# Annotation Game for Textual Entailment Evaluation

Zuzana Nevěřilová

Natural Language Processing Centre, Faculty of Informatics,
Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic

**Abstract.** Recognizing textual entailment (RTE) is a well-defined task concerning semantic analysis. It is evaluated against manually annotated collection of pairs hypothesis–text. A pair is annotated true if the text entails the hypothesis and false otherwise. Such collection can be used for training or testing a RTE application only if it is large enough.

We present a game which purpose is to collect h–t pairs. It follows a detective story narrative pattern: a brilliant detective and his slower assistant talk about the riddle to reveal the solution to readers. In the game the detective (human player) provides a short story. The assistant (the application) proposes hypotheses the detective judges true, false or non-sense.

Hypothesis generation is a rule-based process but the most likely hypotheses that are offered for annotation are calculated from a language model. During generation individual sentence constituents are rearranged to produce syntactically correct sentences.

The game is intended to collect data in the Czech language. However, the idea can be applied for other languages. The paper concentrates on description of the most interesting modules from a language-independent point of view as well as the game elements.

## 1 Introduction

Recognizing Textual Entailment (RTE) is defined in [6, p. 18]: "A text $t$ entails a hypothesis $h$ ($t \Rightarrow h$) if humans reading $t$ will infer that $h$ is most likely true." This definition of entailment is far more relaxed than a mathematical logic definition.

Although RTE seems to be defined loosely ("humans will infer", "most likely"), it is one of the most well defined problems in semantic analysis. RTE systems are evaluated by comparing their outputs with annotated pairs text–hypothesis (h–t pairs). Each pair is annotated either as true (if $t$ entails $h$) or false (if $t$ does not entail $h$).

A collection of h–t pairs can be built manually (similarly to reading comprehension tests for children and for adults[1]) but in natural language processing (NLP) automatic data gathering is preferred.

[1] e.g. OECD PISA `http://www.oecd.org/pisa/`

A. Gelbukh (Ed.): CICLing 2014, Part I, LNCS 8403, pp. 340–350, 2014.

[5] describes four scenarios leading to collecting of h–t pairs in RTE2 challenge[2]:

- IE – texts $t$ were collected using structured template, relations tested in ACE-2004 RDR. Afterwards, hypotheses $h$ were extracted from these texts using IE. These hypotheses have to be evaluated as positive (correct outputs) and negative (incorrect outputs) examples.
- IR – hypotheses $h$ from evaluation datasets TREC and CLEF, texts $t$ were selected from documents retrieved by various search engines.
- QA – transformation of answered questions to affirmations generated hypotheses $h$, original answers (extracted from the web by QA systems) serve as texts $t$.
- text summarization – a sentence occurring in summary was taken as $t$ and simplified by removing sentence parts which leads to $h$.

All these retrieved h–t pairs went through manual annotation. In case of Czech language we cannot apply the same scenario since the appropriate tools are not available, so no evaluation set for recognizing textual entailment currently exist for Czech. The aim of this work is to build a considerable collection without using the above mentioned techniques.

### 1.1 Paper Outline

The paper is organized as follows: in section 2 we discuss the concept of collaboratively created language resources and compare our project with similar ones. In section 3 we present the game, discuss user experience w.r.t. annotation quality, and the game design. Section 4 presents the implementation and several modules that are employed to generate the hypotheses. Even though the game is in operation for a short time we present up to now results in section 5. Section 6 discusses the results and proposes future work.

## 2 Collaboratively Created Language Resources

Together with the rise of Web 2.0 the "collective intelligence" becomes an area of scientific interest. Non-expert users can be involved in many ways into expert tasks. [16] divides the collaboratively created language resources (CCLR) according to several criteria: motivation, annotation quality, setup effort, human participation and task character. The idea of CCLRs is based on collective "human computation" where peoples' brains are used for solving problems difficult for computer programs (such as natural language understanding or image content recognition) and relatively easy for people. Since GWAPs are games, the main motivation for contributors is the fun.

[16] split CCLRs into three categories: mechanized labor (such as Amazon Mechanical Turk), wisdom of the crowds (such as Wikipedia) and games with

[2] http://pascallin2.ecs.soton.ac.uk/Challenges/RTE2

a purpose (or GWAPs). Annotation GWAPs are of three basic kinds: output-agreement, input-agreement or inversion [1]. In all cases GWAPs are games for two (human) players who produce the annotation.

GWAP is a suitable model for demanding NLP tasks. Related works comprise:

- Common Sense Propositions [2] collected by *Verbosity*. One player describes a magic word to the second player whose aim is to guess the magic word only from these descriptions.
- Coreference Annotation [3] where players of *Phrase Detectives* annotate collaboratively coreferences. The game has two modes: annotation (where players select the appropriate coreferent pairs) and validation (where users validate previously annotated data).
- Paraphrase Corpora Collection [4] presents a game *1001 Paraphrases* where the doctors say something and the player has to say the same thing in other words.
- Semantic Relations Collection [14] present a categorization game collecting pairs object–category and a free association game (pairs word–associated word). The three games (*Categorilla*, *Categodzilla* and *Free Associations*) are based on real-life games. The data are available for download in text form. In the data from March 26, 2010 there are 745,030 pairs from the Free Associations and 1,199,235 pairs from Categorilla and Categodzilla.

In our case the players' task is somewhat similar to that in GWAPs. Unlike GWAPs the game is for one player, so no instant human feedback is present. The only case players receive feedback is when a proposition is annotated repeatedly. In this case the player earns points if the annotation corresponds to the majority of previous annotations.

One-player game has a great advantage over two-player games: we can cope with less participants (i.e. registered players). For collecting data in Czech language (spoken by about 10 million people) it is not easy to get a reasonably large worker base.

## 3 The Game

The game narrative refers to a dialogue between a detective and his/her assistant. The purpose of the dialogue is to explain the detective's reasoning to readers. Many players are familiar with this narrative pattern. In addition, the dialogue is always set in a friendly and open atmosphere even if the assistant is baffled. These conditions encourage players to annotate consciously.

The dialogue always starts with a story. The detective (human player) either provides a new story or returns back to a former story. The assistant Watsonson (application) tries to reformulate the story and entails new propositions. Afterwards, the detective can judge assistant's propositions as true or false in the given context. The basic screen with a sample dialog is shown in Figure 1.

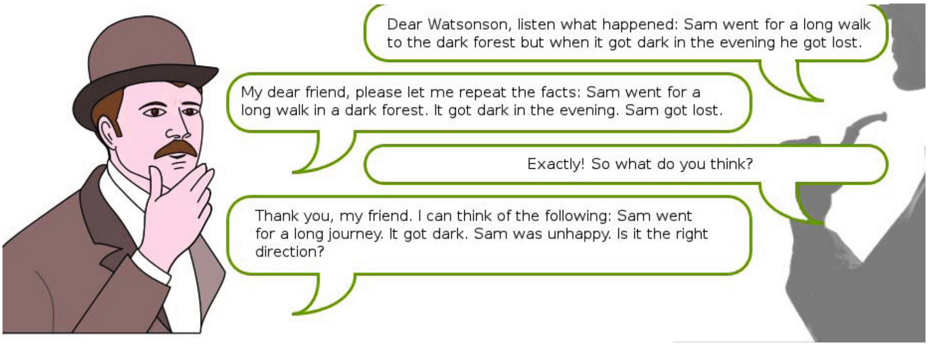
Dear Watsonson, listen what happened: Sam went for a long walk to the dark forest but when it got dark in the evening he got lost.
My dear friend, please let me repeat the facts: Sam went for a long walk in a dark forest. It got dark in the evening. Sam got lost.
Exactly! So what do you think?
Thank you, my friend. I can think of the following: Sam went for a long journey. It got dark. Sam was unhappy. Is it the right direction?

**Fig. 1.** The game environment is a dialogue between the detective and his assistant. N.B. the dialogue was translated in English by the author.

### 3.1 Data Complexity and Annotation Experience

Reading comprehension tests serve to test peoples' understanding capabilities. These tests are often considered difficult. The criticism of the game could confront the difficulty of reading comprehension tests and the lack of annotators' training. However, similarly to further semantic annotation projects, users are encouraged (by the instructions) to use their common sense to decide the annotation value. In addition, as the game advances trickiest entailments are generated. Users thus become experienced by playing the game repeatedly.

The data complexity in relation with CCLRs is discussed in [16, p.10]. According to the authors LR complexity means the data size as well as its characteristics relevant to annotation. In our case the annotation in simple yes/no decision. The data size for each h–t pair is quite small: the text consists usually of a few sentences, hypothesis a one sentence. Players are not forced to annotated every h–t pair. We suppose they prefer to annotate only clear cases.

With all this issues on mind we expect to obtain reasonable-quality annotation.

### 3.2 The Game Design

The game is designed as a conversational game. However, the player does not have to write much. Firstly, s/he enters a new story or obtains an old one then s/he only clicks to annotate or control the dialogue. The player can see the continuous dialogue (as shown in Figure 1) as well as popup windows with individual sentences and annotation buttons ✓, ✗ or ☹.

The player earns points for a new story according to the number of clauses and phrases that have been identified by the syntactic parsing (story score). The player earns further points for every annotation and even more points for agreement with other annotators.

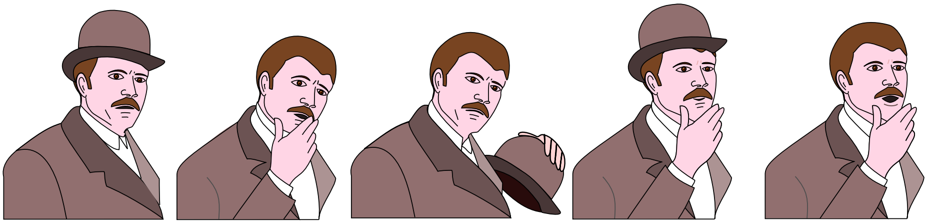

**Fig. 2.** Watsonson's emotions reflect the dialogue flow as well as the story score

Apart points and levels which are typical game elements two other elements are present in order to make the game fun. Firstly, the detective can encourage Watsonson to speak, appreciate him or reproach him. Secondly Watsonson's face reflect his emotions depending on the story score and the dialogue flow: he can be curious, thinking, thinking hard, happy, bored, annoyed, nosy, neutral or sad. Some of the emotions are shown in Figure 2.

## 4 Implementation

The game implementation is based on the integration of existing modules for natural language analysis and generation (such as morphological analyzer and syntactic parser) with new ones. It can be considered as a proof of concept of those existing "universal" software tools for processing the Czech language.

From the RTE's point of view the human player enters a text *t*, the computer player proposes several hypotheses *h* and the human player annotates the h–t pair. The hypothesis *h* vary from simple paraphrases (i.e. syntactic rearrangements) to real entailments (completely new sentences).

When the detective decides to return back to an older story, repeated annotations are obtained. The system encourages beginners to use this option.

### 4.1 Modules

For new hypotheses generation we use several modules from morphological and syntactic level processing (tokenization, disambiguation, parsing) to the semantic level. The modules for phrase re-ordering, synonym and hypernym replacement and verb frame inference are independent and are used in all possible orders to generate more hypotheses.

All semantic modules work on the phrase level (verb phrases, noun phrases, prepositional phrases, adverbial phrases, coordinations) not on word level. The stories and entailments are represented by no particular formalism (such as first order logic). Each clause is a verb phrase and a set of phrases dependent on the verb or with unknown parent (which typically applies to adverbials).

**Table 1.** Story representation: each sentence is divided in clauses, each clause is parsed on phrases. Phrases are marked according their syntactic roles: SUBJ(ect), VERB phrase, OBJ(ect), REFL(exive particle), ADV(erbial).

| Sam šel na dlouhou vycházku do temného lesa, ale když se večer setmělo, ztratil se.<br>Sam went for a long walk in a dark forest but when it got dark in the evening he got lost. | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sam šel na dlouhou vycházku do temného lesa<br>Sam went for a long walk in a dark forest | | | | ono se večer setmělo<br>it got dark in the evening | | | | Sam se ztratil<br>Sam got lost | | |
| Sam | jít | (na) dlouhá vycházka | (do) temný les | on | se | večer | setmět | Sam | se | ztratit |
| Sam | go | (for) long walk | (in) dark forest | it | | in the evening | get dark | Sam | | get lost |
| SUBJ | VERB | OBJ | ADV | SUBJ | REFL | ADV | VERB | SUBJ | REFL | VERB |

**Parsing and Partial Anaphora Resolution.** Players are asked to input a short story. We use syntactic parsing (SET parser [10]) to obtain phrases with known dependencies. The anaphora resolution system Saara [11] supplements unexpressed subjects and replaces demonstrative pronouns with their antecedents. Sentences are generated back from the set-of-phrases representation and they are offered for annotation. All other modules use the set-of-phrases representation. Example of the preprocessing can be viewed in Table 1.

**Word Reordering.** Czech is a (so called) free word order language i.e. nearly all orders of phrases are allowed. For this reason we prefer the term free phrase order. Every sentence is reformulated in all possible phrase orders. Various phrase orders never change the truth value but play a role in text cohesion. Since we generate isolated hypotheses we do not care about text cohesion. We only use the scoring module (see 4.1) to choose the most natural phrase order.

**Synonym and Hypernym Replacement.** We use Czech WordNet [15] for synonym replacement. The module replaces all word expressions found in Czech WordNet by their synonyms. No word sense disambiguation method is used therefore as a result false paraphrases are generated.

Since all transformations originators are recorded we can later discover WordNet synonyms unlikely in stories. For example *pes* has two senses: one corresponds to the synset `dog:1, domestic dog:1, Canis familiaris:1` in Princeton WordNet [7], the other corresponds to `martinet:1, disciplinarian:1, moralist:2`. A search in existing h–t pairs indicates the unlikely occurrence of the second sense

**Table 2.** Synonym replacement using Czech WordNet: "vycházka" (walk) was replaced by "výlet" (trip). N.b. the modifier "dlouhý" (long) had to be modified to fulfil the grammatical agreement with "výlet" (trip) because "vycházka" (walk) is feminine and "výlet" (trip) is masculine.

| | | | |
|---|---|---|---|
| Sam | jít | (na) dlouhá vycházka | (do) temný les |
| Sam | go | (for) long walk | (in) dark forest |
| SUBJ | VERB | OBJ | ADV |
| Sam | jít | (na) dlouhý výlet | (do) temný les |
| Sam | go | (for) long trip | (in) dark forest |

in stories. In fact, none of the hypotheses generated with the replacement *pes–moralista* (moralist) were judged true. An example synonym replacement is shown in Table 2.

Similarly to synonym replacement word expressions are replaced recursively by their hypernyms. In this case two restrictions apply. First, we do not replace word expression by all hypernyms but omit those from the WordNet Top Ontology. Such replacement (e.g. replace "student" by "living entity") will never generate a natural sounding expression. Second, we do not replace by hypernyms in sentences with negative polarity. While in positive sentences (such as "He came in his new coupe.") the hypernym replacement (replace "coupe" by "car") is valid, in negative sentences the replacement results always in false entailments ("He did not came in his new coupe." does not entail "He did not came in his new car."). In Czech double negative is used, so it is easier to detect correctly the sentence polarity in cases like "There was nobody in the classroom."

The hypernym replacement can generate sentences such as "Sam went for a long excursion.", "Sam went for a long journey." and "Sam went for a long travel.".

**Verb Frame Inference.** Word reordering and synonym replacement result in paraphrases while verb frame inference can result into new facts. In this module we take advantage of the Czech verb valency lexicon VerbaLex [9] and use verb valency frames for inferences of three types: equality, effect, precondition.

Verb frame inference is based on correct grammatical case recognition of all sentence constituents dependent on the verb or being without a parent (which applies mostly on adverbials). If the phrases and their cases are recognized correctly, the module obtains the verb plus its syntactic pattern, e.g. *be lost* + nominative:person + adverbial:non-person or *be lost* + nominative:person + *in* locative:non-person.

Subsequently the inference rules are used to transform a syntactic pattern to another pattern, e.g. *be lost* + nominative:person + adverbial:non-person → *be unhappy* + nominative:person. The inference rules were created manually, then augmented automatically using VerbaLex. The process of expansion is described in detail in [12].

For checking the phrase category constraints we use the shallow ontology Sholva. In Sholva [8] currently 154,783 words are classified into eight categories: substance, non-substance, person (incl. institutions), non-person, person-individual, non-person-individual, event, non-event. Each word can be member of more than one class. The annotation of Sholva has been done manually with multiple annotators.

The main advantages of Sholva are two: the size and the simplicity of the data. Using the category constraints we can distinguish verb frames with the same syntactic structure but distinct semantic slot categories. For example we can distinguish cases like *pass somebody on to somebody* (and infer they will communicate) and *pass something on to somebody* (and infer s/he will suffer). In many cases, distinguishing person and non-person is sufficient.

The overall process generates $s$ from $r$ using the following steps:

1. search for the syntactic pattern $s$ in inference rules
2. for all rules $s \rightarrow r$: get syntactic pattern $r_i$
3. fill the sentence constituents from $s$ to appropriate slots in $r_i$
4. check constraints with Sholva
5. if all slots are filled and constraints are satisfied generate a new sentence from $r_i$

An example verb frame inference is shown in Tables 3 and 4.

**Table 3.** The verb frame inference corresponds to the common sense inference "If someone gets lost s/he becomes unhappy"

| | |
|---|---|
| Sam | se ztratil |
| Sam | got lost |
| SUBJ → SUBJ | ztratit se → být nešťastný |
| SUBJ → SUBJ | get lost → to be unhappy |
| Sam | byl nešťastný |
| Sam | was unhappy |

**Table 4.** The verb frame inference corresponds to the common sense inference "If someone gets lost someone else will look for him"

| | | |
|---|---|---|
| Sam | se ztratil | |
| Sam | got lost | |
| SUBJ → OBJ(accusative) | ztratit se → hledat | |
| $\epsilon$ → SUBJ | get lost → look for | |
| někdo | hledal | Sama |
| somebody | looked for | Sam |

**Sentence Generation.** When all transformations (that work with lemmata not with words) are done the generation module finds the appropriate word forms. Czech is a language with rich nominal inflection (different word forms for singular and plural as well as seven grammatical cases[3]). Verb conjugation has further intricacies (two main verb aspects, multi-word verb forms and reflexive particles). Moreover, grammatical agreements are needed between verb in past tense and the subject as well as between noun phrases and their modifiers.

The function of sentence generation module relies on correct recognition of the subject (which is always in nominative). According the subject's gender and number, the appropriate verb form is generated. Afterwards, all noun phrases' and prepositional phrases' modifiers are checked whether they fulfil the agreement on case, gender and number. For generation (i.e. finding a correct word form for a given lemma and a given tag) we use the morphological analyzer/generator majka [13].

**Natural Sounding Sentences.** The application produces tens to hundreds of hypotheses from each input sentence but not all of them are offered for annotation. We use a $n$-gram language model for calculating the most natural sounding sentence. Sentences of highest scores are offered for annotation. Apart from that a random sentence is sometimes selected for annotation to increase the collection diversity.

The appropriate $n$-gram frequencies were calculated using the Czes corpus[4]. The resulting score is calculated according to Equation 1 where $ngram_i$ means the $i$-th $n$-gram normalized frequency and $m$ is the number of tokens. The normalization of each $n$-gram is calculated as shown in Equation 2. Here the raw frequency is normalized by corpus size and 100,000 and divided by raw frequencies of all tokens in the $n$-gram.

$$score = 10^2 \sum_{i=1}^{m-1} 2gram_i + 10^3 \sum_{i=1}^{m-2} 3gram_i + 10^4 \sum_{i=1}^{m-3} 4gram_i + 10^5 \sum_{i=1}^{m-4} 5gram_i \tag{1}$$

$$ngram = \frac{100000 freq_{ngram}}{corpsize \sum_{i=0}^{n} freq_i} \tag{2}$$

## 5 Evaluation

The project is currently in its final testing phase. Two testers inserted 275 reasonably long texts (at least one sentence, at least five words). The system generated 56,872 hypotheses. From these hypotheses 1,784 unique hypotheses were annotated (note the system strongly overgenerates) and 195 were annotated more

[3] Nominative, genitive, dative, accusative, vocative, locative and instrumental.

[4] 465,102,710 tokens in 2013-11-07.

**Table 5.** Multiple annotations

| sum of annotation values | -2 | -1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|
| # hypotheses | 10 | 351 | 184 | 1077 | 125 | 22 | 10 | 2 | 2 | 1 |

than once. The annotations were marked -1 when marked negative, 0 when confused and 1 when positive. The sum of repeated annotation values indicates the correctness of a hypothesis. When annotations of a particular hypothesis oscillate between true and false, the sum converges to 0 which means confused. Table 5 shows how many hypotheses received a particular sum of annotation values. Evidently, positive annotations predominate.

## 6 Conclusion and Future Work

We present an ongoing project of annotation game which aims to create a collection of h–t pairs for future Czech RTE system. The game is similar to GWAPs but it is only for one player. One-player games may be more suitable for collecting data in languages with minor speaker communities. Our outlook is to obtain in a few years a large collection of stories (thousands), hypotheses (tens of thousands) and their annotations as well as information about the way the hypotheses were generated.

The present results have shown which structures are preferred in the short detective stories. Some WordNet synonyms are not used in this kind of text (e.g. dog as martinet), some word orders are not used (verb in the initial position). Our future work will have two main directions. Firstly, we want to gradually reduce the generation from all possible to the most frequently annotated structures. Secondly, we need to keep the game still interesting even for experienced players. In the near future we want to add a knowledge base concerning famous detectives and their cases. We plan prospectively to add more types of inference about time and location.

**Acknowledgements.** This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and by the Ministry of the Interior of CR within the project VF20102014003.

The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme “Projects of Large Infrastructure for Research, Development, and Innovations” (LM2010005) is appreciated.

## References

1. von Ahn, L., Dabbish, L.: Designing games with a purpose. Commun. ACM 51(8), 58–67 (2008), http://doi.acm.org/10.1145/1378704.1378719
2. von Ahn, L., Kedia, M., Blum, M.: Verbosity: a game for collecting common-sense facts. In: CHI 2006: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 75–78. ACM, New York (2006)

3. Chamberlain, J., Kruschwitz, U., Poesio, M.: Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In: Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, People's Web 2009, pp. 57–62. Association for Computational Linguistics, Stroudsburg (2009), http://dl.acm.org/citation.cfm?id=1699765.1699774
4. Chklovski, T.: Collecting paraphrase corpora from volunteer contributors. In: Proceedings of the 3rd International Conference on Knowledge Capture, K-CAP 2005, pp. 115–120. ACM, New York (2005), http://doi.acm.org/10.1145/1088622.1088644
5. Dagan, I., Dolan, B., Magnini, B., Roth, D.: Recognizing textual entailment: Rational, evaluation and approaches. Natural Language Engineering 15(special issue 04), i–xvii (2009), http://dx.doi.org/10.1017/S1351324909990209
6. Dagan, I., Roth, D., Zanzotto, F.M.: Tutorial notes. In: 45th Annual Meeting of the Association of Computational Linguistics. The Association of Computational Linguistics, Prague (2007)
7. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (May 1998); published: Hardcover
8. Grác, M.: Rapid Development of Language Resources. Dissertation, Masaryk University in Brno (2013), http://is.muni.cz/th/50728/fi_d/
9. Hlaváčková, D., Horák, A.: VerbaLex – new comprehensive lexicon of verb valencies for Czech. In: Proceedings of the Slovko Conference (2005)
10. Kovář, V., Horák, A., Jakubíček, M.: Syntactic analysis using finite patterns: A new parsing system for Czech. In: Human Language Technology. Challenges for Computer Science and Linguistics, Poznań, Poland, November 6-8, p. 161 (2011); revised Selected Papers
11. Němčík, V.: Saara: Anaphora resolution on free text in Czech. In: Horák, A., Rychlý, P. (eds.) Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012, pp. 3–8. Tribun EU, Brno (2012)
12. Nevěřilová, Z., Grác, M.: Common sense inference using verb valency frames. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2012. LNCS, vol. 7499, pp. 328–335. Springer, Heidelberg (2012)
13. Šmerk, P.: Towards Computational Morphological Analysis of Czech. Dissertation, Masaryk University in Brno (2010), http://is.muni.cz/th/3880/fi_d/
14. Vickrey, D., Bronzan, A., Choi, W., Kumar, A., Turner-Maier, J., Wang, A., Koller, D.: Online word games for semantic data collection. In: EMNLP 2008: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 533–542. Association for Computational Linguistics, Morristown (2008)
15. Vossen, P.: EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Computers and the humanities. Springer (1998)
16. Wang, A., Hoang, C., Kan, M.Y.: Perspectives on crowdsourcing annotations for natural language processing. Language Resources and Evaluation 47(1), 9–31 (2013), http://dx.doi.org/10.1007/s10579-012-9176-1