.* Amsterdam: John Benjamins Publishing Company, pp. 107-131.

Dobson, W. A. (1974). *A Dictionary of the Chinese Particles: With a Prolegomenon in which the Problems of the Particles are Considered and They are Classified by the Grammatical Functions.* University of Toronto Press.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions.* John Wiley & Sons.

Glušac, M. (2012). Prilozi kao vrsta riječi u hrvatskoj jezikoslovnoj literaturi. In M. Turk & I. Srdoč-Konestra (eds.): *Proceedings of the Fifth Slavistic Congress.*

Rijeka: Filozofski fakultet, pp. 405-413.

Helbig, G. (1988). *Lexikon deutscher Partikeln.* Verlag Enzyklopädie, Leipzig.

Hoekstra, E. (2010). Grammatical information in dictionaries. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress.* Afûk, Ljouwert: Fryske Akademy, pp. 1007-1012.

Hrvatski jezični portal / Croatian Language Portal: http://hjp.znanje.hr. [HJP]

Jojić, L. (Ed.). (2015). *Veliki rječnik hrvatskoga standardnog jezika.* Školska knjiga, Zagreb. [VRH]

Kawashima, S. A. (1999). *A Dictionary of Japanese Particles.* Tokyo: Kodansha International.

Kobozeva, I. M., & Zakharov, L. M. (2004). Types of information for the multimedia dictionary of Russian discourse markers. In *9th Conference Speech and Computer.*

Kordić, S. (2002). *Riječi na granici punoznačnosti.* Hrvatska sveučilišna naklada, Zagreb.

Kunzmann-Müller, B. (1998). Opis sinsemantičkih riječi u rječniku-izazov leksikologiji i leksikografiji. *Filologija*, (30-31), pp. 239-248.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pp. 159-174.

Lang, E. (1989). Probleme der Beschreibung von Konjunktionen im allgemeinen einsprachigen Wörterbuch. In F. J. Hausmann et al. (eds.)*. Wörterbücher, dictionaries, dictionnaires. Ein internationales Handbuch zur Lexikographie*, *1*, pp. 862-868.

Ljubešić, N., & Klubička, F. (2014). {bs, hr, sr}WaC-web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9),* pp. 29-35.

Osswald, R. (2015). Syntax and Lexicography. In Alexiadou, A. & Kiss, T. (eds.). *Syntax – Theory and Analysis. Volume 3, Handbooks of Linguistics and Communication Science, 1963–2000.* De Gruyter. (preprint)

Pintarić, N. (2002). *Pragmemi u komunikaciji.* Zavod za lingvistiku filozofskog fakulteta Sveučilišta u Zagrebu.

Shimchuk, E. G., & Shchur, M. G. (1999). *Slovar'russkikh chastits* [*Dictionary of Russian particles*]. Peter Lang - Europäische Verlag der Wissenschaften: Frankfurt am Main.

Silić, J., & Pranjković, I. (2005). *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta.* Školska knjiga, Zagreb.

Žele, A. (2015). *Dictionary of Slovenian Particles.* Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1128.