Paraphrase and Textual Entailment Generation in Czech

Zuzana Nevěřilová

Natural Language Processing Centre, Faculty of Informatics,
Masaryk University, Brno,
Czech Republic

xpopelk@fi.muni.cz

Abstract. Paraphrase and textual entailment generation can support natural language processing (NLP) tasks that simulate text understanding, e.g., text summarization, plagiarism detection, or question answering. A paraphrase, i.e., a sentence with the same meaning, conveys a certain piece of information with new words and new syntactic structures. Textual entailment, i.e., an inference that humans will judge most likely true, can employ real-world knowledge in order to make some implicit information explicit. Paraphrases can also be seen as mutual entailments. We present a new system that generates paraphrases and textual entailments from a given text in the Czech language. First, the process is rule-based, i.e., the system analyzes the input text, produces its inner representation, transforms it according to transformation rules, and generates new sentences. Second, the generated sentences are ranked according to a statistical model and only the best ones are output. The decision whether a paraphrase or textual entailment is correct or not is left to humans. For this purpose we designed an annotation game based on a conversation between a detective (the human player) and his assistant (the system). The result of such annotation is a collection of annotated pairs text–hypothesis. Currently, the system and the game are intended to collect data in the Czech language. However, the idea can be applied for other languages. So far, we have collected 3,321 H–T pairs. From these pairs, 1,563 were judged correct (47.06 %), 1,238 (37.28 %) were judged incorrect entailments, and 520 (15.66 %) were judged non-sense or unknown.

Keywords. Games with a purpose, paraphrase, textual entailment, natural language generation.

1 Introduction

When reading (and understanding) texts, people routinely derive knowledge that is present in the discourse but not expressed: for example, if people read about *a victim*, they promptly think of an *attack*, maybe they think that *the victim needs help* or they only feel *sympathy*. If a computer program has to infer new information from a text, it needs to process the unexpressed (or implicit) information. [11, p. 149] estimates the ratio of explicit:implicit information to be up to 1:8.22, which means that the vast majority of information is not mentioned in texts. The problem of implicit information or implicit knowledge is known and studied in cognitive science, computational linguistics and artificial intelligence.

In computational linguistics, making implicit information explicit forces syntactic, semantic and pragmatic modules to interact. Firstly, it is necessary to discover "gaps" in the text, secondly, the correct missing entities have to be found, and finally, those entities can be filled in. For example, missing entities at the syntactic level are unexpressed (but obligatory), and such sentence constituents and the gaps are called ellipses. At the semantic level, such missing entities are the unfilled semantic roles [19].

We have built a computer system that is able (to some extent) to fill the gaps at the syntactic and semantic levels. In our approach, the input is a free text in Czech and the result are automatically generated sentences in Czech. We use standard analysis tools (such as a tokenizer, a tagger and a syntactic parser) in order to obtain

Computación y Sistemas Vol. 18, No. 3, 2014 pp. 555–568
ISSN 1405-5546
DOI: 10.13053/CyS-18-3-2040

an inner representation of the input text. From this representation we generate representations of textual entailments and paraphrases. Finally, we use a natural language generation (NLG) module to produce syntactically correct sentences in Czech. The sentences are ranked using a language model and the most successful sentences are offered for annotation.

The contribution of this work is multi-fold: (i) paraphrase and textual entailment generation system can be used in further applications such as question answering, text summarization, plagiarism detection, tutoring systems, and machine translation evaluation, (ii) the annotated collection can be used for a future system for Czech recognizing textual entailment (RTE), and (iii) the agreement on annotations indicates what people consider obvious and easy to recognize and what paraphrases and entailments are rather difficult.

In this article, we will first define textual entailments and paraphrases and then we will describe our paraphrase and textual entailment generation system. We will discuss the concept of collaboratively created language resources in general and briefly describe annotation games for similar projects. The main idea of our annotation game is outlined in [17]; here we present thoroughly the resulting dataset.

2 Textual Entailments and Paraphrases

It seems that introducing unmentioned entities in texts and subsequent inference is something what human communication relies on. From this point of view, textual entailment is essential in the studies of meaning. The author of [1] defines textual entailment as "a relationship between a coherent text T and a language expression H, which is considered as a hypothesis. T *entails* H if the meaning of H, as interpreted in the context of T, can be deduced from the meaning of T." Textual entailment is marked by the arrow symbol: $T \rightarrow H$.

Textual entailments usually apply additional knowledge. For example, to infer from $T =$ "Acme's $14 billion acquisition by Wonderworks Ltd" that $H =$ "Wonderworks Ltd purchased Acme" we need to know that company acquisition means purchase. This additional knowledge is sometimes present in knowledge bases such as WordNet [9] or common sense knowledge bases such as ConceptNet [13]. [7] classified the types of knowledge needed to successfully decide whether T entails H.

Paraphrases typically do not introduce new entities but they convey the same information using different words or syntactic structures. The authors of [2] give the following example:
(1) Wonderworks Ltd. constructed the new bridge.
(2) The new bridge was constructed by Wonderworks Ltd.
(3) Wonderworks Ltd. is the constructor of the new bridge.

Most people would judge all three sentences to be paraphrases. However, sentence (3) differs slightly since it does not state if the bridge has been completed. The authors of [2] remark that people very often ignore these subtle distinctions and therefore they define paraphrase s_2 of sentence s_1 as a sentence that has the same or *almost the same* meaning as s_1 in a given context. A paraphrase also can be seen as a mutual entailment ($s_1 \rightarrow s_2$ and $s_2 \rightarrow s_1$). Paraphrases are constructed using many different manners. The authors of [3] identified 25 classes of English paraphrases and measured that the most common paraphrases are produced by synonym substitution, function words variations, and external knowledge.

3 Paraphrase and Textual Entailment Generation

Figure 1 presents the scheme of the paraphrase and textual entailment generation system. The input sentences are processed by a tokenizer, a tagger, and a syntactic parser. The parse results are enriched by semantic information and partial anaphora resolution in order to fill zero subjects and replace pronouns by their antecedents. We also identify some phrases or subphrases as named entities. Finally, each input text is represented as a list of set of properties (LOSOP). Due to text cohesion, the order of sentences in a story matters significantly. On the other hand, the order of sentence parts does not affect much the correctness of a sentence. Czech is a so called free word order language with the canonical

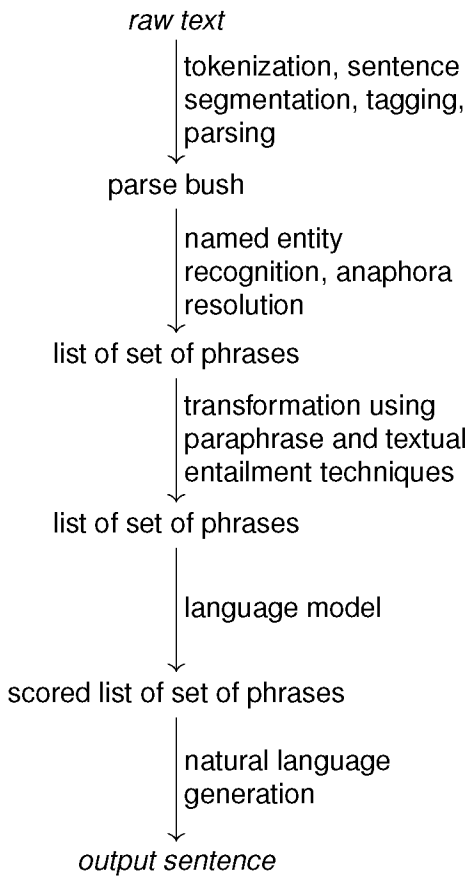

Fig. 1. Overall scheme of the paraphrase and textual entailment generation system

word order subject-verb-object (SVO). Similarly to Spanish, different word orders are possible and usually express subjectivity or put emphasis on some sentence parts.

The inner representation in a form of a LOSOP is then transformed using different paraphrasing and textual entailment techniques. So far, we transform sentences one-to-one, i.e., we do not integrate information from several sentences in order to generate one sentence. We divide the transformations into four groups:

— phrase reordering,
— lexical replacement,
— lexical-syntactic replacement,
— verb frame replacement.

The transformations are independent and are used in all possible orders to generate many hypotheses. Each transformation results in a new LOSOP from which we can generate a syntactically correct sentence in Czech. These new sentences are scored using a corpus-based language model. The sentences with highest scores are then offered for annotation in the annotation game.

3.1 Analysis Phase

We use the syntactic parser SET [15], which is one of the parsers available for Czech. The resulting structure is a dependency syntactic tree but in our project, we work with a *syntactic bush* as defined by [10]. The bush does not contain words in its leaves. Instead, it works at the phrase level (verb phrases, noun phrases, prepositional phrases, adverbial phrases, coordinations are in the leaves) and thus the resulting tree is not very high. Sentences are divided into clauses and each clause is represented as a verb phrase and a set of phrases dependent on the verb or with an unknown parent (which typically applies to adverbials).

Phrases in the parse tree are classified using shallow ontology Sholva [10] that divides words into four classes: `person`, `event`, `substance`, and `person-individual`. Both `person` and `person-individual` classes describe potential agents (or doers) but the former is more general than the latter and can apply e.g. to organizations. Sholva contains 154,783 positive and negative classifications such as *concert* is an `event` and is not a `person`.

We designed a lightweight module for named entity recognition. It is useful mainly in recognizing Sholva classes `person` (person names, organizations, cities), `event` (artworks, dates, holidays) and `person-individual` (person names). The module is based on searching in Freebase[1] data and Czech Wikipedia pages, and pattern matching for recognizing dates, IP addresses, e-mails, etc.

The anaphora resolution module Aara supplements zero subjects and replaces demonstrative pronouns with their antecedents. Antecedent recognition benefits from both syntactic and semantic properties. Czech has masculine animate,

[1] `https://www.freebase.com/`

masculine inanimate, feminine, and neuter genders and two numbers. In past tense, the word forms of the verb differ for each category, for example, in sentences *Girls ran* and *Boys ran*, the verb *run* has different word forms. The grammar agreement in number and gender also applies in predicative complements, e.g., in sentences *Girls were young* and *Boys were young*, the word *young* has different forms. We employ the Czech verb frame lexicon VerbaLex [14] to resolve the ambiguity that cannot be resolved by grammar constraints. For example, the agent (doer) of the verb *to sell* is always a `person` (i.e., a human or an organization). In contrast, it cannot be an `event` or a `substance`. If the constraints are too harsh, the anaphoras are not resolved. In the resulting collection, 84 % of sentences with resolved anaphoras were annotated as correct.

Table 1. Inner representation of a short story (Table adapted from [17])

Sam šel na dlouhou vycházku do temného lesa Sam went for a long walk in a dark forest			
Sam Sam SUBJ +person	jít go VERB	(na) dlouhá vycházka (for) long walk OBJ -person, +event	(do) temný les (in) dark forest ADV -person
ono se večer setmělo it got dark in the evening			
on it SUBJ	se REFL	večer in the evening ADV -person	setmět get dark VERB
Sam se ztratil Sam got lost			
Sam Sam SUBJ +person	se REFL	ztratit get lost VERB	

An example of the analysis can be seen in Table 1. The sentence *Sam šel na dlouhou vycházku do temného lesa, ale když se večer setmělo, ztratil se* (Sam went for a long walk in a dark forest but when it got dark in the evening, he got lost) is divided in clauses, each clause is parsed on phrases. Phrases are marked according to their syntactic roles: SUBJ(ect), VERB phrase, OBJ(ect), REFL(exive particle), ADV(erbial).

3.2 Transformations

In Section 3, we divided the transformations of the inner representation into four groups. In this section, we present each group. The transformations do not work with word forms but with lemmata. Czech is a language with rich nominal inflection: with seven cases[2] and two numbers, many word forms differ in suffixes. A noun lemma is the singular nominative form, an adjective lemma is the positive masculine singular form. A phrase lemma is the same form as the phrase head form. For example, if the phrase head is feminine, then the adjective modifier lemma is singular nominative feminine. The word form ambiguity (e.g., the singular nominative feminine suffix is equal to the plural nominative neuter suffix) complicates automatic inflection in the generation module (see Section 3.2.5).

Each transformation stores its *ancestor*, i.e., the source sentence, and the type of transformation called a *signature*. We can then evaluate not only the resulting sentences but also the successful and unsuccessful transformations.

3.2.1 Phrase Reordering

In Czech, nearly all phrase orders are allowed. For this reason, we prefer the term free phrase order. Every sentence is reformulated in all possible phrase orders. Apparently, various phrase orders do not change the truth value but play a role in text cohesion and subjectivity. Since we generate isolated hypotheses, we do not consider text cohesion.

[2] nominative, genitive, dative, accusative, vocative, locative and instrumental

Table 2. Synonym replacement using Czech WordNet: *vycházka* (*walk*) was replaced by *výlet* (*trip*) (Table adapted from [17])

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

3.2.2 Lexical Replacement

We use Czech WordNet [18] for synonym replacement. Czech WordNet currently contains 28,456 synsets and 43,916 words or word expressions. The module replaces all word expressions found in Czech WordNet by their synonyms. Since no word sense disambiguation method is used, the module sometimes produces false paraphrases, e.g., replacement *head→title* in the context of body parts makes no sense. This disadvantage is partially compensated by the scoring module (see Section 3.2.6).

Since all transformations ancestors are recorded, we can discover WordNet synonyms that are less probable in stories. For example, Czech word *pes* has two senses: one corresponds to the synset `dog:1, domestic dog:1, Canis familiaris:1` in Princeton WordNet [9], another corresponds to `martinet:1, disciplinarian:1, moralist:2`. A search in existing H–T pairs indicates the unlikely occurrence of the latter sense. In fact, 7 of 8 of the hypotheses generated with the replacement *pes–moralista* (moralist) were judged false.

An example synonym replacement is shown in Table 2: in the phrase *dlouhá vycházka* (*long walk*), the head *vycházka* (*walk*) was replaced by the synonym *výlet* (*trip*). The modifier *dlouhý* (*long*) has to be modified to fulfill the grammatical agreement with *výlet* (*trip*) because *vycházka* (*walk*) is feminine and *výlet* (*trip*) is masculine.

Similarly to synonym replacement, phrases are replaced 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 apply hypernym replacement in sentences with negative polarity. While in positive sentences (such as "He came in his new coupe"), the hypernym replacement (replacement coupe→car) is valid, in negative sentences, the same replacement results in false entailments ("He did not came in his new coupe" does not entail "He did not came in his new car"). In Czech, negatives are formed using a prefix. In addition, double negative is used, so it is easier to detect correctly the sentence polarity in cases like "There was nobody in the classroom" than it is in English. Literally, the latter sentence translates as "There was not nobody in the classroom", thus the polarity can be detected from the verb form.

The hypernym replacement of the sentence presented in Table 2 can generate sentences such as "Sam went for a long excursion", "Sam went for a long journey" and "Sam went for a long travel".

3.2.3 Lexical-Syntactic Replacement

We have built a module for modification of the noun or prepositional phrases. We implemented two different modules, the first for generating paraphrases, the second for generating entailments. Both modules are based on morphological derivation for which we use the Czech derivational tool Derivanče[22]. We are aware that several transformations exist (e.g., adverb-adjective associations as in *learn quickly* and *quick learning*), however, we currently lack the corresponding language resources.

Noun-Adjective Associations We assume that the genitive prepositional phrase is equivalent to a derived adjective phrase. An example of such transformation can be seen in Figure 3. We use abbreviations for noun phrase (NP), PREP(osition), ADJ(ective) and GEN(itive).

Table 3. Example of noun-adjective association of the words Afrika–africký (Africa–African)

malý	*africký*	*národ*
small	African	nation
ADJ	ADJ	NOUN
malý	*národ*	*z Afriky*
small	nation	of Africa
ADJ	NOUN	NOUN/GEN

Possessive Adjective–Noun Association We transform the possessive adjectives into nouns. The observation on corpus suggests that possessive adjectives mean mostly possession but if they are in relation with a named entity they mean responsibility in some sense (e.g., the authorship). Following these two observations, we created two patterns:

— X's Y: X owns Y (accusative),

— X's Y: X is the author of Y (genitive).

From these patterns, the system generates analytical entailments (i.e., no knowledge except of language knowledge is needed). Example of these patterns can be seen in Figure 4. The tense of the new sentence depends of the tense of its ancestor.

Table 4. Example of possessive adjective–noun association

POSS		NOUN
Petrovo		*auto*
Peter's		car
NOUN	PRED	ACC
Petr	*vlastní*	*auto*
Peter	owns	a car
POSS		NOUN
Munchův		*Výkřik*
Munch's		The Scream
NOUN	COPULA+COMPL	GEN
Munch	*je autorem*	*Výkřiku*
Munch	is the author of	The Scream

3.2.4 Verb Frame Replacement

Verb frame replacement module transforms verb frames with their slots into different verb frames with the same or new slots. This module produces paraphrases if the verbs are synonyms and the slots remain the same, and entailments if the verbs are not synonyms or the slots differ. We take advantage of the Czech verb valency lexicon VerbaLex [14] that contains 6,244 verb synsets and 19,158 verb frames. We use verb valency frames for inferences of the following types:

— active–passive voice: Y stole X $\rightarrow$ X was stolen by Y,

— passive–active voice: X was stolen by Y $\rightarrow$ Y stole X,

— equality: X comes to Y $\leftrightarrow$ X arrives to Y,

— near-equality: X smokes $\leftrightarrow$ X is a smoker, unlike equality, near-equality is not symmetric,

— precondition: X snores $\rightarrow$ X sleeps,

— effect: X eats $\rightarrow$ X is not hungry.

First, we have to identify correctly all sentence constituents dependent on the verb. If the phrases and their cases are recognized correctly, the verb frame is constructed as the verb together with the syntactic pattern with semantic constraints, e.g., *be lost* + nominative: `person` + *in* locative: `non-person`.

The verb and the pattern are then transformed using the inference rules. The result of the transformation is another verb and a pattern, e.g., *be lost* + nominative: `person` + adverbial: `non-person` $\rightarrow$ *be unhappy* + nominative: `person`. The inference rules for equality and transformations between active and passive voice were generated automatically from VerbaLex, others were created manually. Note that the inference rules form another language resource that supports the paraphrase and entailment generation process.

Using the category constraints from the shallow ontology Sholva, 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).

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

Table 5. The verb frame inference corresponds to the common sense inference "If someone gets lost, they become 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 6. The verb frame inference corresponds to the common sense inference "If someone gets lost someone else will look for them."

Sam	se ztratil	
Sam	got lost	
SUBJ → OBJ(accusative)	ztratit se → hledat	
ϵ → SUBJ	get lost → look for	
někdo	hledal	Sama
somebody	looked for	Sam

1. search for the pattern s in inference rules,
2. for all rules $s \rightarrow r$: get new patterns r_i,
3. fill the sentence constituents from s to appropriate slots in r_i,
4. 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 5 and 6. The former shows a common sense reasoning "When someone gets lost, they become unhappy", the latter shows a reasoning "When someone gets lost, someone else will look for them". Both tables are adapted from [17].

3.2.5 Sentence Generation

Each transformation produces a new LOSOP. In order to produce a grammatically correct sentence, we need to find the appropriate word forms of the corresponding phrase lemmata. Czech nominal inflection was mentioned in Section 3.2, verb conjugation has further intricacies (such as two main verb aspects, multi-word verb forms and reflexive particles). Moreover, grammatical agreements are needed between the verb in past tense and the subject, the copula verb and its predicative complement, and noun phrases and their adjective modifiers. For generation (i.e., finding a correct word form for a given lemma and a given tag), we use the morphological analyzer/generator majka [21].

3.2.6 Natural Sounding Sentences

The system generates tens to hundreds of sentences from each input sentence but only few of them are offered to annotators. We use a statistical n-gram language model to compute the most natural sounding sentence. Only sentences with the highest scores are offered for annotation. Low-score sentences are randomly selected for annotation to increase the collection diversity.

The n-gram frequencies are calculated on the Czes corpus[3]. Due to the rich inflection we count with word n-grams. 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. Each n-gram is normalized as shown in Equation 2 by the corpus size and 100,000 and divided by raw frequencies of all tokens in the n-gram. This formula scores longer sentences higher, which is desirable in our case.

$$CS = \sum_{n=2}^{5} 10^n \sum_{i=1}^{m-n} ngram_i \tag{1}$$

$$ngram = \frac{100000 \times freq_{ngram}}{corp\_size \times \prod_{i=0}^{n} freq_i} \tag{2}$$

We are aware of the fact that people use some transformations more often than others, but unfortunately, we have limited knowledge about "good" or "useful" transformation rules. Similarily, we have no information about "usual" senses of a word (such as the weighted WordNet described in [4]), therefore we cannot e.g. prefer one lexical transformation to another. For this reason, we employed a sentence ranking that is based on previous annotations.

Each generated sentence contains information about its ancestor and the signature (as mentioned in Section 3.2). Obviously, the signatures repeat for

[3] 465,102,710 tokens on 2014-07-29

different sentences. The annotation-based score AS is calculated as a weighted arithmetic average of annotations for a particular signature. If a sentence is annotated as correct, it obtains 1 point, if it is annotated as false, it obtains -1 point, if it is annotated as non-sense, it obtains 0 points. When generating a new sentence, the signature score influences the overall sentence score and thus it influences whether the sentence will be offered for annotation or not.

We expect that the annotation-based score will improve the game since it decreases the probability that a sentence from a "bad" transformation (e.g., *dog* as *martinet*) will appear in the game.

4 Non-expert Annotations

In the previous section, we described several techniques how to generate paraphrases and textual entailments. The crucial question is whether these paraphrases and textual entailments are correct or not. The decision is left completely on humans but creating manually a gold standard is extremely difficult. In this section, we focus on annotation games in general, discuss the appropriateness of a game for the task, and describe our game.

4.1 Collaboratively Created Language Resources

The "collective intelligence" becomes an area of scientific interest with the rise of Web 2.0. Non-expert users are involved in many ways in formerly expert tasks. In [28], collaboratively created language resources (CCLR) are divided by several criteria: motivation, annotation quality, setup effort, human participation, and task character.

CCLRs can be divided into three categories: mechanized labor (such as Amazon Mechanical Turk), wisdom of the crowds (such as Wikipedia) and games with a purpose (or GWAPs). There are three basic kinds of annotation GWAPs: output-agreement, input-agreement,and inversion [26]. In all cases, GWAPs are games for two (human) players who play a game and produce an annotation. Since GWAPs are games, the main motivation for contributors is the fun. Since two humans play, the agreement can be measured.

Apparently, GWAP is a suitable model for NLP tasks concerning semantics. In the following overview (adapted from [17]), we list some games that collect data that are very difficult to obtain automatically:

— Common Sense Propositions [27] 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 [5] where players of *Phrase Detectives* collaboratively annotate 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 [6] 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 [25] 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.

All these games solve NLP tasks that are relatively easy for humans but extremely difficult for computer programs. Paraphrase and textual entailment generation is one of these tasks.

Our game is similar to a GWAP. Unlike GWAPs, the game is for one player, so no instant human feedback is present. Players can receive only moderate feedback when a sentence is annotated repeatedly: in this case, the player earns points if her annotation corresponds to the majority of previous annotations.

One-player games have a great advantage over two-player games: the annotation still works even if we have less participants. For collecting data in the

Czech language (spoken by about 10 million people), it is not easy to get a reasonably large worker base but over time we can obtain a considerable number of annotations.

4.2 Inter-Annotator Agreement

The inter-annotator agreement (IAA) depends strictly on the annotation subject (i.e., what the question is). In the RTE task, the decision is binary, i.e., is H entailed by T or not? In this case, the chance-agreement for two annotators is 50 %. The authors of [20] recognize several entailment phenomena (coreference, simple rewrite rule, lexical relation, implicit relation, factoid, parent-sibling, genitive relation, nominalization, event chain, coerced relation, passive-active, numeric reasoning, spatial reasoning) and extend the annotation task to particular phenomenon identification. In their work, the Cohen's κ vary from 0.412 to 0.847 depending on the entailment phenomena.

In [23], the authors examined the quality of non-expert annotations, particularly Amazon Mechanical Turk annotations, on five tasks. They have shown that the resulting annotation is in high agreement with the gold standard. For the RTE task, the expert IAA has been reported between 91 % and 96 % on the PASCAL RTE-1 dataset [8]. The non-expert annotation have been measured according to a simple *majority voting*. The maximum accuracy 89.7 % was reached averaging over annotations of 10 workers. The authors of [23] reported a reasonable quality of non-expert annotation assuming the task is described as succinct as possible.

In [24], the authors observe that in case of GWAPs, we can measure the agreement as well as the overall number of answers; the agreement measure is considered a better choice since the number of answers can be low and depending on the type of a very unbalanced game (i.e., one unit can have many annotations but another unit can have only one or two annotations). The authors of [24] tested majority measures (relative majority, majorities relatives to different thresholds) and concluded that the best F-score was achieved by relative majority.

5 The Game

The game *Shenlock Holmer meets dr. Watsonson* is based on a well-known scheme: in detective stories, a brilliant detective has to explain his/her deduction methods to some other (less brilliant) character, usually an assistant. The purpose of the dialogue is to explain the detective's reasoning to readers. Such dialogue is usually set in a friendly and open atmosphere even if the assistant is slow. The game narrative follows this literary pattern: the human player plays the role of Shenlock Holmer, the system is in the role of dr. Watsonson.

The dialogue always starts with a story. Shenlock Holmer (the human) either provides a new story or returns back to a former story. His assistant, dr. Watsonson (the system), tries to reformulate the story and to entail new propositions. The detective can judge dr. Watsonson's propositions as true, false or non-sense in the given context. The basic screen with a sample dialog is shown in Figure 2 (Figure reproduced from [17]).

From the point of view of the RTE task, Shenlock Holmer enters a text *T*, dr. Watsonson proposes several hypotheses *H* and Shenlock Holmer annotates the appropriate H–T pair. The hypothesis *H* can be a paraphrase or a textual entailment that reveals new information.

The human players do not always have to type a text. They can "return an older case", so an existing story is used. The system recommends this option to beginners, however, the results show that it is not preferred.

5.1 The Game Design

The game is a dialogue. However, players do not have to write much. They decide either to enter a new story or to get a random previous story. Then, players only click to annotate the sentences or to control the dialogue. The player can see the continuous dialogue (as shown in Figure 2) as well as popup boxes with individual sentences and annotation buttons ✔, ✖ or 😕.

Players earn points for entering a new story according to the number of clauses and phrases that have been identified by the syntactic parsing (story score). Players also earn points for each

Fig. 2. The game environment is a dialogue between the detective Shenlock Holmer and his assistant dr. Watsonson. N.B. the dialogue was translated into English by the author.

Fig. 3. Watsonson's emotions reflect the dialogue flow as well as the story score (Figure reproduced from [17])

annotation and even more points for agreement with other players. Players are encouraged to play more than once by earning badges (such as "experienced detective").

Points and badges are typical game elements (also known as the Points-Badges-Levels or the PBL triad). Apart from that, Watsonson's face reflects 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 of dr. Watsonson are shown in Figure 3.

6 Results

So far, the game has collected 3,321 H–T pairs. From these pairs, 1,563 were judged correct (47.06 %), 1,238 (37.28 %) were judged incorrect entailments, and 520 (15.66 %) were judged non-sense or unknown. The game allows repeated annotations but the results show that players are not much motivated to annotate previous text. Only 456 pairs were annotated more than once. In case of repeated annotations, we count the average of all annotations. The overview of the dataset is shown in Table 7. The presented annotations were collected in 5 months.

Table 7. Parameters of the resulting dataset

	number	%
H–T pairs	3,321	100
correct entailments/paraphrases	1,563	47.06
incorrect entailments/paraphrases	1,238	37.28
non-sense or unknown entailments/paraphrases	520	15.66
single annotations	2,865	86.3
multiple annotations	456	13.7

6.1 Resulting Sentences with respect to the Modules

The quality of a module can be seen from two criteria: (1) how often the module applies, and (2) what the ratio between correct and incorrect (and perhaps non-sense) phrases is. Table 8 shows the respective performance of individual modules.

Table 8. Performance of modules

	correct	incorrect	non-sense	% correct	% incorrect	total
no change	316	8	25	90.54	2.29	349
anaphora resolution	276	127	95	55.42	25.5	498
phrase order	304	88	45	69.57	20.14	437
synonym replacement	434	854	173	29.71	58.45	1,461
hypernym replacement	22	6	19	46.81	12.77	47
verb frame replacement: equivalence	153	142	156	33.92	31.49	451
verb frame replacement: near--equivalence	4	1	1	66.67	16.67	6
verb frame replacement: effect	17	7	0	70.83	29.17	24
verb frame replacement: precondition	20	2	3	80	8	25
possessive-noun replacement	3	0	1	75	0	4
other	14	3	2	73.68	21.4	19
total	1,563	1,238	520	47.06	37.28	3,321

Table 9. The number of sentences with respect to multiple annotations

# of annotations	# of sentences	Fleiss' κ
1	2,865	–
2	329	0.18
3	74	0.44
4	33	0.5
5	9	0.78
6	7	0.3
7	1	-0.17
8	2	0.05
9	1	1

First, we can see that analysis and generation of a sentence is not a self-evident success. About 10 % incorrect or non-sense sentences show that errors occur during morphological analysis, tagging, syntactic parsing, or sentence generation.

We observe that partial anaphora resolution `aara` is used quite frequently with an overall 55.42% success. Even though the results are not fully comparable, note that [16] reported a 60.4% success rate with pronoun resolution tested on the Prague Dependency Treebank [12]. The perspectives on what is a zero subject and what is a clause coordination differ. We illustrate this difference on the sentence from the PDT: *Vítěz skupiny postoupí do bojů o evropský [pohár] a má velmi pravděpodobnou účast na OH 1996 v Atlantě.*[4] (*The winner of the group will advance to the European [Cup] and is very likely to participate in the 1996 Olympics in Atlanta.*). From the *t-layer*, we can see that the sentence is a coordination of two clauses: *will advance* and *is likely to participate*. In our perspective, it is advantageous to understand the sentence as a compound sentence and to divide it in two sentences: *the winner will advance* and *the winner is likely to participate*. Clearly, the resulting application influences strongly the perspective and therefore the anaphora resolution applications (presented by [16] and ours) are designed and evaluated in different ways.

Phrase ordering performs well in most cases. Errors in phrase ordering originate most often from incorrect phrase segmentation and incorrect placement of adverbials.

Synonym replacement is often used but the success rate is not very high (mainly because there is no word sense disambiguation). Hypernym replacement is used less frequently but with

[4] This sentence can be found in the PDT sample data. Its *t-layer* visual representation is available at `http://ufal.mff.cuni.cz/pdt2.0/visual-data/sample/sample1_t_4.htm`.

more success. Verb frame replacements perform better in the case of manually built rules (`near-equivalence`, `effect`, `precondition`) than in the case of automatically generated verb synonyms (relation `equals`).

Lexical-syntactic replacement modules such as possessive-noun transformation are used rarely so we cannot evaluate them yet.

6.2 The Annotation Quality

For testing understanding capabilities of readers, people use reading comprehension tests[5], which are often considered difficult. The criticism of the annotation game could confront the difficulty of such 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 on the annotation value. In addition, as the game advances, more complex entailments are generated. Users thus gain experience by playing the game.

So far, we distinguish players either by their login or by their IP address if they are not logged in. We can tackle potential vandalism by removing contributions of a particular player. On the other hand, we do not plan to rank the annotators.

We measured the IAA using Fleiss' κ. Unlike RTE with only two classes, each sentence can be classified in three classes: true entailment, false entailment and non-sense sentence. The latter case happens mostly when the sentence is misinterpreted by syntactic parsing (or even morphological analysis). For example, if we interpret the sentence "Time flies like an arrow" differently than the annotator, they will annotate the paraphrase "Arrows are liked by time flies" as non-sense.

The results presented in Table 9 show that the majority of sentences is annotated only once. For multiple annotations, the IAA varies a lot but note that for more than 4 annotations we do not have much data. Also, N. B. that Fleiss' κ does not reduce to Cohen's κ when the number of annotators is two. The corresponding Cohen's κ for two annotators is 0.24.

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

7 Conclusion and Future Work

In this article, we presented a new paraphrase and textual entailment generation system for the Czech language and an annotation game that serves as an evaluation method for the system.

The work has several aims: (i) to build a software tool for paraphrase and textual entailment generation, (ii) to discover how good this tool is, and (iii) to gather a collection of H–T pairs. Currently, the collection contains 3,321 H–T pairs from which 47 % were annotated as correct. Such collection can be used for a future Czech RTE system but it is also a valuable object *per se*. We can observe which paraphrases are preferred by language users, what replacements make sense to them, and what entailments are considered easier (with higher agreement) than others.

The system integrates many NLP tasks and the overall performance is influenced by the tagging and parsing accuracy and by the quality of language resources, namely, the verb valency lexicon VerbaLex, Czech WordNet, the Sholva ontology, and the inference rules. Our work is the first contribution to paraphrase and textual generation in Czech language and probably one of the few in the non-English NLP. We would also like to encourage research of this area in the community.

Our future work has two main directions. First, we have to add more paraphrasing and textual entailment techniques, namely, those that are based on knowledge and those that concern time and location. In addition, entailment from more than one sentence at a time will be desirable.

Second, we need to make the game more popular and keep it still interesting even for experienced players. We plan to employ social media and other gamification techniques in order to reach these two goals.

Both the paraphrase and textual entailment generation system and the annotation game are available on the NLPC website[6].

[6] `http://nlp.fi.muni.cz/projects/watsonson/paraphrasing`
`http://nlp.fi.muni.cz/projects/watsonson`

Acknowledgments

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. **Akhmatova, E.** (**2005**). Textual entailment resolution via atomic propositions. In *PASCAL: Proceedings of the First Challenges Workshop on Recognising Textual Entailment*. Southampton, UK, 61–64.
2. **Androutsopoulos, I. & Malakasiotis, P.** (**2009**). A survey of paraphrasing and textual entailment methods. *CoRR*, abs/0912.3747.
3. **Bhagat, R. & Hovy, E.** (**2013**). What is a paraphrase? *Computational Linguistics*, 39(3), 463–472. ISSN 0891-2017. doi:10.1162/COLI_a_00166.
4. **Boyd-Graber, J., Fellbaum, C., Osherson, D., & Schapire, R.** (**2006**). Adding dense, weighted connections to WordNet. In *Proceedings of the Third International WordNet Conference GWC-06*. Masaryk University in Brno, South Jeju Island, Korea, 29–36.
5. **Chamberlain, J., Kruschwitz, U., & Poesio, M.** (**2009**). 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 '09. Association for Computational Linguistics, Stroudsburg, PA, USA. ISBN 978-1-932432-55-8, 57–62.
6. **Chklovski, T.** (**2005**). Collecting paraphrase corpora from volunteer contributors. In *Proceedings of the 3rd International Conference on Knowledge Capture*, K-CAP '05. ACM, New York, NY, USA. ISBN 1-59593-163-5, 115–120. doi:10.1145/1088622.1088644.
7. **Clark, P., Fellbaum, C., & Hobbs, J. R.** (**2006**). The Boeing-Princeton-ISI (BPI) textual entailment test suite. Accessed online 2014-04-14 from `http://www.cs.utexas.edu/~pclark/bpi-test-suite/bpi-rte-knowledge-types.txt`.
8. **Dagan, I., Glickman, O., & Magnini, B.** (**2006**). The PASCAL recognising textual entailment challenge. In **Quiñonero-Candela, J., Dagan, I., Magnini, B., & d'Alché-Buc, F.**, editors, *Machine Learning Challenges*, volume 3944 of *Lecture Notes in Computer Science*. Springer, p. 177–190.
9. **Fellbaum, C.** (**1998**). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press. ISBN 026206197X.
10. **Grác, M.** (**2013**). *Rapid Development of Language Resources*. Ph.D. thesis, Masaryk University, Brno, Czech Republic.
11. **Graesser, A.** (**1981**). *Prose Comprehension Beyond the Word*. Springer-Verlag, New York. ISBN 9783540905448.
12. **Hajič, J., Hajičová, E., Pajas, P., Panevová, J., Sgall, P., & Vidová-Hladká, B.** (**2001**). *Prague Dependency Treebank 1.0 (Final Production Label)*. Linguistic Data Consortium. ISBN 1-58563-212-0. Published: CDROM CAT: LDC2001T10.
13. **Havasi, C., Speer, R., & Alonso, J.** (**2009**). ConceptNet: a lexical resource for common sense knowledge. In **Nicolov, N., Angelova, G., & Mitkov, R.**, editors, *Recent Advances in Natural Language Processing V*, volume 309 of *Current Issues in Linguistic Theory*. John Benjamins, Amsterdam & Philadelphia, 269–280.
14. **Hlaváčková, D. & Horák, A.** (**2005**). VerbaLex – new comprehensive lexicon of verb valencies for Czech. In *Computer Treatment of Slavic and East European Languages*. Slovenský národný korpus. ISBN 80-224-0895-6, 107–115.
15. **Kovář, V., Horák, A., & Jakubíček, M.** (**2011**). Syntactic analysis using finite patterns: A new parsing system for Czech. In *Human Language Technology. Challenges for Computer Science and Linguistics*, volume November 6-8, 2009. Poznań, Poland, 161–171.
16. **Kučová, L. & Žabokrtský, Z.** (**2005**). Anaphora in Czech: Large data and experiments with automatic anaphora resolution. In **Matoušek, V., Mautner, P., & Pavelka, T.**, editors, *Proceedings of 8th International Conference on Text, Speech and Dialogue, TSD 2005*, volume 3658 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. ISBN 978-3-540-28789-6, 93–98. doi:10.1007/11551874_12.
17. **Nevěřilová, Z.** (**2014**). Annotation game for textual entailment evaluation. In **Gelbukh, A.**, editor,

15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part I, volume 8403 of *Lecture Notes in Computer Science*. Springer. ISBN 978-3-642-54905-2, 340–350.

18. **Pala, K. & Smrž, P.** (**2004**). Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 2004(7), 79–88.
19. **Palmer, M. S., Dahl, D. A., Schiffman, R. J., Hirschman, L., Linebarger, M., & Dowding, J.** (**1986**). Recovering implicit information. In *Proceedings of the 24th Annual Meeting on Association for Computational Linguistics*, ACL '86. Association for Computational Linguistics, Stroudsburg, PA, USA, 10–19. doi:10.3115/981131.981135.
20. **Sammons, M., Vydiswaran, V. G. V., & Roth, D.** (**2010**). Ask not what textual entailment can do for you... In *Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, 1199–1208.
21. **Šmerk, P.** (**2010**). *Towards Computational Morphological Analysis of Czech*. Ph.D. thesis, Masaryk University in Brno, Brno, Czech Republic.
22. **Šmerk, P. & Hlaváčková, D.** (**2012**). Derivační analyzátor češtiny Derivanče [Derivational analyzer for Czech]. [software] accessed online 2014-04-09 from http://nlp.fi.muni.cz/projekty/derivance.
23. **Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y.** (**2008**). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08. Association for Computational Linguistics, Stroudsburg, PA, USA, 254–263.
24. **Venhuizen, N., Basile, V., Evang, K., & Bos, J.** (**2013**). Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*. Potsdam, Germany, 397–403.
25. **Vickrey, D., Bronzan, A., Choi, W., Kumar, A., Turner-Maier, J., Wang, A., & Koller, D.** (**2008**). Online word games for semantic data collection. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Morristown, NJ, USA, 533–542.
26. **von Ahn, L. & Dabbish, L.** (**2008**). Designing games with a purpose. *Commun. ACM*, 51(8), 58–67. ISSN 0001-0782. doi:10.1145/1378704.1378719.
27. **von Ahn, L., Kedia, M., & Blum, M.** (**2006**). Verbosity: a game for collecting common-sense facts. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, New York, NY, USA. ISBN 1-59593-372-7, 75–78. doi:http://doi.acm.org/10.1145/1124772.1124784.
28. **Wang, A., Hoang, C., & Kan, M.-Y.** (**2013**). Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47(1), 9–31. ISSN 1574-020X. doi:10.1007/s10579-012-9176-1.

Zuzana Nevěřilová obtained a degree in Computer Science in 2005 from the Faculty of Informatics, Masaryk University, Brno (Czech Republic). She is currently pursuing her Ph.D. degree in Computer Science with specialization in Natural Language Processing. Her broad research interests include semantic analysis, ontologies, and games with a purpose. From 2005, she focused on visualization of ontologies and library data, from 2010, she participated in the EuDML project that aimed to aggregate metadata from European digital mathematics libraries. Presently, she teaches computational linguistics at the Faculty of Arts.

Article received on 07/01/2014; accepted on 01/02/2014.