An intrinsic approach to the evaluation of a conceptual generator

Lucia Specia
Universidade Federal de São Carlos – Centro de Ciências Exatas e de Tecnologia
São Carlos – SP, Brazil, 13565-905
lspecia@dc.ufscar.br

and

Lucia H. M. Rino
Universidade Federal de São Carlos – Centro de Ciências Exatas e de Tecnologia
São Carlos – SP, Brazil, 13565-905
lucia@dc.ufscar.br

Abstract

This work addresses an intrinsic evaluation of conceptual generation (i.e., semantic interpretation) through ConPor, a system that maps syntactic structures from Brazilian Portuguese (BP) onto a semantic representation in UNL (Universal Networking Language). In analyzing ConPor performance we use not only classical measures such as recall and precision, but also semantic proximity. For that, two experiments were devised. In the first, the quality of mapping is assessed in terms of recall and precision taking as reference the manually prepared UNL representations of BP sentences. In the second, performance is assessed with regard to the quality of the output in BP, generated for the UNL representations provided by ConPor. We show that there is no direct implication between high (or low) values in the first experiment and the semantic proximity. We argue that taking semantic proximity into account is complementary to considering the classical recall and precision metrics.

Keywords: Artificial Intelligence, Natural Language Processing, Semantic Interpretation, Conceptual Generation, Intrinsic Evaluation

1 Introduction

ConPor, an automatic CONceptual generator for PORtuguese, is intended to map Brazilian Portuguese (BP) sentences onto Universal Networking Language [20], or UNL, sentences. Such a mapping is based on the syntactic structure of the desired sentence and is carried out by withdrawing the corresponding intra-sentential dependencies, in a process henceforth referred to as *syntactic-conceptual mapping.* Mapping syntactic structures onto conceptual ones along with the process of understanding a sentence is adopted by several researchers in semantic interpretation (or conceptual generation), considering different applications, for example, machine translation, automatic summarization, text analysis, etc. Mitamura and Nyberg [11], for instance, describe a hierarchical model for their *interpretive mapping* in which conceptualization is drawn from a semantic lexicon, resulting in a set of lexical frames indicating different levels of representation. Dorr [7] extends Jackendoff's [8] proposal for semantic structuring, producing her so-called *Lexical Conceptual Structures.* In a controlled environment, Bean et al. [1] define a set of interpretation rules to link syntax and semantics of locative relationships between anatomic entities. Considering a portable model for different domains and applications, Romacker et al. [14] use schemata for mapping dependency graphs onto conceptual ones, through a lexicalized dependency grammar.

Like the above methodologies, ConPor uses a compositional approach to generate the conceptual representations; differently from those, however, it uses UNL as the conceptual language, and addresses BP. To our knowledge, in the field of semantic interpretation, only [1] and [15] address intrinsic evaluation, although in a more limited way than ours, i.e., producing only recall and precision for performance, based on meaning representations. Rosé's [16] evaluation is also considerable, although she uses another methodology for semantic interpretation, interleaving it with parsing. Unlike the others, she focuses only on the comparison of the sentences resulting from the generated representation with the corresponding original ones, aiming at verifying how well the system communicates their main idea. Apart from those proposals, evaluation of semantic interpretation seems to be neglected.

Our assessment of ConPor relates to both viewpoints, comprising two experiments: one that calculates recall and precision on the syntactic-conceptual mapping and another that measures the *semantic proximity* between the sentences pinpointed by ConPor and the original ones. Such a metric is calculated through the comparison between their corresponding meanings. In a way, this is similar to Rosé's assessment; however, she explicitly addresses fluency. We will show that such a feature is implicitly considered in the way semantic proximity is graded when assessing ConPor.

After giving a brief outline of the UNL representation (Section 2) and the ConPor prototype system (Section 3), we report on the ConPor assessment (Section 4), discussing its outcomes in Section 5.

2 The Universal Networking Language

UNL [20] is a conceptual representation language, used initially as the interlingua of the Machine Translation UNL Project [19]. This language is aimed at expressing information conveyed by natural language sentences through binary relations between concepts. Actually, information conveyed by each NL sentence can be represented as a hypergraph whose nodes correspond to concepts and arcs correspond to semantic relations between them. The general syntax of those relations is *RL(UW1,UW2)*, where RL stands for a *Relation Label*, which signals the semantic relation, and UWi, for *Universal Words*, which signal the related concepts. RLs are specified through mnemonics, for example, *agt* for *agent*, *mod* for *modifier*, *obj* for *object*, or *ptn* for *partner*. UWs may be generic, such as *book*, *run*, or *John*, or complex, in which case they indicate meaning variations, for example., *book(icl>room)*, *icl* indicating a hyperonymic relation between *book* and *room*, yielding the meaning *to book a room.* UWs may also carry other information, usually withdrawn from morphosyntactic features that provide further information on the circumstances under which they appear (for example, tense and aspect). Those are signaled by *Attribute Labels*, or ALs. Figure 1 illustrates a UNL encoding for the BP sentence (1). Figure 2 shows the correspondent graph.

(1) Palavras duras destroem as boas intenções.
Literally: Harsh words destroy the good intention.

```
[S]
; palavras duras destroem as boas intenções
agt(destroy.@entry, word.@pl)
mod(word.@pl, harsh)
obj(destroy.@entry, intention.@def.@pl)
mod(intention.@def.@pl, good)
[/S]
```

Figure 1. Sentence (1) UNL encoding

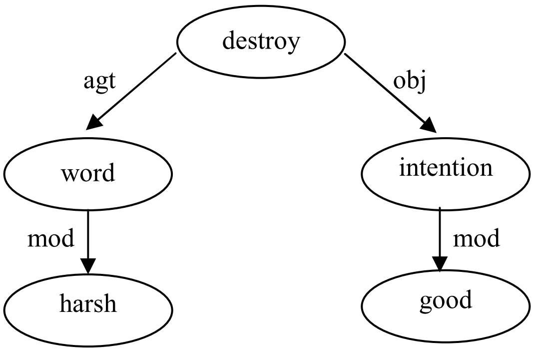

Figure 2. Sentence (1) UNL graph

As we will see in the next section, ConPor identifies semantic roles through syntactic functions, by exploring the functional syntactic structure of a BP sentence. Once identified those roles, it determines the semantic relations, which in turn are expressed through RLs between each pair of UWs (with its respective ALs).

3 The ConPor System

ConPor global environment is shown in Figure 3. Shading boxes show reused resources and processes, while dotted ones signal the actual ConPor components. The former were built at NILC[1]. The parser [10] is still on its beta version and, although broad enough to cover most BP features, it has no semantic processing. Because of this, it may generate multiple syntactic structures for a unique sentence, including incorrect ones. To overcome this and just focus on the conceptual generation, we have not yet embedded the parser into ConPor, which has as input one well-formed hand-selected syntactic structure at each moment. Therefore, this hand selection intends only to eliminate the incorrect syntactic structures. In the case of sentences that have multiple syntactic structures (i.e., syntactically ambiguous sentences), all of the well-formed ones are used as input to the conceptual generation module in ConPor, one-by-one. If one or more of these syntactic structures, although syntactically well-formed, is not semantically adequate, it has no correspondent conceptual structure produced by ConPor. This feature shows the potential of ConPor of manipulating syntactic ambiguity. When we have a complete parser, the hand selection process will no longer be necessary.

The other reused resources in Figure 3, i.e., the BP lexicon [12] and the bilingual Portuguese-UNL Dictionary [6] were customized to the ConPor environment, yielding the so-called *enriched lexicon*. This is a semantic lexicon [5] that currently amounts to 514 lexical items, drawn from a sample corpus of 122 BP horoscope sentences. These were extracted from general-readership daily Brazilian newspapers. Besides reused information, the enriched lexicon also holds specific directions for conceptual mappings, for example, the semantic features of nouns, the argument structures of verbs, and the verb classes.

---

[1] Núcleo Interinstitucional de Lingüística Computacional (www.nilc.icmc.sc.usp.br).

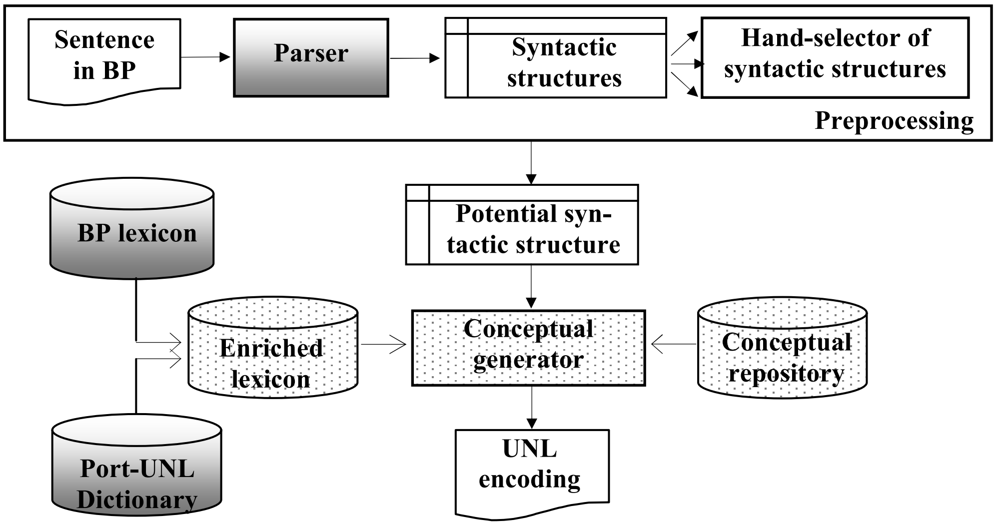

Figure 3. ConPor environment

The conceptual repository contains the syntactic-conceptual mapping models, which feed the conceptual generator process. This is the only specific process of ConPor and is effectively responsible for the syntactic-conceptual mapping. The syntactic-conceptual mapping models include two components: a) a grammar of projection rules that are responsible for mapping the syntactic constituents onto UWs, annotating them with their corresponding semantic roles, producing an intermediate conceptual structure, and b) a set of relationship templates that determine appropriate RLs, in order to connect UWs, according to their semantic roles, into UNL relations, producing the final UNL encoding. Figure 4 details the actual ConPor modules, illustrating the application of the two mapping models components. The identification of the semantic roles is the bulk of the system, when compared to UNL relations determinacy. This process corresponds to a simple rephrasing of semantic roles using the UNL terminology. Because of the nature of the data, it is accomplished in a non-ambiguous way: each pair of semantic roles signals a unique UNL relation.

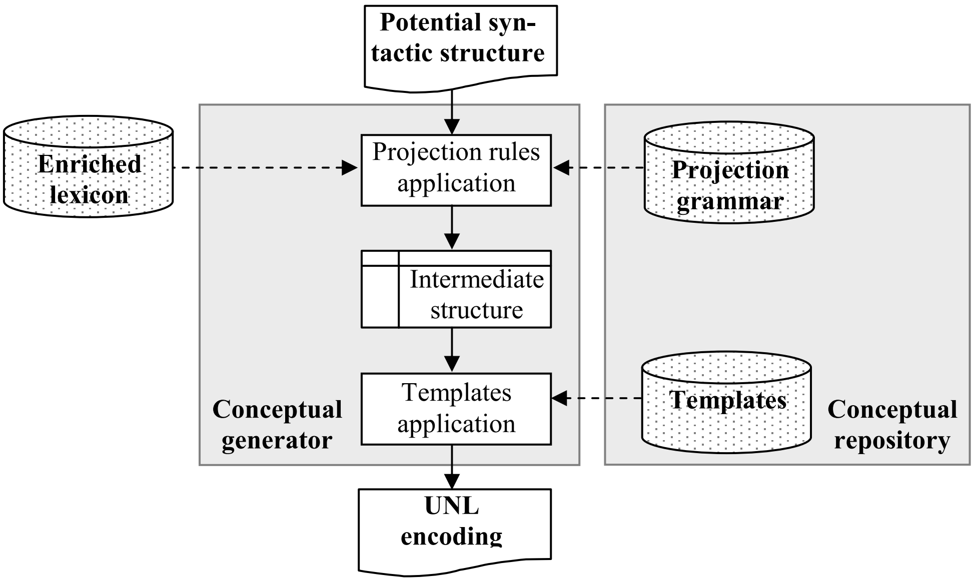

Figure 4. ConPor architecture

Projection rules are compositional and are based on a context-free grammar [4]. They are mainly governed by the argument structure of the verb, in a reading by Borba [2] for BP[2]. Four conceptual classes are considered: *action*, *process*, *state* and *action-process*. Granularity of sentence decomposition is based on the variety of syntactic constructions found in a sample corpus. Correspondingly, determining projection rules is carried out in a stratified way, i.e., from generic rules towards specific ones. The process ends when rules can be no longer decomposed, i.e., when the only step left is to determine UWs for the lexical items. To map, for example, the syntactic structure (Figure 5) of the BP sentence (2), which signals a process, onto the intermediate conceptual structure in Figure 7, some potential projection rules are shown in Figure 6, in a simplified notation[3].

An important aspect of the projection rules is their non-determinism, in that it is possible to combine partial rules in several ways. This makes possible to generate several conceptual structures for a single syntactic structure and shows the potential of ConPor of manipulating semantic ambiguity. This non-determinism also makes possible to estimate ConPor potentialities in dealing with syntactic structures that were not explicitly modeled, but are still licensed.

(2) A lua crescente ocorre amanhã.
Literally: The crescent moon occurs tomorrow.

```
sent([subj(np([mod(det(a)),noum(lua),mod(adj(crescente))])),pred(vp(iv(ocorre))),ap(adv(amanhã)])
```

Figure 5. Syntactic structure for sentence (2)

```
r_ process([SUBJ, PRED, AP],[CE1,CE2,CE3]) :-     pred_process(PRED, SubjRest,CE2),
                                                  subj_process(SUJ, SubjRest,CE1),
                                                  ap(AP,CE3).

pred_process(PRED, SubjRest, CE) :- ivp_process(PRED, SubjRest, CE).

ivp_process(PRED, SubjRest, CE)  :-
                         v(sin(T),can(CanV),[VERBO|_],[]),
                         (sin(_,int),unl(UnlV),cl(processo), rest([suj(SubjRest)]),[CanV|_], []),
                         append(evento:UnlV, .@entry.@, T, CE).
```

Figure 6. Projection rules for sentence (2)

The most general rule in Figure 6, r_process, maps sentence (2) components (SUBJ, PRED and AP) onto their corresponding UNL concepts, along with their semantic roles, by being unfolded into three other rules (not entirely shown here), which help putting together the final conceptual structure. The projection rule for the predicate (pred_process), for example, is decomposed into another rule for intransitive verbal phrase (ivp_process). This, in turn, accesses the enriched lexicon to verify constrains (namely, verb class and transitivity) and retrieve relevant verb features (for example, tense and selectional restrictions for the subject). It also retrieves its UNL concept, adding to it the pertinent ALs (in the example, @entry and @present*)*. The resulting structure is also annotated with the tag 'evento' (event), which signals its semantic role. The complete intermediate structure is shown in Figure 7. In this structure, the concepts corresponding to the syntactic structure constituents are nested in substructures, in order to keep the contextual dependences indicated by the syntactical nesting.

```
[[paciente:moon.@def, modificador:crescent], [[evento:occur.@entry.@present], [tempo:tomorrow]]]
```

Figure 7. Intermediate conceptual structure for sentence (2)

In order to produce each UNL relation, for the final UNL encoding, the relationship templates use the above structure to identify the corresponding RLs. For example, the first substructure above is rewritten as *mod(moon.@def, crescent)*, being the RL *mod* (modifier) identified between the semantic roles 'paciente' (theme) and 'modificador' (modifier). After applying all the possible relationship templates, the final UNL encoding is the one showed in Figure 8.

---

[2] This follows [3] for the English language.

[3] Where: sent = sentence, subj = subject, np = noum phrase, mod = modifier, det = determinant, adj = adjective, pred = predicate, vp = verbal phrase, iv = intransitive verb, ap = adverbial phrase, adv = adverb.

```
[S]
; a lua crescente ocorre amanhã
mod(moon.@def, crescent)
tim(occur.@entry, tomorrow)
obj(occur.@entry, moon.@def)
[/S]
```

Figure 8. Sentece (2) UNL encoding

Differently from other template-based approaches, for example, [21], the incremental use of relationship templates allows for some flexibility: they are not self-contained in ConPor, i.e., there must be a combination of several templates, each one dealing with a specific sentence aspect, to produce the final UNL encoding.

4 ConPor Assessment

For both experiments devised to evaluate ConPor, we defined the following setting [17]: evaluation is *intrinsic* (i.e., it focuses on the very own goals of ConPor, instead of on its use in another task), *blackboxed* (i.e., it considers only the final results of ConPor, to a given input, instead of the results of its internal modules, in order to grade its possible intermediate structures) and *autonomous* (i.e., without comparing the results with the ones of other systems). We focus only on ConPor because there are no other systems that address BP-UNL conceptualization. Similar suggestions have been given by Palmer and Finin [13] for the lack of standards for evaluation in a comparative and intrinsic way.

We used the same *testing corpus* for both experiments. It amounts to 80 BP sentences drawn from 10 distinct magazines and daily newspapers, in the horoscope domain. This corpus is completely independent from the sampling one, although it is limited to those grammatical structures that are addressed by ConPor mapping models.

In what follows we detail both experiments.

4.1 Experiment 1: Assessing the Syntactic-Conceptual Mapping Rules

To evaluate the quality of the syntactic conceptual mapping rules, besides the testing corpus, we built a *reference corpus* of UNL sentences. These resulted from hand-encoding (no BP-UNL automatic encoders are available so far) the sentences of the testing corpus by a UNL specialist. The reference corpus is, thus, our *gold standard* [18].

The experiment was pursued by comparing UNL sentences generated by ConPor with those of the reference corpus. We calculated recall, precision, and f-measure (even weights for recall and precision) for each of the UNL sentences generated by ConPor. Additionally, we also calculated their average. The comparison was carried out by a human, using 3 different symbolic scores, namely: *success* (for satisfactory mapping), *fail* (for over-generated UNL relations), and *miss* (for lacking UNL relations). Like other approaches, for example, [15], an automatic result is considered to be successful only if it is identical to the one of the reference corpus. In evaluating ConPor, this corresponds to verifying the identity between UNL relations.

It is interesting to notice that a *miss* implies the failure of the system in identifying a semantic relation, whilst a *fail* implies the occurrence of an extra relationship, with respect to the reference corpus. However, both miss and fail may apply to a given pair of UWs, in which case two situations may be possible: either the assertion indicates an ambiguous UNL conceptualization, or it is actually an error. Table 1 presents a real example for sentence (3), which is extracted from our testing corpus. In this, the only non-identical UNL relation is evaluated both as a miss (i.e., the system lacks generating the correct relation signaled by *scn*) and a fail (i.e., the system produces an over generated relation, signaled by *plc*). Figures for recall and precision coincide and are equal to 0.67 (so does f-measure).

(3) Dê sugestões nessa area.
Literally: Give suggestions in that area.

Table 1. Reference and automatic encodings for sentence (2)

Reference UNL sentence	ConPor UNL sentence
obj(give.@entry.@imperative, suggestion.@pl)	obj(give.@entry.@imperative, suggestion.@pl)
mod(area, that)	mod(area, that)
scn(suggestion.@pl, area)	plc(suggestion.@pl, area)

The full testing corpus was analyzed in a similar way, yielding 77 sentences with coincident recall and precision rates amounting to 1. The other 3 sentences include: sentence (3) and the 2 that got 0 (zero) for recall and precision because they were not automatically encoded at all. The figures shown for sentence (3) are due to the ambiguity in defining the mapping rules, i.e., due to the impossibility of the system in identifying the difference between a physical and a virtual place (*plc* and *scn*, respectively). Lack of encoding in ConPor is due to the absence of adequate mapping rules for some semantic features appearing in our testing corpus. For example, sentence (4) demands mapping rules that focus on *process*, for the verb 'curtir'. However, the only applicable rules to its corresponding syntactic structure focus on *action-process*, making evident that ConPor does not support the semantics of such an input.

(4) Você poderá curtir a tristeza ou os momentos gratificantes.
Literally: You may experience sadness or gratifying moments.

The average (Av) performance of the mapping rules amount to AvRecall = 0.97, AvPrecision = 0.99, and AvF-measure = 0.97. Such high figures may be explained as follows: (a) mapping rules pose clear decisions to ConPor, since they model constructions that are usually short and simple (this, in turn, is due to the nature of the domain and supposed readership); (b) our testing corpus was built to verify mapping performance for a limited set of linguistic phenomena, yielding only parse trees that can be handled by the ConPor grammar. In fact, the ConPor mapping model does not aim at making operational a theory of computational representation of conceptual structures for any BP linguistic construction, but simply at the mapping of specific constructions (indicated by the sample corpus) onto UNL encodings. Although (a) and (b) may suggest a bias in this experiment, it is worthy noticing that evaluation was carried out in a blackboxed way, totally independent of the intrinsic features of the system.

4.2 Experiment 2: Assessing Semantic Proximity

The criterion under focus here is the semantic similarity between BP sentences resulting from the UNL encodings produced by ConPor and those of the original testing corpus. So, we consider here pairs of BP sentences: authentic ones and those resulting from automatic processing. Similarly to building the UNL reference corpus, we had a human UNL specialist decoding the 78 UNL sentences onto BP ones, who had no access to the original BP sentences.

Semantic proximity is viewed in a subjective way: 27 human judges, native speakers of BP, pinpointed it. We calculated both, the average of the judges' scores for each sentence and the general average, i.e., the total average score for all the sentences. Three scores were considered in the judgment, as shown in Table 2. We consider that, once highly graded, semantic proximity signals a good expressiveness of the UNL sentences generated by ConPor, with relation to the meaning of the original sentences. Implicitly, it means that ConPor performs well in mapping their parsing structures onto UNL sentences.

Table 2. Semantic proximity evaluation options

Semantic proximity	Score
No proximity at all (very distant meanings)	1
More or less proximity (relatively near meanings)	2
Complete proximity (completely near meanings)	3

As it can be noticed, this experiment is less strict than the other, although more subjective, because semantic proximity relies on the judges' intuition in identifying sentences that, although non-identical, may still be semantically close. This allows partially inadequate UNL encodings generated by ConPor to be effective, when compared to source data.

In our assessment, 49 out of the 78 sentence pairs under evaluation are identical and were directly scored 3. The 29 remaining pairs resulted in 10.3% scoring 3, 75.8% between score 2 and 3, 13.8% under score 2, and 6.8% scoring 1 (those two cases for which ConPor could not produce any UNL sentence). The figures for all the sentences, including those identical pairs and those which were not processed by ConPor are shown in Table 3.

Table 3. Figures for semantic proximity of the 80 sentences

Score	Number of sentences	Percentage of sentences
1 a 1,9	6	7%
2 a 2,9	22	28%
3	52	65%

Most of the worst cases refer to problems with linking verbs in BP, i.e., 'ser' and 'estar'. Both verbs correspond to the English *to be*, but the latter has a temporary connotation, whilst the former does not. For example, the pair composed by (5), automatically generated and hand-decoded, and (6), appearing in the original testing corpus, was graded 1.7. Both sentences are expressed in UNL through the RL *aoj* (attribute of object), i.e., by the binary relation *aoj(charming, you)*. Since there are no means in UNL to stress this difference, ConPor cannot solve this problem and ambiguity arises in decoding such a UNL relation onto BP.

(5) Você é charmoso.
Literally: You are charming..

(6) Você está charmoso.
Literally: You are charming..

The best cases usually involve ordering of sentence components, mostly when they are conjoined. Interchangeability, in BP, justifies why ordering in our assessment poses no problem for semantic proximity. Average cases seem to be the most complex ones. Sometimes, they also involve 'ser' and 'estar', but they often flag the ordering of adverbs or gender variations during decoding. Gender variations consist of a problem in BP, since gender is not represented in UNL, resulting in an obscure decoding. Adverb ordering was judged problematic for some BP adverbs. For example, 'muitas vezes' (lit.: very often) in the original sentence (7) provided a UNL sentence whose decoding resulted in (8). This pair was graded 2.22. In BP, the rationale for this could be that in (7), the adverb implies frequency, whilst in (8) it implies quantity (many times). Such a variation occurs because UNL binary relations are considered independent from each other, in the sentence context.

(7) Muitas vezes você só mostrará alegrias.
Literally: Very often you will only show happiness.

(8) Você só mostrará alegrias muitas vezes.
Literally: You will only show happiness very often.

It is important to notice that a sentence that is scored low in Experiment 1 may be adequate from the semantic proximity viewpoint. This is the case of sentence (3), which, although not very precise (precision rate = 0.67), was averaged 3 by human judges. This means that even when ConPor produces UNL encodings that would not be considered by human encoders, or suggested by the UNL specification, their corresponding decodings into BP still can allow identifying their correspondence to the original sentences. In the example for sentence (1), the variations are very subtle and could be handled by ConPor if we added extra semantic features to the related lexical items. However, this decision is questionable: it should be considered only if there were many problematic cases. According to the current data, the burden of recategorizing the whole lexicon would not be worthy.

The general average score obtained in Experiment 2, considering the whole testing corpus, was 2.77 (maxim= 3). This shows that ConPor accomplishes well the task of producing results that are semantically close to the corresponding sources. Like for the Experiment 1, this high figure is certainly due to the simplicity of the input syntactic constructions. Again, we should stress that evaluation was intuitively carried out by humans, who had no access to UNL representations and had no idea of ConPor strategies.

5 Discussion and Conclusions

In this paper we describe two experiments aiming at assessing ConPor, which maps BP sentences into UNL ones. We focus only on intrinsic evaluation, analyzing both the UNL sentences and their proximity to the alleged meaning of the source sentences. Although Experiment 1 allows verifying the performance of both the projection rules and heuristics, it does not ensure that the conceptual structures generated by ConPor convey the same, or a near, meaning to the original sentences. It is possible that a *correct* mapping signaled by Experiment 1 yields an unsatisfactory semantic proximity. Conversely, an incorrect syntactic-conceptual mapping caused, for example, by the impossibility of ConPor in recognizing the already mentioned semantic subtleties, not necessarily implies contradictory BP utterances. So, there is no direct implication between high (or low) values in Experiment 1 and those of Experiment 2. Because of this, semantic proximity is shown to complement other means to evaluate semantic interpretation. More importantly, it means that producing only recall and precision distributions for syntactic-conceptual mappings of any natural language does not suffice the most important goal of an automatic system: guaranteeing that automatically produced conceptual representations actually convey most of the original meaning of its input sentences.

Although Rosé has also shown the importance of considering the semantic interrelationship to convey the same meaning as the original one, her "perfect rate" for quality embeds both fluency and the communication of the original idea. However, it is not always the case that a non-fluent sentence implies no semantic proximity at all. So, if we wish to emphasize the quality of semantic interpretation, fluency should be seconded with respect to meaning correspondence. Adding this to the above reasons, we make evident that our Experiment 2 poses a contribution to the evaluation of semantic interpretation.

Semantic proximity has also been addressed in other fields, for example, Automatic Summarization, to calculate the semantic load of both, the original text and the automatic summaries. Semantic load resembles semantic informativeness, i.e., the amount of content information that is common to both, the summaries and related source texts [9]. Using this same idea for semantic interpretation, we should count the agreements between sentence segments in order to properly convey the corresponding semantic informativeness. By doing so, we would have a discrete strategy of matching and scaling. Clearly, although this evaluation technique resembles ours, they deeply differ in that we allow for evaluating semantic proximity at the conceptual level in a less strict way, as already emphasized. Sentences (7) and (8) are good examples of this: although their content segments are entirely coincident and, thus, they would score the highest for semantic informativeness, the average obtained through human judgment in Experiment 2 shows that they are not semantically equivalent. This deep distinction is due to the influence of ordering, or structural variations, in grading semantic interpretation, which may not be relevant in Automatic Summarization, but is proved to be crucial in semantic interpretation.

The current ConPor version is limited in that it focuses only on the syntactic-conceptual mapping and in that it has a syntactic coverage relatively small. These certainly are influential factors for the high figures shown before. Both problems should be tackled as soon as possible for robustness. In spite of having such a degree of simplicity in ConPor, our proposal brings about the difficulty in specifying sound procedures to evaluate conceptual generation. The related metrics would be equally applicable to more refined approximations to semantic interpretation and would also allow exploring larger sets of linguistic phenomena than those embedded in our testing corpus, addressing longer and more complex utterances.

References

[1] Bean, C., Rindflesch, T.C. and Sneiderman, C.A. Automatic Semantic Interpretation of Anatomic Relationships in Clinical Text. In *Proceedings of the AMIA Annual Fall Symposium*, (Orlando, 1998), pp. 897-901.

[2] Borba, F.S. *Dicionário gramatical de verbos do português contemporâneo do Brasil*. São Paulo: Fundação Editora Unesp. 1990.

[3] Chafe, W. *Meaning and the Structure of the Language*. Chicago: University of Chicago Press. 1970.

[4] Chomsky, N. *Aspects of the Theory of Syntax*. Cambridge: The MIT Press. 1965.

[5] Cruse, A.D. *Lexical Semantics*. Cambridge: Cambridge University Press. 1986.

[6] Dias-da-Silva, B.C., Sossolote, C.R.C., Zavaglia, C., Montilha, G., Rino, L.H.M., Nunes, M.G.V, Oliveira Jr., O.N. and Aluísio, S.M. The design of a Brazilian Portuguese machine tractable dictionary for an interlingua sentence generator. In Lima, V.L.S. (ed.), *Proceedings of the III Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, (Porto Alegre, 1998), pp. 71-78. Download: http://www.nilc.icmc.usp.br/nilc/publications.htm.

[7] Dorr, B.J. The Use of Lexical Semantics in Interlingual Machine Translation. *Machine Translation*, Vol. 7, No. 3, (1992), pp. 135-193.

[8] Jackendoff, R. *Semantic Structures*. Cambridge: The MIT Press. 1990.

[9] Mani, I. *Automatic Summarization*. Philadelphia: John Benjamins Publishing Company. 2001.

[10] Martins, R.T., Hasegawa, R. and Nunes, M.G.V. *Curupira: um parser funcional para o português*. Série de Relatórios Técnicos do NILC, NILC-TR-02-26, (São Carlos, Dezembro, 2002). Download: http://www.nilc.icmc.usp.br/nilc/publications.htm.

[11] Mitamura, T. and Nyberg, E.H. Hierarchical Lexical Structure and Interpretive Mapping in Machine Translation. In *Proceedings of COLING-92*, (Nantes, 1992).

[12] Nunes, M.G.V., Vieira, F.M.C., Zavaglia, C., Sossolote, C.R.C. and Hernandez, J. *A Construção de um Léxico da Língua Portuguesa do Brasil para suporte à Correção Automática de Textos*. Série de Relatórios Técnicos do ICMC, (Tech.Rep. 42), (São Carlos, 1996). Download: http://www.nilc.icmc.usp.br/nilc/publications.htm.

[13] Palmer, M.S. and Finin, T. Workshop on the Evaluation of Natural Language Processing Systems. *Unisys Center for Advanced Information Technology*, Pennsylvania (1990).

[14] Romacker, M., Markert, K., Hahn, U. Lean semantic interpretation. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, (Stockholm, 1999), pp. 868-875.

[15] Romacker, M. and Hahn, U. An Empirical Assessment of Semantic Interpretation. In *Proceedings of the 6th Applied Natural Language Processing Conference & the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, (Seattle, 2000), pp. 327-334.

[16] Rosé, C.P. A Framework for Robust Semantic Interpretation. In *Proceedings of the 6th Applied Natural Language Processing Conference & the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, (Seattle, 2000), pp. 311-318.

[17] Sparck-Jones, K. and Galliers, J.R. *Evaluating Natural Language Processing Systems: an Analysis and Review*. Springer-Verlag, 1996.

[18] Teufel, S. and Moens, M. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In Mani, I. and Maybory, M.T. (eds.), *Advances in Automatic Text Summarization*. Cambridge: The MIT Press. 1999.

[19] Uchida, H., Zhu, M. and Senta, T.D. *The UNL, a Gift for a Millennium*. Tokyo: UNU/IAS/UNL Center. 1999.

[20] UNL. *The Universal Networking Language Specifications*. Tokyo: UNU/IAS/UNL Center. 2001.

[21] Wilensky, R. and Arens, Y. PHRAN - A Knowledge Based Natural Language Understander. In *Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics*, (Philadelphia, 1980), pp. 117-121.