CroaTPAS: A Survey-based Evaluation

Costanza Marini
[1]University of Pavia, [2]University of Bergamo
[1]University of Pavia, Dipartimento di Studi Umanistici, Strada Nuova 65, 27100 Pavia
[2]University of Bergamo, Dipartimento di Lingue, Letterature e Culture Moderne, via Salvecchio 19, 24129 Bergamo

costanza.marini@unibg.it

Abstract

The Croatian Typed Predicate Argument Structures resource (CroaTPAS, Marini and Ježek, 2019) is a Croatian/English bilingual digital dictionary of corpus-derived verb valency structures, whose argument slots have been annotated with Semantic Types labels following the CPA methodology (Hanks, 2013). CroaTPAS is tailor-made to represent verb polysemy and currently contains 180 Croatian verbs for a total of 683 different verbs senses. In order to evaluate the resource both in terms of identified Croatian verb senses, and of the English descriptions explaining them, an online survey based on a multiple-choice sense disambiguation task was devised, pilot tested and distributed among respondents following a snowball sampling methodology. Answers from 30 respondents were collected and compared against a yardstick set of answers in line with CroaTPAS's sense distinctions. Jaccard similarity index was used as a measure of agreement. Since the multiple-choice items respondents answered to were based on a representative selection of CroaTPAS verbs, they allowed for a generalization of the results to the whole of the resource.

Keywords: Croatian, pattern, survey

1. Introduction

CroaTPAS (Marini and Ježek, 2019) is a Croatian/English bilingual digital dictionary focusing on representing verb polysemy. It currently contains the semantically annotated corpus-derived verb valency structures of a selection of 180 Croatian verbs, which will be made freely available online by the end of 2022.

In order to evaluate the overall goodness of the resource, both in terms of adequacy of the identified Croatian verb senses, and of the English descriptions explaining them, an online survey mainly consisting of a multiple-choice sense disambiguation task was devised, pilot tested and later distributed among candidate respondents using a snowball sampling methodology.

The items respondents were presented with contained a selection of verbs deemed representative of the whole resource, thus allowing for a generalization of the results to the whole of CroaTPAS.

2. The resource

The Croatian Typed Predicate Argument Structures resource (CroaTPAS, Marini and Ježek, 2019) is a digital lexicographic resource containing a collection of corpus-derived Croatian verb valency structures, whose argument slots have been manually annotated with a hierarchy of semantic labels called System of Semantic Types (Ježek 2019).

Like its Italian sister project T-PAS (Ježek et al., 2014), CroaTPAS is primarily conceived for representing verb polysemy, since each semantically typed verb argument structure in its inventory – henceforth called *pattern* – corresponds to a different verb sense. In its inventory, the resource currently contains 180 Croatian verbs, for a total of 683 different patterns.

2.1 Generative Lexicon Theory

According to Generative Lexicon Theory, which is the shared theoretical framework both resources rely on (Pustejovsky, 1995; Pustejovsky and Ježek, 2008), verb meaning is conceived as "contextually generated" by the interaction between the semantics of the verb and that of its arguments (Figure 1).

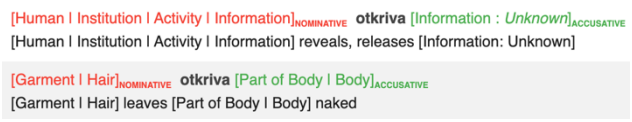

Figure 1: CroaTPAS patterns encoding two of the meanings of the verb otkrivati (Eng. 'to reveal')

For instance, all the corpus lines linked to the first pattern of the Croatian verb *otkrivati* (Eng. 'to reveal') above contain direct objects that may be classified as unknown pieces of [Information], thus generating the meaning of "releasing that information". On the other hand, all the corpus lines containing a [Garment] or [Hair] as subject and a direct object typed as [Part of Body] or [Body] generate the meaning of "leaving that body part naked".

2.2 CPA Methodology

The resource methodology is a customized version of Corpus Pattern Analysis (Hanks, 2013), a lexicographic methodology resting on the idea that meaning should be mapped onto its prototypical contexts of use.

CPA usually requires the following four steps: 1) 250 corpus lines are randomly sampled for each verb from a reference corpus, in this case, the Croatian Web as Corpus (Ljubešić and Klubička, 2014), a web-crawled reference corpus of standard Croatian containing 1.2 billion tokens; 2) the different verb senses are identified by the lexicographer; 3) pattern strings are created in a pattern editing environment labelling argument slots with the appropriate Semantic Types and, finally, 4) numbers are assigned to the corpus lines exemplifying each identified pattern, so that each semantically tagged valency structure is justified by corpus evidence.

In CroaTPAS, underneath each pattern string, users will also be presented with an English definition of the verb meaning portrayed above, as you can see in Figure 1. These definitions go by the name of "sense descriptions" and contain the same Semantic Types used in the corresponding

pattern string. They were written in English in order to make CroaTPAS a bilingual online resource available to Croatian language learners.

3. The survey

To evaluate the overall goodness of both the identified verb senses stored in CroaTPAS, and the English sense descriptions elucidating them, it was decided to administer an online multiple-choice questionnaire aimed at native speakers of Croatian with a good command of the English language, as well as to individuals with native-like Croatian proficiency.

In the multiple-choice section of the survey, respondents had to carry out a verb sense disambiguation task on a selection of 91 corpus examples with GDEX values (Kilgarriff et al. 2008) higher than 0.8 extracted from the Croatian Web as Corpus via the Sketch Engine (Kilgarriff et al. 2014). GDEX is an algorithm able to identify Good Dictionary EXamples by assigning corpus sentences a score ranging from 0 to 1 based on their lexical and syntactic complexity.

Each example featured one of the recorded verb senses of the CroaTPAS verb under scrutiny and was followed by English alternative sense explanations to choose from, corresponding to the array of English sense descriptions available in CroaTPAS underneath the patterns of that specific verb. Here is an example of multiple-choice item.

(0) Kako *podnijeti* ljetne vrućine, a osjećati se ugodno?[1]

o [Human | Human Group] can stand, endures [Anything: Negative]
o [Human | Institution] submits, files [Document | Request | Offer]

Before starting data collection, the questionnaire was briefly piloted by a group of two respondents, whose feedback contributed to improving the survey (see § 3.3).

As for the sampling method, the choice fell upon *snowball sampling* (Johnstone Young 2016: 169, Dörnyei 2007: 98), which consists in contacting a small group of good candidate participants, who are then asked to generate a chain reaction forwarding the survey to other appropriate candidate participants among their contacts. Given this choice of method, the evaluation survey was presented as a Google Form, i.e., a free online survey which can be built using Google Suite and easily forwarded via link.

All instructions were given in English to ensure that respondents did realize the need to be proficient not only in Croatian but also in English to be able to carry out the verb sense disambiguation task that constitutes the main bulk of the survey.

It was also decided not to mention the name of the resource in any part of the survey, nor to go into technical details when it comes to verb polysemy, so as not to distract respondents from the task at hand.

Special attention was devoted to thanking and reassuring respondents of the confidentiality and anonymity of their answers, as well as of the availability of the author to answer any possible question concerning the project. To comply with the ethical principle of informed consent, respondents were explicitly asked to submit the form only if they accepted that their anonymous answers were going to be used for research purposes

3.1 Background Information

In light of the fact that asking for demographic information at the start of a questionnaire can be off-putting (Fife-Schaw 2006), background questions were asked after the multiple-choice section. Questions included both open questions, multiple-choice items, and three sentence completion items involving semantic differential scales (Dörnyei 2007: 105). Two of the latter were designed to let respondents complete statements concerning their language proficiency in English and Croatian by marking a 5-step continuum between two polar adjectives, namely *basic* and *excellent*, in order "to elicit a more meaningful answer than a simple question" (*ibidem*, 107).

3.2 Verb Selection

Out of the 180 verbs in CroaTPAS, 32 were excluded since they only feature one sense and could thus not be used in a sense disambiguation task such as the one devised for the survey. To provide respondents with corpus examples from a representative selection of CroaTPAS verbs, the 148 remaining entries were divided by pattern number as well as aspect, and percentages were subsequently calculated for each verb class.

To guarantee a verb selection representative of the whole resource, we decided to keep the percentages fixed and determine how many verbs would have to be chosen for each class given an arbitrary total of 20 verbs. Given their paucity, three biaspectual verbs were included by default in the poll to guarantee their evaluation. Table 1 provides a complete overview of the final selection of verbs included in the survey after pilot testing it.

3.3 Pilot testing

Following Johnstone Young (2016: 176), the questionnaire was piloted before beginning with data collection. The pilot group included two respondents, who were asked to complete the draft survey and reflect on its design, the wording of items and the clarity of the example sentences.

The items from the background information section were deemed clear and able to capture the background of both respondents.

Both appreciated the presence of a non-binary gender option, and both agreed that asking for participants who had either a native or "a native-like proficiency of Croatian" was a good way to include not only foreigners, but also Serbian, Bosnian and Montenegrin native speakers.

The most important amendment made after the pilot testing phase was eliminating all multiple-choice items based on verbs with 7 and 11 senses, since both participants found that skimming through multiple-choice lists containing that many senses took too hard a toll on their attention levels. Moreover, since the survey was deemed quite long, it was decided to remove the items for one of the 5-pattern imperfective verbs, too.

The following Table offers an overview of the final 19 verbs included in CroaTPAS's evaluation survey after pilot testing, corresponding to a total of 91 items.

[1] "How to *endure* summer heat and feel comfortable?"

Perfective	**N**	**verbs**
2P	2	podnijeti, prekinuti
3P	2	isključiti, sletjeti
4P	2	otkriti, ubiti
5P	1	prodati
6P	1	popiti
Imperfective	**N**	**verbs**
2P	2	gostiti, željeti
3P	2	čitati, kupovati
4P	2	osnovati, slati
5P	1	voziti
6P	1	žderati
Biaspectual	**N**	**verbs**
2P	1	informirati
3P	1	napredovati
5P	1	kontrolirati

Table 1: The final selection of verbs included in CroaTPAS's evaluation survey after pilot testing

Following the respondents' feedback, several of the sentences included in the multiple-choice items were also discarded and replaced with shorter and simpler sentences. Despite their high GDEX scores, in fact, these sentences were identified as problematic since they either contained anaphoric pronouns pointing at referents outside sentence limits, thus taking away the readers' focus from verb meaning, or were deemed syntactically too complex, for example by featuring the verb under scrutiny only at the end of the sentence.

4. Results

In a period of approximately 2 months, we were able to collect answers from 30 respondents, which was deemed a reasonable sample to carry out the evaluation on.

4.1 Respondent Sample

The average age of our 30 respondents is 35.4 years: 12 (40%) are in their 20s, 7 (23.3%) in their 30s, 7 (23.3%) in their 40s, 3 (10%) in their 50s and one (3.3%) in her 60s. Gender-wise, 20 respondents identify as female, 9 as male and 1 as non-binary.

As for educational level, 83.3% of the participants in the study has attended or is currently attending university, while 16.7% holds a secondary school diploma. Of the university-trained respondents, those who decided to specify their field of interest, 87% have a Humanities background (Croatian language and literature, Linguistics, Foreign Languages, Social studies, Political Sciences, Theatre) and 13% Hard sciences (IT, Engineering and Chemistry).

All respondents except one consider Croatian as one of their native languages. All of them grew up in either Croatia or Bosnia and Herzegovina except for four, who were either born in or moved to an English-speaking country quite early on. Two of these four respondents still live abroad (included the only non-native speaker of Croatian), while the rest lives in Croatia.

4.2 Jaccard Index of Similarity

For what concerns the multiple-choice section of the survey, each of the respondents' answer sheets was compared against a yardstick set of answers in line with CroaTPAS's sense distinctions. To provide a measure of how similar each of the 30 survey answers was to the yardstick, we calculated the Jaccard index of similarity.

The Jaccard similarity between two sets A and B is defined as "the ratio of the number of elements in the intersection of A and B over the number of elements in the union of A and B" (Zumel and Mount 2014: 184).

Given that respondents were presented with 91 multiple-choice items, each of the 30 survey answer sheets was assigned a similarity score ranging from 0 to 91 depending on the number of multiple-choice answers in line with the answers from CroaTPAS's yardstick answer sheet. That number was then divided by 91 and multiplied by 100, thus returning a normalised Jaccard index expressing the similarity score (%) between each collected answer sheet and CroaTPAS's annotation.

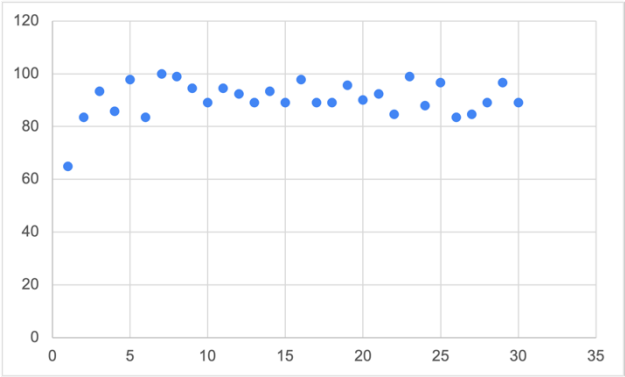

Figure 2: Dispersion plot of the similarity scores (%) of the 30 survey answers sheets against CroaTPAS's yardstick

As you can see from the dispersion plot above, all survey answer sheets but one range between 100% and 83.51% similarity. The only answer sheet scoring a lower similarity value (64.84%) was identified by Rosner's Test[2] as a possible outlier both at 5% and 1% significance and subsequently discarded.

Therefore, since the mean similarity score of the remaining 29 survey answer sheets stands at 91.36% (± 5.12) and data sets with a normalised Jaccard similarity above 85% can be considered highly stable (Zumel & Mount 2014: 184), we can conclude that the collected survey answers form a proper cluster showing a high level of agreement with the yardstick answer representative of CroaTPAS's sense distinctions.

4.3 Similarity Scores and Polysemy

After assessing the overall similarity scores of the collected survey answer sheets with the yardstick, we decided to group together the individual multiple-choice answers in five classes (2P, 3P, 4P, 5P and 6P) according to the degree of polysemy (two, three, four, five or six senses) expressed by the verb they portray and then calculate five new sets of similarity scores comparing them to the five corresponding

[2] Rosner's Test was run using the statistical software ProUCL 5.0

groups of yardstick answers representing the sense distinctions made in CroaTPAS.

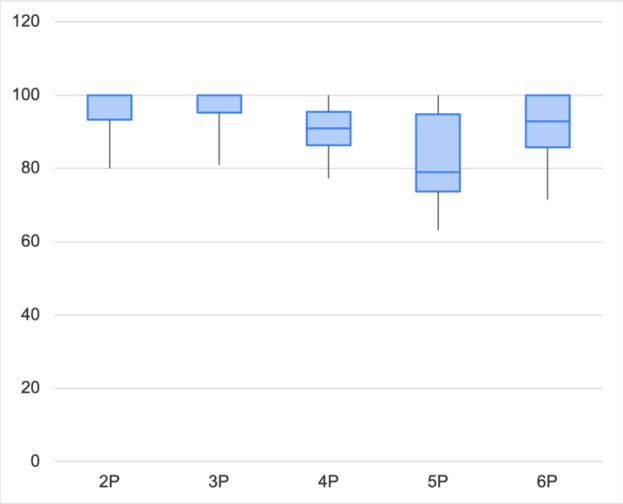

Figure 3: Box plots showing the different distribution of similarity scores (%) in survey answers referring to verbs with a different degree of polysemy

As you can see from the box plot in Figure 3, the similarity score of the survey answers does vary according to the number of senses expressed by the verbs they refer to. Participants tend to be more in line with the yardstick answers when it comes to less polysemous verbs, scoring a mean similarity value of 95.5% (± 5.94) in the 2P answer class and 95.89% (± 5.2) similarity in the 3P answer class. On the other hand, when the verb is more polysemous, the mean similarity scores of the answers decrease slightly to 91.38% (± 6.8) for answers to items containing verbs with four senses, 83.3% (±11.43) for answers to items on five-sense verbs and 91.13% (± 8.89) for those to items containing six-sense verbs.

Bearing in mind that mean similarity scores for all answer classes remain higher than 80%, thus showing a high level of agreement with the yardstick annotation regardless of verb polysemy, we might venture at tracing this difference back to the fact that disambiguating meanings when given more options is more demanding than when one is given fewer options to choose from.

4.4 Similarity Scores and Gender

To provide further support to CroaTPAS's evaluation, we divided the similarity scores of the survey answers by gender. Given that only one respondent identified as non-binary, only two similarity score distributions were compared, namely the one corresponding to the answers given by female respondents (19) and the other corresponding to the answers given by male participants (9). Two box plots were drawn for to provide a graphical representation of each population (Figure 4).

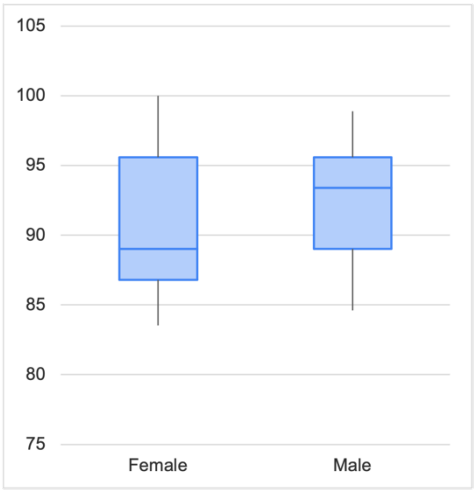

Figure 4: Box plots showing the different distribution of similarity scores (%) in survey answers by gender

As you can see, the two box plots are quite similar: the mean similarity score for survey answers provided by women respondents is 90.80 % (± 5.46), while the mean similarity score of male respondents 92.80 % (± 4.60). However, to assess the possible presence of a statistically significant gender bias, we ran a t-Test for two independent means, after making sure that both populations qualify as normally distributed using the Kolmogorov-Smirnov Test of Normality.

As it turns out, there is no significant difference for gender between the two populations, since the computed t-value is 1.02785, which is lower than 2.0555, the critical value for 26 degrees of freedom and 10% level of significance (5% in each tail).

4.5 Similarity Scores and English Level

As in the case of gender, it was decided to investigate the possible influence of the English language proficiency on the recorded similarity scores. This was the reason why participants were asked to rate their level of English in the first place.

The box plots in Figure 5 show the different distribution of similarity scores according to the different levels of English language skills respondents declare to possess. Only one participant gave themselves 1/5 on the semantic differential scale provided in the online survey and was thus discarded. Having already excluded the outlier, the remaining 28 respondents distribute on three distinct self-assessed language levels: 5 on level 3, 12 on level 4 and 11 on level 5. The mean similarity scores for the three levels are all quite high, standing at 92.53% (± 5.46), 91.21% (± 5.62) and 91.71% (± 4.46), respectively.

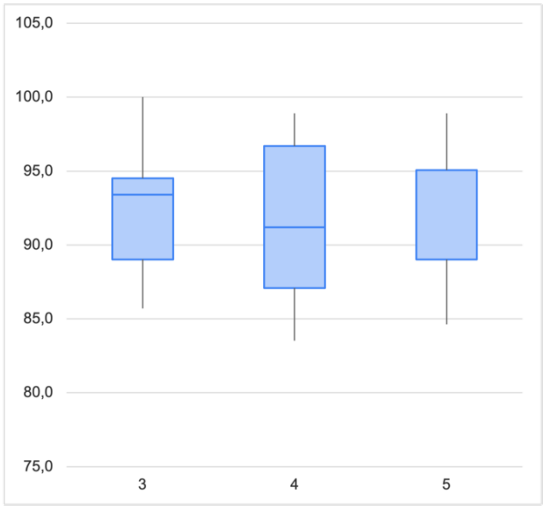

Figure 5: Box plots showing the different distribution of similarity scores (%) by self-assessed level of English

After ascertaining that all populations qualify as normally distributed, t-Tests were run between the similarity scores associated to levels 3 and 4, levels 3 and 5 and levels 5 and 4, which returned the following t-values: 0.44395, 0.31856 and 0.23454. The corresponding critical values of t for 15, 14 and 21 degrees of freedom at 10% level of significance (5% in each tail) are 2.1314, 3.1448 and 2.0796.

Since in all three cases, computed t-values are well below the corresponding critical values, we can conclude there is no statistically significant influence in terms of the self-assessed English language skills possessed by respondents on their sense-disambiguation task results.

This may either mean that asking respondents to self-evaluate their English language proficiency is not a good indicator of their actual English knowledge or that the English sense descriptions provided as multiple-choice options in the sense-disambiguation section of the survey were sufficiently clear to guarantee an effective meaning disambiguation regardless of the respondents' English language skills.

5. Conclusions

In conclusion, the attempt at evaluating the CroaTPAS resource generalising on the results of an online multiple-choice Google Form survey devised on a selection of verbs representative of the whole resource gave very good results.

In a period of approximately two months, 30 answer sheets were collected through a snowball sampling methodology. To provide an agreement metric between the respondents' answers and CroaTPAS's verb sense distinctions, the participants' answers were compared against a yardstick set of answers in line with CroaTPAS and a normalised Jaccard index of similarity was subsequently calculated.

After discarding one respondent, the mean similarity score of the remaining 29 was calculated at 91.36% (± 5.12). Since data sets with a Jaccard similarity above 85% can be considered highly stable (Zumel & Mount 2014: 184), the collected survey answers qualify as a single cluster with a high level of agreement with CroaTPAS's annotation of sense distinctions.

The distribution of similarity scores is not found to vary depending on gender nor on the respondents' English language skills.

6. Bibliographical References

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics: Quantitative, Qualitative and Mixed Methodologies*. Oxford: Oxford University Press.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: The MIT Press.

Fife-Schaw, C. (2006). Questionnaire Design. In Breakwell, G. M. et al. *Research Methods in Psychology* (3rd ed.). London: Sage.

Ježek, E., Magnini, B., Feltracco, A., Bianchini, A., and Popescu, O. (2014). T-PAS: A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing. In *Proceedings of the 9th conference on International Language Resources and Evaluation (LREC)*. Reykjavik, Iceland.

Ježek, E. (2019). Sweetening Ontologies Cont'd: Aligning bottom-up with top-down ontologies. In *Proceedings of CREOL 2019*. Graz, Austria.

Johnstone Young, T. (2016). Questionnaires and Surveys. In Zhu, H. (Ed.) Research Methods in Intercultural Communication : A Practical Guide, 165-181.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., and Rychlý, P. (2008). GDEX: automatically finding good dictionary examples in a corpus. *Proceedings of the 13th EURALEX International Congress*. Barcelona, Spain.

Kilgarriff, A., Baisa, V., Busta, J., Jakubícek, M., Kovár, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), 7-36.

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

Marini, C. and Ježek, E. (2019). CROATPAS: Resource of Corpus-derived Typed Predicate Argument Structures for Croatian. In *Proceedings of the 6th Italian Conference on Computational Linguistics (CLiC-it).* Bari, Italy.

Pustejovsky, J. (1995). *The Generative Lexicon.* Cambridge: The MIT Press.

Pustejovsky, J. and Ježek, E. (2008). Semantic Coercion in Language: Beyond Distributional Analysis. *Italian Journal of Linguistics*, 20: 181-214.

Zumel, N., and Mount, J. (2014). *Practical data science with R*. Shelter Island, NY: Manning Publications Co.