The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016

EDITORIAL BOARD

Published twice a year by Charles University (Prague, Czech Republic)

Editorial office and subscription inquiries:
ÚFAL MFF UK, Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
E-mail: pbml@ufal.mff.cuni.cz

ISSN 0032-6585

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016

CONTENTS

Articles

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 5–20

Against the Argument–Adjunct Distinction in Functional Generative Description

Adam Przepiórkowski[a b]

[a] Institute of Computer Science, Polish Academy of Sciences
[b] Institute of Philosophy, University of Warsaw

Abstract

The aim of this paper is to critically examine the tests used to distinguish arguments from adjuncts in Functional Generative Description (Sgall et al., 1986) and to question the general usefulness of this distinction. In particular, we demonstrate that neither of the two tests used in FGD to distinguish *inner participants* from *free adverbials* (i.e. tests based on iterability and specificity) stands up to scrutiny, and we also point out practical problems with the application of the dialogue test, used to distinguish *semantically obligatory* and *semantically optional* dependents. Since these tests are among the most frequently cited tests for the Argument–Adjunct Distinction in contemporary linguistics, these results cast a shadow on the general validity of this dichotomy.

1. Introduction

Probably all modern linguistic theories assume some form of the Argument–Adjunct Distinction (AAD), which may be traced back to Lucien Tesnière's (1959) distinction between *actants* and *circumstants*: the former are elements of valency frames of respective predicates, while the latter are not. In case of the Functional Generative Description (FGD; Sgall et al. 1986), the relevant classes were at one point (Panevová, 1974, 1978) called *inner participants* and *free adverbials*, respectively, but in later FGD works these terms adopted different meanings (see below), so we will use the widespread cross-theoretical terms *arguments* and *adjuncts* to distinguish valency dependents from non-valency dependents. Obviously, all these terms are relative to a particular occurrence of a particular predicate: a phrase which is an adjunct to a predicate in one utterance (e.g. *that day* in *I saw John that day*; cf. Larson 1985, p. 595)

 Corresponding author: adamp@ipipan.waw.pl
Cite as: Adam Przepiórkowski. Against the Argument–Adjunct Distinction in Functional Generative Description. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 5–20. doi: 10.1515/pralin-2016-0008.

may be an argument of another predicate or even of another occurrence of the same predicate: either a complement (*I saw that day in my mind*) or the subject (*That day saw the coronation of Charles IV*).

The aim of this paper is to critically examine the tests used to distinguish arguments from adjuncts in FGD, as presented in Panevová 1974, 1978. In particular, we will point out the weaknesses of particular tests used when making this distinction, and of the way they are combined. Some attention will be devoted to the so-called "dialogue test" (Sgall and Hajičová, 1970; Panevová, 1974, 1978); we will show that tests aiming at the same property of dependents have been used in another linguistic tradition since around the same time.

2. AAD in FGD

Two basic dichotomies play a role in FGD in distinguishing arguments from adjuncts: 1) that between – to use the current FGD terminology – *inner participants* (called *complements* in Panevová 1974, 1978) and *free adverbials* (called simply *adverbials* in Panevová 1974, 1978), and 2) that between *semantically obligatory* and *semantically optional* dependents.

Two tests are used in case of the first dichotomy, between inner participants and free adverbials (Panevová, 1974, p. 11):

(1) "Can the given type of participant [i.e. *dependent* in our terminology; AP] depend on every verb?"

(2) "Can the given type of participant [i.e. *dependent*; AP] depend more than once on a single verb token…?"

In the literature on arguments and adjuncts, the former test is known as the *specificity* criterion (e.g. Koenig et al. 2003), and the latter – as the *iterability* criterion (e.g. Williams 2015, pp. 69–70). For example, time (TWHEN), location (LOC) and manner (MANN) dependents seem to be allowed to occur with almost any verb and they may be iterated, as in the famous example from Bresnan 1982b, p. 164:[1]

(3) Fred *deftly* [MANN] handed a toy to the baby *by reaching behind his back* [MANN] *over lunch* [TWHEN] *at noon* [TWHEN] *in a restaurant* [LOC] *last Sunday* [TWHEN] *in Back Bay* [LOC] *without interrupting the discussion* [MANN].

The assumption in Panevová 1974 is that these two tests go hand in hand, with the exception of the type of dependent called *actor* in FGD (roughly, deep subject), which may in principle occur with almost any verb (like free adverbials do) but is not iterable (just like inner participants). While Panevová 1978, pp. 232–233, plays down the iterability criterion (2) and instead relies only on the specificity test (1), iterability is still mentioned in later FGD work, so we will include it in the discussion below.

[1]The original annotations Manner, Temp and Loc are substituted here with the FGD functors MANN, TWHEN and LOC, assuming the repertoire of functors listed in Žabokrtský 2005, pp. 117–118.

The other dichotomy used to distinguish arguments from adjuncts concerns semantic obligatoriness, as verified by the dialogue test.[2] This test may be illustrated on the basis of the verb ARRIVE and used to decide whether the possible ablative (where from) and adlative (where to) dependents are semantically obligatory (and, hence, arguments in the sense used in this paper), even though both are syntactically optional. Let us imagine that A said *John arrived*. If the dialogue continues by B asking *Where from?* and A answering *I don't know*, there is nothing particular about the dialogue. However, if B asks *Where?* and A answers *I don't know*, there is something funny about it: how could have A said *John arrived* if he cannot answer the question where John arrived? Perhaps a different verb should have been used by A. Hence, according to this test, the adlative dependent, unlike the ablative dependent, is semantically obligatory.

Given these two distinctions: inner participants vs. free adverbials and semantically obligatory vs. semantically not obligatory, arguments are in practice – although not in these exact words – defined in Panevová 1974, 1978 as the set-theoretic sum of inner participants and semantically obligatory dependents. That is, arguments naturally fall into three sub-classes: semantically obligatory inner participants (including the deep subject), semantically optional inner participants, and semantically obligatory free adverbials.

The picture is actually a little more complicated in Panevová 1978, where another criterion is taken into account, namely, "semantic unity of modification". This leads to 8 theoretical possibilities (given that iterability is played down in that publication), summarised in Table 1. Panevová 1978, p. 234, states that "the combination of features

	1 obligatoriness	2 limited number of governing verbs	3 semantic unity of a modification
1	+	+	+
2	+	+	−
3	+	−	+
4	+	−	−
5	−	+	+
6	−	+	−
7	−	−	+
8	−	−	−

Table 1. Three criteria for AAD (from Panevová 1978, p. 233)

in the lines 4, 6 remain unrealized" and that "only such participants characterized in

[2]Our description of this test is based on that of Przepiórkowski et al. 2016.

the matrix at least by two positive [i.e. +; AP] features belong to inner participants [i.e. arguments; AP]". Given these statements, it is not necessary to understand the notion of "semantic unity of modification" to see that it is redundant for the purposes of AAD: after removing lines 4 and 6, the lines with at least two + values are exactly the lines where at least one of columns 1 and 2 contains a +, i.e. the lines corresponding to the sum of semantically obligatory dependents and dependents passing the specificity test on inner participants. Hence, in the remainder of this paper we will not be concerned with the "semantic unity of modification", which also seems to play no role in FGD literature after Panevová 1978.

3. Iterability

Bresnan 1982b, p. 165, contrasts example (3) above with the following example,[3] purportedly showing that instruments cannot be iterated:

(4) *John escaped from prison *with dynamite* [MEANS] *with a machine gun* [MEANS].

However, there is a clear difference between the repeated functors in the two examples: in (3), they refer to the same entity, while in (4), they refer to two different entities. In particular, *dynamite* and *machine gun* necessarily denote two different instruments, while both *in a restaurant* and *in Back Bay* refer – with different granularities – to the single location of the event. Similarly, there is a single time of the event described in (3), which is referred to via phrases *over lunch*, *at noon* and *last Sunday*, and – arguably – there is a single manner of handing a toy to the baby which may be variously characterised as *deftly*, *by reaching behind his back* and *without interrupting the discussion*.

Once this difference between (3) vs. (4) is acknowledged, it becomes less clear that there is any iterability contrast between different functors. For example, Goldberg (2002, pp. 334–335, 341) argues that instrumental phrases may be iterated as long as they "concentrically" refer to the same entity, and supports this claim with the following examples:[4]

(5) With a slingshot he broke the window with a rock.

(6) The robot opened the door with a key with its robotic arm.

Another – perhaps more convincing – example of iterated arguments is mentioned in Zaenen and Crouch 2009, p. 646:

[3] Again, the original tag Inst is replaced here with the corresponding FGD functor MEANS.

[4] Within FGD, instruments are supposed to be freely iterable, as they are treated as free adverbials, but the only example of such iterability we have come across is not convincing. Panevová 2003, p. 2, provides the following Russian example, but it is controversial that both phrases marked as *Instrument* should bear the same semantic role (functor); *na rojale*, lit. 'on piano', should rather be classified as one of the core arguments of the verb IGRAT' 'play':

(i) Ivan umeet igrat' na rojale (*Instrument*) tol'ko pravoj rukoj (*Instrument*).
[Ivan can play a piano only with his right hand.]

(See also Sgall et al. 1986, p. 161, fn. 58, on the possibility of partitioning such a general Instrument role into more specific roles.)

(7) I count on you, on your kindness.

As shown in (8), taken from Urešová 2011, p. 148, the PDT-Vallex valency dictionary of Czech (Hajič et al., 2003) based on FGD treats SPOLÉHAT (NA), the Czech for 'count (on)', as taking a prepositional patient:

(8) **spoléhat** ACT(1) PAT(na+4;↓že)

There seems to be no reason to assume that the corresponding phrases in (7) should bear a different functor, so this example involves a repetition of the patient functor, and the Czech and Polish facts are similar, as shown in (9)–(10), which are literal translations of (7).

(9) Spoléhám na vás, na vaši laskavost. (Czech)

(10) Liczę na was, na waszą życzliwość. (Polish)

Hence, in all three languages, the sentences (7) and (9)–(10) should be analysed as exhibiting iteration of the (semantically obligatory) patient, i.e., a clear case of an inner participant (and a prototypical argument).

It is easy to construct examples of other iterated inner participants, for example, an iterated actor, as in the following Polish example, where the three nominative NPs are understood as referring to the same person:

(11) Ważny urzędnik wczoraj przyszedł, dyrektor departamentu,
important.NOM official.NOM yesterday came director.NOM department.GEN
bardzo wysoko postawiona osoba…
very highly placed person
'An important official came yesterday: the director of a/the department, a very high-ranking person.'

It could be argued that (7) and (9)–(10), and maybe also (11), should be analysed as some special construction, perhaps a type of apposition. Perhaps so. But whatever the analysis of such examples of iterated inner participants, the burden is on the shoulders of the proponents of the dichotomy to show that this analysis does not carry over to examples of iterated free adverbials. Since we are not aware of such an argument, we conclude then that iterability, as currently understood, fails to distinguish inner participants from free adverbials and, hence, does not seem relevant for the Argument–Adjunct Distinction.

4. Specificity

Taken literally, the specificity test also gives undesirable results, as very few of the intended members of the class of free adverbials may really "depend on every verb". For example, McConnell-Ginet 1982, p. 166, notes that WEIGH fails to combine with many typical adverbials:

(12) *Annie weighs 120 pounds {heavily / beautifully / quickly / elegantly}.

(13) *Annie weighs 120 pounds {for her mother / with a fork / in an hour / toward Detroit}.

Even such prototypical types of free adverbials as TWHEN (time) and LOC (location) are subject to exceptions. As shown in Koenig et al. 2003, p. 80, where an experiment consisting in the manual examination of 3909 English verbs is reported, 0.2% (i.e. 8) of them do not combine with dependents of type TWHEN and 1.8% (i.e. as many as 70) do not combine with LOC. Matters are much worse in case of most other dependent types claimed (Panevová, 1974, p. 12) to occur with all verbs.

It is also clear that the results of this test depend on the granularity of functors. For example, simplifying a little, Koenig et al. 2003 treat as arguments those dependents which may occur with up to 30% of all verbs, and as adjuncts – those which may occur with at least 90% of all verbs. It seems then that agents should count as typical adjuncts. However, Koenig et al. 2003 avoid this conclusion by splitting this dependent type into more fine-grained semantic roles, as proposed in Dowty 1989, 1991, and showing that each of them occurs with less than 30% of the examined verbs.

Similar reasoning may be applied to durative dependents (THL, i.e., "temporal how long") – typical free adverbials. It is probably true that they may modify all or almost all verbs, including the class of verbs which Laskowski (1998) calls atemporal. However, not all durative dependents are alike, and it has been shown that prepositional durative phrases such as *for two hours* have different distribution and semantic implications than bare nominal durative phrases such as *two hours* even in languages such as English (Morzycki, 2005). Also in Slavic languages, the distribution of the two kinds of duratives differs, as the following Polish examples illustrate:

(14) a. Janek tańczył przez dwie godziny.
Janek.NOM danced.IMPERF for two.ACC hours.ACC
'Janek was dancing for two hours.'

b. Janek ?*(tylko raz) zatańczył przez dwie godziny.
Janek.NOM only once.ACC danced.PERF for two.ACC hours.ACC

c. Janek ?(ani razu) nie zatańczył przez dwie godziny.
Janek.NOM not once.GEN NEG danced.PERF for two.ACC hours.ACC
'For two hours, Janek didn't dance.'

(15) a. Janek tańczył dwie godziny.
Janek.NOM danced.IMPERF two.ACC hours.ACC
'Janek was dancing for two hours.'

b. *Janek (tylko raz) zatańczył dwie godziny.
Janek.NOM only once.ACC danced.PERF for two.ACC hours.ACC

c. *Janek (ani razu) nie zatańczył dwie godziny.
Janek.NOM NEG danced.PERF two.ACC hours.ACC

The three examples in (14) show that prepositional PRZEZ+NP[ACC] duratives may combine with both imperfective and perfective verbs, although their semantic contribution to the sentence differs in these two cases. In (14a), which involves an imperfective verb, the natural understanding of the sentence is that the event of dancing lasted for two hours. This meaning is absent in (14b–c), which involve the perfective

counterpart of that verb: they cannot mean that dancing lasted (in b.) or not lasted (in c.) for two hours. Rather, the durative PPs set the time frame, during which an event is said to occur once (however long it lasted – perhaps only a few minutes) or not occur at all. For this reason, the naturalness (14b–c) is greatly enhanced by dependents with meanings such as 'only once', 'a number of times' or 'not even once' – this is especially true about (14b).

On the other hand, as shown in (15), bare NP[ACC] seem not to have this time frame meaning, and may only refer to the duration of the event expressed by the verb – this explains the ungrammaticality of (15b–c), even in versions with added dependents meaning 'only once' or 'not even once'. Since the two types of durative dependents may contribute to the meaning of the sentence in different ways and, hence, have different distributions, they should be represented by different functors in a fully precise and explicit generative description; let us call these functors THL-PP ("prepositional temporal how long") and THL-NP ("bare nominal temporal how long"). Now, while THL-PP may still be claimed to be able to occur with all or almost all verbs (but with different meaning contribution, depending on the broadly understood aspectual characteristic of the verb – either as eventuality[5] duration or time frame duration), THL-NP is confined to imperfective verbs, as well as delimitatives (Piernikarski, 1969; Bogusławski, 2004) and perhaps a limited number of other verbs, and always denotes eventuality duration. Adopting the prevailing view that in Slavic languages such as Czech or Polish aspect is lexicalised, i.e. that the imperfective and the perfective variants are two different lexemes, this means that such dependents will only combine with perhaps more than half of the verbs, but certainly very far from all of them, and yet they are considered typical free adverbials in FGD.

A similar point can also easily be made on the basis of various functors whose presence depends on whether the verb requires an agent (as is well known, many verbs do not, e.g. weather predicates or psych-verbs), e.g. the INTT (intent) functor and perhaps MEANS (instrument). The problem that many intended free adverbials do not really combine with various classes of verbs is duly noted in a footnote (fn. 6 in Panevová 1974 and fn. 13 in Panevová 1978), where it is stated that "it appears as a rule that such a combination is not grammatically excluded but is unusual due to cognitive or ontological reasons" (Panevová, 1978, p. 252). Unfortunately, this view makes the test largely unusable in practice, as there is no operational procedure of distinguishing "grammatical unacceptability" from "cognitive or ontological unacceptability". Moreover, it is not clear that such a distinction is justified at all; as shown in Levin 1993, grammatical behaviour of verbs (their diathesis patterns) strongly correlates with their meaning (which may be hard to distinguish from "cognitive or ontological" aspects).

In summary, very few classes of free adverbials, if indeed any, "can depend on every verb", and attempts to distinguish reasons for not satisfying this criterion have

[5] We use Emmon Bach's (1986) term here, which generalises events and states.

never, to the best of our knowledge, been translated into an operational test, so the specificity criterion simply does not do the job it was supposed to do.

5. Semantic obligatoriness

The dialogue test has an interesting history. While it was first presented in Panevová 1974, pp. 17–19, it was apparently inspired "by the method of 'given and new information', which was briefly sketched by Sgall and Hajičová (1970, § 3.1)" (Panevová, 1974, p. 16).

Sgall and Hajičová (1970, p. 17) critically discuss the view that clauses have a possibly implicit "time of the clause" semantic dependent, so that sentences such as *He wrote all these books* might be understood as talking about (at least) three entities: the agent, the patient and the time. But if this were so, the questions *Who wrote them?* and *When did he?* should be fully analogous: the *wh*-word should refer to an entity already introduced in the discourse. In other words, these questions should have the following meanings: *Who do you mean?* and *What time do you mean?*. However, while *Who wrote them?* may be understood as *Who do you mean?* in the context of the previously uttered *He wrote all these books*, the question *When did he?* cannot be paraphrased as *What time do you mean?*; for such a paraphrase to work, the original utterance should have been *He wrote all these books then*, or so. Hence, "time of the clause" is not an implicit semantically obligatory dependent in *He wrote all these books*. This should be contrasted with examples such as *John returned*. Here, *Where from?* also cannot be understood as *What origin of the journey do you mean?*, but *Where to?* may indeed be understood as *What destination of the journey do you mean?*.

On the other hand, it has remained hardly noticed that the property targeted by the dialogue test is also discussed in another thread of linguistic work, starting with Fillmore 1969.[6] The following sentences, among others, are discussed there, with some semantic roles (Fillmore's "deep cases") unrealised syntactically in the last two sentences (Fillmore, 1969, pp. 118–119):

(16) The boys blamed the girls for the mess.

(17) The boys blamed the girls.

(18) The girls were blamed for the mess.

In comparison with the complete sentence (16), the offence is missing in (17), and the accuser is absent in (18). However, these two implicit dependents have different kinds of interpretations in the two sentences. The last sentence, (18), is "a syntactically complete sentence, in the sense that it can appropriately initiate a discourse (as long as the addressee knows who the girls are and what the mess is). In this case the speaker is merely being indefinite or non-committal about the identity of the accuser" (Fillmore, 1969, p. 119). Hence, this sentence may be paraphrased as *The girls were blamed*

[6] Interestingly, Fillmore 1969 is cited in Sgall and Hajičová 1970 and in Panevová 1974, 1978, but in a different context.

for the mess by someone. On the other hand, sentence (17) is "one which cannot initiate a conversation and one which is usable only in a context in which the addressee is in a position to know what it is that the girls are being blamed for" (ibid.). That is, it cannot be paraphrased as *The boys blamed the girls for something*, but rather as *The boys blamed the girls for it*. "The two situations correspond, in other words, to definite and indefinite pronominalization" (ibid.).

Fillmore 1969 does not have much more to say about this phenomenon, but the discussion in Fillmore 1986 makes it clear that the definite interpretation of the implicit dependent in (17) concerns the same phenomenon as the semantic obligatoriness picked out by the dialogue test. For example, when comparing the verb EAT, which allows for an indefinite implicit dependent, with FIND OUT, whose implicit dependent must be definite, Fillmore 1986, p. 96 says: "It's not odd to say things like, 'He was eating; I wonder what he was eating'; but it is odd to say things like 'They found out; I wonder what they found out'". This test very closely resembles the dialogue test, and it gives the same results. For example, in case of ARRIVE, it would be natural to say *He arrived; I wonder where he arrived from*, but the following sounds odd: *He arrived; I wonder where he arrived*.

This distinction between the two classes of implicit dependents has been widely discussed in the literature; some of this discussion is summarised in Williams 2015, ch. 5, where the relation to the dialogue test is alluded to (pp. 100–101). Such implicit dependents are carefully analysed in Recanati 2002, 2007 (his *unarticulated constituents*), where a position similar to that of FGD is adopted: definite implicit dependents, i.e. those classified as semantically obligatory by the dialogue test, are claimed to be present in the semantic structure of respective sentences, while the existential implicit dependents, i.e. those classified as optional by the dialogue test, are claimed to be absent from the semantic representation. On the other hand, according to Recanati (2002, 2007), such absent dependents may be added to the argument structure of the predicates via essentially pragmatic – context-dependent – processes. On this analysis, given that the a. and b. sentences below are synonymous, there is no difference between the direct object (and, hence, a prototypical argument in most theories) of EAT and the locative dependent of DANCE (a prototypical adjunct); in both cases the a. sentences have only one argument (the subject), and the b. sentences have two arguments:

(19) a. John is eating.
b. John is eating something or other.

(20) a. John is dancing.
b. John is dancing somewhere or other.

The situation is markedly different in case of verbs such as NOTICE and ARRIVE, where the b. sentences below are not synonymous with the a. sentences; better paraphrases are given in c.:

(21) a. John noticed.
b. John noticed something or other.
c. John noticed it / the thing.

(22) a. John arrived.
b. John arrived somewhere or other.
c. John arrived there / at the destination.

Note that, when talking about arguments, Recanati (2002, 2007) completely disregards the distinction between inner participants and free adverbials.

The importance of Recanati 2002, 2007 for the current considerations lies in the discussion of difficulties in applying the dialogue test.[7] At first sight it might seem that location is a semantically obligatory argument of RAIN, as the dialogue in (23) seems to pattern with the awkward (24) rather than with the natural (25):

(23) A: It is raining.
B: Where is it raining?
A: I have no idea.

(24) A: John has arrived.
B: Where has he arrived?
A: I have no idea.

(25) A: John has danced.
B: Where has he danced?
A: I have no idea.

However, Recanati (2002, p. 317) carefully constructs a context that makes a dialogue such as (23) sound natural:

> *I can imagine a situation in which rain has become extremely rare and important, and rain detectors have been disposed all over the territory… In the imagined scenario, each detector triggers an alarm bell in the Monitoring Room when it detects rain. There is a single bell; the location of the triggering detector is indicated by a light on a board in the Monitoring Room. After weeks of total drought, the bell eventually rings in the Monitoring Room. Hearing it, the weatherman on duty in the adjacent room shouts: 'It's raining!' His utterance is true, iff it is raining (at the time of utterance) in some place or other.*

Translated to the dialogue test, this renders (Recanati, 2007, p. 129):

(26) A (the weatherman): It is raining!
B: Where?
A: I have no idea — let's check.

Hence, Recanati (2002, 2007) concludes that, contrary to the standard view in the kind of (philosophically inclined) literature he cites, RAIN has no semantically obligatory location argument; in case location is expressed in the sentence (as in *It is raining in Paris*), such an argument is added via a pragmatic process proposed in Recanati 2002. But in order to reach this conclusion, the first impression given by a straight-

[7] Other subtleties and "thorny questions" regarding the practical use of the dialogue test are discussed in Panevová 2001, § 4. The fact that it is not always easy to apply the dialogue test when constructing a valency dictionary is also mentioned in Urešová 2006, p. 95, and in Przepiórkowski et al. 2016, p. 14.

forward application of the dialogue test had to be rejected and a very specific context had to be constructed.

In fact, the discussion in the philosophical literature on the applicability of the dialogue test remains inconclusive, as it seems that – by constructing sufficiently unusual contexts – all implicit arguments should be considered semantically optional. In particular, Recanati 2010, p. 117, cites an anonymous *Linguistics and Philosophy* referee as providing the following context, which suggests that the object of NOTICE, apparently semantically obligatory on the straightforward application of the dialogue test, is in fact semantically optional:

> *Consider a scenario with a patient who has been in a semi-coma, and a technician in another room is reading the output of an EEG or whatever it is that measures brain activity in various areas of the brain. It seems to me that a trained technician could know when brain activity signals 'noticing', and since for the semi-coma patient, the fact that he's noticing (something) is all that's important, one might imagine the technician being able to shout 'He's noticing!' without being in any position to know or say what it is that the patient is noticing.*

These considerations open the possibility that the dialogue test does not really distinguish between semantically obligatory and semantically optional constituents, and that the perceived obligatoriness is a perhaps graded function of context: in case of some dependents of some verbs it is easier to construct a context in which the dependent is understood existentially (i.e. as semantically optional), and in case of other it is more difficult to construct such a context, but perhaps it is always possible. In any case, in order to be truly operational, the dialogue test and the conditions in which it may be used should be further elaborated.

6. Does FGD need AAD?

Let us take stock. Functional Generative Description proposes two orthogonal classifications of dependents: into inner participants and free adverbials, and into semantically obligatory and semantically optional. The product of these classifications gives four classes, three of which – with the exception of semantically optional free adverbials – together constitute arguments, i.e. valency elements. This in itself is not unreasonable and it is interesting to note that a very similar idea is apparently independently proposed in Goldberg 2002, pp. 344–346, within the framework of Construction Grammar (Goldberg, 1995).

The distinction between semantically obligatory and semantically optional dependents, even if sometimes difficult to test, is widely assumed in contemporary linguistics. However, by itself this distinction does not correlate with the common understanding of the AAD, as it distinguishes between the traditional complements of NOTICE and DEVOUR on one hand (they are semantically obligatory), and the traditional complements of EAT and SPEAK (to somebody) on the other (they are semantically optional). The other distinction, that between inner participants and free adverbials,

while very close to AAD, is not operational as it stands: prototypical inner participants (including patients and actors) seem to be as iterable as prototypical free adverbials, and there are many exceptions to the intended results of the specificity test and no procedure of distinguishing "grammatical unacceptability" from "cognitive or ontological unacceptability" is in sight. Hence, also the combination of these two distinctions, used to differentiate between arguments and adjuncts in FGD, is not operational.

Problems with the binary AAD have been noticed in various theories, and the most common escape strategy is to posit an additional intermediate class, between arguments and adjuncts. Probably the best known example of an application of this strategy is the class of "argument adjuncts" (a-adjuncts) of Grimshaw 1990, ch. 4, encompassing optional dependents corresponding to subjects in active voice of verbs: *by*-phrases with passive verbs and certain possessive dependents of nominals. A more recent example is Needham and Toivonen 2011, which extends the intermediate class of "derived arguments" to various other types of dependents, including instruments and benefactives. The common feature of such "derived arguments" is that they seem to satisfy some of the tests for argumenthood and fail other such tests. The same approach is proposed in Panevová 2003 (see also Lopatková and Panevová 2006). An obvious problem with this strategy is that it replaces one vague distinction with two even vaguer distinctions.

Some theories have AAD hard-wired into their formalisms. This is the case with those versions of transformational theories (especially, the Government and Binding theory of 1980s; Chomsky 1981) that distinguish between arguments and adjuncts tree-configurationally (where, roughly, arguments are sisters to X heads and adjuncts are sisters to X′ projections, assuming Jackendoff's (1977) $\bar{X}$ syntax); this is also the case with Lexical Functional Grammar (Bresnan, 1982a; Dalrymple, 2001), which distinguishes argument grammatical functions from adjuncts within functional structures.[8] However, as far as we can tell, nothing within FGD seems to depend on this dichotomy. In particular, instead of distinguishing between arguments and adjuncts, all dependent types (FGD's functors) are treated uniformly in the formal FGD definitions of the *basic component* and the *tectogrammatical representation* (Sgall et al., 1986, pp. 150–153), where only the distinction between obligatory and optional dependents is implicit in the definitions of obligatory expansion rules. Also, in the discussion of the role of valency frames in Sgall et al. 1986, pp. 158–159, semantic obligatoriness and iterability are referred to separately, the notion of argument apparently being superfluous. Similarly, no contentful reference to this dichotomy is made in the discussion of the *systemic ordering* (Sgall et al., 1986, pp. 194–203); in particular the ordering proposed for Czech (pp. 198–199) has argument functors intermingled with functors typical of adjuncts. Further, no reference to arguments or adjuncts as a class is made in ch. 4 of Sgall et al. 1986, where correspondences between linguistic levels are dis-

[8] But see Przepiórkowski 2016 for a voice of dissent.

cussed; etc. So it seems that the benefit of maintaining the AAD is purely practical: when describing the potential of particular predicates to combine with various kinds of dependents, some combinations seem more idiosyncratic or perplexing than other, so some dependents (let us call them arguments) should be explicitly mentioned in the lexical entry of a given predicate, and other (let us call them adjuncts) may be assumed to be sufficiently predictable to be omitted from such lexical entries.

But it is not clear that, when carefully examined, *any* types of dependents are sufficiently regular to be excluded from the lexicon: as mentioned above, out of 3909 English verbs carefully examined by two independent annotators, some 1.8% (about 70) apparently do not combine with one of the most prototypical types of adjuncts, namely, that of event location; the existence of various exceptions of this kind is also mentioned – but played down – in FGD work on valency; some types of traditional adjuncts seem to depend on the grammatical or lexical aspect of the verb (in Slavic); etc. Hence, the current approaches to valency lexicons may be viewed as only first approximations of future full-fledged valency dictionaries containing information about all dependent types (i.e. functors): whether they are possible at all (some predicates will not combine with some types of dependents at all), whether they are semantically obligatory, to what extent they are iterable, under what conditions they may accompany the predicate, etc. Obviously, developing such a dictionary would require much more work, as all functors would have to be examined in case of each predicate, not just those that spring to mind as specific to this predicate. Let us imagine that such a dictionary exists, organised just as the Czech FGD valency dictionary PDT-Vallex but not assuming any fundamental distinction between two classes of dependents. We believe that this dictionary would still count as an FGD dictionary and that it would not violate any fundamental FGD principles. If so, FGD does not really need the ill-defined Argument–Adjunct Distinction and would be a more precise and parsimonious theory without it; after all, one of the two fundamental working principles of FGD is that (Sgall et al., 1986, p. 101):

> *the number of elementary units on the tectogrammatical level should be as small as possible, so that clear reasons can be given for every newly recognized unit or distinction.*

As argued above, such a clear reason is lacking for the Argument–Adjunct Distinction.

Acknowledgements

Many thanks to Jadwiga Linde-Usiekniewicz, Jarmila Panevová, Agnieszka Patejuk and Zdeňka Urešová for their comments on previous versions of this paper, and to Magdalena Danielewiczowa, Leonid Iomdin, Václava Kettnerová, Jadwiga Linde-Usiekniewicz, Jarmila Panevová and Alexandr Rosen for their questions and comments after the presentation of this paper at a valency conference in Warsaw in June 2016. Thanks are also due to two anonymous reviewers of PBML, whose comments have led to numerous improvements. This research is partially supported by

the Polish Ministry of Science and Higher Education within the CLARIN ERIC programme 2015–2016 (http://clarin.eu/) and by the IC 1207 COST Action PARSEME (http://www.parseme.eu/).

Bibliography

Bach, Emmon. The Algebra of Events. *Linguistics and Philosophy*, 9:5–16, 1986.

Bogusławski, Andrzej. Small is Beautiful. A Note on Verbs of Small Events. In *Tipologiceskie obosnovanija v grammatike. K 70-letiju professora V.S. Chrakovskogo*, pages 61–75. Znak, Moscow, 2004.

Bresnan, Joan, editor. *The Mental Representation of Grammatical Relations*. MIT Press Series on Cognitive Theory and Mental Representation. The MIT Press, Cambridge, MA, 1982a.

Bresnan, Joan. Polyadicity. In *The Mental Representation of Grammatical Relations* Bresnan (1982a), pages 149–172.

Chomsky, Noam. *Lectures on Government and Binding*. Foris, Dordrecht, 1981.

Dalrymple, Mary. *Lexical Functional Grammar*. Academic Press, San Diego, CA, 2001.

Dowty, David. On the Semantic Content of the Notion of 'Thematic Role'. In Chierchia, Gennero, Barbara H. Partee, and Raymond Turner, editors, *Properties, Types and Meaning: II*, pages 69–129. Kluwer, Dordrecht, 1989.

Dowty, David. Thematic Proto-roles and Argument Selection. *Language*, 67(3):547–619, 1991.

Fillmore, Charles J. Types of Lexical Information. In Kiefer, Ferenc, editor, *Studies in Syntax and Semantics*, pages 109–137. Reidel, Dordrecht, 1969.

Fillmore, Charles J. Pragmatically Controlled Zero Anaphora. In *Proceedings of the Twelfth Annual Meeting of the Berkeley Linguistics Society*, pages 95–107, Berkeley, 1986. Berkeley Linguistics Society.

Goldberg, Adele E. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago University Press, Chicago, IL, 1995.

Goldberg, Adele E. Surface Generalizations: An Alternative to Alternations. *Cognitive Linguistics*, 13(4):327–356, 2002.

Grimshaw, Jane. *Argument Structure*. Linguistic Inquiry Monographs. The MIT Press, Cambridge, MA, 1990.

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In Nivre, Joakim and Erhard Hinrichs, editors, *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Norway, 2003.

Jackendoff, Ray. $\bar{X}$ *Syntax: A Study of Phrase Structure*. The MIT Press, Cambridge, MA, 1977.

Koenig, Jean-Pierre, Gail Mauner, and Breton Bienvenue. Arguments for Adjuncts. *Cognition*, 89:67–103, 2003.

Larson, Richard. Bare NP adverbs. *Linguistic Inquiry*, 16:595–621, 1985.

Laskowski, Roman. Kategorie morfologiczne języka polskiego – charakterystyka funkcjonalna. In Grzegorczykowa, Renata, Roman Laskowski, and Henryk Wróbel, editors, *Gramatyka współczesnego języka polskiego: Morfologia*, pages 151–224. Wydawnictwo Naukowe PWN, Warsaw, 2nd edition, 1998.

Levin, Beth. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, 1993.

Lopatková, Markéta and Jarmila Panevová. Recent Developments in the Theory of Valency in the Light of the Prague Dependency Treebank. In Šimková (2006), pages 83–92.

McConnell-Ginet, Sally. Adverbs and Logical Form: A Linguistically Realistic Theory. *Language*, 58(1):144–184, 1982.

Morzycki, Marcin. *Mediated Modification: Functional Structure and the Interpretation of Modifier Position*. Ph.D. Thesis, University of Massachusetts, Amherst, 2005.

Needham, Stephanie and Ida Toivonen. Derived Arguments. In Butt, Miriam and Tracy Holloway King, editors, *The Proceedings of the LFG'11 Conference*, pages 401–421, Stanford, CA, 2011. CSLI Publications. URL `http://cslipublications.stanford.edu/LFG/16/lfg11.html`.

Panevová, Jarmila. On Verbal Frames in Functional Generative Description. Part 1. *The Prague Bulletin of Mathematical Linguistics*, 22:3–40, 1974.

Panevová, Jarmila. Inner Participants and Free Adverbials. *Prague Studies in Mathematical Linguistics*, 6:227–254, 1978.

Panevová, Jarmila. Valency Frames: Extension and Re-examination. In Chrakovskij, Viktor S., Maciej Grochowski, and Gerd Hentschel, editors, *Studies on the Syntax and Semantics of Slavonic Languages. Papers in Honour of Andrzej Bogusławski on the Occasion of his 70th Birthday*, pages 325–340. BIS, Oldenburg, 2001.

Panevová, Jarmila. Some Issues of Syntax and Semantics of Verbal Modifications. In *Proceedings of MTT 2003 – First International Conference on Meaning-Text Theory*, 2003. `http://meaningtext.net/mtt2003/proceedings/13.Panevova.pdf`.

Piernikarski, Cezar. *Typy opozycji aspektowych czasownika polskiego na tle słowiańskim*. Ossolineum, Polska Akademia Nauk, Komitet Słowianoznawstwa, Warsaw, 1969.

Przepiórkowski, Adam. How *not* to Distinguish Arguments from Adjuncts in LFG. Paper delivered at HeadLex16 (Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar), to appear in the CSLI on-line proceedings, 2016.

Przepiórkowski, Adam, Jan Hajič, Elżbieta Hajnicz, and Zdeňka Urešová. Phraseology in two Slavic Valency Dictionaries: Limitations and Perspectives. *International Journal of Lexicography*, 29, 2016. URL `http://ijl.oxfordjournals.org/content/early/2016/02/22/ijl.ecv048.abstract?keytype=ref&ijkey=jWNJn7Cxf7WJRhD`. Forthcoming.

Recanati, François. Unarticulated Constituents. *Linguistics and Philosophy*, 25:299–345, 2002.

Recanati, François. It is raining (somewhere). *Linguistics and Philosophy*, 30:123–146, 2007.

Recanati, François. *Truth-Conditional Pragmatics*. Clarendon Press, Oxford, 2010.

Sgall, Petr and Eva Hajičová. A "Functional" Generative Description. *The Prague Bulletin of Mathematical Linguistics*, 14:9–37, 1970.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, 1986.

Šimková, Mária, editor. *Insight into Slovak and Czech Corpus Linguistics*. Veda, Bratislava, 2006.

Tesnière, Lucien. *Éléments de Syntaxe Structurale*. Klincksieck, Paris, 1959.

Urešová, Zdeňka. Verbal Valency in the Prague Dependency Treebank from the Annotator's Viewpoint. In Šimková (2006), pages 93–112.

Urešová, Zdeňka. *Valenční slovník Pražského závislostního (PDT-Vallex)*. Ústav formální a aplikované lingvistiky, Prague, 2011.

Williams, Alexander. *Arguments in Syntax and Semantics*. Cambridge University Press, 2015.

Zaenen, Annie and Dick Crouch. OBLS Hobble Computations. In Butt, Miriam and Tracy Holloway King, editors, *The Proceedings of the LFG'09 Conference*, pages 644–654, Trinity College, Cambridge, UK, 2009. CSLI Publications. URL `http://cslipublications.stanford.edu/LFG/14/index.shtml`.

Žabokrtský, Zdeněk. *Valency Lexicon of Czech Verbs*. Ph.D. Thesis, Charles University, Prague, 2005.

Address for correspondence:
Adam Przepiórkowski
`adamp@ipipan.waw.pl`
Instytut Podstaw Informatyki PAN
ul. Jana Kazimierza 5
01-248 Warszawa
Poland

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 21–30

In favour of the Argument–Adjunct Distinction (from the Perspective of FGD)

Jarmila Panevová

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

In this article, the critical remarks of Adam Przepiórkowski concerning the Argument–Adjunct distinction in Functional Generative Description printed in this issue are discussed and some counterarguments are given.

1. Introduction

In his article *Against the Argument–Adjunct Distinction in Functional Generative Description* (Przepiórkowski, 2016) printed in this issue, Adam Przepiórkowski (henceforth AP) presents his critical remarks on the theory of valency used in the Functional Generative Description (henceforth FGD). Any serious discussion on this topic poses a significant challenge to the author responsible for the discussed texts to check her own formulations and their background. I was stimulated to go back to my first attempt at coping with valency issues within FGD and to compare my "historical" and contemporary views with AP's opinion. My comments here are based on the conviction that linguistic research based on some defined criteria is better than research based on intuition and ad hoc classification. We do not expect, however, that all language phenomena fit into the boxes determined by any criteria because of the vagueness belonging to the language itself, because of unclear boundaries between ontological content and language meaning and other phenomena typical for natural languages (such as ellipsis, metaphoric meanings etc.).

 Corresponding author: panevova@ufal.mff.cuni.cz
Cite as: Jarmila Panevová. In favour of the Argument–Adjunct Distinction (from the Perspective of FGD). The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 21–30. doi: 10.1515/pralin-2016-0009.

2. Argument–Adjunct Distinction in FGD

Let me start here with the last AP's objection given at the end of his article (in Sect. 6), because there AP's alternative solution is proposed:[1] He suggests that the discussed dichotomy between what he calls arguments (corresponding to the actants or inner participants and semantically obligatory adverbials in our current terminology) and adjuncts (corresponding to non-valency free adverbials or free modifications in our terminology) is not needed for FGD at all. Instead, he proposes to construct a "full-fledged lexicon" capturing all information about all possible dependents of a verb with all requirements and constraints for their application. This idea, however, has as its theoretical consequence in a shift of grammar (or at least its great part) to the lexicon. Restrictions such as incompatibility of temporal modifications with the meaning "how long" with perfective aspect of verb in Slavic languages (see e.g. AP's examples (14a), (14b), (14c) and (15a), (15b), (15c))[2] are treated in FGD by the grammatical component similarly to other types of (non-valency) constraints (e.g. incompatibility of adverb *yesterday* with the future tense of its governing verb etc.). Many valency issues concerning the grammar-lexicon interface as the relation between valency members and their alternations or deletability on the surface due to aspect, diathesis, generalization, lexicalization etc. are discussed in the papers and books on valency and lexicon in FGD (Panevová, 2014; Panevová and Ševčíková, 2014; Lopatková et al., 2008, e.g.). In addition, the lexicon proposed by AP will not satisfy the basic principle applied for the tectogrammatical level of FGD, namely to introduce as small number of elementary units as possible (quoted by AP from Sgall et al., 1986, p. 101), as it implies overloading the lexical entries with the information repeated many times with different lexical items.

AP considers the criteria used for the distinction between valency and non-valency members formulated for FGD in Panevová (1974, 1975) as ill-formed. He has in mind the criteria (1) and (2), quoted from Panevová (1974, p. 11) and slightly modified here by me as (I) and (II):

(I) Can the given modification depend on every verb? (AP calls this criterion with a contemporary term "specificity").

(II) Can the given modification be repeated with a single governor? (called "iterability" by AP).

Modifications which satisfy neither (I) nor (II) are classified as valency members in FGD and they are called inner participants (their governors need to be listed and they are not repeatable).[3] The rest of modifications which are positively classified

[1] In the continuation of my reply, I follow the structure of AP's article.

[2] (14a) *Janek tańczyl dwie godziny.*, (14c) **Janek zatańczyl dwie godziny.* For a more detailed comment see Section 4 below.

[3] The special position of Actor is taken into consideration in both discussed texts (mine and AP's as well).

according to (I) and (II) are called free adverbials and their free compatibility with any verb is assumed.

AP is right that for some types of modifications classified according to the criteria (I) and (II) as free adverbials, there are some "ontological" conditions blocking their compatibility with some semantic classes of verbs. Some types of unexpected combinations are illustrated and discussed in Panevová (1994, p. 227, examples (1)–(5) there), e.g. adverbial of purpose (AIM) connected with the verb of the change of state is illustrated by (1), the adverbial of instrument (MEANS) with unconscious event is illustrated by (2). On the other hand, the necessity for each verb to list a set of its inner participants called either Patient or Objective (or otherwise) seems to be obvious, because an omission of such dependents usually leads to ungrammaticality.

(1) John fell ill in order to be punished for his sins.

(2) He missed the target with two out of five arrows.

Though such combinations exemplified by (1) and (2) are rare, to exclude the possible combinations from the system in advance means to reduce the potentiality of the language system as a living organism.

Table 1 presented in Panevová (1978) and included by AP in Section 2 of his article followed by his critical remarks is really a bit misleading in the context of the basic criteria for the determination of valency and non-valency dependents. Actually, the table was motivated by my considerations about granularity vs. unity within the set of dependents and it was inspired by Fillmore's (1971) notion of "hypercase" and "crossed brackets" which I "played down" later.[4] Instead of the table a simpler scheme was used:

	Obligatory	Optional
Inner participant	+	+
Free adverbial	+	−

This scheme representing the structure of the verbal valency frame indicates that the class of modifications determined as inner participants always enter the valency frame as an obligatory or optional part of it (marked +),[5] while adverbials are included in the valency frame only if they were determined as (semantically) obligatory.[6]

3. Iterability

As for the criterion (II), iterability, AP is right that it is not quite easy to find natural examples of repeating or splitting some types of modifications.

[4] The same motivation is behind considerations in Panevová 1977, p. 55f and Panevová 1978, p. 243ff.

[5] According to the results of our empirical studies Actor, Patient, Addressee, Origin, and Effect belong to inner participants.

[6] Free adverbials belong to the valency frame either if the sentence is grammatically incorrect without them (**Jan směřuje* [*John aimed]) or if – in case of their absence on the surface – the "dialogue" test supports their semantic obligatoriness (see Section 5 below).

However, if I understand the author well, he admits the iterability of such modifications which refer to the same entity. He supports this view by Bresnan's and Goldberg's examples quoted by him as (3), (5) and (6) with multiple occurrence of modifications of the same type. AP demonstrates the ungrammaticality of two instruments (MEANS) referring to the different entities by Bresnan's sentence quoted as (4). Our examples (3) and (4) below[7] with two instruments (MEANS) referring to the different entities could be understood as counterexamples to Bresnan's (and AP's?) view.

(3) Jan napsal gratulaci matce poraněnou levou rukou perem.
'John wrote the congratulation to his mother by his injured left hand by the pen.'

(4) Silnici opravili pomocí těžkých strojů štěrkem a pískem.
'They repaired the road with the help of heavy machines by gravel and sand.'

In Panevová (1974, p. 15), we have presented also examples with iterated cause and iterated condition without reference to the identical entity or event.

In our approach the issue of iterability of dependents is not influenced by their identical or different reference; according to our view, such aspects are related to the layer of the cognitive content. With regard to the language structure three locatives and one apposition relation are present in the ex. (5):

(5) V Praze, hlavním městě České republiky, bydlí Jan na Vyšehradě v malém bytě.
'In Prague, the capital of the Czech Republic, John lives at Vyšehrad in a small flat.'

The idea about iterativity of free modifications and the variety of examples supporting it in the quoted papers on FGD illustrate again the power of the language as a recursive system with potential infiniteness.

AP's examples (7) (with its modifications (8), (9), (10)) and ex. (11) which should document the iterability of inner participants could be hardly interpreted as the repetition of Patient (Objective or so) and Subject (Actor), respectively. The necessity of separation of the two parts of the supposed "iteration of the inner participant" by a comma for the "second occurrence of the participant" is sufficient for a hesitation about their status. AP admits that they could be classified as appositions.[8] Moreover, one can doubt about acceptability of (11) with splitted "apposition".[9] Thus we still believe that iterability of inner participants is not possible.

[7] According to AP's opinion (in his Footnote 4) the dependent in my Russian example *na rojale* 'on piano' is a "core argument" of the verb *igrat´* 'to play' instead of Instrument as it is presented in Panevová (2003). In both Czech valency lexicons VALLEX (Lopatková et al., 2008) and PDT-Vallex (Urešová, 2011) the verbal frame of the corresponding meaning of *hrát* 'to play' has an optional accusative PAT and the dependent expressed as *na* + *Local* is characterized as a free adverbial with functor of MEANS (*hrál tu skladbu (na piano)*) 'he played this composition (on the piano)'.

[8] Apposition is a different type of syntactic relation than dependency, the core of our discussion here.

[9] See an analogy given by Eyende Van and Kim (2016): "Separating the appositions from the anchor, as in (5b), yields an ill-formed result:" (5a) Sarajevo, the capital of neighboring Bosnia, is where the World

4. Specificity

AP presents McConnell-Ginet's (1982) examples to illustrate the limitation of free compatibility of the verb *to weigh* with "many typical adverbials" which he quotes as examples (12) and (13). However, many typical or untypical adverbials modifying this verb could be given, see (6), (7):

(6) Annie weighs surprisingly/obviously 120 pounds.

(7) Annie weighed 120 pounds in the last year/for the last 5 years/to her mother's disappointment/at hospital/in Detroit.

As for the position of aspectual pairs in Slavonic languages, this issue has been discussed in linguistics for several last decades. In FGD the aspectual pairs are understood as a single lexeme (represented by different verb lemmas distinguished with a different morphological features), see Panevová et al. (1971, p. 28ff), Sgall et al. (1986, p. 165), Lopatková et al. (2008, p. 14). The choice of the appropriate form of the temporal modifier with regard to the aspectual meaning of the verb is treated in the grammatical component of FGD. It concerns the compatibility of the verb with the functors "how long" and "for how long",[10] as well as the possibility to omit a Patient (Objective) with the imperfective aspect of particular classes of verbs and its obligatory occurrence with the perfective aspect (e.g. *psát/napsat* [to write], *číst/přečíst* [to read], *počítat/spočítat* [to count] etc.).[11]

5. Semantic obligatoriness

AP puts under scrutiny the "dialogue" test used to determine whether a particular modification is (semantically) obligatory for a tested lexical item. This test was proposed as a tool for the decision about whether a modification absent in the surface form of a grammatically correct sentence is obligatory or not from the point of view of the deep (tectogrammatical) structure. According to the opinion of the users of this test (e.g. the authors of valency dictionaries,[12] annotators etc.), some doubts about acceptability of the answer "I don't know" occur especially in cases where the tested modification could be "generalized" in the given context.[13]

War began.
(5b) *Sarajevo is where the World War began, the capital of neighboring Bosnia.

[10] These two meanings are understood as contextual variants of a single functor in Panevová et al. (1971, p. 75). In the Prague Dependency Treebank, the solution to introduce them as two different functors was applied, see Mikulová et al. (2006, p. 484).

[11] Czech and Polish probably differ as to possibilities of the omission of object: the presence of object is required in Czech counterparts of AP's ex. (14b) and (14c).

[12] Lopatková et al. (2008), Urešová (2011).

[13] The notion of generalization of participants is analyzed esp. in Panevová (2001, 2004). See below, e.g. ex. (9).

AP is right that in this sense the dialogue test has common features with Fillmore's approach and with his examples (quoted by AP in his article as (16), (17), (18)) illustrating the presence of the "indefinite implicit argument". I have discussed Fillmore's examples recalled by AP in this context in Panevová (1994, ex. (14)–(18), my example (14) is repeated here as (8)).

(8) a. John bought a dozen roses.[14]
b. John paid Harry five dollars.

Fillmore (1977, Section 5) considers two possible solutions:

> "one typical way of dealing with conceptually obligatory but superficially optional elements in a sentence is to claim that these elements are present in the deep structure but deleted or given zero representation on the surface structure",

but he prefers the solution

> "to say that a word like *buy* or *pay* activates the commercial event"

[containing money and buyer articulated in (8)b. – JP], however

> "it may not be necessary to believe that everything that is included in our understanding of the sentence *[the whole commercial scene in this case – JP]* is necessarily a part of underlying grammatical structure of the sentence".

The users of our dialogue test face the same problems formulated in the quoted part of Fillmore's paper: What is the "complete scene" and which part of it must be reflected in the underlying sentence structure (and therefore included in the valency frame).

In Panevová (2001, Section 2) and in Panevová (2014, Section 4), several experimental dialogues were construed in order to use the dialogue test in an attempt to determine the boundaries of "generalized" semantically obligatory valency members (on the one hand) and semantically optional participants and non-valency free adverbials (on the other hand). Some of them are illustrated by (9), (10) and (11) below:

(9) a. Sue sells at Bata store.
b. What does she sell?
c. To whom does she sell?

The answer "I don't know" given by the speaker of (9a) after (9b) and (9c) is not appropriate, though he/she can list neither the sold objects nor the set of buyers. A more appropriate response to these odd questions would sound like (9d) and (9e), respectively; it indicates the obligatory but generalized valency slots (Patient and Addressee) for the verb *to sell* with this lexical meaning. Further difficulties for the testing are caused by the polysemy of lexical units. In (9), we have to do with the meaning "to

[14] Ex. (14) in Panevová (1994) repeated here as (8) corresponds to Fillmore's (1977) ex. (12a), quoted now from the reprinted version 2003, p.192). John is used instead of I as the identity of the speaker with the Actor is not suitable for testing by the proposed dialogue test.

be a shop-assistant", while in (10) a usual transfer action of an object is presupposed. In the latter meaning the Addressee (*buyer*) seems to be an optional participant:

(9) d. She sells goods *typical* for Bata stores (shoes, bags etc.)
e. She sells to the *typical* customers of the shop.

(10) a. John finally got rid of his old car, he sold it.
b. To whom?
c. I don't know (who the buyer was).

In ex. (11a) there is a complete scene for the verb *to speak*, Addressee and Patient are expressed. However, the question (11d) addressing (11b) could be naturally answered "I don't know", it indicates that for the Czech verb *mluvit* [*to speak*] the Patient is an optional rather than an obligatory participant, while the question (11e) addressing (11d) hardly could be answered "I don't know", but rather "with the listeners present there", which is the formulation typical of generalized members.

(11) a. John spoke with his teacher about the good-bye party.
b. I saw John in the corridor as he spoke with his teacher.
c. John spoke in the corridor about his leaving for the USA.
d. What did they speak about?
e. With whom did John speak?

These dialogues, if they are accepted as well-articulated dialogue tests, support a valency frame with an obligatory Addressee and an optional Patient for the verb *mluvit* [to speak] with the given meaning (corresponding to the meaning 1 in VALLEX).

As to the verb *to rain*, I can imagine a simpler context for AP's ex. (23) quoting Racanati's considerations:

(12) A: There will be a lot of mushrooms this year, because it was raining.
B: Where was it raining?
A: I don't know (exactly), everywhere.[15]

AP is right that the idea to use the dialogue test for identifying valency was inspired by the article Sgall and Hajičová (1970) where a set of possible questions related to the sentence tested by them *He wrote all these books.* is proposed. Actually, this set of questions was presented there in order to test relationships between form, meaning and content. However, one of the questions formulated there was for me more stimulating for the application to Slavonic languages exhibting the pro-drop character: The Czech example (13) justifies more transparently the application of the dialogue test for the purposes of testing obligatoriness. The odd feeling of the dialogue in (13) is obvious:

[15] In both valency dictionaries (VALLEX and PDT-Vallex) the verb *pršet* [to rain] has an empty frame (i.e. it has no valency dependents). I am not sure whether the reply "everywhere" in dialogue test is equivalent for the notion of generalized complement rather than for a reply "I don't know". However, the question B in (12) sounds really odd, if B does not ask for adding of specific information (which is excluded from the dialogue test, see Panevová (1978, p. 229)).

(13) a. Včera už Marii knihu vrátil.
[Yesterday *(he)* already gave the book back to Mary.]
b. Kdo ji Marii/jí vrátil? [Who did give it back?]
c. *Nevím [* I don't know.]

During the long period of its application the dialogue test, described in many papers and used for compilation of valency dictionaries, appeared to work well for most cases. In problematic cases the result depends on particular users – as Fillmore says – on his/her particular knowledge of the scene.

6. Does FGD need AAD?

According to AP the dichotomy valency vs non-valency dependents does not play any role in the formalism of FGD with a single exception included in the definitions of obligatory expansion rules during the sentence generation. However, the application of these rules for a particular verb is one of the crucial prerequisites to generating grammatically correct outputs.[16] AP correctly states that "no contentful reference to this dichotomy *(i.e. argument - adjunct, JP)*" is made in the discussion of systemic ordering relevant to the word order of verbal dependents in (Sgall et al., 1986, Chapter 3, pp. 198–199). However, this fact cannot serve as a counterargument to the given dichotomy. It means that the dichotomy simply concerns an aspect of the sentence structure independent of communicative function. In the specification of the systemic ordering the labels of the functors are used as notions defined in the Chapter 2 of the quoted monograph.

Summarizing the approach used for FGD I believe that:

- By the application of criteria (I) and (II) from Section 2, we have provided a classification of the proposed list of semantic units (functors) into two classes: inner participants and free adverbials.
- The dialogue test proposed as a tool for constituting valency frames in the cases of surface absence of a position pretending to be included in the valency frame makes it possible to distinguish between semantically obligatory and optional modifications. The issue (I) reflects a paradigmatic dimension for the list of possible modifications, while with (II) a syntagmatic dimension (verb + its dependents) is taken into account.

We have tried to present here a realistic view on the criteria (I) and (II) and on the dialogue test, admitting some empirical difficulties connected with different pragmatic attitudes of specific speakers. However, first of all we want to defend the necessity of the argument (valency dependent) and adjunct (non-valency dependent)

[16] Thus the expansion of the verb by an obligatory PAT is stated once in the basic component generating tectogrammatical representations, the lexical data (which verbs are concerned) being extracted from the lexicon.

dichotomy for the theoretical framework based on the cooperation of two modules: lexicon and grammar.

Acknowledgements

The work on this paper was supported by the project GA ČR 16-02196S "Valenční slovník českých substantiv založený na korpusu".

Bibliography

Eyende Van, F. and J. B. Kim. Loose Apposition. *Functions of Languge*, 23(1):17–39, 2016.

Fillmore, C. J. Some Problems for Case Grammar. In O'Brien, R. J., editor, *22nd annual round table. Linguistics: Development of the sixties – viewpoints of the seventies*, volume 24, pages 35–56. Georgetown University Press, Washington, D. C., 1971.

Fillmore, C. J. The Case for Case Reopened. In Cole, P. and J. M. Sadick, editors, *Syntax and Semantics, Grammatical Relations 8*, pages 59–81. Academic Press, New York – San Francisco – London, 1977.

Lopatková, Markéta, Zdeněk Žabokrtský, and Václava Kettnerová. *Valenční slovník českých sloves*. Nakladatelství Karolinum, Praha, 2008. ISBN 978-80-246-1467-0.

McConnell-Ginet, Sally. Adverbs and Logical Form: A Linguistically Realistic Theory. *Language*, 58(1):144–184, 1982.

Mikulová, Marie, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 30, Prague, Czech Rep., 2006.

Panevová, Jarmila. Verbal frames in Functional Generative Description. Part 1. *The Prague Bulletin of Mathematical Linguistics*, 22:3–40, 1974.

Panevová, Jarmila. Verbal frames in Functional Generative Description. Part 2. *The Prague Bulletin of Mathematical Linguistics*, 23:17–52, 1975.

Panevová, Jarmila. Verbal Frames Revisited. *The Prague Bulletin of Mathematical Linguistics*, 28: 55–72, 1977.

Panevová, Jarmila. Inner Participants and Free Adverbials. *Prague Studies in Mathematical Linguistics*, 6:227–254, 1978.

Panevová, Jarmila. Valency Frames and the Meaning of the Sentence. In Luelsdorff, Ph. L., editor, *The Prague School of Structural and Functional Linguistics*, pages 223–243. Benjamins Publ. House, Amsterdam-Philadelphia, 1994.

Panevová, Jarmila. Valency Frames: Extension and Re-examination. In Chrakovskij, Viktor S., Maciej Grochowski, and Gerd Hentschel, editors, *Studies on the Syntax and Semantics of Slavonic Languages. Papers in Honour of Andrzej Bogusławski on the Occasion of his 70th Birthday*, pages 325–340. BIS, Oldenburg, 2001.

Panevová, Jarmila. Some Issues of Syntax and Semantics of Verbal Modifications. In *Proceedings of MTT 2003 – First International Conference on Meaning-Text Theory*, 2003. http://meaningtext.net/mtt2003/proceedings/13.Panevova.pdf.

Panevová, Jarmila. Všeobecné aktanty očima Pražského závislostního korpusu (PZK). In *Korpus jako zdroj dat o češtině. Sborník konference ve Šlapanicích*, 2004.

Panevová, Jarmila. *Contribution of Valency to the Analysis of Language*, chapter 1, pages 1–18. Studies in Language Companion Series, 158. John Benjamins Publishing Company, Amsterdam, The Netherlands, 2014. ISBN 9789027259233.

Panevová, Jarmila and Magda Ševčíková. *Delimitation of information between grammatical rules and lexicon*, volume 215 of *Linguistik Aktuell / Linguistics Today*, pages 33–52. John Benjamins Publishing Company, Amsterdam, The Netherland, 2014. ISBN 978-90-272-5598-3.

Panevová, Jarmila, Eva Benešová, and Petr Sgall. *Čas a modalita v češtině*. Univ. Karlova, 1971.

Przepiórkowski, Adam. Against the Argument–Adjunct Distinction in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*, 106:5–20, 2016.

Sgall, Petr and Eva Hajičová. A "Functional" Generative Description. *The Prague Bulletin of Mathematical Linguistics*, 14:9–37, 1970.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, 1986.

Urešová, Zdeňka. *Valenční slovník Pražského závislostního (PDT-Vallex)*. Ústav formální a aplikované lingvistiky, Prague, 2011.

Address for correspondence:
Jarmila Panevová
panevova@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 31–44

Predicting the Performance of Parsing with Referential Translation Machines

Ergun Biçici

ADAPT Research Center, School of Computing, Dublin City University, Ireland

Abstract

Referential translation machine (RTM) is a prediction engine used for predicting the performance of natural language processing tasks including parsing, machine translation, and semantic similarity pioneering language, task, and domain independence. RTM results for predicting the performance of parsing (PPP) in out-of-domain or in-domain settings with different training sets and types of features present results independent of language or parser. RTM PPP models can be used without parsing using only text input and without any parser or language dependent information. Our results detail prediction performance, top selected features, and lower bound on the prediction error of PPP.

1. Predicting Parsing Performance with Referential Translation Machines

Training parsers and parsing can be computationally costly and labeled data scarce or expensive to obtain. Predicting the performance of parsing (PPP) can be useful for parsing technology, for filtering sentences in noisy domains such as informal text or speech, for estimating the effort for understanding text, for determining whether a sentence is well-formed and meaningful enough to send to other natural language processing (NLP) tasks such as machine translation in an NLP pipeline. PPP involves finding a function f:

$$\mathtt{f}(\mathcal{M}_\mathtt{P}, \mathcal{D}_{\text{train}}, \mathtt{S}\ [, \mathtt{S_P}']) \quad \approx \quad \mathtt{eval}(\mathtt{S_P}', \mathtt{S_P}) \tag{1}$$

where

- $\mathcal{M}_\mathtt{P}$ is a parsing model built using $\mathcal{D}_{\text{train}}$ for training,
- $\mathcal{D}_{\text{train}}$ is the set of training sentences and $\mathcal{D}_{\text{test}}$ is test data,

 Corresponding author: ergun.bicici@yahoo.com
Cite as: Ergun Biçici. Predicting the Performance of Parsing with Referential Translation Machines. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 31–44. doi: 10.1515/pralin-2016-0010.

• $S_P{}'$ refers to parsing output obtained on $S \in \mathcal{D}_{\text{test}}$ and its reference is S_P,
• `eval` returns the bracketing F_1 score by EVALB (Sekine and Collins, 1997) implementing the PARSEVAL F_1 measure,
• the performance of $\mathcal{M}_P$, which use $\mathcal{D}_{\text{train}}$, is being predicted for input S,
• `f` predicts the value of the `eval` function to approximate the performance without the reference S_P given a training set and a test set not necessarily after training a parsing model or parsing.

Ravi et al. (2008) predict the performance of Charniak and Johnson (CJ) parser (Charniak and Johnson, 2005) using text-based and parser-based features, and additional parser output (Bikel parser (Bikel, 2002)). Additional parser output is used as a reference to obtain a feature with bracketing F_1 score. In Section 3.3, we achieve better results using only textual features and obtain similar results without any parser or label dependent information or without an additional parser or its output.

Each referential translation machine (RTM) (Biçici and Way, 2015) model is a data translation prediction model between the instances in the training set and the test set, and translation acts are indicators of the data transformation and translation. RTM effectively judges monolingual and bilingual similarity while identifying translation acts between any two data sets with respect to a reference corpus. RTM allows development of prediction models specially prepared for a given training and test set pair. RTM PPP models are built for each task emerging from training set, test set, and label set obtained from a parser. RTMs achieve top results in machine translation performance prediction (MTPP) in quality estimation task (Biçici et al., 2015b; Biçici, 2016), can achieve better results than open-source MTPP tool QuEst (Shah et al., 2013; Biçici and Specia, 2015), and can achieve top results in semantic similarity prediction tasks (Biçici and Way, 2015). We provide a current picture on PPP detailing prediction performance, top selected features, and lower bound on prediction error of PPP.

RTMs judge the quality or the semantic similarity of texts by using relevant retrieved data close to the task instances as interpretants, selected preferably from the same domain. RTM PPP use parallel and monolingual sentences as interpretants, which provide context and data for MTPP system (MTPPS) (Biçici and Way, 2015) to derive features measuring the closeness of the test sentences to the training data, the difficulty of translating them, and the presence of the acts of translation for building prediction models. RTMs present an accurate and language-independent model for NLP performance prediction and provide a parser-independent model, which enables the prediction of the performance of any parser in any language. Figure 1 depicts the workflow for a general RTM model and explains the model building process. Given a training set `train`, a test set `test`, and some corpus $\mathcal{C}$, preferably in the same domain, the RTM steps are:

1. $\text{select}(\texttt{train}, \texttt{test}, \mathcal{C}) \rightarrow \mathcal{I}$
2. $\text{MTPP}(\mathcal{I}, \texttt{train}) \rightarrow \mathcal{F}_{\texttt{train}}$
3. $\text{MTPP}(\mathcal{I}, \texttt{test}) \rightarrow \mathcal{F}_{\texttt{test}}$
4. $\text{learn}(M, \mathcal{F}_{\texttt{train}}) \rightarrow \mathcal{M}$
5. $\text{predict}(\mathcal{M}, \mathcal{F}_{\texttt{test}}) \rightarrow \hat{y}$

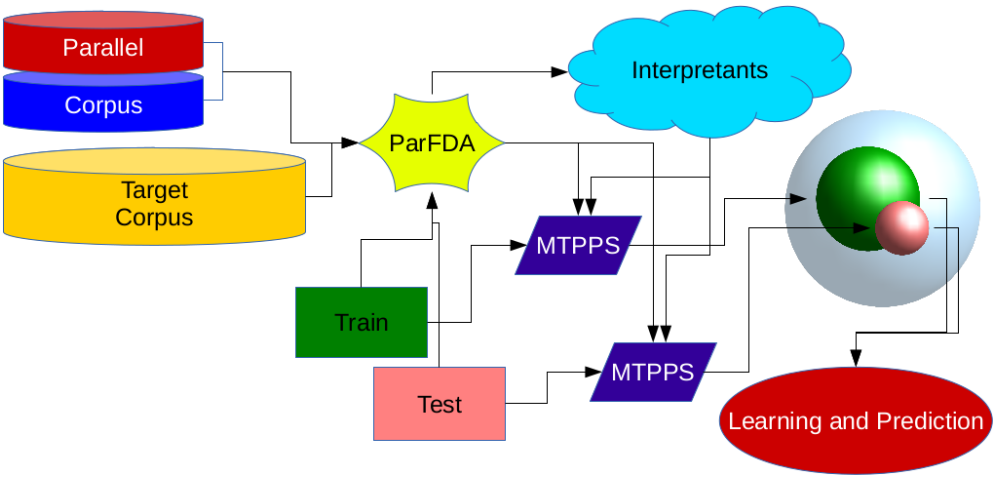


Figure 1. RTM workflow: ParFDA selects interpretants close to the training and test data using parallel corpus in bilingual settings and monolingual corpus in the target language or just the monolingual target corpus in monolingual settings; an MTPPS use interpretants and training data to generate training features and another use interpretants and test data to generate test features in the same feature space; learning and prediction takes place taking these features as input.

RTM PPP models use MTPPS to generate features and parallel feature decay algorithms (ParFDA) (Biçici et al., 2015a) for instance selection. The modularity of RTM enables additional knowledge sources to be retrieved by ParFDA, which can be used for deriving additional features to be included before learning and prediction.

2. Statistical Lower Bound on Prediction Error

We evaluate the prediction performance with correlation (r), root mean squared error (RMSE), mean absolute error (MAE), and relative absolute error (RAE). Given that $\hat{\mathbf{y}}, \mathbf{y} \in \mathbb{R}^n$ are the prediction of F_1 and the target F_1 respectively:

$$\mathrm{MAE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \qquad \mathrm{RAE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{\sum_{i=1}^{n} |\bar{y} - y_i|} \tag{2}$$

	WSJ24		WSJ02-21	
n	1346		6960	
μ	0.7095		0.7145	
s	0.1636		0.1633	
d	0.0087		0.0038	
RAE	$\hat{d}$	$\hat{n}$	$\hat{d}$	$\hat{n}$
1%	0.0013	57335	0.0013	58164
5%	0.0067	2296	0.0066	2329
10%	0.0134	576	0.0133	584
20%	0.0268	146	0.0265	148
30%	0.0402	66	0.0398	67
40%	0.0536	38	0.0531	39
50%	0.0670	25	0.0664	26
75%	0.1004	13	0.0995	13
80%	0.1071	12	0.1062	12
85%	0.1138	11	0.1128	11

Table 1. Estimated d *and* $\hat{d}$ *and* $\hat{n}$ *required for the noise levels based on RAE for PPP with bracketing* F_1.

We also use relative MAE (MAER) and RAE (MRAER) (Equation (3)) (Biçici and Way, 2015). We use MAER and MRAER for easier replication and comparability with relative errors for each instance. Evaluation with MRAER can help identify which tasks and subtasks require more work by design and RTM PPP results reaching 0.75 MRAER in Section 3 are in line with performance in semantic textual similarity in English and easier than MTPP (Biçici and Way, 2015). MAE treats errors equally whereas RMSE is giving more weight to larger errors and can become dominated by the largest error. Therefore, MAE and RAE and their relative versions MAER and MRAER are better metrics to evaluate the performance.

$$\mathrm{MAER}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\sum_{i=1}^{n} \frac{|\hat{y_i} - y_i|}{\lfloor |y_i| \rfloor_\epsilon}}{n} \qquad \mathrm{MRAER}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\sum_{i=1}^{n} \frac{|\hat{y_i} - y_i|}{\lfloor |\bar{y} - y_i| \rfloor_\epsilon}}{n} \tag{3}$$

We obtain expected lower bound on the prediction performance and the number of instances needed given a RAE level. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ represent the target sampled from a distribution with mean μ and standard deviation σ, then the variance of $\sum_{i=1}^{n} y_i$ is $n\sigma^2$ and of the sample mean, $\bar{y}$, is $\frac{\sigma^2}{n}$ with the standard deviation becoming $\frac{\sigma}{\sqrt{n}}$. From a statistical perspective, we can predict the number of training instances we need for learning to increase the signal to noise ratio, $\mathrm{SNR} = \frac{\mu}{\sigma}$, or the ratio of the mean to the standard deviation. Increasing the number of instances leads

to decrease in the noise and increase SNR. We want to find a confidence interval, $[\bar{y} - t\frac{s}{\sqrt{n}}, \bar{y} + t\frac{s}{\sqrt{n}}]$, where t is found by the Student's t-distribution for $n-1$ degrees of freedom with confidence level α and s is sample standard deviation. True score lies in the interval with probability $1-\alpha$:[1]

$$P(\bar{y} - t\frac{s}{\sqrt{n}} \leq \mu \leq \bar{y} + t\frac{s}{\sqrt{n}}) = 1 - \alpha. \quad (4)$$

The absolute distance to the true mean or the width of the interval, d, is empirically equal to MAE and the relationship between RAE and MAE is as follows:

$$RAE = \frac{nMAE}{\sum_{i=1}^{n} |\bar{y} - y_i|} \quad (5)$$

$$d = \frac{ts}{\sqrt{n}} \Rightarrow n = \frac{t^2 s^2}{d^2} \quad (6)$$

Using Equation 5, we can derive the MAE or $\hat{d}$ for a given RAE as an estimate of d. With $\alpha = 0.05$ and $p = 0.95$, we confidently estimate $\hat{d}$ and the corresponding $\hat{n}$ to reach the required noise level for the prediction tasks given a possible RAE level using Equation 6. Statistical lower bound on PPP error lists how many training instances to use for PPP (Table 1).

Table 1 presents the d possible for the bracketing F_1 score distribution and the training set sizes required for reaching a specified noise level based on RAE. We achieved top results in MTPP using RTMs (Biçici et al., 2015b) with a RAE level of 0.84 when predicting HTER, which is a score in the range $[0, 1]$. We also achieved good results in MTPP with RTMs as Biçici (2016) presents with a RAE level of 0.82 when predicting HTER.

Table 4 from Section 3.3 presents similar RAE levels in in-domain PPP and with only 12 labeled instances for PPP, we can reach the top prediction performance, which achieves 0.84 RAE. Figure 2 samples from normal n-gram F_1 (Biçici, 2011) distributions with $\mu = 0.2316$ from MTPPDAT (Biçici, 2014) for different σ and shows that prediction error decrease by: (i) increasing n; (ii) decreasing s.[2]

3. Experiments

We use the Wall Street Journal (WSJ) and Brown corpora distributed with Penn Treebank version 3 (Marcus et al., 1993, 1999). WSJ02-21 refers to WSJ sections in range 2–21, WSJ24 refers to section 24, WSJ23 refer to section 23, and WSJ0-1-22-24 refer to

[1] This forms the basis for many statistical significance tests in machine translation (Biçici, 2011).

[2] MTPPDAT contains document and sentence translation experiments collected from 4 different settings: tuning, no tuning, multiple perspective learning, and adaptation (Biçici, 2015).

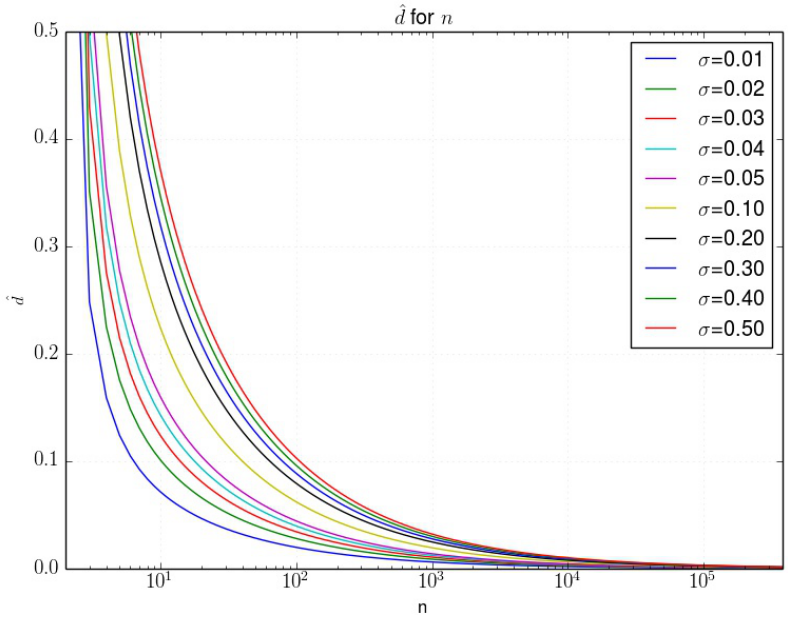


Figure 2. $\hat{d}$ *decrease by: (i) increasing* n*; (ii) decreasing* s*.*

sections 00, 01, 22, and 24 combined. BTest refers to the test set formed by selecting every 10th sentence from the Brown corpus (Ravi et al., 2008). WSJ02-21 contains 39832 sentences in total and WSJ0-1-22-24 contains 6960 sentences. We obtain the raw format for the Penn Treebank starting from the parse annotated sentences.

3.1. Parsers

CCL: CCL (Seginer, 2007) is an unsupervised parsing algorithm, which allows equivalent classes with reciprocal links between words (link structures).

PCFG: Plain PCFG (probabilistic context free grammar) parser uses the Stanford supervised parser (Klein and Manning, 2003). PCFG model is unlexicalized; it has context-free rules conditioned on only the parent nodes; it does not have language dependent heuristics for unknown word processing; and it selects the left-most category as the head of the right hand side of a rule.

CJ: Charniak and Johnson (Charniak and Johnson, 2005) develop a parser achieving the highest performance by reranking 50 best parses with a maximum entropy reranker.

		Test	# sents	CCL	PCFG	CJ
PPP	train	WSJ0-1-22-24	6960	0.5508	0.6978	0.9139
		WSJ24	1346	0.5489	0.6915	0.9057
	test	WSJ23	2416	0.5501	0.6933	0.9141
		BTest	2425	0.5646	0.6773	0.8561

Table 2. Baseline performance in terms of bracketing F_1.

Corpus	numB	depthB	avg depthB	R/L	avg R/L
WSJ02-21	46.4	11.1	0.2678	6.46	6.68
WSJ23	45.6	11.0	0.2728	6.36	6.66
$S_{P_{CCL}}{}'$	38.6	9.3	0.2829	6.14	6.14
$S_{P_{PCFG}}{}'$	41.6	10.0	0.2735	6.11	5.72
$S_{P_{CJ}}{}'$	42.6	11.0	0.2887	5.96	6.27
BTest	38.1	9.6	0.3060	6.09	5.50
$S_{P_{CCL}}{}'$	31.8	8.8	0.3551	6.77	6.04
$S_{P_{PCFG}}{}'$	35.1	9.1	0.3165	7.05	5.25
$S_{P_{CJ}}{}'$	35.6	9.7	0.3248	6.63	5.50

Table 3. Tree structure statistics: number of brackets (numB), depth (depthB), average depth per node (avg depthB), numB on the right branches over the numB on the left (R/L), and average right to left branching over all internal tree nodes (avg R/L).

All parsers use WSJ02-21 for training and Table 2 lists the baseline performances of the parsers in terms of bracketing F_1 over all sentences in the test sets along with the number of sentences in each.[3]

3.2. Features and Learning Settings

We use WSJ24 or WSJ0-1-22-24 and WMT datasets (Bojar et al., 2015) and LDC English Gigaword (Parker et al., 2011) for building RTM PPP models. We use features from three categories where detailed feature descriptions can be found in (Biçici and Way, 2015): (i) Textual features (Text), which contain coverage and diversity features

[3]The number of instances are the same as in (Bacchiani et al., 2006) and in (Kummerfeld et al., 2012) for WSJ23. The number of sentences reported in (Ravi et al., 2008) are lower. CCL lowercases input text and outputs lowercased trees; hence its performance is independent of casing. The output CCL tree is composed of text without labels and to be able to use the EVALB bracketing scores, we label each node with 'NP' and enclose them with brackets. We could use any tag instead of NP since we are not calculating tag accuracy.

about how well test features are found in the training set, language model features, distributional similarity features, translation features, information retrieval related features, character n-grams, and sentence length related features; (ii) link structure based (+CCL), which contain Text features over CCL from CCL parser, which can be used in all learning settings since CCL is unsupervised; (iii) tree structure based (+Tree) features, which contain the number of brackets used (numB), depth (depthB), average depth per node (avg depthB), number of brackets on the right branches over the number of brackets on the left (R/L),[4] and average right to left branching over all internal tree nodes (avg R/L).

We select up to 100 features from the most frequent parse tree structures and add 10 base tree statistical features for source and target. This feature set is called TreeF in (Biçici and Way, 2015). Parse tree branching statistics for WSJ2-21, WSJ23, and BTest together with the parser outputs obtained with different parsers are in Table 3. CCL output parse trees tend to have fewer branches and less depth. However, CCL outputs trees with closer R/L and avg R/L to the test set than PCFG. CJ outputs trees with closest numB and depthB to the test sets. PCFG achieves the closest avg depthB. Table 3 indicates that right branching dominates English. We observe that CCL's performance slightly increases on BTest whereas supervised parsers perform worse.

We present RTM PPP model results for in-domain (WSJ23) and out-of-domain (BTest) test sets in three different feature settings (Text, Text+CCL, Text+CCL+Tree). For each combination of training set, test set, and training and test labels obtained from a parser, we build an RTM model; thus the total number of RTM models we build is 12. Training set is used for optimizing parameters of the predictor with k-fold cross validation. The learning model is selected based on the performance on the training set and it is either bayesian ridge regression (BR) (Tan et al., 2015) or support vector regression (SVR) after feature selection (FS), partial least squares (PLS), or PLS after FS (Biçici et al., 2015b).

3.3. In-domain Results

In-domain PPP results are in Table 4 where dim is the actual number of features used for each row (e.g. after removing non-informative features, after FS, after PLS). Using more training data improves the performance and we need only 15 feature dimensions for reaching top MRAER performance with SVR model with FS+PLS in setting Text. Previous work (Ravi et al., 2008) obtains 0.42 for r and 0.098 for RMSE when predicting the performance of CJ on in-domain PPP. We obtain lower r and close RMSE values however, we do not use any parser or label dependent information or a top performing reference parser whose performance is close to CJ's. Ravi et al. (Ravi et al., 2008) also do not present separate results with the feature sets they use. The top

[4] For nodes with uneven number of children, the nodes in the odd child contribute to the right branches.

Train	Setting	Parser	Model	dim	r	RMSE	MAE	RAE	MAER	MRAER
WSJ24	Text	CCL	SVR	305	0.47	0.135	0.1074	0.87	0.226	0.83
	Text	PCFG	FS+PLS-BR	5	0.31	0.162	0.1265	0.95	0.275	0.88
	Text	CJ	FS-SVR	16	0.26	0.104	0.0699	0.88	0.107	0.78
	Text+CCL	CCL	FS-BR	16	0.47	0.135	0.1084	0.88	0.223	0.84
	Text+CCL	PCFG	SVR	331	0.3	0.163	0.1241	0.93	0.292	**0.85**
	Text+CCL	CJ	FS-SVR	16	**0.27**	0.104	0.0698	0.88	0.107	0.78
	Text+CCL+Tree	CCL	SVR	384	0.47	0.135	0.1071	0.87	0.225	0.83
	Text+CCL+Tree	PCFG	FS+PLS-SVR	15	0.26	0.17	0.1295	0.97	0.291	0.95
	Text+CCL+Tree	CJ	SVR	386	**0.27**	**0.103**	0.0699	0.88	0.107	0.78
WSJ0-1-22-24	Text	CCL	SVR	310	0.49	0.133	0.1052	0.85	0.221	**0.82**
	Text	PCFG	SVR	310	0.37	0.16	0.1224	**0.91**	0.272	0.88
	Text	**CJ**	**FS+PLS-SVR**	**15**	0.25	0.108	**0.0675**	**0.85**	0.106	**0.75**
	Text+CCL	CCL	SVR	336	0.49	0.133	0.1052	0.85	0.221	**0.82**
	Text+CCL	PCFG	SVR	336	0.37	0.16	0.1222	**0.91**	**0.271**	0.87
	Text+CCL	CJ	PLS-SVR	90	0.26	0.107	0.0678	**0.85**	0.106	**0.75**
	Text+CCL+Tree	**CCL**	**SVR**	387	**0.5**	**0.132**	**0.1041**	**0.84**	**0.219**	**0.82**
	Text+CCL+Tree	**PCFG**	**FS-SVR**	248	**0.38**	**0.159**	**0.122**	**0.91**	**0.271**	0.87
	Text+CCL+Tree	CJ	PLS-SVR	80	**0.27**	0.106	0.0677	**0.85**	**0.105**	**0.75**

*Table 4. RTM top predictor results with in-domain test set WSJ23. Using more training data improves the performance. Text reach top MRAER performance with only 15 dimensions. Best result for each metric is in **bold**.*

r they obtain with their text-based features is 0.19, which is lower than our results in setting Text.

A high RAE indicates that PPP is hard and currently, we can only reduce the error with respect to knowing and predicting the mean by about 16%. CJ parsing output is the easiest to predict as we see from the MRAER results. The MAE we achieve for PPP of CJ is 0.0675 and it is about 7.4% of the 0.9141 overall F_1 score for CJ on WSJ23. This error percentage is 17.6% and 18.9% for PCFG and CCL respectively. Figure 3 lists plots about the top RTM PPP predictor's performance in terms of absolute error and absolute error relative to the magnitude of the target in WSJ23 where instances are sorted according to the magnitude of the target F_1 scores.

3.4. Out-of-domain Results

Out-of-domain parsing decreases the performance of supervised parsers (Table 2) but not the the performance of CCL, which is unsupervised, since it uses limited domain dependent information and CCL's performance is actually slightly increased. RTM results in out-of-domain PPP are lower than in in-domain (Table 5). Adding Tree features in out-of-domain improves the performance more compared with the improvement in in-domain. Previous work (Ravi et al., 2008) obtains 0.129 RMSE for

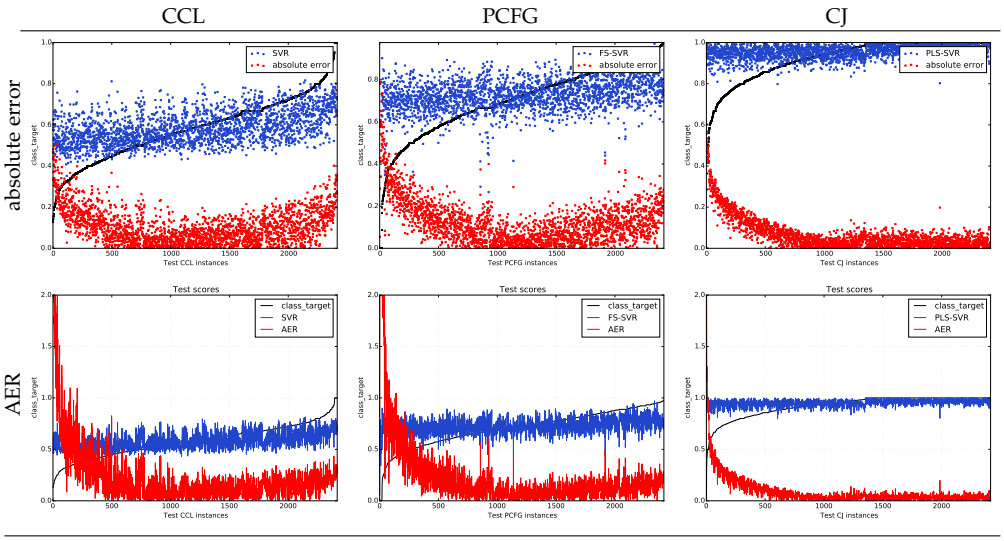


Figure 3. RTM top predictor performance using WSJ0-1-22-24 training set, WSJ23 test set, and Text+CCL+Tree setting. At the top are plots with prediction vs. the absolute error (distribution below) and at the bottom are plots with prediction vs. AER or absolute error relative to the magnitude of the target (distribution below).

CJ in out-of-domain PPP. RTM obtains about 36% larger RMSE but without using an additional parser output or parser specific features. We also note that the number of sentences reported in (Ravi et al., 2008) for datasets WSJ23, WSJ24, and BTest is less than the official datasets released as part of Penn Treebank (Marcus et al., 1993). RTM for CJ achieves better MRAER than top sentence MTPP with 0.84 MRAER (Biçici et al., 2015b). Figure 4 lists plots from the top RTM predictor's performance in BTest.

3.5. Feature Selection Results

We select features with recursive feature elimination (RFE) (Guyon et al., 2002; Pedregosa et al., 2011), which iteratively removes least informative features according to their weights provided by a learning model and this removal process provides their ranking. We use the following abbreviations: GM is the geometric mean between the precision and recall and T is used for target; $\langle P(T|S), b1 \rangle$ is the backward 1-gram log probability of the translation probability of target translation T given source sentence S and $\langle P(S,T), 2, 5 \rangle$ is the average joint logprob of the joint translation probability over 2-grams among top 5 selected instances; avgD20 is a relative entropy distance measure over the top 20 instances; and bpw is the bits per word. We observe that translation

Train	Setting	Parser	Model	dim	r	RMSE	MAE	RAE	MAER	MRAER
WSJ24	Text	CCL	SVR	305	0.45	0.144	0.1153	0.91	0.221	**0.9**
	Text	PCFG	FS+PLS-BR	8	0.25	0.182	0.1414	0.95	0.342	0.87
	Text	CJ	SVR	305	0.23	0.168	0.1043	0.87	0.244	0.77
	Text+CCL	CCL	FS-SVR	16	0.44	0.145	0.1161	0.91	0.223	0.92
	Text+CCL	PCFG	FS+PLS-BR	7	0.31	**0.177**	0.1388	0.94	**0.329**	0.87
	Text+CCL	CJ	FS+PLS-SVR	**3**	**0.25**	**0.167**	**0.1031**	**0.86**	**0.242**	**0.76**
	Text+CCL+Tree	CCL	SVR	383	0.45	0.143	0.115	0.91	0.221	0.91
	Text+CCL+Tree	PCFG	SVR	386	0.27	0.183	0.1376	0.93	0.352	**0.85**
	Text+CCL+Tree	CJ	SVR	386	0.23	0.168	0.1042	0.87	0.244	0.77
WSJ0-1-22-24	Text	CCL	SVR	310	0.45	0.143	0.1143	0.9	0.22	**0.9**
	Text	PCFG	PLS-SVR	70	0.29	0.182	0.1376	0.93	0.344	0.87
	Text	CJ	PLS-SVR	35	0.24	0.174	0.1045	0.88	0.248	0.79
	Text+CCL	CCL	SVR	336	**0.46**	**0.142**	0.1138	0.9	**0.219**	**0.9**
	Text+CCL	PCFG	SVR	336	**0.35**	**0.177**	**0.1351**	**0.91**	0.335	**0.85**
	Text+CCL	CJ	FS-SVR	21	0.24	0.175	0.105	0.88	0.249	0.8
	Text+CCL+Tree	CCL	SVR	386	**0.46**	**0.142**	**0.1135**	**0.89**	**0.219**	**0.9**
	Text+CCL+Tree	PCFG	SVR	394	0.32	0.181	0.1359	0.92	0.344	0.86
	Text+CCL+Tree	CJ	FS-SVR	22	0.24	0.175	0.1048	0.88	0.249	0.8

*Table 5. RTM top predictor results with out-of-domain test set BTest. Text+CCL reach top MRAER performance with only 3 dimensions. Best result for each metric is in **bold**.*

features dominate in the ranking of the top 2 features after FS for each PPP setting (Table 6) with only 7 out of 36 entries are not translation features.

4. Contributions

RTM PPP models work without training a parser or without parsing with it or without any parser dependent information by using only text input. We have contributed to the state-of-the-art in prediction science with results for PPP with RTM system and with expected lower bound on the prediction performance and the number of instances needed for prediction given a RAE level. RTM results on PPP allow better setting of expectations for each task and domain. Our results show that to obtain the top performance we only need 12 labeled instances and we can reach the top performance in a 15 dimensional space. Ability to predict outcomes enables preparation and savings in computational effort and can reduce costs in industrial settings.

Acknowledgments

This work was supported in part by SFI for the "Monolingual and Bilingual Text Quality Judgments with Translation Performance Prediction" (13/TIDA/I2740) project and in part by the ADAPT research center (www.adaptcentre.ie, 07/CE/I1142) at Dublin

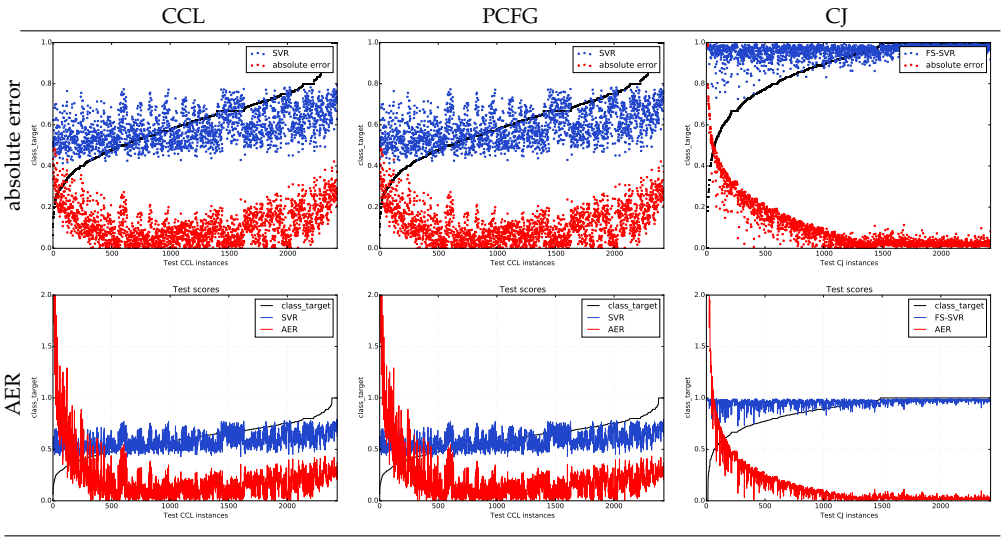


Figure 4. RTM top predictor performance using WSJ0-1-22-24 training set, BTest test set, and Text+CCL+Tree setting. At the top are plots with prediction vs. the absolute error (distribution below) and at the bottom are plots with prediction vs. AER or absolute error relative to the magnitude of the target (distribution below).

City University. We also thank the SFI/HEA Irish Centre for High-End Computing (ICHEC, `www.ichec.ie`) for the provision of computational facilities and support.

Bibliography

Bacchiani, Michiel, Michael Riley, Brian Roark, and Richard Sproat. MAP adaptation of stochastic grammars. *Computer Speech & Language*, 20(1):41–68, 2006. doi: 10.1016/j.csl.2004.12.001.

Biçici, Ergun. *The Regression Model of Machine Translation*. PhD thesis, Koç University, 2011. Supervisor: Deniz Yuret.

Biçici, Ergun. MTPPDAT: Machine Translation Performance Prediction Dataset, 2014. URL `https://github.com/bicici/MTPPDAT`.

Biçici, Ergun. Domain Adaptation for Machine Translation with Instance Selection. *The Prague Bulletin of Mathematical Linguistics*, 103:5–20, 2015. ISSN 1804-0462. doi: 10.1515/pralin-2015-0001.

Biçici, Ergun. Referential Translation Machines for Predicting Translation Quality and Related Statistics. In *Proc. of the First Conference on Statistical Machine Translation (WMT16)*, Berlin, Germany, 8 2016. Association for Computational Linguistics.

Test	Parser	Text		+CCL		+Tree	
WSJ23	CCL	$\langle P(S,T),2,2\rangle$ bpw	-0.007	$\langle P(S,T),1,10\rangle$	0.372	$\langle P(T\|S),2,2\rangle$ bpw	-0.203
		$\langle P(S,T),2,5\rangle$ bpw	-0.006	$\langle P(S,T),1,3\rangle$	0.372	$\langle P(T\|S),2,1\rangle$ bpw	-0.222
	PCFG	$\langle P(S,T),2,5\rangle$ bpw	-0.082	$\langle P(S,T),2,5\rangle$ bpw	-0.082	avgD20	-0.176
		$\langle P(S,T),2,2\rangle$ bpw	-0.089	$\langle P(S,T),2,2\rangle$ bpw	-0.089	T tree num	-0.162
	CJ	$\langle P(S,T),1,5\rangle$ bpw	-0.001	$\langle P(S,T),1,1\rangle$ bpw	-0.001	$\langle P(T\|S),1,2\rangle$ bpw	-0.095
		$\langle P(S,T),1,1\rangle$ bpw	-0.001	$\langle P(T\|S),1,2\rangle$ bpw	-0.095	$\langle P(S,T),1,1\rangle$ bpw	-0.001
BTest	CCL	$\langle P(T\|S),2,5\rangle$ bpw	-0.218	$\langle P(T\|S),2,10\rangle$ bpw	-0.183	$\langle P(T\|S),2,5\rangle$ bpw	-0.218
		$\langle P(T\|S),2,10\rangle$ bpw	-0.183	$\langle P(T\|S),1,2\rangle$ bpw	-0.3	$\langle P(T\|S),2,10\rangle$ bpw	-0.183
	PCFG	T 1&2gram GM	-0.142	$\langle P(T\|S),b2\rangle$	0.181	avgD20	-0.036
		1&2gram GM	-0.142	T $\langle P(T\|S),b2\rangle$	0.181	coverscore	0.14
	CJ	$\langle P(S,T),1,1\rangle$ bpw	-0.048	$\langle P(S,T),1,1\rangle$ bpw	-0.048	$\langle P(S,T),1,1\rangle$ bpw	-0.048
		$\langle P(T\|S),3,10\rangle$ bpw	-0.074	1&2gram prec	-0.205	$\langle P(T\|S),1,2\rangle$ bpw	-0.107

Table 6. RTM PPP model top 2 features for SVR with training set WSJ0-1-22-24.

Biçici, Ergun and Lucia Specia. QuEst for High Quality Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 103:43–64, 2015. ISSN 1804-0462. doi: 10.1515/pralin-2015-0003.

Biçici, Ergun and Andy Way. Referential translation machines for predicting semantic similarity. *Language Resources and Evaluation*, pages 1–27, 2015. ISSN 1574-020X. doi: 10.1007/s10579-015-9322-7.

Biçici, Ergun, Qun Liu, and Andy Way. ParFDA for Fast Deployment of Accurate Statistical Machine Translation Systems, Benchmarks, and Statistics. In *Proc. of the EMNLP 2015 Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, 9 2015a. Association for Computational Linguistics. URL `http://aclanthology.info/papers/parfda-for-fast-deployment-of-accurate-statistical-machine-translation-systems-benchmarks-and-statistics`.

Biçici, Ergun, Qun Liu, and Andy Way. Referential Translation Machines for Predicting Translation Quality and Related Statistics. In *Proc. of the EMNLP 2015 Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, 9 2015b. Association for Computational Linguistics. URL `http://aclanthology.info/papers/referential-translation-machines-for-predicting-translation-quality-and-related-statistics`.

Bikel, Daniel M. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proc. of the second international conference on Human Language Technology Research*, HLT '02, pages 178–182, San Francisco, CA, USA, 2002. URL `http://dl.acm.org/citation.cfm?id=1289189.1289191`.

Bojar, Ondrej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Pavel Pecina, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proc. of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September 2015.

Charniak, Eugene and Mark Johnson. Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, Ann Arbor, Michigan, June 2005. doi: 10.3115/1219840.1219862.

Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1-3):389–422, 2002. doi: 10.1023/A:1012487302797.

Klein, Dan and Christopher D. Manning. Accurate unlexicalized parsing. In *Proc. of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA, 2003. doi: 10.3115/1075096.1075150.

Kummerfeld, Jonathan K., David Leo Wright Hall, James R. Curran, and Dan Klein. Parser Showdown at the Wall Street Corral: An Empirical Investigation of Error Types in Parser Output. In *EMNLP-CoNLL*, pages 1048–1059, 2012.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Marcus, Mitchell P., Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. Treebank-3, Linguistic Data Consortium, 1999.

Parker, Robert, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English Gigaword Fifth edition, Linguistic Data Consortium, 2011.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Ravi, Sujith, Kevin Knight, and Radu Soricut. Automatic prediction of parser accuracy. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 887–896, Stroudsburg, PA, USA, 2008. URL `http://dl.acm.org/citation.cfm?id=1613715.1613829`.

Seginer, Yoav. *Learning Syntactic Structure*. PhD thesis, Universiteit van Amsterdam, 2007.

Sekine, Satoshi and Michael J. Collins. Evalb – Bracket Scoring Program, 1997. URL `http://cs.nyu.edu/cs/projects/proteus/evalb`.

Shah, Kashif, Eleftherios Avramidis, Ergun Biçici, and Lucia Specia. QuEst - Design, Implementation and Extensions of a Framework for Machine Translation Quality Estimation. *The Prague Bulletin of Mathematical Linguistics*, 100:19–30, 2013. doi: 10.2478/pralin-2013-0008.

Tan, Liling, Carolina Scarton, Lucia Specia, and Josef van Genabith. USAAR-SHEFFIELD: Semantic Textual Similarity with Deep Regression and Machine Translation Evaluation Metrics. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 85–89, 2015. URL `http://aclweb.org/anthology/S15-2015`.

Address for correspondence:
Ergun Biçici
`ergun.bicici@yahoo.com`
ADAPT Research Center, School of Computing, Dublin City University, Ireland

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 45–68

RealText-lex: A Lexicalization Framework for RDF Triples

Rivindu Perera, Parma Nand, Gisela Klette

Auckland University of Technology

Abstract

The online era has made available almost cosmic amounts of information in the public and semi-restricted domains, prompting development of corresponding host of technologies to organize and navigate this information. One of these developing technologies deals with encoding information from free form natural language into a structured form as RDF triples. This representation enables machine processing of the data, however the processed information can not be directly converted back to human language. This has created a need to be able to lexicalize machine processed data existing as triples into a natural language, so that there is seamless transition between machine representation of information and information meant for human consumption. This paper presents a framework to lexicalize RDF triples extracted from DBpedia, a central interlinking hub for the emerging Web of Data. The framework comprises of four pattern mining modules which generate lexicalization patterns to transform triples to natural language sentences. Among these modules, three are based on lexicons and the other works on extracting relations by exploiting unstructured text to generate lexicalization patterns. A linguistic accuracy evaluation and a human evaluation on a sub-sample showed that the framework can produce patterns which are accurate and emanate human generated qualities.

1. Introduction

Central to the entire discipline of Web of Data, is the concept of data representation using Resource Description Framework (RDF) triple form (Auer et al., 2007). An RDF triple is a subject, predicate, and object construct which makes data easily interlinked. The subject must always be a Uniform Resource Identifier (URI), so that it can be linked. The predicates are formed based on a clearly specified ontology. According to W3C recommendation on Linked Data (Lassila et al., 1998), an object can be a literal

 Corresponding author: rivindu.perera@aut.ac.nz
Cite as: Rivindu Perera, Parma Nand, Gisela Klette. RealText-lex: A Lexicalization Framework for RDF Triples. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 45–68. doi: 10.1515/pralin-2016-0011.

instead of a URI. However, where possible, the objects must be represented in URI form as well.

As RDF triples open a lot of possibilities in Web data representation, it opens up opportunities to utilize this data in a wide range of application areas. One such area is Natural Language Generation (NLG). In essence, given RDF triples for an entity, there should be a possibility of generating a readable and meaningful description for that entity which has a wide variety of applications in diverse domains. For instance, an information kiosk in a museum can retrieve information from an open domain Linked Data resource (e.g., DBpedia) and transform this information to natural text to present to a user as a description with the possibility to expand or limit the amount of presented information according to the user needs. Such flexibility in amount of content presented is possible only because of the existence of the middle tier framework to transform the information to natural text, so that the information selection process is completely independent of the presentation.

The aim of the RealText project[1] is to generate readable, accurate, and informative descriptions from Web of Data (i.e., RDF triples) for Question Answering over Linked Data (QALD). RealText project consists of four major frameworks; namely,

- $\text{RealText}_{\text{lex}}$: lexicalization framework
- $\text{RealText}_{\text{agg}}$: aggregation framework
- $\text{RealText}_{\text{reg}}$: referring expression generation framework
- $\text{RealText}_{\text{rel}}$: surface realization framework

In this paper we limit the scope of our discussion to $\text{RealText}_{\text{lex}}$- lexicalization framework. This framework utilizes the DBpedia as the Linked Open Data resource to generate lexicalization patterns to transform triples into natural language sentences. In following sections we discuss the framework in detail.

The rest of the paper is organized as follows. Section 2 discusses what we mean by lexicalization in the context of Linked Data framework. In Section 3, we provide an introduction to DBpedia (Bizer et al., 2009; Lehmann et al., 2014), the RDF store used for the project and motivation for utilizing it. Section 4 discusses the related works in the area and compares our approach with them. In Section 5, we discuss the framework in detail. Section 6 discusses the experimental framework and provides an analysis of the results including some comparisons. Section 7 concludes the paper with an outlook on future work.

2. Lexicalization in RealText

Before we provide algorithmic details on our lexicalization framework, we first define the lexicalization in terms as used in Linked Data context.

[1] http://people.aut.ac.nz/~rperera/projects/realtext.html

Triple	Lexicalization Pattern
$\langle$*Steve Jobs, founder, Apple Inc*$\rangle_{\mathrm{T}}$	$\langle$*S?, is the founder of, O?*$\rangle_{\mathrm{L}}$
$\langle$*Klaus Wowereit, party, Social Democratic Party*$\rangle_{\mathrm{T}}$	$\langle$*S?, is a member of, O?*$\rangle_{\mathrm{L}}$
$\langle$*Canada, currency, Canadian dollar*$\rangle_{\mathrm{T}}$	$\langle$*O?, is the official currency of, S?*$\rangle_{\mathrm{L}}$
$\langle$*Canada, capital, Ottawa*$\rangle_{\mathrm{T}}$	$\langle$*O?, is the capital city of, S?*$\rangle_{\mathrm{L}}$
$\langle$*Rick Perry, birth date, 1950-03-04*$\rangle_{\mathrm{T}}$	$\langle$*S?, was born on, O?*$\rangle_{\mathrm{L}}$

Table 1. Example lexicalization patterns

According to Reiter and Dale (2000), lexicalization is the process of choosing words to represent abstract data in natural a language. This essentially focuses on selecting the content word to represent the same meaning.

The way that lexicalization is considered in RealText project is more sophisticated than the aforementioned definition. We consider the lexicalization as a process of finding patterns that can transform a given triple to the basic natural language form. To explain this further we have provided some examples in Table 1.

As shown in Table 1, RealText$_{\mathrm{lex}}$ module is simply not looking for a lexical choice; it is meant to construct a syntactically correct and semantically appropriate pattern which can transform the triple into a natural language segment.

3. DBpedia: an interlinking hub for Linked Data

We utilize DBpedia as our RDF store for retrieving triples. The experiments to demonstrate lexicalization in this paper are specific to DBpedia due to three main reasons:

- sheer of volume
- as an interlinking hub
- open access

DBpedia is currently the fastest growing Linked Data resource that is available freely. Table 2 depicts relevant statistics illustrating its growth over five major releases. In Table 3 we compare the DBpedia against two other leading RDF stores. The results clearly shows that DBpedia has become a crystallization point in the Linked Data area hosting a vast amount of knowledge in triple form.

The nature of Linked Data is that the data is essentially interlinked. The amount of links (both incoming and outgoing) from the Linked Data resource enables it to be referenced from other similar resources. Table 4 summarises the interlinking for both incoming and outgoing links. The numbers show that DBpedia has become a central interlinking hub for Linked Data. Due to this high interlinkage, a framework that is based on DBpedia triples also indirectly contributes to the rest of the Linked Data cloud as well. This is possible because of the knowledge representation nature

Release version	Entities (in millions)	Triples (in billions)	Ontology classes
2014	4.58	3.00	685
3.9	4.26	2.46	529
3.8	3.77	1.89	359
3.7	3.64	1.00	320
3.6	3.50	0.67	272

Table 2. DBpedia growth rate in last 5 releases. Number of entities, triples and ontology classes are considered.

Triple store	Entities (in millions)	Triples (in billions)	Ontology classes	Query language
DBpedia	4.58	3	685	SPARQL
Freebase	44	2.4	40616	MQL
YAGO	10	0.12	451708	SPARQL

Table 3. Comparison of DBpedia statistics with Freebase and Yago

of Linked Data which enabled it to be reused without significant redefinition. This was one of the main motivations that influenced us to employ DBpedia for our lexicalization framework.

4. Related work

Duma and Klein (2013) introduce the LOD-DEF system, which focuses on sentence template based verbalization for Linked Data. The approach is based on selecting a sentence where subjects and objects of triples are mentioned and then removes them from the sentence to make that sentence a template. These templates can be later reused given similar triples. However, this slot filling exercise shows a very naive approach towards the lexicalization of Linked Data and cannot be employed for individual triples. Duma and Klein (2013) do not take additional steps to further abstract the sentence template to generalize it. If the template contains certain adjectives and adverbs which were related to the training triples, then these are propagated to the test phase which ultimately makes the template inaccurate. Additionally, they do not employ preprocessing steps such as co-reference resolution. It is rather hard to find sentences which do not contain co-references to main subject and therefore we can confidently assume that when applied on a wide scale text collection, LOD-DEF

Property	Incoming links	Outgoing links
Total links	39 million	4.9 million
Number of datasets	181	14
Top 5 resources	Linked Open Colours	Freebase
	DBpedia Lite	Flickr Wrapper
	Flickr Wrapper	WordNet
	Freebase	GeoNames
	YAGO	UMBEL

Table 4. Statistics on DBpedia interlinking

can end up extracting patterns with unnecessary information corresponding to co-references.

An approach which closely resembles our framework that can be found in literature is the Lemon Model (Walter et al., 2013). In this approach a sentence collection is employed to extract patterns to lexicalize triples, which is similar to our approach. However, the pattern extraction process uses typed dependency paths between subject and object values to derive the pattern. In essence, given a typed dependency-parsed sentence which contain the subject and object, the shortest path between the subject and object is searched and the sub-string is extracted. This sub-string is considered as a template to lexicalize the triple. Although the approach is linguistically sound, this method has several challenges. Firstly, the sentence collection is used as a raw set without preprocessing. This means that sentences having co-references to an entity are not considered as candidate sentences. Furthermore, the extracted pattern is not further processed to make it cohesive by removing adverbs and adjectives which can make the pattern specific to a triple. The framework proposed in this paper, addresses these issues. Instead of dependency parsing, we use a state-of-the-art relation extraction mechanism to extract cohesive patterns from natural text followed by a series of alignment phases in an effort to improve the accuracy.

Ell and Harth (2014) propose yet another verbalization model based on maximal sub-graph patterns. The main focus of this study is the transformation of multiple triples represented in natural language into a graph form. In contrast, our framework is focused on how lexicalization patterns can be generated to transform individual triples to natural language sentences. We are more interested in this specific objective so that the framework is as widely generalizable as possible, hence would be able to support integration with rest of the modules in RealText framework as introduced in Section 1. In addition, Ell and Harth (2014) do not carry out any additional processing for further realization of the extracted pattern. The idiosyncrasy of any natural language including the English, means that there has to be additional post-processing of the noise within unstructured text. This is achieved by the post-processing realiza-

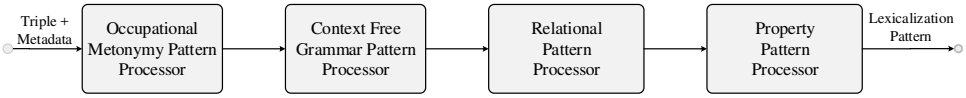


Figure 1. Schematic representation of the lexicalization framework

tion operations which helps transform the noise text into accurate readable text which emanates human produced qualities.

5. RealText$_{lex}$ framework

The schematic representation of the RealText$_{lex}$ is shown in Fig. 1. The framework is composed of four main modules, all of them targeted towards generating lexicalization patterns. Although, they are placed in a pipeline, if one of the module finds a lexicalization pattern for a given triple, then the remaining modules will not be executed.

The following sections describe the modules shown in Fig. 1 in detail. Section 5.1 and Section 5.2 discuss the process of preparing input data (i.e., triples with metadata). Section 5.3 to Section 5.6 explain individual modules depicted in Fig. 1.

5.1. Triple retrieval

DBpedia SPARQL endpoint is available to extract triples through the published Web API[2]. However, due to availability issues our module uses the local cloned version of the DBpedia which provides uninterrupted resource access on demand. In particular, when requested for a particular RDF file of an entity, the retrieval module first checks the local repository and if not available downloads it from the DBpedia automatically and adds to the repository for any future use.

Currently, DBpedia RDF files contain two types of triples. The DBpedia properties which are extracted in raw form and provided under *dbprop schema*, do not contain a unique naming convention throughout the whole DBpedia entities. To address this drawback DBpedia *OWL property schema* was introduced by mapping *dbprop schema* to a consistent schema using community effort. This research utilizes the DBpedia *OWL property schema*.

We employed Jena RDF parser[3] to extract triples from the RDF files. However, the extracted triples contain some triples that are not appropriate for verbalization. These triples contain links to Wikipedia (e.g., Wikipedia page external link), identifiers (e.g.,

[2]http://dbpedia.org/sparql

[3]https://jena.apache.org/

viaf ID) and other properties (e.g., map caption, picture caption, and picture size) which were appropriately removed.

5.2. Metadata Embedding for Triples

The triples retrieved from the above step are associated with metadata, which pattern processing modules need to consult when generating lexicalization patterns. We provide below, a discussion on metadata and the methodology for deriving them if not available directly. Fig. 2 illustrates the proto-phrase representation of a triple to illustrate the use of the meta data.

Triple Verbalizations: The triples essentially do not contain verbal representation of data. In essence, the subject of a triple is a URI to an entity and predicates are represented as properties of a designed ontology. The only exception is that objects in a triple can contain literal values which are already verbalized, and in many occasions objects also contain URIs to other entities. The objective of triple verbalization is to transform the triple to derive the verbal form by addressing the aforementioned representations. Initially, we only address the above issues when verbalizing triples. Then in Section 5.5.3 we discuss further verbalization extensions for the triple object value, specific to relational pattern processing.

Ontology Class Hierarchy: The lexicalization patterns that are extracted for triples can be specific to the ontology class that they belong to. For an instance, consider two triples, $\langle$*Skype, author, Janus Friis*$\rangle_{\mathrm{T}}$ and $\langle$*The Scream, author, Edvard Munch*$\rangle_{\mathrm{T}}$, which are retrieved from DBpedia. Both of these triples contain the same predicate "*author*", however the entities described here belong to two different ontology classes, "*Software*" and *"Art Work"* respectively. The first triple can be lexicalized as "*Skype is developed by Janus Friis*", while the second triple will be generally lexicalized as "*The Scream is painted by Edvard Munch*". This differentiation is due to the fine ontology class that the subjects of the two entities belong to. This illustrates that associating the ontology hierarchy with the lexicalization pattern is critical when searching for a matching pattern for a new triple.

Predicate Information: We also tag each triple if the predicate requires a date value, measured numbers, or a normal numbers as the object. This is mainly to support the relational pattern extraction process and will be further explained in Section 5.5. To identify whether a predicate needs a date value, XML schema definitions associated (if any) with objects are consulted. The current data representation in DBpedia provides only the XML schema definition with the predicate representing numerical (e.g., double, integer) or temporal (e.g., date/time) properties. The predicates which require measurement unit in the real world are not associated with measurement unit information. This causes a severe issue when transforming these predicates into natural language. For example, to transform the triple $\langle$*Michael Jordan, height, 1.98*$\rangle_{\mathrm{T}}$ to natural language, we need the measurement unit for height. To address this , a measurement unit database was created which provides details on predicates which re-

Predicate	Ontology URI	Measurement Unit	
		Short Name	Long name
height	http://dbpedia.org/ontology/height	m	meter
budget	http://dbpedia.org/ontology/budget	USD	US Dollars
areaTotal	http://dbpedia.org/ontology/areaTotal	m^2	square meter
populationDensity	http://dbpedia.org/ontology/populationDensity	$ppkm^2$	persons per square kilometre

Table 5. Sample set of records from the measurement unit database

quire measurement units. Table 5 depicts sample set of selected records from this database.

Natural Gender: Natural gender of a subject is another property that affects lexicalization pattern not generalizable across all entities that associate a particular predicate. For instance consider the two triples, $\langle$*Barack Obama, spouse, Michelle Obama*$\rangle_{\mathrm{T}}$ and $\langle$*Michelle Obama, spouse, Barack Obama*$\rangle_{\mathrm{T}}$. Although they have the same predicate and both subjects belong to the same fine ontology class, a lexicalization pattern generated for the first triples such as $\langle$*S?, is the husband of, O?*$\rangle_{\mathrm{L}}$ cannot be used for the second triple as the natural gender of subjects are different. Due to this fact, the framework also associates the natural gender of the subject with the retrieved triple. To find natural gender we consult the DBpedia NLP (Natural Language Processing) dataset (Mendes et al., 2012) as a primary resource and missing records are added.

Object Multiplicity: Some triples contain the same subject and predicate with different objects. These triples with multiple objects require different natural language representation compared to another predicate with single object. For example consider triples related to *Nile River*, $\langle$*Nile, country, Egypt*$\rangle_{\mathrm{T}}$, $\langle$*Nile, country, Rwanda*$\rangle_{\mathrm{T}}$, and $\langle$*Nile, country, Uganda*$\rangle_{\mathrm{T}}$ which describe the countries that Nile River flows through. However, the same information is represented for *East River* as $\langle$*East River, country, USA*$\rangle_{\mathrm{T}}$ which describes that *East River* is located in USA. These two scenarios need two different lexicalization patterns such as $\langle$*S?, flows through, O?*$\rangle_{\mathrm{L}}$ and $\langle$*S?, located in, O?*$\rangle_{\mathrm{L}}$ respectively. This shows that object multiplicity plays a crucial role in deciding the most appropriate lexicalization pattern for a given triple. Therefore, each triple is associated with a property which describes the multiplicity computed by analysing the whole triple collection.

The triples with aforementioned metadata are passed to the pattern extraction modules (explained in Section 5.3 to Section 5.6).

Subject_{Raw}	Steve_Jobs
Predicate_{Raw}	birthDate
Object_{Raw}	1955-02-24
$\text{Subject}_{Verbalized}$	Steve Jobs
$\text{Predicate}_{Verbalized}$	birth date
$\text{Object}_{Verbalized}$	1 February 24, 1955 2 24 February 1955 3 .....
OntologyClasses	1 Agent 2 Person
Predicate(RequireDate)	True
Predicate(DateInfo)	Type Single Format YMD
Predicate(RequireNormalNumber)	False
Predicate(RequireMeasuredNumber)	False
Predicate(MeasurementUnitInfo)	Short name Null Long name Null
NaturalGender	Male
Multiplicity	False

Figure 2. Example proto-phrase specification of a triple

5.3. Occupational Metonym Patterns

Metonym is a single word or phrase which is referred to not by its own name, but by a name that is associated with the meaning of it (Kövecses and Radden, 1998). A well understood and highly used metonym is "Hollywood", which is used to denote the USA film industry. In the same way, there exist several metonyms which are created based on the occupations. Some of them are "commander", "owner", and "producer" which are used, respectively, to denote someone who gives commands to one or more people, someone who owns something, and someone who produces something.

5.3.1. Morphological Formation

Fig. 3 shows the classification hierarchy of English morphology and highlights under which category occupational metonyms are classified. Based on this classification, it is clear that occupational metonyms are nominalization of verbs.

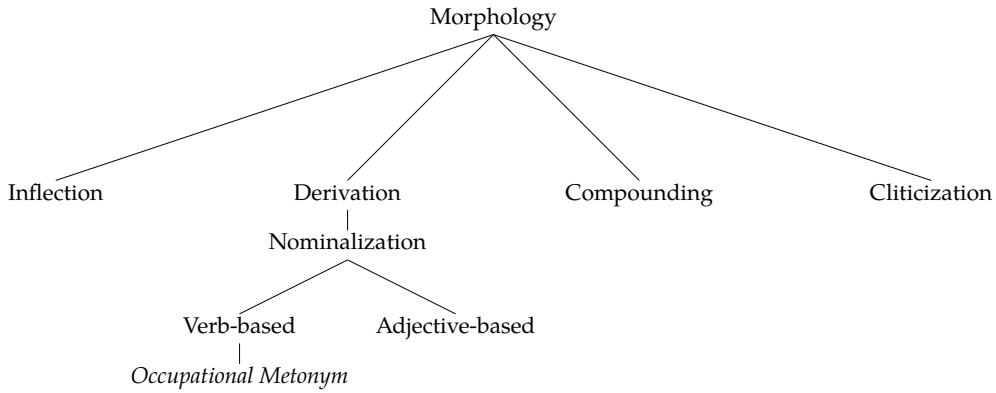


Figure 3. Classification hierarchy of English morphology

Two widely used forms of nominalization for occupation metonyms is the affixing of so-called agentive nominals; *-er* and *-or* nominals. These nominalizations can be directly applied on a base verb as well as can be applied on top of other morphological inflections. For example, Fig. 4(a) and Fig. 4(b) show two different occupational metonym forms in different granularity of applying nominalizations to form occupational metonyms.

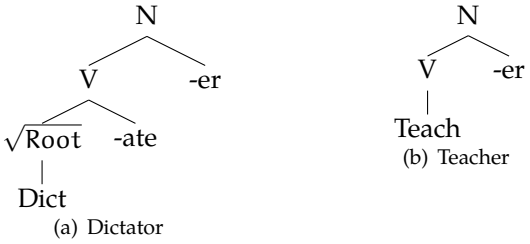


Figure 4. Two different occupational metonym formation applying -er nominals

Although it is possible to develop an unsupervised lexicon by nominalizing verbs, idiosyncrasy of English makes it rather hard. In some cases, the nominalized noun may also refer to non-agentive nominals (Schäfer, 2011) as in the two examples below.

- scratcher – a scratched lottery ticket
- broiler – a special part of a stove for cooking meat or fish

The occupational metonym lexicon used for this research is built under supervision by carefully considering accurate occupational metonyms.

There are multiple occasions where the aforementioned occupational metonyms appear as predicates of the triple. For example, the triple $\langle$*Batman Begins, publisher, Christopher Nolan*$\rangle_{\mathrm{T}}$ contains the "publisher" as the predicate which is an *-er* nominalized form the verb "publish". Since the base verb of the nominalization indicates the verb related to the profession, we can specify a straightforward lexicalization as $\langle$*Christopher Nolan, published, Batman Begins*$\rangle_{\mathrm{LT}}$. However, a further realization of the pattern can be formed by a passivized version as $\langle$*Batman Begins, is published by, Christopher Nolan*$\rangle_{\mathrm{LT}}$.

5.4. Context Free Grammar Patterns

Context Free Grammar (CFG) is considered dual purpose in NLP. This means that it can be used to understand the language as well as to generate language, based on a given grammar rules. For instance, Busemann (2005) describes the TG/2 shallow NLG system, which uses CFG rules and associated templates to generate natural language text. Furthermore, Stribling et al. (2005) demonstrated the SCiGen program which generates scientific papers using handwritten CFG rules. However, a burden associated with CFG is that the grammar rules need to be specified in advance, either as handwritten rules or as phrase structure trees derived from a seed corpus.

Due to the burdens associated with CFG based language production, our system does not use CFG as the main source. Only certain predicates which satisfy a predetermined constraint are associated with a CFG pattern. The constraint is that the predicate must either be a verb in past tense (e.g., influenced) or a predicate that is provided in passive form (e.g., maintained by). The CFG basic grammar form ($\mathcal{L}_0$) for single sentence level construction can be illustrated as follows:.

$S \rightarrow NP\ VP$

$NP \rightarrow NNP$

$VP \rightarrow VBD\ NP$

where *S* denotes a sentence. *NP*, *NNP*, *VP*, and *VBD* represent a noun phrase, proper noun, verb phrase, and verb in past tense, respectively.

The CFG patterns are applied to the triples with predicates which are identified as verbs in past tense and if the identified verb has a frame NP↔VP↔NP. For an example, the triple $\langle$*Socrates, influenced, Plato*$\rangle_{\mathrm{T}}$ can be lexicalized as its predicate satisfies the above CFG rule (i.e., NP↔VP↔NP); in essence the verb "*influence*" has the required frame. In addition, to these types of triples, CFG pattern processing module also covers the predicates which are passive form verbs (e.g., $\langle$*Aristotle, influencedBy, Parmenides*$\rangle_{\mathrm{T}}$). Besides the methodology, CFG pattern processing also needs a verb frame database to identify whether verb contains the required frame. To accomplish this, we have built a verb frame database based on the VerbNet (Kipper et al., 2008), and this database also provides all the forms of the verb (past, present, and past participle).

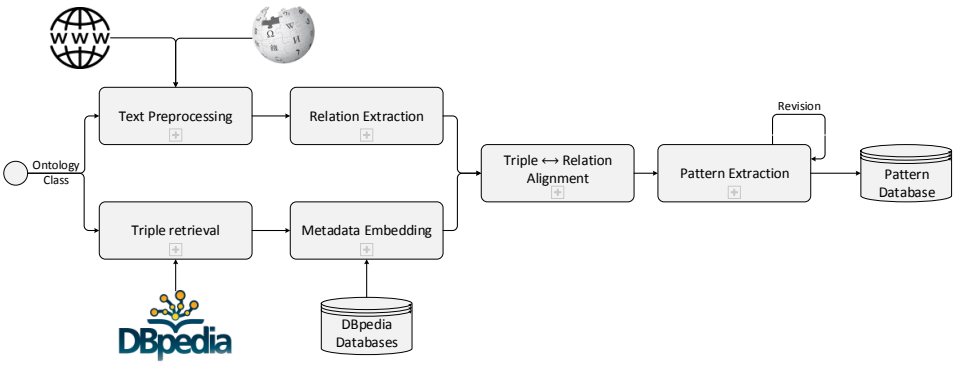


Figure 5. Relational pattern extraction process

Ontology class hierarchy	Agent, Person	Agent, Organisation, Company
Entities	Jimmy Wales Larry Sanger Natalie Portman	Google Intel Microsoft

Table 6. Sample input to the relational pattern extraction module. The example shows two ontology class hierarchies and associated entities. The actual input contains a series of class hierarchies and their associated entities.

5.5. Relational Patterns

Relational patterns are lexicalization patterns which are derived from the unstructured text using relation extraction. In brief, we process large number of unstructured text resources related to the triples and extract relations using Open Information Extraction (OpenIE) (Etzioni et al., 2008). These relations are then aligned with the triples to extract lexicalization patterns. Fig. 5 depicts the schematic representation of the relational pattern extraction process.

The module is initiated with ontology class hierarchy and associated entity collection. Table 6 depicts a sample input to the framework.

The module takes the aforementioned input and then moves to a parallel process of relation extraction and triple retrieval. During this process, it collects text related to each of the entities provided and then extract relations from the collected text. On the other hand triples related to these entities are retrieved from the DBpedia and enriched with metadata which is needed for the latter processes. The relations are

then aligned with triples to extract relational patterns. We explain the process in detail in the following sections.

5.5.1. Text Preprocessing

We first retrieve unstructured text related to the entities from Wikipedia as well as from web based text resources. Since DBpedia contains information extracted from Wikipedia (i.e., Wikipedia Infoboxes which contain the unstructured data are converted to Linked Data), it is considered as a primary resource for text to be extracted. Articles extracted from Wikipedia are wrapped in a HTML boilerplate and this causes a serious issue when extracting pure text representation of the article. To address this the module employs the Boilerpipe (Kohlschütter et al., 2010), a shallow text feature based boilerplate removal algorithm.

However, Wikipedia itself is not sufficient to build a text corpus to extract wide range of relations. Therefore, we extract text from other web resources when building the text corpus.

What we expect from this text is a description related to a particular entity. Also sentences in the description should discuss information related to the entity. However, the text extracted from this step can contain co-references to already mentioned entities. Such conferences cannot be resolved once the relation extraction is performed. Therefore, as a preprocessing task we resolve the co-references by applying the entity full name. For example a paragraph like,

"*Abraham Lincoln* is regarded as one of America's greatest heroes. *He* is a remarkable story of the rise from humble beginnings to achieve the highest office in the land."

will be converted to,

"*Abraham Lincoln* is regarded as one of America's greatest heroes. *Abraham Lincoln* is a remarkable story of the rise from humble beginnings to achieve the highest office in the land."

We utilized the Stanford CoreNLP (Manning et al., 2014) for this task. However, manual corrections are done where necessary to stop propagating preprocessing errors to the latter modules which perform relation extraction and triple-relation alignment.

The result of this process, co-reference resolved set of sentences, is passed to the relation extraction process.

5.5.2. Relation Extraction

The task of relation extraction is to extract relation triples from the co-reference resolved text. The approaches towards relation extraction can be categorized into two camps, Closed Information Extraction (ClosedIE) (Moens, 2006) and Open Information Extraction (OpenIE) (Etzioni et al., 2008).

The ClosedIE, which is the traditional approach towards the relation extraction, attempts to extract natural language relations between two mentioned entities. This

approach relies on rule based methods, kernel methods and sequence labelling methods. These methods have several key drawbacks compared to ClosedIE, such as, the need for hand-crafted rules, the need for hand-tagged data, and difficulties in domain adaptability.

For the purpose of applying relation extraction in this project, we looked at a domain independent technique, which looks at the linguistic structure of the sentence to extract relations. The recently proposed OpenIE was chosen for this purpose because it can handle a large scale open domain corpus such as the web (web as a corpus). OpenIE approach for relation extraction deviates significantly from the traditional relation extraction process. OpenIE identifies relations using relational phrases. A relational phrase is a natural language phrase that denotes a relation in a particular language. The identification of such relational phrases makes the system scalable by extracting arbitrary number of relations without tagged data. Furthermore, as relational phrases are based on linguistic knowledge and do not involve domain knowledge, OpenIE can work in multiple domains with minimum configurations.

We used Ollie (Mausam et al., 2012) OpenIE system in this module. Ollie has several advantages over the other two analysed systems, ClauseIE (Del Corro and Gemulla, 2013) and Reverb (Fader et al., 2011). ClasueIE is a clause based OpenIE module which performs on a pre-specified set of clauses derived from dependency parsing. Due to this specification, ClasueIE is unable to find many linguistic structures outside its scope. As Ollie is trained on large number of instances, it can extract several relations which are not covered by ClauseIE. On the other hand, Ollie is the successor of Reverb, and hence Ollie has significant improvements over Reverb.

5.5.3. Triple Relation Alignment

Once the relations are extracted using the OpenIE, we then align each relation with the triple to identify candidate relations which can be considered as lexicalization patterns. The aligner is mainly focused on mapping the subject and object of a triple with the arguments of a relation. To accomplish this mapping we employ the word overlapping measure. In particular, we employ the Phrasal Overlap Measure (POM) calculated according to (1).

$$\mathrm{sim}_{overlap,phrase}(s_1, s_2) = \tanh\left(\frac{overlap_{phrase}(s_1,s_2)}{|s_1| + |s_2|}\right) \tag{1}$$

where, $s_1$and s_2 are two text strings and $overlap_{phrase}(s_1,s_2)$ is calculated using (2).

$$overlap_{phrase}(s_1, s_2) = \sum_{i=1}^{n} \sum_{m} i^2 \tag{2}$$

where, m is a number of i-word phrases that appear in two text strings.

The overlapping is calculated based on the exact textual representation. However, there can be scenarios where the object of a triple has more than one representation. For example, a date can be represented by multiple formats in natural language. Therefore, when calculating the overlap between the triple object and the relational argument phrase, all possible formats and verbalizations of the triple object must be consulted. The list below shows the verbalizations carried out to support phrasal overlap matching.

Date: The dates are verbalized for phrase matching by converting the date form to 7 different formats.

Measured Values: Triple objects which are measured values can be represented in multiple ways by associating them with different measurement units. However, the challenge here is that DBpedia does not provide the measurement unit of the original triple object value. To overcome this, a database is created which maps triple objects (only measured ones) to the measurement units.

Normal Numbers: Normal numbers are transformed to different scales as well as to verbal representation.

5.5.4. Pattern Extraction

The pattern extraction process elicits a lexicalization pattern from the aligned relation by substituting them with expressions. In essence we represent the subject as S? and object as O?.

A naive replacement of subject and object cannot be accomplished here due to several reasons.

Firstly, relation arguments can be mapped with one of the verbalizations instead of a triple object. If the relation object is aligned with one of the verbalizations of the object value, then direct replacement can cause information loss of unnecessary information being included in the pattern. To avoid this, the module searches for each verbalization in the triple argument and then replace them with required token.

Secondly, triple object can be mapped with a compound token from a relation argument. Consider the below example where a triple and an argument are provided, which has an acceptable alignment score.

Triple: $\langle \textit{Barack Obama, spouse, Michelle Obama} \rangle_{\mathrm{T}}$

Relation: $\langle \textit{Barack Obama, was married to, Michelle LaVaughn Robinson Obama} \rangle_{\mathrm{R}}$

In the above scenario, the triple object is mapped to the relation arg_2, which is expressive. A partial substitution of the triple object is possible in such scenarios, but they result in inaccurate data leaving some tokens unaffected. To solve this issue we introduce the dependency tree based compound token substitution. We first aggregate the relation segments, so that it is transferred to a natural language sentence. This sentence is then dependency parsed and universal typed dependencies (de Marneffe et al., 2014) are extracted for the relation argument. An example scenario of dependency parsed aggregated sentence for the relation $\langle \textit{Barack Obama, is married to, Michelle}$

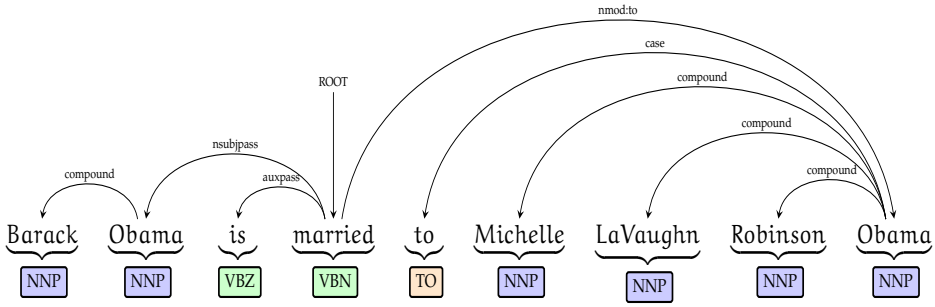


Figure 6. Compound noun identification based on the compound dependency relation

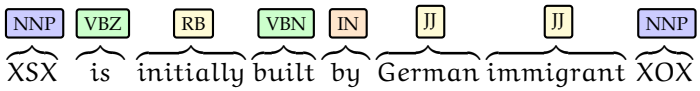


Figure 7. POS tagged transformed sentence

LaVaughn Robinson Obama$\rangle_{\mathrm{R}}$ is shown in Fig. 6. Typed dependencies which represent the compound relations are transformed back to multi-word phrases and other tokens are kept as separate. Next, each multi-word phrase is checked whether it contains the triple object tokens in full. In the occasion of such a scenario, the multi-word phrase is substituted with the triple object value (e.g., $\langle$*S?, is married to, O?*$\rangle_{\mathrm{L}}$).

Additionally, a post-processing revision step is designed to extract cohesive patterns which can be generalized regardless of the entity it is actually associated with. This cohesion is focused on filtering out adjectives and adverbs from the text. The extracted pattern is first transformed to natural language sentences by aggregating them and replacing subject and object expressions (S? and O?) with proper nouns (XSX and XOX) to avoid parser misclassification by taking the punctuations of the expressions into account. Fig. 7 depicts an example scenario where presence of adjectives make the patterns specific to a single entity. The example pattern is extracted from the sentence "Brooklyn Bridge is initially built by German immigrant John A. Roebling" for the triple $\langle$*Brooklyn Bridge, architect, John A. Roebling*$\rangle_{\mathrm{T}}$. However, this pattern cannot be generalized due to the adjectives and adverbs. Therefore, a further cleaning is done on patterns to remove adjectives and adverbs.

In addition to the aforementioned tasks, relational pattern extraction needs a threshold point to select a lexicalization pattern. This is because relational patterns come with different alignment scores. In the research we set this value to 0.21 as this value is corresponds to the single token matching in the alignment. In essence, if subject

Pattern	Predicate	Triple	Resulting lexicalized triple
$\langle$*S?'s P?, is, O?*$\rangle_{\mathrm{L}}$	height	$\langle$*Michael Jordan, height, 1.98*$\rangle_{\mathrm{T}}$	$\langle$*Michael Jordan's height, is, 1.98*$\rangle_{\mathrm{LT}}$
$\langle$*S?, has, O? P?*$\rangle_{\mathrm{L}}$	championships	$\langle$*Rubens Barrichello, championships, 0*$\rangle_{\mathrm{T}}$	$\langle$*Rubens Barrichello, has, 0 championships*$\rangle_{\mathrm{LT}}$
$\langle$*S?, is, O?*$\rangle_{\mathrm{L}}$	occupation	$\langle$*Natalie Portman, occupation, actress*$\rangle_{\mathrm{T}}$	$\langle$*Natalie Portman, is, an actress*$\rangle_{\mathrm{LT}}$
$\langle$*P? in S?, is, O?*$\rangle_{\mathrm{L}}$	largestCity	$\langle$*Australia, largesCity, Sydney*$\rangle_{\mathrm{T}}$	$\langle$*Largest city in Australia, is, Sydney*$\rangle_{\mathrm{LT}}$
$\langle$*S?, P?, O?*$\rangle_{\mathrm{L}}$	isPartOf	$\langle$*Delft, isPartOf, South Holland*$\rangle_{\mathrm{T}}$	$\langle$*Delft, is part of, South Holland*$\rangle_{\mathrm{LT}}$

Table 7. Property patterns with examples

and object are composed of one token each the normalized overlap measure of both subject and object equals to the 0.5 and hyperbolic tangent value of this is 0.45. Therefore, the multiplication of subject and object alignment equals to 0.2116. Since all other alignments are greater than this value of alignment, the single token alignment is considered as a cut-off point for relational lexicalization patterns.

5.6. Property Patterns

Property patterns specify a limited lexicon where certain predicates are associated with a pre-specified list of templates as lexicalization patterns. Five such patterns are specified, which will be applied only to the predetermined list of predicates. Table 7 list the 5 patterns with some examples of lexicalization when applied to triples with predetermined predicates. As shown in Table 7, the pattern may contain all three triple expressions which will be replaced by their verbalized form during the lexicalization. The module is designed in such a way that it can be scaled with newly defined property patterns without additional effort.

Since property pattern module is at the end of the pattern processing sequence, some of the triples may still use the pattern determined in a previous module instead of the property pattern thus making the property pattern to be ignored. This setting is arranged if majority of the triples are lexicalized with the property patterns, then the linguistic variation is negatively affected by having more similar sentences throughout a passage. Since language variety is one of the fact that make language naturalize, the framework attempts to maintain the variety to a level that it can achieve with the current settings.

Another important factor to notice in property patterns is that they are not associated with the ontology class of the subject. This is intentionally left in order to generalize the property patterns and apply them in a wide scale thus providing at least a basic lexicalization for majority of the triples.

6. Evaluation

In this section we discuss the evaluation of the lexicalization framework. Section 6.1 introduces the evaluation settings and present the acquired results. In Section 6.2, we discuss these results in detail and explain the limitations of the proposed framework.

6.1. Evaluation settings and results

Table 8 depicts a sample set of triples and some of the lexicalization patterns generated by the framework that can be associated with those triples. The table also depicts the pattern source of each lexicalization pattern and in case if the source is a relational pattern, the alignment score is also provided.

The evaluation of the framework is two fold. We first carried out an author evaluation on the linguistic accuracy of the extracted patterns and appropriateness to triples. The second evaluation was based on a survey where a group of participants were asked to rate the lexicalization patterns for their linguistic accuracy and appropriateness. Since human evaluation is resource expensive, the second evaluation considered only a randomly selected set of triples and associated lexicalization patterns from a pool.

6.1.1. Linguistic accuracy evaluation

This evaluation phase analysed lexicalization patterns selected for 400 triples from 28 entities categorized under 25 ontology classes. Since the framework described here is part a of a larger project which utilizes the Linked Data in Question Answering (QA), the source of triples is a collection of triples associated with entities in a Linked Data based QA dataset, QALD-2 (Unger, 2011) test dataset.

We evaluated each lexicalization pattern for their syntactic and semantic accuracy and the results are shown in Fig. 8. According to the figure it is clear that framework was able to generate grammatically correct patterns for 283 triples from the complete 400 triple data collection. The framework was unable to associate any pattern for 70 triples and generated incorrect patterns for 47 triples. Except for one entity (E-3), for all other entities, the framework was able to associate more than 50% of the triples with accurate lexicalization patterns.

During the analysis, we found several factors that affect a triple to be left without a lexicalization pattern since most of them need a relational pattern if the other modules

Triple	Pattern	Source	Score
$\langle$*Marlon Fernandez, birth place, London*$\rangle_T$	$\langle$*S?, was born in, O?*$\rangle_L$	Relational	0.8192
$\langle$*Marlon Fernandez, birth date, 2001-11-09*$\rangle_T$	$\langle$*S?, was born on, O?*$\rangle_L$	Relational	0.9028
$\langle$*K2, first ascent person, Achille Compagnoni*$\rangle_T$	$\langle$*S?, was climbed by, O?*$\rangle_L$	Relational	0.4182
$\langle$*Battlestar Galactica, network, Syfy*$\rangle_T$	$\langle$*S?, is aired on, O?*$\rangle_L$	Relational	0.2910
$\langle$*United Kingdom, currency, Pound sterling*$\rangle_T$	$\langle$*O?, is the official currency of, S?*$\rangle_L$	Relational	0.3852
$\langle$*Battlestar Galactica, creator, Glen Larson*$\rangle_T$	$\langle$*S?, was created by, O?*$\rangle_L$	Metonym	-
$\langle$*Rick Perry, successor, Bill Ratliff*$\rangle_T$	$\langle$*O?, succeeded, S?*$\rangle_L$	Metonym	-
$\langle$*Ottawa, population total, 883391*$\rangle_T$	$\langle$*S?'s population total, is, O?*$\rangle_L$	Property	-
$\langle$*Lisbon, highest region, Benfica*$\rangle_T$	$\langle$*highest region in S?, is, O?*$\rangle_L$	Property	-
$\langle$*Aristotle, influenced, Jewis Philosophy*$\rangle_T$	$\langle$*S?, influenced, O?*$\rangle_L$	CFG	-
$\langle$*Microsoft, founded by, Bill Gates*$\rangle_T$	$\langle$*S?, is founded by, O?*$\rangle_L$	CFG	-

Table 8. Sample set of triples, lexicalization patterns, and the pattern source. S? and O? denote subject and object respectively.

are incapable of assigning predefined lexicalization pattern. One of the main reasons for the relational pattern processing not being able to capture all possible scenarios is due to the lack of text available for entities used to extract patterns. This is mainly due to two reasons: firstly, some entities (e.g., Rubens Barrichello, Marlon Fenandez) do not have enough information recorded on the web, and secondly, the information is available but cannot be extracted due to technical limitations in the presentation layer (e.g., dynamic content). These two limitations will be further investigated and expanded as our future work.

Another aspect of lexicalization is that some entities used for relational pattern mining might have acronyms which are frequently used. For example, the entity Secret Intelligence Service is called MI6 in text collection. This affects the alignment of relations with the triples since triples use the full name while the relation which is extracted from text which uses the acronym. At this level of research, we did not focus on acronym resolution, however, it is one of the obvious future tasks.

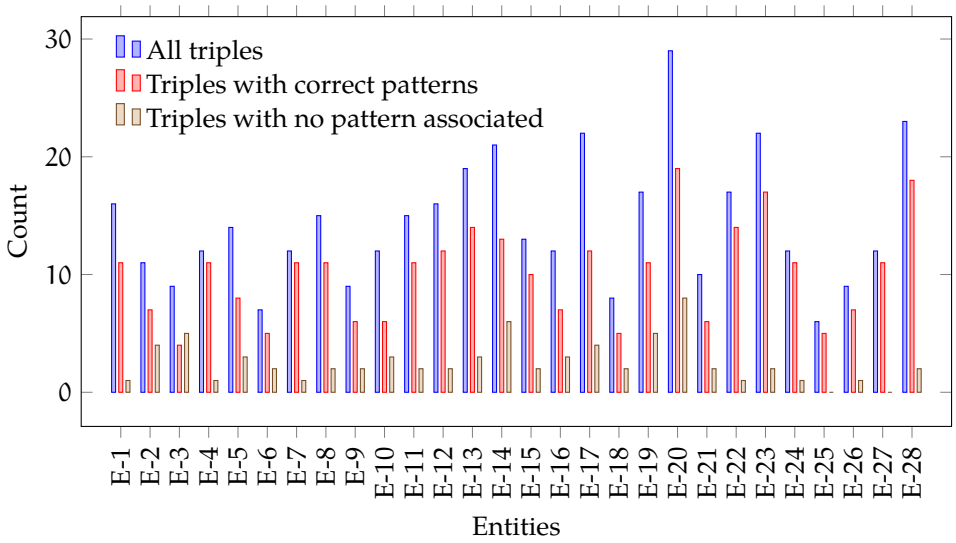


Figure 8. Lexicalization pattern evaluation for linguistic accuracy. Entities are denoted as E-1 to E-28.

6.1.2. Human evaluation with ratings

We hired 5 postgraduate students to evaluate a 40 randomly selected lexicalization patterns from the whole pattern pool which contained 400 lexicalization patterns generated for the QALD-2 test dataset. The participants were briefed with the task by providing them with three examples showing how each lexicalization pattern should be ranked according to the criteria provided. The inter-annotator agreement measured in Cronbach Alpha resulted with values 0.866 and 0.807 for readability and accuracy respectively.

Table 9 shows the results of this evaluation phase where lexicalization patterns are classified into the weighted average of rating values. In addition, a shallow analysis revealed a possible existence of a correlation between the readability and accuracy rating values. To further study this two tailed Spearman correlation analysis was performed which resulted in 0.828 correlation coefficient (p<0.001). This strong positive correlation reveals that accuracy of a lexicalization patterns and its readability are closely related.

6.1.3. Comparison with Lemon model

Among the three related works described in Section 4, the Lemon model (Walter et al., 2013) which has similar objective to ours, focussing on generating lexicaliza-

Weighted average of ratings	Accuracy	Readability
1.0 - 2.0	1	1
2.1 - 3.0	0	0
3.1 - 4.0	11	10
4.1 - 5.0	28	29

Table 9. Human evaluation results of the lexicalization patterns. The weighted average of ratings are categorized into four classes.

	RealText	Lemon
Accuracy (Full Automatic)	70.75%	37%
Accuracy (Semi automatic)	-	76%
DBpedia classes	25	30

Table 10. A comparison between RealText and Lemon. Note that semi automatic accuracy is not mentioned for our framework (RealText) as it is fully automatic.

tion patterns for individual triples rather than the whole graph. However, as Lemon model is not available for evaluation and has not released the evaluation dataset, this comparison limited to the results shown in (Walter et al., 2013).

According to the results shown in Table 10, it is clear that RealText has performed with a much higher accuracy than Lemon model in full automatic mode. Furthermore, human intervention between the process has boosted the Lemon model accuracy by 105.4%. Using human intervention in triple databases with millions of triples is not feasible as it may need excessive human resources. In this paper we showed a cost effective and scalable lexicalization framework. The framework is scalable in two ways. Firstly, the pattern mining modules connected through a loosely coupled architecture makes it possible to plug additional pattern mining modules. Secondly, utilizing OpenIE and universal typed dependencies make it possible to apply our framework in another language with minimum redesign.

6.2. Observations and discussions

The linguistic accuracy evaluation revealed that the framework was able to generate 283 accurate lexicalization patterns for 400 triples. This means that the framework achieved an accuracy level of 70.75%. The most similar system available for comparison, Lemon model, was able to achieve only 37% accuracy in its full automatic model. This shows that our approach has produced lexicalization patterns with much higher

accuracy compared to the latest state-of-the-art model. This was further attested by the human evaluation where more than 70% of the lexicalization patterns are rated between values 4.1 and 5 for both accuracy and readability. In addition, more than 90% of the lexicalization patterns were rated above the average rating values for both accuracy and readability. This again confirms the quality of the lexicalization patterns achieved by our framework.

Our post analysis on human evaluation by calculating the correlation between readability and accuracy revealed that the two have a strong positive correlation. Similar evidence can be found in a previous research carried out by Reiter and Belz (2009) in a different domain.

7. Conclusion and future work

This paper presented a lexicalization framework for RDF triples. The framework centred on mining patterns to transform RDF triples using four pattern mining modules. The evaluation of the framework concentrated on both linguistic accuracy evaluation and human evaluation. Both evaluations showed that the framework can generate accurate and readable lexicalization patterns and the results are far better compared to the most similar existing lexicalization module, Lemon model.

In future we plan to expand the framework to other Linked Data resources and well to show the scalability of the framework. In addition we will also be applying the framework in practical applications to assess the applicability of the designed framework. Much of the background work for this had already taken place. As the first application we have planned to integrate a biography generator which selects triples from DBpedia and employ the lexicalization framework to generate a textual biography.

Acknowledgements

The work reported in this paper is part of the RealText project funded by Auckland University of Technology.

Bibliography

Auer, S, C Bizer, G Kobilarov, and J Lehmann. Dbpedia: A nucleus for a web of open data. In *6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, pages 722–735, Busan, Korea, 2007. Springer-Verlag. URL http://link.springer.com/chapter/10.1007/978-3-540-76298-0{_}52.

Bizer, C, J Lehmann, and G Kobilarov. DBpedia-A crystallization point for the Web of Data. *Web Semantics: science …*, 2009. URL http://www.sciencedirect.com/science/article/pii/S1570826809000225.

Busemann, Stephan. Ten Years After : An Update on TG/2 (and Friends). *Proceedings 10th European Workshop on Natural Language Generation*, 2, 2005.

de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. Universal Stanford Dependencies: A cross-linguistic typology. In *9th International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, 2014. ISBN 978-2-9517408-8-4. URL papers3://publication/uuid/D4B7BB39-4FFB-4AA6-B21E-701A91F27739.

Del Corro, Luciano and Rainer Gemulla. ClausIE: clause-based open information extraction. pages 355–366, may 2013. URL http://dl.acm.org/citation.cfm?id=2488388.2488420.

Duma, Daniel and Ewan Klein. Generating Natural Language from Linked Data: Unsupervised template extraction. In *10th International Conference on Computational Semantics (IWCS 2013)*, Potsdam, 2013. Association for Computational Linguistics.

Ell, Basil and Andreas Harth. A language-independent method for the extraction of RDF verbalization templates. In *8th International Natural Language Generation Conference*, Philadelphia, 2014. Association for Computational Linguistics.

Etzioni, Oren, Michele Banko, Stephen Soderland, and Daniel S. Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, dec 2008. ISSN 00010782. doi: 10.1145/1409360.1409378. URL http://dl.acm.org/ft{_}gateway.cfm?id=1409378{&}type=html.

Fader, Anthony, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Empirical methods in Natural Language Processing*, pages 1535–1545, 2011. ISBN 978-1-937284-11-4. doi: 10.1234/12345678. URL http://dl.acm.org/citation.cfm?id=2145432.2145596$\delimiter"056E30F$nhttp://dl.acm.org/citation.cfm?id=2145596.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40, 2008. ISSN 1574020X. doi: 10.1007/s10579-007-9048-2.

Kohlschütter, Christian, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate Detection using Shallow Text Features. In *ACM International Conference on Web Search and Data Mining*, pages 441–450, 2010. ISBN 9781605588896. doi: 10.1145/1718487.1718542. URL http://portal.acm.org/citation.cfm?id=1718542.

Kövecses, Zoltán and Günter Radden. Metonymy: Developing a cognitive linguistic view. *Cognitive Linguistics (includes Cognitive Linguistic Bibliography)*, 9(1):37–78, 1998.

Lassila, Ora, Ralph R Swick, et al. Resource Description Framework (RDF) model and syntax specification. 1998.

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Soren Auer, and Christian Bizer. DBpedia - A Large-scale , Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web journal*, 5(1):1–29, 2014.

Manning, Christopher, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *The 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, 2014. Association for Computational Linguistics.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. Open language learning for information extraction. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, Jeju Island, jul 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390948.2391009.

Mendes, Pablo N., Max Jakob, and Christian Bizer. DBpedia for NLP: A Multilingual Cross-domain Knowledge Base. In *International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2012.

Moens, Marie Francine. *Information extraction: Algorithms and prospects in a retrieval context*, volume 21. 2006. ISBN 1402049870. doi: 10.1007/978-1-4020-4993-4.

Reiter, Ehud and Anja Belz. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, 35(4):529–558, dec 2009. ISSN 0891-2017. doi: 10.1162/coli.2009.35.4.35405. URL http://dl.acm.org/citation.cfm?id=1667988.1667994.

Reiter, Ehud and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, United Kingdom, jan 2000. ISBN 9780511551154. URL http://www.cambridge.org/us/academic/subjects/languages-linguistics/computational-linguistics/building-natural-language-generation-systems.

Schäfer, Florian. Naturally atomic er-nominalizations. *Recherches linguistiques de Vincennes*, 40 (1):27–42, 2011. ISSN 0986-6124. doi: 10.4000/rlv.1303. URL http://rlv.revues.org/351.

Stribling, Jeremy, Max Krohn, and Dan Aguayo. SciGen, 2005. URL https://pdos.csail.mit.edu/archive/scigen/.

Unger, Christina. Question Answering over Linked Data: QALD-1 Open Challenge. Technical report, Bielefeld University, Bielefeld, 2011.

Walter, Sebastian, Christina Unger, and Philipp Cimiano. A Corpus-Based Approach for the Induction of Ontology Lexica. In *18th International Conference on Applications of Natural Language to Information Systems*, pages 102–113, Salford, 2013. Springer-Verlag.

Address for correspondence:
Rivindu Perera
rivindu.perera@aut.ac.nz
Software Engineering Research Laboratory (WT-704),
School of Engineering, Computer and Mathematical Sciences,
Auckland University of Technology,
Private Bag 92006,
Auckland 1142, New Zealand

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 69–124

Linguistically Annotated Corpus as an Invaluable Resource for Advancements in Linguistic Research: A Case Study

Jan Hajič, Eva Hajičová, Jiří Mírovský, Jarmila Panevová

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

A case study based on experience in linguistic investigations using annotated monolingual and multilingual text corpora; the "cases" include a description of language phenomena belonging to different layers of the language system: morphology, surface and underlying syntax, and discourse. The analysis is based on a complex annotation of syntax, semantic functions, information structure and discourse relations of the Prague Dependency Treebank, a collection of annotated Czech texts. We want to demonstrate that annotation of corpus is not a self-contained goal: in order to be consistent, it should be based on some linguistic theory, and, at the same time, it should serve as a test bed for the given linguistic theory in particular and for linguistic research in general.[1]

1. Introduction

It is now quite easy to have access to large corpora for both written and spoken language. Corpora have become popular resources for computationally minded linguists and computer science experts developing applications in Natural Language Processing (NLP). Linguists typically look for various occurrences of specific words

[1] The present contribution is based in part on our previous summarizing study on annotation (Hajič et al., 2015), and also on studies concerning some particular linguistic phenomena quoted in the respective Sections below. We are grateful to our colleagues for providing us their material and expertise. Most importantly, we owe our thanks to Markéta Lopatková for her careful reading of the prefinal version of this paper and for her invaluable comments. The authors highly appreciate the comments and suggestions given by the two anonymous reviewers and have tried to take them into account when preparing the final version of the paper. All the responsibility, however, rests with the authors of the present paper.

 Corresponding author: hajicova@ufal.mff.cuni.cz
Cite as: Jan Hajič, Eva Hajičová, Jiří Mírovský, Jarmila Panevová. Linguistically Annotated Corpus as an Invaluable Resource for Advancements in Linguistic Research: A Case Study. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 69–124. doi: 10.1515/pralin-2016-0012.

or patterns, computational specialists construct language models and build taggers, parsers, and semantic labelers to be used in various applications.

It has also already been commonly accepted in computational and corpus linguistics that grammatical, lexical, or semantic, etc. annotation does not "spoil" a corpus, if the annotation is done in such a way that it does not remove substantial information about the raw corpus, such as spacing etc. (ideally, as stand-off annotation). On the contrary, annotation may and should bring an additional value to the corpus. Necessary conditions for this aim are:

(i) its scenario is carefully (i.e. systematically and consistently) designed, and
(ii) it is based on a sound linguistic theory.

This view is corroborated by the existence of annotated corpora of various languages: Penn Treebank (English; Marcus et al., 1993), its successors as PropBank (Kingsbury and Palmer, 2002), NomBank (Meyers et al., 2004) or Penn Discourse Treebank (Prasad et al., 2008), Tiger (Brants et al., 2002) and Salsa (German; Burchardt et al., 2006), Prague Dependency Treebank (Czech; Hajič et al., 2006; Bejček et al., 2013), and many others.

The aim of our contribution is to demonstrate, on the basis of our experience with the annotated corpus of Czech, the so-called Prague Dependency Treebank (PDT), how annotation process and its results help to test a linguistic theory, to develop it further and also to compare it with other theories, that is, how such work may contribute to a better understanding of the language system.

We first present a brief account of PDT in its current state (Sect. 2), passing over to a layer-by-layer description of individual cases which may serve as examples of phenomena for the understanding of which the annotated treebank was instrumental (Sections 3 and 4). In Sect. 5 we add some statistical information on PDT data and on the tools available as well as some remarks on the annotation process as such. We sum up our observations in Sect. 6 highlighting first in which points the existing theoretical framework has been complemented and adding one particular aspect the study of which has been made possible by the consistent and systematic annotation.

2. The Prague Dependency Treebank in a nutshell

The Prague Dependency Treebank is an effort inspired by the PennTreebank; the work started as early as in the mid-nineties and the overall scheme was already published in 1997 (see Hajič et al., 1997 and Hajič, 1998). The basic idea was to build a corpus annotated not only with respect to the part-of-speech tags and some kind of (surface) sentence structure, but also capturing the syntactico-semantic, deep structure of sentences.

The annotation scheme of PDT is based on a solid, well-developed theory of an (integrated) language description, the so-called Functional Generative Description (FGD) (see, e.g., Sgall, 1967; Sgall et al., 1969; Sgall et al., 1986); at the time of the devel-

opment of the annotation scheme this theory had already been applied to an analysis of multifarious linguistic phenomena, mostly concentrated on Czech but also in comparison with English, Russian or some other (mainly Slavonic) languages. The principles of FGD were formulated as a follow-up to the functional approach of the Prague School and with due respect to the strict methodological requirements introduced to linguistics by N. Chomsky. The FGD framework was formulated as a generative description that was conceived of as a multi-level system proceeding from linguistic function (meaning) to linguistic form (expression), that is from the generation of a deep syntactico-semantic representation of the sentence through the surface syntactic, morphemic and phonemic levels down to the phonetic shape of the sentence. From the point of view of formal grammar, both syntactic levels were based on the relations of dependency rather than constituency. The main focus was laid on the account of the deep syntactic level, called "tectogrammatical" (the term borrowed from Putnam's (1961) seminal paper on phenogrammatics and tectogrammatics). On this level, the representation of the sentence has the form of a dependency tree, with the predicate of the main clause as its root; the edges of the tree represent the dependency relations between the governor and its dependents. Only the autosemantic (lexical) elements of the sentence attein the status of legitimate nodes in the tectogrammatical representation; functional words such as prepositions, auxiliary verbs and subordinate conjunctions are not represented by separate nodes and their contribution to the meaning of the sentence is captured within the complex labels of the legitimate nodes (see below on the characteristics of the tectogrammatical level in PDT). An important role in the derivation of sentences is played by the information on the valency properties of the governing nodes, which is included in the lexical entries: the valency values are encoded by the so-called functors, which are classified into arguments and adjuncts. It is assumed that each lexical entry in the lexicon is assigned a valency frame including all the obligatory and optional arguments appurtenant to the given entry; the frame also includes those adjuncts that are obligatory with the given entry; in accordance with the frame, the dependents of the given sentence element are established in the deep representation of the sentence and assigned an appropriate functor as a part of their complex label. The representation of the sentence on the tectogrammatical level also captures the information structure of the sentence (its topic–focus articulation) by means of the specification of individual nodes of the tree as contextually bound or non-bound and by the left-to-right order of the nodes. Coordination and apposition is not considered to be a dependency relation as they cannot be captured by the usual binary directional dependency relation. Coordinated sentence elements (or elements of an apposition) introduce a non-dependency, "horizontal" structure, possibly n-ary and/or nested, but still undirectional, where all elements have (in the standard dependency sense) a common governor (the only exception is formed by coordinated main predicates which naturally have no common governor). The coordinated (or appended) elements can also have common dependent(s). All the depen-

dency relations expressed in a sentence with coordination(s) and/or apposition(s) can be extracted by "multiplying" the common dependency relations concerned.

The design of the annotation scenario of PDT (see, e.g., Hajič, 1998; Böhmová et al., 2003; Hajič et al., 2006; Bejček et al., 2011; Bejček et al., 2013) follows the above conception of FGD in all of the fundamental points:

(i) it is conceived of as a multilevel scenario, including the underlying semantico-syntactic layer (tectogrammatical),
(ii) the scheme includes a dependency based account of syntactic structure on both (surface and deep) syntactic levels,
(iii) the scheme also includes the basic features of the information structure of the sentence (its topic–focus articulation) as a component part of the underlying syntax, and
(iv) from the very beginning, both the annotation process and its results have been envisaged, among other possible applications, as a good test of the underlying linguistic theory.

PDT consists of continuous Czech texts, mostly of the journalistic style (taken from the Czech National Corpus) analyzed on three levels of annotation (morphology, surface syntactic structure, and underlying syntactic structure). At present, the total number of documents annotated on all the three levels is 3,165, amounting to 49,431 sentences and 833,193 (occurrences of) word forms and punctuation marks (tokens). PDT, Version 1.0 (with the annotation of the first two levels) is available from the Linguistic Data Consortium, as is Version 2.0 (with the annotation of the third, underlying level). PDT Version 2.5 (with some additions) as well as the current PDT Version 3.0 are available from the LINDAT/CLARIN repository.[2]

The original annotation scheme has the following multilevel architecture:

(a) **morphological layer**: all tokens of the sentence get a lemma and a (disambiguated) morphological tag,
(b) **analytical layer**: a dependency tree capturing surface syntactic relations such as subject, object, adverbial; a (structural) tag reflecting these relations is attached to the nodes as one of the component parts of their labels,
(c) **tectogrammatical layer** capturing the underlying ("deep") syntactic relations: the dependency structure of a sentence on this layer is a tree consisting of nodes only for autonomous meaningful units (function words such as prepositions, subordinating conjunctions, auxiliary verbs etc. are not included as separate nodes in the structure, their contribution to the meaning of the sentence is cap-

[2] http://www.lindat.cz

tured by the complex labels of the autonomous units). Every node of the tectogrammatical representation is assigned a complex label consisting of:[3]

- the lexical value of the word (for verbs and certain nouns, with a reference to its sense captured in the corresponding valency lexicon entry),
- its '(morphological) grammatemes' (i.e. the values of morphological categories such as Feminine, Plural etc. with nouns, Preterite, etc. with verbs),
- its 'functors' (such as Actor, Patient, Addressee, Origin, Effect and different kinds of circumstantials (adjuncts), with a more subtle differentiation of syntactic relations by means of subfunctors, e.g. 'in', 'at', 'on', 'under', 'basic', 'than', etc.), and
- the topic–focus articulation (TFA) attribute containing the values for contextual boundness, on the basis of which the topic and the focus of the sentence can be determined. Pronominal coreference is also annotated.

In addition to the above-mentioned three annotation layers in PDT, there is also one non-annotation layer representing the "raw-text". In this layer, called the "word layer", the text is segmented into documents and paragraphs and individual tokens are recognized and associated with unique identifiers. Figure 1 displays the relations between the neighboring layers as annotated and represented in the data. Thus, for example, the Czech sentence *Můžete to vysvětlit například?* (lit.: "Can-you it explain on-example", E. translation: "Could you explain it with an example? ") contains a modal verb, a pronoun, a content verb, and a prepositional phrase (with a typo).

One methodological comment should be made. Though partitioned into layers, the annotation scheme of the Prague Dependency Treebank was built as a complex one: we have annotated all the language phenomena on the same collection of texts rather than to select only some phenomenon or phenomena of a particular layer without taking into account other phenomena of the same layer. At the same time, however, each layer of annotation is accessible separately, but with a possible explicitly annotated link to the other layers of annotation. The relations between the layers are in part captured in the associated valency lexicon for verbs and their arguments, along the lines suggested in (Hajič and Honetschläger, 2003; Hajič and Urešová, 2003).

In the process of the further development of PDT, additional information has been added to the original in the follow-up versions of PDT, such as the annotation of basic relations of textual coreference and of discourse relations in the Prague Discourse Treebank (PDiT), multiword expressions etc.

---

[3] In Fig. 1 there is only a very simplified tectogrammatical representation of the given sentence as the Figure is meant to illustrate the interlining of layers of annotation rather than to bring a full annotation of the sentence on each of the layers. On the tectogrammatical layer (t-layer), the modal verb *můžete* [can you] does not obtain a node of its own and the modal meaning is captured by an index attached to the lexical verb *vysvětlit* [explain], which is however not displayed in the Figure, and also the morphological categories are omitted. (The index 'inter' stands for interrogative mood, and, e.g., #Gen is a label of a node representing a "general" participant, ADDR standing for Addressee.)

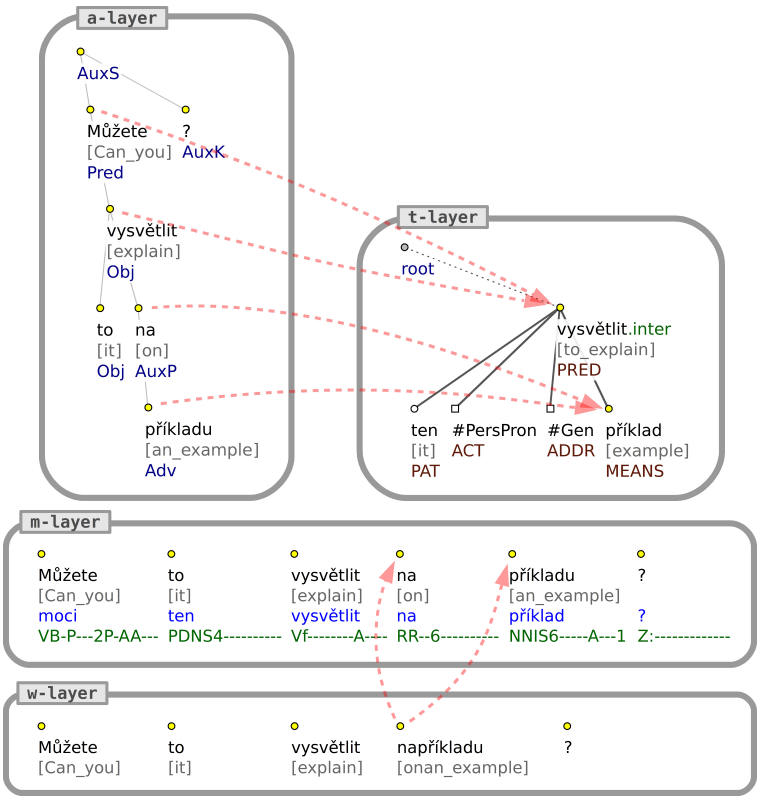


Figure 1. An example of the interlinking of the PDT layers in the Czech sentence "Můžete to vysvětlit na-příkladu?" [Lit.: Can-you it explain on-an-example?] [Can you explain it on-an example?]. The arrows represent non-1:1 relations among nodes on different layers of annotation; square nodes in the tree on the t-layer represent 'newly' generated nodes – nodes without a surface counterpart.

3. Case studies I: Morphology, surface and deep syntax

The contribution of corpus annotation for the theoretical description of language was greater than we expected at the beginning of the process. There are two phases in which this contribution comes out: In the first, preparatory decision-making phase, the proposed scheme was tested before a definite scenario with detailed instructions was approved for the build-up of the annotated corpus; at that point, the development of the annotation scenario itself reflected the state-of-the-art of the object of annotation. The tuning of the scenario (and the theoretical reflections there-off) was mainly

based on the annotators' feedback. The second phase started when the annotated corpus was ready for its exploitation for studies of theoretical issues at the end of the creation of the annotated corpus. The collection of data annotated according to the consistent scenario opened new horizons for the theoretical study of the particular phenomena on the basis of rich, real material not readily available before the time of corpus linguistics.

In the following subsections, we present an analysis of some grammatical issues based on the annotation process which has stimulated a modification of the theoretical framework of FGD or has made it necessary to supplement the existing handbooks of Czech grammar. For this purpose we have selected the following issues: Sect. 3.1 presents arguments for the necessity of an introduction of new morphological grammatemes constituting the category of diathesis, in Sect. 3.2 the peculiarities of counting objects occurring typically in pairs or groups and their morphological consequences are discussed, while in Sections 3.3 and 3.4 issues connected with valency are analyzed and an introduction of the "quasi-valency" class of modifiers is substantiated. Selected types of deletions are described in Sect. 3.5. In Sect. 3.6 variants of nominative subjects are analyzed. Sect. 4 is devoted to issues the analysis of which has brought modifications of the FGD approach to some particular aspects of the information structure of the sentence (Sect. 4.1) or of phenomena that concern the domain of discourse, which go beyond the domain of grammar and as such have been out of the scope of FGD interests (discourse relations in Sect. 4.2 and associative and coreference relations in Sect. 4.3).

As the empirical material of our analysis is Czech, we accompany the Czech example sentences with their translations to English, in most cases both literal and free. When necessary, we add (simplified) glosses capturing the information on relevant morphological and syntactic features of the Czech forms.[4]

3.1. Diathesis[5]

The morphological meanings of verbs connected with the verbal voice were usually limited to the opposition active – passive. Our analysis of the Czech data has demonstrated that there are other constructions productive enough in Czech to be considered as members of the same category as active and passive. Due to their productivity and due to the consequences they have for the syntactic structure we proposed to assign these analytical verb forms a new morphological category (grammateme)[6] called

[4] It should be noted that in order to make the glosses easier to survey we accompany them only by those features that are necessary for the understanding of the point under discussion. We assume that the abbreviations in the glosses are self-explaining and correspond to the Leipzig glossing rules; if not, we add a commentary in the text or in a footnote.

[5] In many cases the analytical passive diathesis and simple resultatives seems to be ambiguous, but some formal criteria how to distinguish their diathesis values are given in Panevová and Ševčíková (2013).

[6] For the notion of grammateme, as applied in FGD, see Sect. 2 above.

"diathesis" with the values active, analytical passive, resultative diathesis (simple and possessive) and recipient passive.

Our classification slightly differs from the Czech traditional descriptions, in which these constructions are analyzed as a special verbal tense, with a certain analogy to perfect tenses in other languages (Mathesius, 1925) or as a special verbal category called "resultative state" ("výsledný stav" in Czech, see Hausenblas, 1963)[7]. Our analysis (Panevová et al., 2014) supports the idea about the position of this construction within the diathesis paradigm. These types of diathesis are expressed by different morphemic forms, they have different morphological meanings, and they influence the syntactic structure, which is different from their unmarked active counterparts. The active sentences (1) and (4) have the following counterparts differentiated by the value of diathesis: (2) and (3) for (1), (5) for (4),

(1) Dcera už připravila matce oběd.
daughter-NOM-sg already prepare-3-sg-PST mother-DAT-sg lunch-ACC-sg
[The daughter has already prepared lunch for her mother.]

(2) Oběd už je připraven.
lunch-NOM-sg-M-Sbj already be-AUX-3-sg-PRS prepare-PTCP-PASS-sg-M
[Lunch is already cooked.]

(3) Matka už má oběd připraven.
mother-NOM-sg-F-Sbj already have-AUX-3-sg-PRS lunch-ACC-sg-M prepare-PTCP-PASS-sg-M
[lit. Mother has her lunch already prepared.]

(4) Nakladatelství zaplatilo autorovi honorář včas.
Publishing_house-NOM-sg-Sbj pay-3-sg-PST author-DAT-sg-M fee-ACC-sg-M in_time
[The publishing house has paid the fees to the author on time.]

(5) Autor dostal od nakladatelství honorář zaplacen včas.
author-NOM-sg-M-Sbj receive-AUX-3-sg-PST-M from-PREP publishing_house-GEN-sg fee-ACC-sg-M pay-PTCP-PASS-ACC-sg-M in_time
[The author has received his fees from the publishing house in time.]

[7] A detailed analysis of resultative constructions in contemporary Czech from theoretical and empirical view is presented in Giger (2003).

In (2) and (3) the action of the preparation of lunch is presented from the point of view of the result, while actions in the active constructions are presented from the point of view of the Actor. In the simple resultative (ex. (2)) the result (*oběd* [lunch]) of the action (Patient of the verb) is shifted to the position of surface subject and the Actor is omitted. In the possessive resultative constructions (ex. (3)) two kinds of restructuring are possible: in (3) mother could be understood to be the actor of the lunch preparation, but the meaning that somebody else has prepared a lunch (for mother) is also possible. In (3) the verb *mít* [have] is used in possessive resultative as an auxiliary and this construction enters the verbal paradigm;[8] since there exist verbs for which this type of diathesis is not applicable, the feature *+res_poss* indicating the possible participation of the given verb in this kind of diathesis is included in the lexicon.

Example (5) represents a less frequent diathesis where the verb *dostat* [receive] is used as an auxiliary. Contrary to its unmarked counterpart (4), the Addressee of the action in (5) is shifted into the position of the surface subject; the Actor of the action could be optionally expressed (here by the prepositional phrase *od nakladatelství* [from the publishing house]).[9]

As a result of these observations and analysis the original set of morphological categories was rearranged and extended in the modified version of the theoretical framework of FGD and in PDT (Urešová, 2011a).

3.2. Number of nouns

In the category of number the Czech nouns enter a basic opposition: singular (sg) and plural (pl). However, this category exhibits some peculiarities, especially with nouns denoting pairs or typical groups (such as *boty* [shoes], *rukavice* [gloves], *sirky* [matches], *klíče* [keys]). With other nouns we use the class of basic numerals, see *jedna kniha* [one book-sg], *dvě knihy* [two books-pl], *tři knihy* [three books-pl], etc. For counting the objects denoted by pair and group nouns, the set numerals are obligatorily used instead of the basic numerals. Rich material provided by the PDT supported an introduction of a new morphological category called pair/group meaning. Thus, we work with two paradigmatic patterns of the meaning of number: the former is connected with counting single objects, the latter with counting pairs of them or the typical sets of them (e.g. *jedna bota, tři boty* [one-*basic numeral* shoe-sg, three-*basic numeral* shoes-pl], *jeden klíč, pět klíčů* [one-*basic numeral* key-sg, five-*basic numeral* keys-pl] vs. *jedny* [set numeral] *boty, troje* [set numeral] *boty* [one pair of shoes, three pairs of shoes]; *jedny* [set numeral] *klíče, patery* [set numeral] *klíče* [one set of keys, five sets of

[8] The grammaticalization of this category indicates that Czech belongs to the class of "habere" languages (see Clancy, 2010).

[9] The syntactic diathesis (deagentization, dispositional constructions and reciprocals) has been implemented in PDT 3.0 and was described from the theoretical point of view in Panevová et al. (2014).

keys]). The differences between Czech and English demonstrate that in Czech the pair and set meaning of the nouns is grammaticalized, since a special type of compatibility with numerals is required.

If nouns occurring typically in groups or sets occur in a plural form without a numeral the sentences are often ambiguous. In (6a) the regular plural form (unmarked as for pair/group meaning) of the noun *rukavice* [glove] is used. For (6b), (7a) and (7b) several interpretations are possible; their English translations reflect their preferred meanings chosen on the basis of world knowledge or a broader context. In (7b), e.g., the knowledge of the habits used in this office would help for disambiguation if the charwoman has a single key belonging to each office or if for any office a set of keys were needed.

(6a) Často něco ztrácím, teď mám doma několik levých
often something loose-1-sg-PRS just-now have-1-sg-PRS at-home several left
rukavic.[10]
glove-pl
[I usually lose my things, just now I have at home several left gloves.]

(6b) Musím si koupit nové rukavice.
need-1-sg-PRS REFL-DAT buy-INF new glove-sg-PAIR/GROUP
[I have to buy a new pair of gloves.]

(7a) Ztratila jsem klíče od
Loose-1-sg-PST be-AUX-1-sg-PRS key-sg-PAIR/GROUP from-PREP
domu.
house-GEN-sg
[I have lost my keys from my home.]

(7b) Uklízečka má klíče od všech
Charwoman-NOM-sg have-3-sg-PRS key-ACC-pl from-PREP all
pracoven.
office-GEN-pl
[The charwoman has keys from all of the offices.]

The introduction of the new morphological category pair/group meaning is based first of all on the requirement of economy of the description of these nouns in the lexicon: A single lexical entry is sufficient for the nouns referring either to a single (particular) object, or to a typical pair, or a typical set of these objects. The compatibility of the members of the opposition +pair/group vs. -pair/group meaning with a different class of numerals is also a strong argument in favour of the introduction of

[10] In order to explain the pair/group meaning as a new unit we use in the glosses for (6) and (7) the meanings of the number rather than their forms.

Noun lemma	# of plural forms	# of pl. forms with the pair/group meaning	Percentage
dvojče [twin]	5	5	100.0%
pouto [tie]	5	5	100.0%
ledvina [kidney]	7	7	100.0%
vlas [hair]	11	11	100.0%
kopačka [football shoe]	5	5	100.0%
ucho [ear]	9	9	100.0%
lyže [ski]	13	13	100.0%
schod [stair]	6	6	100.0%
ruka [hand, arm]	81	77	95.1%
prst [finger/toe]	10	9	90.0%
oko [eye]	89	80	89.9%
rameno [shoulder]	9	8	88.9%
rukavice [glove]	8	7	87.5%
kolej [rail]	16	14	87.5%
noha [foot, leg]	20	17	85.0%
kulisa [scene]	6	5	83.3%
koleno [knee]	5	4	80.0%
bota [shoe]	30	24	80.0%
klíč [key]	8	5	62.5%
zub [tooth]	14	8	57.1%
rodič [parent]	87	37	42.5%
křídlo [wing]	17	5	29.4%
doklad [document]	35	8	22.9%
cigareta [cigarette]	17	3	17.6%
lék [medicine]	16	2	12.5%
brambor [potato]	9	1	11.1%
těstovina [pasta]	7	0	0.0%
Total	**618**	**414**	**67.0%**

Table 1. Noun lemmas with five or more plural occurrences in the PDT 2.0

a special category assigned to forms used for the meanings of the noun number. The choice between the values proposed here was checked manually in the data of PDT 2.0 by two annotators; the plural forms of nouns suspected for their use typically in the pair/group meaning with the frequency equal and higher than 5 were selected and the task of the annotators was to make choice between three possibilities: “one pair/group”, “several pairs/groups”, “undecided between preceding two groups”.

Table 1 lists noun lemmas with five or more plural occurrences in the PDT 2.0 data arranged according to the percentage of occurrences assigned the pair/group meaning out of all plural occurrences of these nouns in the final annotation.

3.3. Valency in the lexicon and in the sentence

The theoretical framework for verbal valency was elaborated within FGD in the 1970's (see Panevová, 1974–75, 1977, 1994 and others) and it was based partially on Tesnière's approach, partially on Fillmore's case grammar. The lexicographical aspects as the other obligatory part of valency description was a challenge for building valency dictionaries; the FGD theory was applied in the VALLEX dictionary (Lopatková et al., 2008). The framework for verbal valency was based on the division of verbal modifications into the class of participants (actants, arguments) and free modifications (adjuncts, circumstantials). The modifications determined by the empirical tests as participants enter the valency frame (for the tests, see the publications quoted above). For the labeling of the 1st and 2nd participants a modified Tesnière's approach is applied: if the verb has one participant, it is the Actor; if it has two participants, they are labeled as Actor and as Patient. In labeling the 3rd and other participants their semantics is taken into account. Valency frame is defined as a set of modifications classified as valency slots of the lexical item. Every modification satisfying the criteria for the participants enter the valency frame of the respective verb: they fill either an obligatory position (*vyžadovat co-ACC* [to require sth], *věřit komu-DAT* [to believe sb], *vzpomínat na koho-Prep-ACC* [to remember sb/sth] or an optional position[11] (*koupit někomu-DAT* něco [to buy *sb/sth* to somebody], *požadovat* něco *od někoho-Prep-GEN* [to ask sb *for sth*], *překvapit* někoho *něčím-INS* [to surprise sb *by sth*]).

In addition, the valency frame also contains such adjuncts that were determined by the above mentioned test as obligatory with the particular lexical item (*směřovat někam* [to lead up somewhere], *trvat jak dlouho* [to last how long], *tvářit se nějak* [to look somehow]). According to one of the important theoretical principles of this valency theory, an occurrence of the same lemma with different valency signals the ambiguity of the given lemma. This principle caused some difficulties for annotators during the annotation procedure. To overcome these difficulties the valency dictionary PDT-VALLEX (Hajič et al., 2003; Urešová, 2011b,a) was built as an on-line tool helping the annotators to check the existing valency frames and/or to add a new valency frame.

Some other additions needed to account for the complexity of the valency theory were stimulated by practical problems within the process of annotation. One of them is connected with the types of omissions of valency members on the surface without an influence on grammaticality.[12]

[11] Optional positions are denoted by italics.

[12] Here we do not have in mind an omission of a valency member conditioned by the textual deletions occurring esp. in dialogues.

An omission of a valency member has different reasons:[13]

(i) The participant is marked in the valency dictionary as optional and as such can be omitted.
(ii) The participant is obligatory, but its lexical setting is generalized.

The notion of generalization is interpreted as a group of persons/objects/circumstances typical/usual for this position. In (8a), (8b) and (9a), (9b) the differences between (a) and (b) sentences are connected with a filled valency position and a generalized valency position expressed by a null, respectively, and the verbs concerned represent one class of verbs with the deletion of an obligatory participant under special conditions. In (8b) and (9b) the generalized participants with a null form on the surface are interpreted as: *this dog does not bite anybody, Paul is able to read everything/any text*, respectively. In the tectogrammatical (deep) representation the positions of Patient (in (8b)) and Effect (in (9b)) are filled by the lemma #Gen, and in (8a) and (9a) all positions prescribed for the verbs *kousat* [bite] and *číst* [read] by their respective valency frames are used. In general, this phenomenon is known and described in linguistics as "an intransitive usage of transitive verbs", but a full description of the morphosyntactic conditions allowing for an omission of the participant is not usually taken into account. Perfective aspect of the verbs concerned[14] excludes the omission of a valency member as demonstrated by ex. (9c). The generalization of the valency member is supported by the morphological categories of gnomic present tense (often connected with the ability mood) and imperfective aspect (as in ex. (9b)).

(8a) Tenhle pes hodné lidi nekouše.
this dog-NOM-sg good people-ACC-pl not_bite-3-sg-PRS-IPFV
[This dog does not bite good people.]

(8b) Tenhle pes-NOM-sg nekouše.
this dog not_bite-3-sg-PRS-IPFV
[This dog does not bite.]

(9a) Pavel čte všechny nové romány.
Paul read-3-sg-PRS-IPFV all new novel-ACC-pl
[Paul reads all new novels.]

[13] We also leave aside here the zero subject position which is typical for Czech as a pro-drop language, because the comparison of overt and missing subjects represents a separate empirically non-trivial problem which we discuss elsewhere. For a detailed, theoretically based as well as empirically tested typology of the so called null subject languages, see Camacho (2013).

[14] In FGD and in VALLEX the aspectual pairs of verbs are understood as morphological forms of the same lexical unit.

(9b) Pavel už dobře čte.
Paul already well read-3-sg-PRS-IPFV
[Paul already reads well.]

(9c) *Pavel už dobře přečte.
Paul already well read-3-sg-PFV
[Paul already reads well.][15]

Examples of the difficulties connected with the annotation procedure representing another subclass of verbs allowing for generalization (in this case of Addressee) are given in (10). In (10b) and (10c) the noun expected as the filler of the valency position of Addressee is "generalized"; generalization of the Addressee is acceptable be this verb in the perfective as well as in the imperfective aspect. The realizations (10a), (10b), (10c) correspond to the verbal frame of the lexical entry for *prodat/prodávat* [sell-PFV/ sell-IPFV]: ACT (NOM), PAT_{Gen} (ACC), $ADDR_{Gen}$ (DAT). In the deep structure of (10b) the position of ADDR has the lemma #Gen. The lower index *Gen* assigned to the participants in the valency frame used here to demonstrate that the possibility to generalize this valency slot must be treated in the dictionary. In ex. (10c) both PAT and ADDR can be generalized (see Fig. 2), because they satisfy the conditions prescribed for a possible deletion if the verb is used in the form of gnomic present and imperfective aspect.[16]

(10a) Jan prodal auto sousedovi.
John-NOM sell-3-sg-PST-PFV car-ACC-sg neighbour-DAT-sg
[John sold his car to his neighbour.]

(10b) Jan prodává auto.
John-NOM sell-3-sg-PRS-IPFV car-ACC-sg
[John is selling his car.]

(10c) Lucie prodává v supermarketu.
Lucy-NOM sell-3-sg-PRS-IPFV in-PREP supermarket-LOC-sg
[Lucy sells in a supermarket.]

The missing Patient and Addressee in (10c) are understood as goods usually sold in the supermarkets to the usual customers of the supermarket, respectively. The generalized members are again filled into the deep syntactic representation with the lexical label #Gen.

[15] Strictly speaking, no translation can be assigned to (9c) different from that for (9b) because in English there is no equivalent of the perfective form of the Czech verb.

[16] An alternative solution would be the introduction of a new lexical unit for *prodávat* [sell] with the meaning *být prodavačem* [to be a shop assistant].

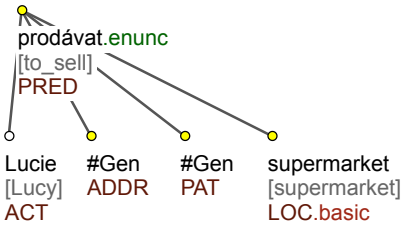


Figure 2. Sentence (10c): Lucie prodává v supermarketu.

The class of verbs with which the generalization of Patient is not limited to an imperfective form can be exemplified by (11), (12), though their valency frames contain an obligatory Patient.[17]

(11) Pokojská uklidila.
Chambermaid-NOM-sg clean-3-sg-PST-PFV
[The chambermaid has (already) cleaned.]

(12) Každé ráno ustelu a vyvětrám.
every morning make_a_bed-1-sg-PRS-PFV and ventilate-1-sg-PRS-PFV
[Every morning I make the bed and ventilate.]

Generalization is present also in the constructions with the possessive resultative, see (13) and (14), where the obligatory Patient in both sentences is generalized.

(13) Dnes máme vyprodáno.
today have-AUX-1-pl-PRS sell_out-PTCP-N-sg
[Today we are sold out.]

(14) Už mám zaplaceno.
already have-AUX-1-sg-PRS pay_for-PTCP-N-sg
[I have already paid my bill.]

The examples (8) through (14) illustrate that even though the theory of valency applied was formulated thoroughly, an extension of the theory is needed because of many empirical problems: the omissions (either connected with generalized valency positions or with other empirical issues) need an account of the restrictions on the morphological meanings available for the rules for deletion which influence the treatment of lexical entries as well as the syntactic structure.

[17] For a remark on the verb *vyvětrat* [ventilate], see (iii) below.

(iii) The omission of a verbal participant (usually the Patient) occurs also with verbs where the construction without the given valency member is close to the domain of phraseology or at least to the lexicalization of the verb, see (15), (16) and (17), where an explicit insertion of the Patient is either impossible or it does not bring novel information.

In variants (a) the Patient (*pivo* [beer], *prostor* [space] and *cigareta* [cigarette], respectively) is expressed overtly, in (b) PAT is missing. Ex. (15b) differs from the other two examples in the degree of lexicalization: the only reading of (15b) is *Janův otec je opilec* [John's father is a drunk]. In (16b) and (17b) the empty position for the valency member can be easily filled by a noun, which is semantically restricted excluding a free choice of a filler for the Patient.

(15a) Janův otec pije hodně pivo.
John's father-NOM-sg drink-3-sg-PRS very_much beer-ACC-sg
[John's father drinks beer very much.]

(15b) Janův otec hodně pije.
John's father-NOM-sg very_much drink-3-sg-PRS
[John's father drinks a lot.]

(16a) Po požáru vyvětrali všechny prostory.
After-PREP fire-LOC-sg ventilate-3-pl-PST all space-ACC-pl
[After the fire they have ventilated all spaces.]

(16b) V pokoji bude příjemněji, až
In-PREP room-LOC-sg be-3-sg-FUT pleasant-ADV-ALL after-CONJ
vyvětráš.
ventilate-2-sg-FUT-PFV
[It will be more pleasant in the room after you ventilate.]

(17a) Pohodlně se usadila a zapálila si
comfortably REFL-ACC sit_down-3-sg-PST-F and light-3-sg-PST-F REFL-DAT
cigaretu.
cigarette-ACC-sg
[She sat down comfortably and lighted a cigarette.]

(17b) Zapálila si a začala vyprávět své
light-3-sg-PST-F REFL-DAT and start-3-sg-PST-F relate-INF her
zážitky.
experience-ACC-pl
[lit. She lighted and started to relate her experiences.]

	Total	Generalized	
ACT(or)	87,118	6,910	7.9%
PAT(ient)	68,030	2,574	3.8%
ADDR(ressee)	10,150	3,640	35.9%
EFF(ect)	7,207	178	2.5%
ORIG(in)	847	30	0.4%

Table 2. Frequencies of participants and their generalization

Examples (15) through (17) point again to the necessity of cooperation between the lexical and the syntactic modules of the language description. Any type analyzed here needs a subtle treatment in the lexicon in order to offer a solid basis for sentence generation. The technical implementation of the new results reflecting the conditions for deletions of valency positions is in progress.

In Table 2, we present the frequency of particular participants depending on a verb as attested in PDT 3.0. Numbers in the first column correspond to all occurrences of the participant with a verbal head, in the second and third columns their generalized position is indicated.

3.4. Introduction of the notion of "quasi-valency"

During the extended studies of empirical data relevant for valency we have come across modifications that have properties typical for the class of participants ((i) through (iii)) as well as those typical for the class of free modifications ((iv) and (v)):

(i) they occur with a limited class of verbs
(ii) their morphological forms are given by their head
(iii) they cannot be repeated with a single verb occurrence
(iv) they have a specific semantics, contrary to the Actor, Patient and Effect (the semantics of which is usually heterogeneous)
(v) they are mostly optional

On the basis of these properties new functors were introduced: the modifiers Obstacle (OBST) and Mediator (MED) represent a more subtle division of the general modification of Means/Instrument.

(18a) Jan zakopl nohou o stůl
John-NOM stumble-3-sg-PST leg-INS-sg over-PREP table-ACC-sg
[John stumbled over the table *with his leg*.]

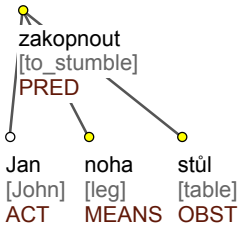


Figure 3. Sentence (18a): Jan zakopl nohou o stůl.

(18b) Matka se píchla nůžkami.
Mother-NOM-sg REFL-ACC prick-3-sg-PST scissors-INS-sg-PAIR/GROUP
[Mother pricked herself *with the scissors*.]

(18c) Šípková Růženka se píchla o trn.
Sleeping Beauty-NOM REFL-ACC prick-3-sg-PST by-PREP thorn-ACC-sg
[Sleeping Beauty pricked herself *by a thorn*.]

In (18a) *noha* [the leg] is a proper Means (Instrument), while the construction *o stůl* [over the table] is rather an Obstacle (see Fig. 3). Similar considerations concern the construction *o trn* [by a thorn] in (18c), which is also classified as an Obstacle. In (18b) *nůžky* [scissors] functions as an Instrument in the proper sense, its semantics implies the semantics of handling this instrument (which implies its movement). In (18b) a manipulation with scissors is supposed, while in (18a) and (18c) the referent of the noun stays fixed. The feature of an unconscious action is typical of (18a) and (18c), while in (18b) the action can be either conscious or unconscious.

Up to now, we have found only one Czech verb with an obligatory Obstacle (*zavadit* [to brush against]); otherwise with verbs listed in the dictionary as compatible with OBST this modification is optional.

Another semantic specification of the Instrument is expressed in Czech by the prepositional group *za + ACC*; we proposed to call it Mediator (see ex. (19)).

(19) Jan přivedl psa za obojek.
John-NOM bring-3-sg-PST dog-ACC-sg by-PREP collar-ACC-sg
[John brought the dog by its collar.]

In example (20), the ear of the boy is understood to be an object that mediates the contact between father's left hand and the boy. A supportive argument for the distinction to be made between the "classical" Instrument (*ruka* [hand]) and the Mediator

(*ucho* [ear]) is the fact that the Instrument and the Mediator *ucho* [ear] can co-occur in a single sentence.[18]

(20) Otec chytil kluka levou rukou za
Father-NOM-sg catch-3-sg-PRF boy-ACC-sg left-INS-sg hand-INS-sg by-PREP
ucho.
ear-ACC-sg
[Father has caught the boy's *ear* by his left *hand*.]

Because of the introduction of the class of quasi-valency modifiers into the formal framework of FGD the list of functors (semantic relations) originally used was checked and as the consequence of these observations a new list of valency members was provided: the modifiers of Intention (INTT) and Difference (DIFF) were shifted from the list of free modifiers into the list of quasi-valency members. For the modifier of Intention, see (21) and for the modifier of Difference, see (22):[19]

(21) Jan jel navštívit svou tetu.
John-NOM went-3-sg-PST visit-INF his-POSS aunt-ACC-sg
[lit. John left *to visit* his aunt.]

(22) Náš tým zvítězil o dvě branky.
our-POSS team-NOM-sg win-3-sg-PST by-PREP two-ACC goal-ACC-pl
[Our team won *by two goals*.]

3.5. Selected types of deletions

According to the annotation scenario for the surface layer of annotation in PDT only elements present on the surface are represented by separate nodes in the dependency tree. However, there are elements obviously missing for the complete meaning of the sentence. The following technical solution for such cases was proposed for the surface (analytical) layer of annotation: if the governor of some member of the sentence is not present, the syntactic function of this member receives the value ExD (with meaning "extra-dependency"). The nodes with this value are an excellent challenge for the studies of deletions (ellipsis) which must be reconstructed in the deep (tectogrammatical) structure.

In this section we present only selected types of grammatical deletions conditioned or even required by the grammatical system of language.[20] One special type of dele-

[18] See also Fillmore (1977), quoted from Fillmore (2003, p. 189): "A reason for feeling sure that two roles are distinct is that the same two nouns, preserving their case roles, can also occur together … in a single sentence."

[19] A detailed analysis and argumentation for these modifiers is given in Panevová et al. (2014).

[20] For a detailed discussion on the reconstruction of deletions, see Hajič et al. (2015) and Hajičová et al. (2015).

tions, namely the surface deletion of valency members, was analyzed in more details above in Sect. 3.3. Here we want to comment upon some complicated cases of deletions.

Comparison structures are a very well known problem for any language description aiming at a representation of the semantic (deep/underlying) structure. These considerations concern the comparison with the meaning of equivalence (introduced usually by the expression *jako* [as]; the subfunctor we use has the label 'basic') and the comparison with the meaning of difference (introduced usually by the conjunction *než* [than]; the subfunctor is called 'than').[21]

There are some comparison structures where the restoration of elements missing on the surface seems to be easy enough from the point of view of semantics (see (23a) and its restored version (23b), but most comparisons are more complicated, see (24) through (26):

(23a) Jan čte stejné knihy jako jeho
John-NOM read-3-sg-PRS same-ACC-pl book-ACC-pl as-CONJ his-POSS
kamarád.
friend-NOM-sg
[John reads the same books as his friend.]

(23b) Jan čte stejné knihy jako (čte)
John-NOM read-3-sg-PRS same-ACC-pl book-ACC-pl as-CONJ (read-3-sg-PRS)
(knihy) jeho kamarád.
(book-ACC-pl) his-POSS friend-NOM-sg
[John reads the same books as his friend (*reads books*).]

The introduction of the deleted elements into (24a) seems to be as easy as in (23b), however, for the expansion of "small clauses" expressing comparison illustrated by ex. (24b) such a solution is not sufficient: (24b) is not synonymous with (24a). More complicated expansion for (24b) is proposed and exemplified by (24c) as its deep structure counterpart.

(24a) Jan žije na vesnici stejně pohodlně
John-NOM live-3-sg-PRS in-PREP village-LOC-sg same-ADV comfortably-ADV
jako jeho rodiče.
as-CONJ his-POSS parents-NOM-sg
[John lives in the country as comfortably as his parents.]

[21] More simple comparative structures expressed by secondary prepositions with nouns (such as *na rozdíl od* [in contrast to], *ve srovnání s* [in comparison with], *proti* [against], e.g. in *Ve srovnání s minulým rokem je letos úroda brambor vyšší* [Lit. In comparison with the last year the crop of potatoes is in this year higher] are left aside here.

(24b) Jan žije na vesnici stejně pohodlně
John-NOM live-3-sg-PRS in-PREP village-LOC-sg same-ADV comfortably-ADV
jako u svých rodičů.
as-CONJ with-PREP his-POSS parents-GEN-sg
[John lives in the village comfortably as well as with his parents.]

(24c) Jan žije na vesnici stejně pohodlně
John-NOM live-3-sg-PRS in-PREP village-LOC-sg same-ADV comfortably-ADV
jako (Jan) (žít) (nějak) u svých
as-PREP (John-NOM) (live-3-sg-PRS) (some way-ADV) with/PREP his-POSS
rodičů.
parents-GEN-sg
[John lives in the village comfortably as well as he lives (somehow) with his parents.]

The compared members in (24a) and (24b) are not apparently of the same sort: the two modifications are collapsed in a single "small clause". This phenomenon contradicts the notation used in dependency based representations in FGD: the two functions (comparison and location) could not be assigned to the single node introduced by the comparison construction.

Though some extensions of the embedded predication (e.g. (24c), (26b)) do not sound natural, they represent only a theoretical construct required by the shortened surface shape (for details, see Panevová and Mikulová 2012). In the deep structure of (24c), the inserted node *žít* [to live] is labeled as CPR (comparison) and the node *rodiče* [parents] bears the functor LOC [location] depending on the restored node governing the comparison (*žít* [to live] in this case). While in (24a) the way of John's life in the country is compared with the identical way of his parents' life there, in (24b) John's life in the country is compared with the way of his (respective) life with his parents. John's way of life is presented as comfortable in the main clause, so his life with his parents may be assumed to be comfortable as well, however this assumption is not expressed explicitly. Therefore the adverbial specifying the way of life in the reconstructed representation is denoted by the underspecified artificial node *nějak* [in some way] rather than by a repetition of the lexical value *pohodlně* [comfortably]. In the tectogrammatical (deep) structure of (24c) the inserted node *žít/žije* [live] is labeled as comparison (CPR) and depends on the lexically identical predicate of the main clause, while *u rodičů* [with the parents] and *nějak* [in some way] are its children labeled as location (LOC) and manner (MANN), respectively.[22]

Examples (25) and (26) support the arguments presented for (24): (i) expansion of the surface shape of comparison structure is necessary, and (ii) fuzzy artificial lemmas

[22] For *nějak* [in some way] the artificial lemma #Some is used in PDT.

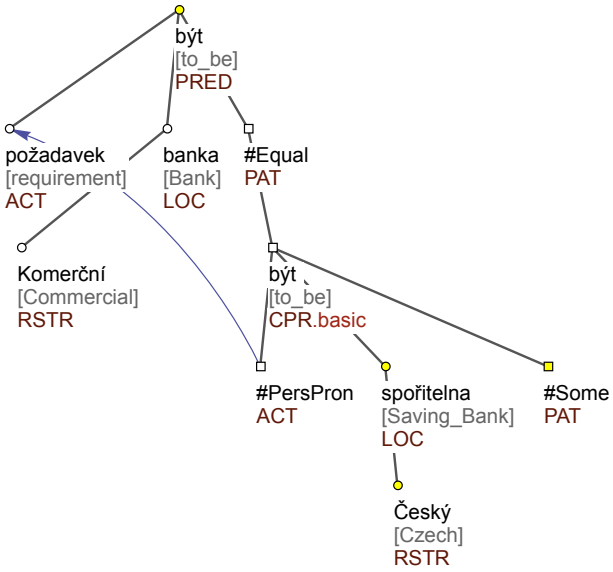


Figure 4. Sentence (25b): Požadavky u Komerční banky jsou jako u České spořitelny.

are introduced because of their lexical underspecification. Two types of comparison (identity in (25) and diference in (26)) are exemplified by the (25) and (26) as well.

(25a) Požadavky u Komerční banky jsou jako
requirement-NOM-pl in-PREP Commercial Bank-GEN-sg be-3-pl-PRS as-CONJ
u České spořitelny.
in-PREP Czech Saving Bank-GEN-sg
[lit. The requirements in Commercial Bank are as in Czech Saving Bank.]

(25b) Požadavky u Komerční banky jsou (stejné)
requirement-NOM-pl in-PREP Commercial Bank-GEN-sg be-3-pl-PRS (same)
jako (jsou požadavky) u
as-CONJ (be-3-pl-PRS requirement-NOM-pl) in-PREP
České spořitelny (nějaké-#Some)
Czech Saving Bank-GEN-sg (some-#Some)
[lit. The requirements in Commercial Bank are (the same) as (are the requirements) in Czech Saving Bank.]

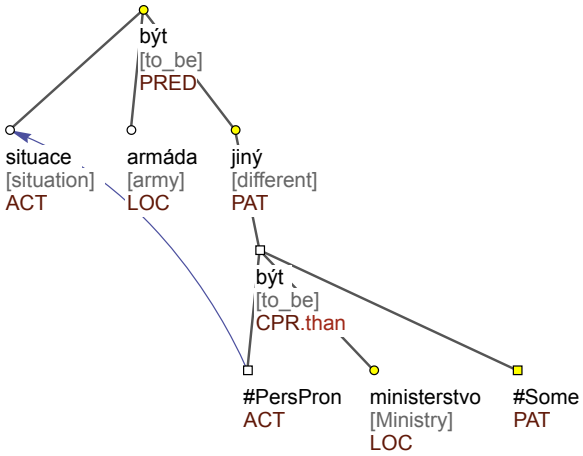


Figure 5. Sentence (26b): Situace v armádě je jiná než na ministerstvu.

(26a) Situace v armádě je jiná
situation-NOM-sg-F in-PREP army-LOC-sg be-3-sg-PRS different-NOM-sg-F
než na ministerstvu.
than-CONJ at-PREP Ministry-LOC-sg
[lit. The situation in the army is different than at the Ministry.]

(26b) Situace v armádě je jiná
situation-NOM-sg-F in-PREP army-LOC-sg be-3-sg-PRS different-NOM-sg-F
než (je situace) na ministerstvu
than-CONJ (be-3-sg-PRS situation-NOM-sg-F) at-PREP Ministry-LOC-sg
(nějaká-#Some)
(some-#Some)
[lit. The situation in the army is different than (the situation) at the Ministry is (some).]

Also the analysis of other types of adverbials points to the possibility to restore in the deep structure representation a whole embedded predication (e. g. adverbial phrases introduced by the expressions *kromě* [except for, besides], *místo* [instead of]). In the surface structure there are again two types of modifications, see (27a), where the adverbial of direction is embedded into the adverbial of substitution sharing the same predicate on the surface. As we have mentioned above when analyzing complex comparisons, the FGD framework does not allow for an assignment of more than a single

function to a single sentence member. The extension of this predication on the underlying level is illustrated by (27b):

(27a) Místo do Prahy přijel Jan do
instead-of-PREP at-PREP Prague-GEN arrive-3-sg-PST John-NOM at-PREP
Vídně.
Vienna-GEN
[Instead of arriving at Prague, John arrived at Vienna.]

(27b) Místo toho, aby přijel do
instead-of-PREP that-PRON AUX-3-sg-COND arrive-3-sg-PST at-PREP
Prahy, přijel Jan do Vídně.
Prague-GEN arrive-3-sg-PST John-NOM at-PREP Vienna-GEN
[Instead of arriving at Prague, John arrived at Vienna.]

From the point of view of their surface form, these deletions are not as transparent as e.g. dropped subject or comparison constructions, but the difficulties the annotators had during the annotation procedure stimulated a more detailed analysis sketched briefly above and presented in detail in Panevová et al. (2014).

3.6. Non-nominative subjects

The non-nominative subjects are the topic of many typologically oriented studies (see e. g. Bhaskararao and Subbarao, 2004). The fact that in some languages the subject is expressed by the dative, genitive and other forms is well known. Such marginal forms are present in Czech as well, where prototypical subjects have the form of nominative. The traditional term "dative subject" is applicable for the Czech examples as (28) and (29), in the deep structure of which the dative is understood as the Actor; a similar structure is assigned to the sentences where nominative is present, but it is not understood as an Actor, see (30), due to the semantic parallel structure with different formal exponents (see (31)).

This solution corresponds the theory of valency used in FGD: Any verb has in its valency frame in the lexicon a slot for the Actor (1st actant according to Tesnière, 1959). Actor is prototypicaly expressed by Nominative, however there are two types of exceptions: either the verb has an unprototypical patterning of its valency complementations (see ex. (28) – (31)), or the participant of Actor is stylistically or semantically modified (see ex. (32) – (35); semantic modifications of Actor are represented by the subfunctors of the Actor.

(28) Je mu smutno.
be-3-sg-PRS he-DAT-M-sg-Sbj sad-ADV
[He is sad.]

(29) V Praze se rodičům líbí.
in Prague REFL-ACC parents-DAT-Sbj like-3-sg-PRS-ACT
[My parents like Prague.]

(30) Bolí mě hlava.
ache-3-sg-PRS-ACT I-ACC head-NOM-Sbj
[I have a headache.]

(31) Bolí mě v krku.
ache-3-sg-PRS-ACT I-ACC-Sbj in-PREP throat-LOC-sg
[I have a sore throat.]

The genitive subject occurs in Czech sentences as a stylistic variant of the nominative, see (32) and (33), where the negative genitive and partitive genitive, respectively, are used. The genitive forms, however, carry some additional semantic information with respect to the unmarked nominative form, but in contemporary Czech they are accepted as a little bit archaic, therefore they are rare in the PDT. The semantic contribution to the unmarked nominative forms is expressed by the introduction of "subfunctors" (rendering semantic variations of the nominative subject/Actor); in addition new semantic shades of the construction (formally rendered by the value of subfunctors) are expressed by some prepositions, see (34) displayed in Fig. 6 and (35).

(32) Z vyhlazovacích táborů nebylo úniku.
from-PREP extermination camp-GEN-pl not_be-3-sg-PST-N escape-GEN-sg-Sbj
[From the extermination camps there was no escape.]

(33) Přibývá podnikatelů, kteří nemají kancelář a
increase enterpreneur-GEN-pl-Sbj who not_have-3-pl-PRS office-ACC-sg and
podnikají doma.
do_business-3-pl-PRS home
[The number of entrepreneurs who have no office and who do business from their homes increases.]

(34) Své expozice bude mít okolo 60
their exposition-ACC-pl be-3-sg-FUT have-INF approximately-PREP 60
stavebních firem.
building firm-GEN-pl-Sbj
[Approximately 60 building firms will have their own expositions.]

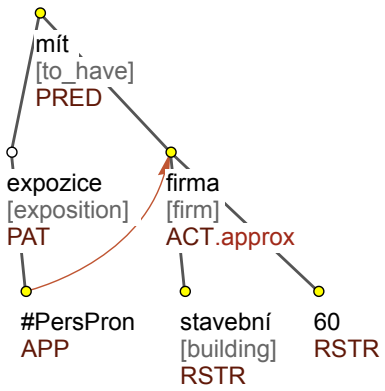


Figure 6. Sentence (34): Své expozice bude mít okolo 60 stavebních firem.

(35) Za každou stranu přišlo po pěti
for-PREP every party-ACC-sg come-3-sg-PST-N by-PREP five-LOC
delegátech.
deputy-LOC-pl-Sbj
[Every party was represented by five deputies.]

An approximate amount of firms is expressed in (34) by the preposition *okolo/kolem, na* + accusative [around], the distributive meaning is expressed in (35) by the prepositional case *po* + Locative [by]. The subfunctors *approximity, distributivity* corresponding to these meanings were introduced into the list of subclassified meanings of the main syntactic functors (Actor in this case).

In Sections 3.1 to 3.6 we presented examples of grammatical phenomena which either were not yet described explicitly or were not described at all. Some of these issues are known, but their consequences for a consistent description have not yet been fully considered. Some of these results are reflected in the annotation guidelines and all of them enriched the theoretical description of Czech grammar.

4. Case studies II: Information structure of the sentence, discourse relations and coreference

4.1. Topic–focus annotation in the Czech corpus

In the theoretical account of topic–focus articulation (TFA in the sequel, see e.g. Sgall, 1967; Sgall et al., 1973, 1980; Hajičová et al., 1998) within the framework of the Functional Generative Description, the dichotomy of topic and focus – which divides the sentence into what the sentence is about (its topic) and what it says about the topic

(its focus) – is understood as based on the primary notion of contextual boundness. The TFA information/feature is an integral part of the representation of sentences on the underlying (tectogrammatical) sentence structure, since TFA is semantically relevant.[23] Every node of the tectogrammatical dependency tree carries – in addition to other characteristics such as the tectogrammatical lemma, the type of dependency or (morphological) grammatemes – an index of contextual boundness: a node can be either contextually bound or non-bound. This feature, however, does not necessarily mean that the entity is known from a previous context or new but rather how the sentence is structured by the speaker as for the information structure. Thus, for example, in the second sentence of the discourse segment *When we walked around the town, we met Paul and his wife. I immediately recognized HIM, but not HER* (capitals denoting the intonation center, if the sentences are pronounced), both Paul and his wife are mentioned in the previous context, but the sentence is structured as if they are a piece of non-identifiable information, i.e, marked as contextually non-bound. Contrary to that, the above segment from the information structure point of view can also be structured in a different way, which, in the surface form of the sentence in English, would involve a different placement of the intonation center: *When we walked around the town, we met Paul and his wife. I immediately RECOGNIZED him*. In this segment, both Paul and his wife are also introduced in the first sentence, but in the second sentence, it is the event of recognizing which is structured as bringing 'new' (non-identifiable) information, while Paul – being referred to by a non-stressed pronoun – is taken as contextually bound (identifiable). In Czech, the two situations would be expressed by a different word order and different forms of the pronoun corresponding to English *him*, namely *jeho* vs. *ho*: *Hned jsem poznal JEHO* versus *Hned jsem ho POZNAL.*

The annotation of PDT follows the theoretical description rather closely: to each node of the dependency tree on the tectogrammatical layer of PDT a special attribute of TFA is assigned which may obtain one of the three values: *t* for a non-contrastive contextually bound node, *c* for a contrastive contextually bound node and *f* for a contextually non-bound node. [24]

The left-to-right dimension of a tectogrammatical tree serves as the basis for the specification of the scale of communicative dynamism: communicative dynamism is specified as the deep word order, with the dynamically lowest element standing in the

[23] The semantic relevance of TFA has been documented in the writings quoted above and elsewhere by such sentences differing only in their TFA structure as *I work on my dissertation on SUNDAYS.* vs. *On Sundays, I work on my DISSERTATION.*, or *English is spoken in the SHETLANDS.* vs. *In the Shetlands ENGLISH is spoken.*, or *Dogs must be CARRIED.* vs. *DOGS must be carried.*, etc. The difference in TFA is expressed, in the surface structure, either by word order (as in Czech), or by a different position of the intonation centre denoted here by capitals (this holds both for Czech and for English) or even by some specific sentence structure (e.g., cleft constructions in English).

[24] There are 206,537 tectogrammatical nodes annotated as contextually bound, out of them 30,312 are contrastively contextually bound. Further, 354,841 nodes are contextually non-bound and for 38,493 nodes (and for 2,849 technical roots), contextual boundness is not annotated (e.g., for coordinating nodes).

leftmost position and the most dynamic element (the focus proper of the sentence) as the rightmost element of the dependency tree.

4.1.1. The identification of the boundary between topic and focus

For the identification of the dichotomy of topic and focus (which is supposed to be very important especially for the specification of the scope of negation) on the basis of the information on contextual boundness for each node, a rather strong hypothesis was formulated, namely that the topic-focus distinction can be made depending on the status of the main verb (i.e. the root) of the sentence and its immediate dependents: Basically, 1. if the verb is contextually bound (*t*, *c*) then the verb and all the contextually bound nodes depending immediately on the verb and all nodes subordinated to these nodes constitute the topic, the rest of the sentence belonging to its focus; 2. if the verb is contextually non-bound (*f*), then the verb and all the non-bound nodes immediately depending on it and all nodes subordinated to these nodes constitute the focus, the rest of the sentence belonging to its topic; 3. if both the main verb and all nodes immediately depending on the main verb are contextually bound, then follow the rightmost edge leading from the main verb to the first node(s) on this path that are contextually non-bound; this/these node(s) and all the nodes subordinated to it/them belong to focus (see the definition of topic and focus by Sgall, 1979, see also Sgall et al., 1986, 216f).

To test this hypothesis on the PDT data, we have proceeded in three steps:

(i) a minor modification and implementation of the algorithm so that it can be applied to the data of the whole PDT,
(ii) manual parallel annotation of the control raw data as for the topic and focus of the individual sentences,
(iii) comparison of the values obtainded from the manual annotation with the automatically assigned Topic–Focus bipartition and evaluation of the results.

The results of the implementation of the modified algorithm indicate that a clear division of the sentence into topic and focus according to the hypothesized rules has been achieved in 94.28% of sentences to which the procedure has been applied; 4.41% of sentences contained the so-called proxy focus (itself a part of topic but a part that has the focus subordinated to it).[25] The real problem of the algorithm then rests with

[25] More exactly, proxy focus is a node A such that A is contextually bound, A differs from the main verb and the focus of the sentence is subordinated to A. The introduction of the notion of proxy focus was invoked to handle cases where the focus of the sentence is so deeply embedded that it does not include the verb or any of its immediate dependents (see Hajičová et al., 1998). Thus in *I met the teacher of CHEMISTRY* as an answer to *Which teacher did you meet yesterday?* the focus *chemistry* depends on a head (*teacher*) that has a specific status, it is a proxy focus: it is contextually bound and thus does not belong to the focus; however, it is the only part of the upper subtree of the sentence that lies on the path from the root of the tree (the verb) to the focus.

the case of ambiguous partition (1.14%) and cases where no focus was recognized (0.11%) as the assumption of the TFA theory is that all sentences should contain focus (though there may be topicless sentences, e.g., those that bring hot news: *KENNEDY was assassinated*!) but this is a very small part of the data analyzed.

However, in order to validate the hypothesis it is necessary to compare the results achieved by the automatic identification of topic and focus with the judgements of Czech speakers (step (ii) above). For the control annotation, PDT documents comprising a total of 11,000 sentences have been analyzed manually, most of them in three parallel annotations (about 10,000 sentences), and about 600 sentences in six parallel annotations (a detailed description of the project is given in Zikánová et al., 2007; we present here a brief summary of the methodology used and the results). The annotators were mostly high school students, having some (common sense) basic idea of the dichotomy of topic and focus (as "the aboutness relation") but were not familiar with the theoretical framework TFA is based on. They worked with the raw texts (i.e. without any annotation) and were instructed to mark – according to their understanding – every single word in the sentence as belonging either to topic or to focus; they were supposed to take nominal groups as an integrated element and they were also told that they may assign all the elements of the sentences to the focus. At the same time, they were supposed to mark which part of the sentence they understand as topic and which part as focus. In subordinated clauses and in coordinated constructions they were asked to mark each clause separately. One of the important subtasks of this project was to follow annotators' agreement/disagreement. The disagreement in the assignments of the two parts of the sentence as a whole was rather high and indicates that the intuitions concerning the division of the sentence into its topic and focus parts may dramatically differ. However, it is interesting to note that the annotators' agreement in the assignments of individual words in the sentences to topic or to focus was much higher (about 75% in both the three and six parallel analyses compared to 36% of the assignments of the topic and the focus as a whole) than the assignments of the topic–focus boundary.

The work on the step (iii) is still in progress. It is a matter of course that in that step, the variability of manual solutions must be taken into considerations; the annotators were asked to assign a single, most plausible TFA annotation, different annotators for the same text may have chosen a different interpretation. We are aware of the fact that while we get only a single, unambiguous result from the automatic procedure, more ways of interpretation could be possible. This mostly occurs with the assignment of the verb: actually, it is the assignment of the verb to topic or to focus, in which the annotators differed most frequently.[26]

[26] See the discussion in K. Rysová et al. (2015a). It should be added that no machine learning methods for TFA assignment have been considered so far.

4.1.2. Systemic ordering as the order of elements in the focus

The empirical study of Czech texts has led to the assumption (Sgall et al., 1980, p. 69) that the ordering of the elements in the focus part of the sentence is primarily given by the type of the complementation of the verb. This assumption resulted in a rather strong hypothesis called systemic ordering of the elements in the focus of the sentence. The hypothesis was empirically tested pairwise (i.e., successively for two of the complementation types) and it was also supported by several psycholinguistic experiments (Sgall et al., 1980, p. 72ff; Preinhaelterová, 1997). The following ordering has been established for Czech:

Actor – Temporal (when – since when – till when – how long) – Location (where) – Manner – Extent – Measure – Means – Addressee – From where – Patient – To where – Effect – Condition – Aim – Cause

Even at the time of the formulation of the hypothesis, several accompanying assumptions were taken into account:

(i) It was assumed that systemic ordering is a universal phenomenon and that at least in most European languages the order of the principle verb complementations (such as Actor – Addressee – Patient) is the same, which was also attested by experiments for English and German; at the same time it was clear that languages may differ in the (underlying) order of the particular elements.
(ii) It was understood that there are several factors that may influence the underlying order in focus such as the rhythmical factor (short complementation before the longer one), or the lexical meaning of some verbs which may be associated more closely with a certain type of complementation (e.g., the verb *pay* in construction with Patient: *pay the debts*); such a construction may have a character of a phraseological expression (*to pave the way*, *to make claims*, etc.).
(iii) In the original formulation no difference was made between sentential and non-sentential structures expressing the given complementation. This difference certainly influences the ordering and has to be taken into account.
(iv) The question has remained open as for the character of the ordering: does each complementation have a separate position in the scale or is it the case that more than a single type of complementation occupy a given position on this scale?
(v) It was clear from the very beginning that the hypothesis of systemic ordering is very strong and that in spite of the fact that it was based on the examination of hundreds of examples, further investigation based on a much broader material is needed, which may lead to a more precise specification or modification(s), as is the case with all empirical statements.

The material of the Prague Dependency Treebank opened the possibility to validate the hypothesis. After the first attempts made by Zikánová (2006), a deeper and a more complex analysis is presented by K. Rysová (2014a), who arrives at several interesting

and important observations summarized in the sequel. 1. First of all, she confirms that the sentential character of a complementation is a very important factor in that there is a tendency of a contextually non-bound element expressed by a clause to follow the non-sentential element (which is apparently connected with the 'weight' of the element mentioned above in point (iii)). 2. She also points out the influence of the form of the complementation: the assumed order Manner – Patient is more frequent if the complementation of Manner is expressed by an adverb and the complementation of Patient by a nominal group.[27] 3. When examining the position of the Actor on the scale, a substantial number of counterexamples of the original hypothesis (with the position of Actor at the beginning of the scale) concern cases for which the outer form of the Actor plays an important role: in sentences with the verb *být* (to be) in structures of the type *je nutné* (PAT) *přiznat* (ACT) (*it is necessary to acknowledge*), where Actor is expressed by infinitive, Patient precedes Actor, while the hypothesized order Actor – Patient is attested to if both complementations are expressed by nominal groups.

Rysová's analysis (using the PDT material with the manual annotation) is based on examples where there are two complementations in the focus of the sentence; her analysis confirms that there is a considerable tendency that in such pairs one ordering prevails over the other, which, as a matter of fact, was the starting point of the postulation of the systemic ordering hypothesis. However, with some pairs, such as Patient and Means, there was a balance between the frequency of the two possible orders, which may indicate that for some particular complementations more than a single complementation occupy one position on the scale (see point (iv) above). She also mentions the possibility that the order might be influenced by the valency characteristics of the verbs, namely by the difference in the optional/obligatory character of the given complementations: she assumes that there is a tendency that obligatory complementations seem to follow the optional ones, but she admits that this tendency is not a very influential word order factor.

Rysová observes that in some cases the decisions of the annotators are not the only possible ones and that this fact has to be taken into consideration when drawing conclusions. This observation is confirmed also by the data on the annotators' agreement/disagreement, see also Veselá et al. (2004) or Zikánová (2008) and below in Section 5.

4.1.3. Rhematizers (focusing particles, focalizers)

A specific function of certain particles from the point of view of a bipartitioning of the sentence was noted first by Firbas (1957) in connection with his observation of a specific rhematizing function of the adverb *even.* It should also be mentioned at this point

[27] As one example for all, let us mention a combination of a node with the functor MANN and a node with functor PAT, both contextually non-bound and directly depending on a node with a verbal semantic part of speech. There are 1,111 such cases, in 933 out of them, MANN precedes PAT in the surface order (in agreement with the systemic ordering), in 174 cases MANN follows PAT

that a semantic impact of the position of several kinds of adverbials and quantifiers was substantiated already by Sgall (1967), who exemplifies the semantic relevance of topic/focus articulation on the English quantifier *mostly*. Sgall's argumentation is followed by Koktová (1999, but also in her previous papers), who distinguishes a specific class of adverbials called attitudinal.

The same class of words was studied later in the context of formal semantics by Rooth (1985) in relation to the prosodic prominence of the words that followed them; he called this class 'focalizers'.

Both terms – rhematizer and focalizer – refer to the apparent function of these particles, namely as being 'associated' with the focus of the sentence; the position of the focalizer (and the accompanying placement of the intonation center) indicates which reading of the sentence is being chosen from the set of alternatives. However, the assumption of such an exclusive function of these particles has been found to be too simplistic, an analogy with a semantic analysis of negation was claimed to be a more adequate approach (Hajičová, 1995). A distinction has been made between 'the (global) focus' of the sentence and 'the focus' of the focalizer (specified as the part of the sentence that follows the focalizer) by Hajičová et al. (1998). Comparing the analysis of the semantic scope of negation and the analysis of the function of focalizers, it is necessary to also consider the possibility of a secondary interpretation of the position of the focalizers. This issue was demonstrated in examples such as *JOHN criticized even Mother Teresa as a tool of the capitalists*. This sentence may occur in a context illustrated by the question *Who criticized even MOTHER TERESA as a tool of the capitalists?* The predicate of the indicative sentence *criticized even Mother Teresa as a tool of the capitalists* is repeated from the question; the only part of this sentence that stands in the focus is *John* (with a paraphrase 'the person who criticized even Mother Teresa as a tool of capitalists was John'). Such an understanding would compare well with the sometimes indicated recursivity of topic/focus articulation.

Based on the observations on the scope of focalizers as reflected in PDT and a similarly based annotation of English in the so-called Prague English Dependency Treebank (see Cinková et al., 2009), some complicated (and intricate) cases have been singled out, concerning first of all the occurrence of focalizers with a restricted freedom of position, with a distant placement of focalizers and their possible postpositions, and the semantic scope of focalizers. The function and the diversity of expressions originally called rhematizers has been studied in detail by Štěpánková (2013).

It is interesting to notice that contrary to the general characteristics of Czech as a language with a relatively "free" word order (i.e. without grammatical word-order restrictions), in the placement of the focalizer *only* English is more flexible than Czech is: this particle can be placed either immediately before the element it is 'associated with' or between the subject and the verb in English.

In Czech, a backward scope of focalizers is not that frequent as in English, but it is also possible. For example, the intonation center in the sentence quoted here from the Prague Czech–English Dependency Treebank as (36), if pronounced, would

be placed on the word *inflation* (as indicated here by capitalization); the postposited focalizer *only* having its scope to the left. In the Czech translation of this sentence, the focalizer *jen* (only) has to be placed in front of the focused element. It is interesting to note that there was a single example of a backward scope of a rhematizer in the whole of the PDT.

(36) Scénář 1, známý jako „konstantní zmrazení dolaru", nahrazuje Pentagonu výdaje jen kvůli INFLACI.
[Scenario 1, known as the "Constant Dollar Freeze", reimburses the Pentagon for INFLATION only.]

Štěpánková's comprehensive and detailed analysis (Štěpánková, 2013) based on the PDT material demonstrates that the class of focalizers is larger than originally (and usually) assumed; properties similar to those of 'prototypical' focalizers *only*, *even*, *also* are evident also with *alone*, *as well*, *at least*, *especially*, *either*, *exactly*, *in addition*, *in particular*, *just*, *merely*, *let alone*, *likewise*, *so much as*, *solely*, *still/much less*, *purely*, and several others (prototypical Czech rhematizers are *pouze*, *jen*, *jenom*, *zejména*, *zvláště*, *především*, *obzvlášť*, *hlavně*, *jedině*, *například*, *toliko*, *ne*, *ano*, *výhradně*, *výlučně*). Even more importantly, her material provides evidence that according to the context in which they are used, these elements are ambiguous and may obtain functions other than that of a focalizer. Table 3 quoted from Štěpánková's dissertation (Štěpánková, 2013) based on the Czech data from PDT illustrates the ambiguity of a rhematizer obtaining also a function that is classified as a free modification (adverbial modifier).

Expressions that function in some contexts as rhematizers may also obtain – in other contexts – an attitudinal function, especially in cases when the given expression relates to the whole sentence irrespective of the position in which it occurs in the surface shape of the sentence, see the difference between (37) and (38). In (37), the expression *třeba* functions as an adverbial of attitude (ATT) (translated to E. *maybe*), in (38) the same expression obtains the function of a rhematizer (translated to E. *for instance*).

(37) Třeba.ATT Honza se tam bude nudit.
[Maybe Honza will feel bored.]

(38) Třeba.RHEM HONZA se tam bude nudit.
[For instance HONZA will feel bored.]

Examples of such ambiguous expressions in Czech are *to*, *leda*, *též*, *rovněž*, *také*, *taktéž*, *zároveň*, *prakticky*, *spíše*, *třeba* (in English a similar homonymy concerns expressions such as *only*, *at best*, *also*, *at the same time*, *practically*, *rather*, *maybe*, …).

One specific issue connected with the analysis of constructions with rhematizers is the scope of rhematizers. Since the scope is relevant for the meaning of the sentence, it must be possible to derive it on the basis of tectogrammatical representations. One possibility is to represent the scope of rhematizers on the basis of the indication of the

Expression	Used in the function of a rhematizer	Function of an adverbial	Used in the function of an adverbial
nejvýše nanejvýš	*I would have given him at most a home prison.*	EXT – specification of a numeral	*It cost at most one hundred crowns.*
už již	*Already KOMENSKÝ spoke about it.*	TWHEN – meaning "now"	*The time has already come to go to bed.*
zrovna právě teprve	*Exactly THIS I have told him.*	TWHEN – meaning "now", or EXT – "exactly"	*He has just left the car. Invite just one hundred people.*
až	*It looked too bad.*	EXT – meaning "up to", "almost"	*The meeting will be attended by up to 100 people.*
zase	*I am bad and Jim for his part well.*	TWHEN	*I will come again.*
přímo	*He was quite amazing.*	DIR2 – meaning "directly" MANN	*The road went directly to the village. Tell me downright.*
zvlášť	*Take care especially of the kids.*	MANN	*We will pay separately.*
hned	*He took right away three apples.*	TWHEN	*I will come back immediately.*
naopak	*George on the contrary ran away.*	MANN – meaning: in an opposite way, contrary to	*He did everything contrary to what they TOLD him.*

Table 3. Ambiguity of Czech rhematizers obtaining also a function of a free modification

topic–focus articulation, namely on the contextual boundness of individual nodes of the tree and the boundary between topic and focus. The rhematizer that signals the focus of the sentence has in its scope all the contextually non-bound items that follow it in the surface shape of the sentence; the scope of the rhematizer signaling the contrastive topic is basically the first element with the value of contrastive contextually bound element (together with its dependents) that follow it. If the rhematizer is the

only contextually non-bound element of the sentence, it is assumed to have a backward scope. However, these basic assumptions have not yet been put under a detailed scrutiny and wait for their validation on the PDT material.

To sum up, the analysis based on the PDT material has confirmed that there is a special class of particles that have a specific position in the TFA of the sentence and that these particles have some common features with negation. It has also been demonstrated that these particles called in linguistic literature rhematizers, focalizers or focussing particles need not be restricted to a position indicating the focus (rheme) of the sentence, rather, they can also occur in the topic of the sentence; also, there can be more than a single rhematizer in the sentence. In the theoretical description, these observations lead to the conclusion that it is necessary to distinguish between the focus of the whole sentence and the focus of a focalizer. Finally, we have observed that the scope of a focalizer has important consequences for the semantic interpretation of the sentence.

4.1.4. Contrastive study of TFA based on a parallel corpus

The existence of parallel corpora equipped with basically the same scheme of annotation offers an invaluable material for contrastive linguistic studies and thus for a re-evaluation of existing hypotheses. Let us quote here one of the numerous examples based on the comparison of a particular phenomenon in Czech and English.

A similarly based annotation as in PDT, though not covering all the features captured by the Czech corpus, exists for English in the so-called Prague Czech–English Dependency Treebank 2.0 (PCEDT; Hajič et al., 2011, see also K. Rysová et al., 2015b)[28] comprising an annotation of Czech and English parallel texts (almost 50 thousand sentences for each part) along the lines of PDT. This material has allowed for a more detailed contrastive analysis of tectogrammatical (underlying syntactic) sentence structures also concerning the topic–focus structure of Czech and English sentences. As an example, we present here the results of a case study concerning the use of the indefinite article with the subject of an English sentence.

Basically, in both languages a common strategy in communication is to proceed from retrievable, identifiable information to an unretrievable one. This strategy can be documented for Czech by the fact that in PDT, there is only a small portion of cases in which a contextually bound item in the topic of the sentence does not provide a coreferential link (i.e., it does not serve as an anaphor; it should be noted that the coreference annotation in PDT captures so far only relations of a nominal group to an antecedent, see below). As for English, a good indicator of such a rare situation is the appearance of the indefinite article in the subject position of sentences, if one assumes the unmarked position of the intonation center at the end of the sentence. Such cases

[28] http://ufal.mff.cuni.cz/pcedt2.0/en/index.html

are rather rare and can be explained by an interaction of other factors as documented on the material from the Czech–English corpus in Hajičová et al. (2011).

We started from the hypothesis that one of the possibilities how to "topicalize" Patient (Object) in English is passivization. In PCEDT, with the total number of 49,208 sentences (and, for comparison, with the total number of 54,304 predicates – roughly: clauses) there were 194 cases of an occurrence of a nominal group with an indefinite article in the function of a subject of a passive construction. These cases were compared with their Czech counterparts and can be classified into four groups as follows:

(a) Most frequent constructions contain a General Actor, not expressed in the surface (see Sect. 3.3 above)

These sentences are translated into Czech with the subject (expressing the Patient) at the end of the sentence (in Focus!); in English, the postposition of the subject into the final position is not possible due to the grammatically fixed English word-order, see (39) with the assumed position of intonation centre denoted by capitals:

(39) (Preceding context: Soviet companies would face fewer obstacles for exports and could even invest their hard currency abroad. Foreigners would receive greater incentives to invest in the U.S.S.R.) Alongside the current non-convertible ruble, a second CURRENCY would be introduced that could be freely exchanged for dollars and other Western currencies.
[Czech equivalent: Zároveň se současným nekonvertibilním rublem bude zavedena druhá MĚNA, která by mohla být volně směnitelná za dolary a další západní měny.]

(b) The indefinite article is used with the meaning "one of the", see (40):

(40) A seat on the Chicago Board of Trade was sold for $ 390,000, unchanged from the previous sale Oct. 13.
[Czech equivalent: Členství (meaning: membership, e.g., the status of a member) v Chicagské obchodní radě bylo prodáno za 390 000 dolarů, což je nezměněná cena od posledního prodeje 13. října.]

(c) Interesting though few cases involve a contrast in the topic part, see (41), with the assumed intonation center (in focus) on the year 1984 and a contrastive accent (in topic) on *faster*:

(41) (Preceding context: The "Designated Order Turnaround" System was launched by the New York Stock Exchange in March 1976, to offer automatic, high-speed order processing.) A faster version, the SuperDot, was launched in 1984.
[Czech translation (in the indicated context, with the same intonation contour): Rychlejší verze SuperDot byla spuštěna v roce 1984.]

4.2. Annotation of discourse relations

The annotation of the textogrammatical layer of PDT also serves as the starting point of the annotation of discourse relations and the basic relations of textual coreference. Though we do not consider these relations to belong to the underlying layer of language description as understood in the theoretical framework of Functional Generative Description, however, technically, the annotation of these phenomena is based on the tectogrammatical layer of PDT. As claimed in Mírovský et al. (2012), Nedoluzhko and Mírovský (2013), and Jínová et al. (2012), such an approach has its advantages: the annotators (and, eventually, an automatic preprocessing procedure) can take into account the information relevant for discourse relations that is already present in the underlying representation of the sentence (e.g., the dependency relation between the governing clause and its dependent clauses in case of the relation of cause and admission); in addition, the tectogrammatical representations contain a "reconstruction" of the deleted items in the surface structure (see Sect. 3.5 above), which is very important for the identification of coreference relations, but also relevant for the establishment of certain discourse relations.

The annotation of discourse relations in PDT 3.0 (present also in the Prague Discourse Treebank, PDiT, see Poláková et al., 2013) is based on the annotation scenario applied to the annotation of English texts in the Pennsylvania Discourse Treebank (Prasad et al., 2008). In the process of annotation, the annotators identify so-called connectives and for each of the connective they look for its so-called arguments, i.e., pieces of the text that are connected by some kind of discourse relation indicated by the connective. In this approach, it is assumed that there should be always two arguments connected by one connective.[29]

Fig. 7 exhibits the annotation of a discourse relation between the sentences: *Slovenská elita byla zklamána politickou volbou Slovenska.* [The Slovak elite were disappointed by the political choice of Slovakia.] and *Proto většina kvalitních odborníků zůstala v Praze*. [Therefore, most of the good specialists stayed in Prague.]. A discourse relation between the trees is marked with a thick curved arrow; the type of the relation (reason) is displayed next to the tectogrammatical functor of the starting node. The connective assigned to the relation (*proto* [therefore]) is also displayed at the starting node, as well as the range of the arguments entering the relation (range: 0 -> 0, indicating that in this case, only the two mentioned trees (clauses) enter the relation).

As indicated above, discourse annotation in PDT 3.0 is focused on an analysis of discourse connectives, the text units (or arguments) they connect and on the semantic relation expressed between these two units. A discourse connective is defined as a predicate of a binary relation – it takes two text spans (mainly clauses or sentences) as its arguments. It connects these units and signals to a semantic relation

[29] It should be noted that while the annotation of the discourse relations in the Pennsylvania Discourse Treebank was carried out on running texts, in case of PDiT the disscourse relations are annotated on the tree structures (of the PDT tectogrammatical layer).

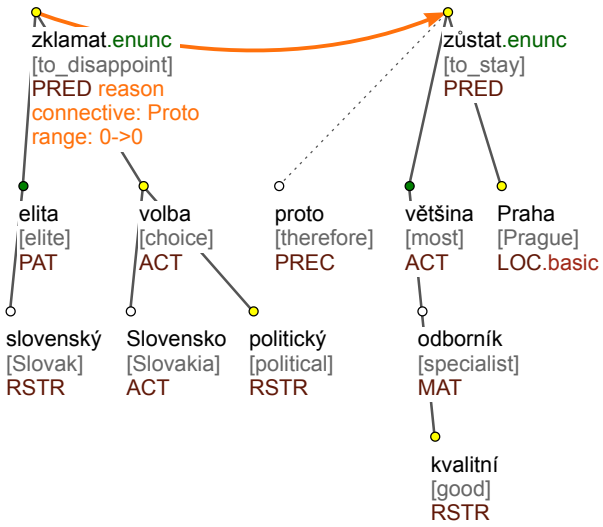


Figure 7. Annotation of a discourse relation between the sentences: Slovenská elita byla zklamána politickou volbou Slovenska. Proto většina kvalitních odborníků zůstala v Praze. [The Slovak elite were disappointed by the political choice of Slovakia. Therefore, most of the good specialists stayed in Prague.]

between them at the same time. Discourse connectives are morphologically inflexible and they never act as grammatical constituents of a sentence. Like modality markers, they are "above" or "outside" the proposition. They are represented by coordinating conjunctions (e.g. *a* [and], *ale* [but]), some subordinating conjunctions (e.g. *protože* [because], *pokud* [if], *zatímco* [while]), some particles (e.g. *také* [also], *jenom* [only]) and sentence adverbials (e.g., *potom* [afterwards]), and marginally also by some other parts-of-speech – mainly in case of fixed compound connectives like *jinými slovy* [in other words] or *naproti tomu* [on the contrary]. The annotation is focused only on discourse relations indicated by overtly present (explicit) discourse connectives – the relations not indicated by a discourse connective were not annotated in the first stage of the project.[30]

The taxonomy of discourse relations in PDT 3.0 is based on the taxonomy used in the Penn Discourse Treebank[31] but it is modified by taking into account the theory of the Functional Generative Description and the tradition of the Czech studies (e.g., the

[30] There are 18,161 discourse relations annotated in the data, out of them 5,538 relations are inter-sentential, 12,623 relations are intra-sentential.

[31] See The Penn Discourse TreeBank 2.0 Annotation Manual (Prasad et al., 2007).

addition of the relation of gradation and explication). The taxonomy contains four basic groups of relations: temporal, contingency, contrast, and expansion.

Within these main groups several subgroups are being distinguished, namely synchronous and asynchronous with temporal relations, reason – result, condition, explication, and purpose in the contingency group; confrontation, opposition, concession, correction, and gradation in the group of contrast, and conjunction, exemplification, specification, equivalence ,and generalization in the group of relations of semantic extension.

In addition to the discourse relations proper, some other types of information have been included in our annotation scheme as well as the appurtenance of the text into the so-called genres (Poláková et al., 2014). This complex annotation makes it possible to search in the annotated corpus for the combination of the deep syntactic structure, information structure, coreference, and genre information.

The process of manual checking of the consistency of annotation that was carried out after the whole treebank was annotated has led not only to a necessary unification of the understanding of some relations but also to interesting observations concerning the complexity of some relations, or to an analysis of multiword connectives, of multiple coordination, etc.

The analysis of annotated data (see Table 4) helped us observe which types of relations are more frequently expressed within the frame of a single sentence and which hold rather between complexes of sentences (divided by final punctuation marks). Up to now, this distribution could be only approximated on the basis of language intuition.

The largest proportion of occurrences within a single (complex) sentence is documented for the relation of purpose, condition, and disjunctive alternative. These relations only rarely occur between two independent sentences. On the basis of these calculations, a preliminary hypothesis can be formulated that the semantic content expressed by the arguments of the above relations are more closely bound together than with the other relations. Also, the relatively high position of the conjunction relation is surprising as one would expect a more balanced distribution, perhaps similar to that found with opposition.

In the course of the annotation, it came out that some connective means connect implicatures or deletions hidden in the text rather than arguments expressed explicitly in the text. To capture these relations, a category called “pragmatic” relations has been introduced, see (42), where the second sentence containing a connective *však* [however] does not express an opposition to the fact that several orders are in progress but an opposition to the unexpressed implication that to have many orders means a large income for the firm.

Group of relation	Type of relation	Intra-sentential	Inter-sentential
Contingency	Purpose	100%	0%
	Condition	99%	1%
	Pragmatic condition	93%	7%
	Reason–result	61%	39%
	Explication	43%	57%
	False reason–result	31%	69%
Contrast	Correction	73%	27%
	Concession	70%	30%
	Confrontation	53%	47%
	Gradation	52%	48%
	Pragmatic opposition	46%	54%
	Opposition	43%	57%
	Restrictive opposition	37%	63%
Expansion	Disjunctive alternative	95%	5%
	Specification	82%	18%
	Conjunction	81%	19%
	Conjunctive alternative	79%	21%
	Equivalence	40%	60%
	Exemplification	19%	81%
	Generalisation	9%	91%
Temporal	Synchronous	77%	23%
	Asynchronous	70%	30%

Table 4. Ratio of types of discourse relations occurring intra-sententially and inter-sententially

(42) Podle vedoucího výroby Miloše Přiklopila má Seba rozpracovanou celou řadu zakázek. Zákazníci však vyvíjejí velký tlak na snižování cen tkanin.
[According to the production manager M.P. several orders are in process in SEBA. The customers, however, make a big pressure on the lowering of the price of the material.]

It will be a matter of future research to see, which relations are more frequently indicated by explicit connectives and which can be easily implied implicitly. Such research may bring interesting results when based on parallel corpora; the fact that we also have at our disposal a parallel Czech–English treebank makes such research possible.

Another research topic relates to the fact that an analysis of discourse relations cannot be restricted to a study based on a rather narrowly defined class of connective devices. Similarly as in the Penn Discourse Treebank (see Prasad et al., 2008), in the current stage of discourse annotation we have focused on the so-called alternative lexicalizations (AltLex, or secondary connectives, see M. Rysová, 2012), that is expressions connecting discourse arguments but not belonging to a narrow class of connectors; the structure of these expressions ranges from a single word to a whole sentence. The first attempt to single out these secondary connectives resulted in a list of 1,201 occurrences of relations signaled by them in PDT. Contrary to the annotation phase that worked only with primary connectives that related clausal arguments, secondary connectives may relate also nominalizations (in 310 cases out of the 1,201 relations rendered by these secondary connectives, e.g., *He was absent because he was ill*. vs. *The reason for his absence was his illness*.). Similarly as is the case of primary connectives, also secondary connectives are not typically component parts of the arguments, they, as it were, stand outside them. However, the situation is different with some verbs of saying (such as *to add, to complement, to continue*) where the verb behaves as a secondary connective but also represents the second argument: its meaning includes also the information that somebody said, wrote, etc., something before (see M. Rysová, 2014b).

4.3. Annotation of coreferential and associative relations

In addition to discourse relations, the annotation scheme of PDT has been enriched by the annotation of coreferential and associative relations. As for coreference, we distinguish between grammatical and textual coreference. Grammatical coreference is restricted to certain syntactic relations within a sentence and the antecedent can be determined in principle on the basis of grammatical rules of the given language. Table 5 shows basic types of grammatical coreference distinguished in PDT.

As for textual coreference, it can be expressed not only by grammatical or lexical means (such as pronominalization, grammatical agreement, repetition of lexical units, use of synonyms, paraphrasing, use of hyponyms, or hypernyms within lexical cohesion) but it can also follow from the context and pragmatics; in contrast to grammatical coreference, textual coreference often goes beyond sentence boundaries.[32]

Two types of textual coreference between nominal groups can be distinguished, namely that specific and generic reference, see (43) as an example of the former type and (44) as an example of the latter type.

(43) **Marie** a Jan spolu odjeli do Izraele, ale **Marie** se musela vrátit kvůli nemoci.
[**Mary** and John left together for Israel, but **Mary** had to return because of illness.]

[32] Textual coreference can be found in 84,306 cases, grammatical coreference in 20,624 cases. There are 30,470 examples of bridging anaphora.

Type of relation	Example
Coreference of reflexive pronouns	*Dcera se musela dlouho přesvědčovat, aby pokračovala v tréninku.* [in the reading of: *The daughter* had to persuade *herself* to continue in training.]
Coreference of relative means (*který*, *jenž*, *což* etc.)	*Za informační dálnici se považuje světová telekomunikační síť, po níž lze přenášet zvuk, data i obraz a která tak otevírá přístup k množství informatických služeb.* [The information motorway is such a world wide telecommunication network, which … and which …]
Relation of "control" present with a specific group of verbs, e.g. *začít* [begin to], *dovolit* [allow to], *chtít* [want to], *dokázat* [prove] to etc.)	*Vedení sekce plánuje vyklidit knihovnu.* [The management of the department plans to empty the library.] (the unexpressed subject of the infinitive *to empty* is in a coreference relation to the Actor of the main clause: *the management*)
Coreference with a complement with so-called "double dependency"	*Honza zastihl Hanku běhat kolem rybníka.* [Honza found Hana *to run* round the pool.] (coreference of the unexpressed subject of the verb to run with the Patient of the verb *found* – Hana)

Table 5. Types of grammatical coreference in PDT

(44) **Psi** štěkají. To je způsob, jak [**oni**] vyjadřují své emoce.
[**Dogs** are barking. This is the way [**they**] express their emotions.]

The border line between these two types is not always clearcut and the interpretation may be rather subjective, see (45), where the expression *hospoda* [restaurant] may have either a specific reference (the concrete enterprise) or a generic one (restaurant as a type of enterprise).

(45) Začal jsem provozováním **hospody**, která byla mnohokrát vykradena. [… 2 sentences follow …] **Hospoda** byla jen startem, polem k podnikání s masem a masnými výrobky.
[I started with opening a **restaurant**, which was many times visited by thieves. [… 2 sentences follow…] The **restaurant** was just a start, an opportunity to deal with meat and meat products …]

We are fully aware that coreference relations may exist not only between nominal groups but also between verbs which denote events. For the time being, however,

our scheme captures only cases where a verb appears as an antecedent of a nominal group. This phenomenon is referred to in literature as a textual deixis.

Side by side with coreference, several other textual relations contribute to the cohesion of text and help the addressee to understand a certain expression as referring to a known entity even if the two entities are not in a strict coreferenctial relation. We call such relations associative anaphora (Nedoluzhko, 2011); in English oriented studies, they are called *bridging anaphora/relation*, *indirect anaphora*, *associative anaphora* etc.

These relations can be classified in several ways, according to the purpose of their description. In our scheme, we concentrate on the linguistic meaning of anaphora and therefore our classification is rather a detailed one. At the same time, however, we have tried to define the types rather strictly, in order to keep the consistency of the annotation. In PDT 3.0, the following types of associative relations are distinguished:

(a) Relation between a whole and a part (*a house – its roof*)
(b) Relation between a set and its subset or member (*a class – several pupils – a pupil*)
(c) Relation between an object and a function defined on that object (*a coach – a team*)
(d) Relation of a pragmatic and semantic contrast (*last year – this year – next year*).
(e) Non-coreferential anaphoric relation, in case of an explicit anaphoric reference to an non-coreferential antecedent (often accompanied by expressions *such as*, *the same*, *similar*, etc.)

In addition to these types, we also distinguish some other specific relations, such as family relations (*father – son*), place – inhabitant (*Prague – Praguians*), author – piece of work (*Rodin – Thinker*), a thing – its owner, event – argument (*enterprise – enterpreneur*), an object and a typical instrument (*copybook* – pen).

Contrary to the domains exemplified in Sections 3.1 through 3.6 above and in 4.1, in the analysis of which we could build upon our well-established theory, and in Sect. 4.2, in which we could modify or complement an existing scenario proposed by another team working on a similar project (namely the Penn Discourse Treebank), we have not found any consistent, uniform and well-developed scheme that would suit our purpose to integrate both the aspects – discourse relations and coreference in broad sense – into the overall system of PDT. In this sense, any step or proposal of a taxonomy of coreferential (and associative) relations within PDT was in itself a contribution to the development of a suitable and consistent approach to the description of these aspects of text coherence resulting in a basic annotation scenario for the phenomena concerned.

5. Some corpus statistics: Inter-annotator agreement

The strength of an annotated corpus lies not only in the quality of the underlying linguistic theory and in its contribution to this theory but also in three other aspects:

– the quality of the annotation process
– the size of the annotated data
– the quality of a search tool for the corpus

The quality of the annotation process can be measured by agreement between the annotations of the same data performed by two or more annotators. As annotation is an expensive process, usually the data are annotated only by one annotator and only a small part of the data is annotated in parallel by two annotators, just for the purpose of measuring the inter-annotator agreement. In this Section, we report on inter-annotator measurements in PDT or other corpora of the Prague dependency family.

The size of the annotated data is also very important, as a small corpus might not offer enough material for a sufficient analysis of scarce phenomena. We have included some figures concerning the size of the data and the frequency of some of the phenomena in the sections above at places for which these figures were relevant.[33] The size and complexity of a corpus are also in a close relation to the possibility to retrieve relevant examples from the data, which is a task for the search tool. For PDT (and other corpora using the same data format), a powerful and user-friendly querying system exists called PML Tree Query (PML-TQ; Pajas and Štěpánek, 2009).

Since the first years of annotation of PDT, the inter-annotator agreement has been measured for many individual annotation tasks. The measurements and the analysis of the disagreements help detect errors in the annotations, improve the annotation guidelines, and find phenomena difficult from the annotation point of view. We present numbers measured on PDT or on the Czech part of Prague Czech-English Dependency Treebank (PCEDT), which uses the same annotation scenario and annotates a similar type of data (journalistic texts).

For classification tasks (tasks where the places to be annotated are given, i.e., identifying such places is not a part of the annotator's decision) we use simple agreement ratio, i.e. percentage of the places where the annotators assigned the same value; sometimes we also mention Cohen's κ (Cohen, 1960), a measure that shows how much better the inter-annotator agreement is compared with the agreement by chance. For more complex tasks, where the identification of the place to be annotated is also a part of the annotator's decision, we use F1-measure, which is the harmonic mean of precision and recall.[34]

On the **morphological layer**, disambiguation of the automatic morphological analysis was done in parallel by pairs of annotators on the whole PDT data. The inter-annotator agreement on the assignment of the correct **morphological tag** to words

[33] If not stated otherwise, the numbers reported come from 9/10 of the whole PDT data, as the last tenth of the data is designated to serve as evaluation test data and as such should not be observed or used in any way other than testing. In these 9/10 of PDT (used as train and development test data), there are 43,955 sentences in 2,849 documents.

[34] http://en.wikipedia.org/wiki/F1_score

with an ambiguous morphological analysis was 95% (Bémová et al., 1999); if the unambiguous words are also counted, the agreement is 97% (Hajič, 2005). Note that in Czech, there are approx. 4.7 thousand different morphological tags.[35]

For the **analytical layer** in PDT, as far as we know, no measurements of the inter-annotator agreement have been published.

On the **tectogrammatical layer**, there are many annotation tasks. The measurements were performed during the annotation of PDT (the numbers for PDT on the tectogrammatical layer, unless specified otherwise, come from Hajičová et al., 2002) and the Czech part of PCEDT (numbers come from Mikulová and Štěpánek, 2010).

(i) The agreement on **linking the tectogrammatical nodes to** their counterparts from **the analytical layer** in PCEDT was 96% for the lexical counterparts and 93.5% for the auxiliary nodes.

(ii) The agreement on assigning **sentence modality** for 268 complex cases of coordinated clauses in PDT (ver. 3.0) was 93.7% with Cohen's κ 89% (Ševčíková and Mírovský, 2012).

(iii) The agreement on establishing the correct **dependency** between pairs of nodes (i.e. the establishment of dependency links together with the determination which member of the pair is the governor) was 91% (64 differences in 720 dependency relations) in PDT, and 88% in PCEDT.

(iv) The agreement on assigning the correct type to the dependency relation (the tectogrammatical **functor**) was 84% (112 differences in 720 relations) in PDT, and 85.5% in PCEDT.

(v) The agreement on assigning the correct value to individual nodes in the annotation of **topic-focus articulation** (i.e. the assignment of the values 'contextually bound' or 'contextually non-bound' within the TFA attribute; 'correct' here means 'as judged by the author of the manual', i.e. the agreement is measured pairwise between each annotator and the arbiter) was approx. 82% (81%, 82%, 76%, and 89% for different annotators) (Veselá et al., 2004).

(vi) In the task of marking **multiword expressions** in the data (which was done on top of the tectogrammatical layer for PDT 2.5), the authors used their own version of weighted Cohen's κ (with adjusted upper agreement bound) and report the agreement above chance of 64.4% (Bejček and Straňák, 2010).

The mismatches between annotators were carefully studied. A comparison of the agreement figures given in (iii) and (iv) indicates that annotators were more confident of their judgements when building the dependency structure rather than when labeling the nodes by functors. This observation indicates that it was not difficult to decide which node is the governor and which is the dependent. Discrepancies between an-

[35] For comparison with other projects, let us mention the inter-annotator measurement during the annotation of the German corpus NEGRA, as reported by Brants (2000). Their agreement in the part-of-speech annotation was 98.57%. However, the size of their part-of-speech tagset was only 54 tags.

notators were found in the decisions on the type of dependency relation, i.e. on the labels for valency members as well as for these of free modifications. This fact demonstrates that the boundaries between some pairs of functors are rather fuzzy, or perhaps they were not defined in an exhaustive way. The functor MEANS (Instrument) and EFF (Effect) were often interchanged as well as the functor BEN (Beneficient) and ADDR (Addressee), though the former member of the pair belongs to the class of free modifications and the latter to the class of valency members. These mismatches are connected with a more or less effective application of the criteria for obligatory positions in the valency frame of the corresponding items. However, there are only few mismatches which are systematic, most of discrepancies are subjective/individual.

Among the **phenomena crossing the sentence boundary**, we have measured the inter-annotator agreement in PDT for the extended (nominal) textual coreference, bridging anaphora and discourse relations. To evaluate the inter-annotator agreement in these annotations, we used several measures:

(i) The connective-based F1-measure (Mírovský et al., 2010) was used for measuring the agreement on the recognition of a **discourse relation**, the agreement was 83%.
(ii) The chain-based F1-measure was used for measuring the agreement on the recognition of a **textual coreference** or a **bridging anaphora**, the agreement was 72% and 46%, respectively.
(iii) A simple ratio and Cohen's κ were used for measuring the agreement on the type of the relations in cases where the annotators recognized the same relation, the agreement was 77% (Cohen's κ 71%) for **discourse**, 90% (Cohen's κ 73%) for **textual coreference**, and 92% (Cohen's κ 89%) for **bridging anaphora** (Poláková et al., 2013).[36]

The numbers of the inter-annotator agreement for the phenomena crossing the sentence boundary reveal some simple observations: it is quite clear that recognizing the presence of a textual coreference relation is easier than that of a bridging relation. For both textual coreference and bridging anaphora, it is more difficult to find the existence of a relation rather than to select its type – once the presence of the relation is agreed upon, the annotators are able to assign its type with high accuracy. For discourse relations, on the contrary, an assignment of the type of a relation seems to be more difficult than recognition of its presence.

As mentioned above, the nature of the tasks required to apply for the different annotation tasks different measures for the inter-annotator agreement. Although the numbers expressing different measures of evaluation are not – strictly speaking – directly comparable (especially Cohen's κ cannot be compared with other measures),

[36] For comparison, the simple ratio agreement on types of discourse relations in Czech (77%) is the closest measure to that of measuring the inter-annotator agreement used on subsenses (second level in their sense hierarchy) in the Penn Discourse Treebank 2.0, reported in Prasad et al. (2008).

they confirm the general idea that the deeper we go in the abstraction of the language description, the more difficult it is to achieve high values of the inter-annotator agreement.

Measuring the inter-annotator agreement and studying discrepancies between annotators repeatedly proved to be an indispensable part of the annotation process of PDT and other corpora. Not only is it necessary for ensuring a high quality annotation (for reasons mentioned above) but it may even reveal shortcomings in the underlying linguistic theory. It is the only way to establish and enumerate the difficulty of a given annotation task and to set a higher boundary for the accuracy we can expect from automatic methods of annotation.

6. Summary and outlook

6.1. Contributions of the annotation to the theory

In the present paper, we have presented several selected case studies based on the Prague Dependency Treebank Version 3.0 that are supposed to document the importance of corpus annotation at different linguistic layers for a verification of established linguistic hypotheses and for their eventual modifications, or, as the case may be, for making the linguistic description of some phenomena more precise.

The basic ideas of the theoretical Framework of the FGD were formulated before the large language resources were available and as such, they were applied in the design of the original annotation scenario of PDT. During the process of annotation of the raw texts the hypotheses formulated on the basis of the theory were tested and by testing them the accuracy of the theory itself was furthermore accessed, and the gaps within the list of morphological meanings, syntactic and semantic units have been identified. These gaps, however, should not be understood as errors in the original proposal since many of the phenomena concerned had not been noticed before by any reference grammar of Czech.[37] In the present contribution several of these improvements have been discussed at the end of each Section: the necessity of the two levels of syntax (surface and deep/underlying levels, called tectogrammatics) is supported by the introduction of the category of diathesis (see 3.1), by the new grammateme pair/group number (see 3.2) and by the restoration of elements missing on the surface structure and required by the deep representation (see 3.5). Also a new class of valency members (called quasivalency) was introduced (see 3.4). While in the classical version of the FGD the issues of lexicon were not in the focus of our attention, the introduction of new categories (functors, subfunctors, grammatemes) opened new aspects of the interplay between grammar and lexicon which were analyzed in particular case studies above and became a source of extension of the theoretical framework.

[37] The notion of "reference grammar" is not commonly used in Czech linguistics but the Mluvnice češtiny [Czech Grammar] (Komárek et al., 1986, Daneš et al., 1987) is supposed to be a standard source of references, and, as for Czech syntax, the monograph by Šmilauer (1947) is most frequently used in this sense as well.

In the domain of information structure, the annotated data helped us to develop in more detail the hypotheses concerning the deep word order (so-called systemic ordering) as documented in 4.1.2 and to achieve a more detailed analysis of the special class of particles called in linguistic literature rhematizers, focussing particles, or focalizers. The analysis of rich corpus material has indicated that the position of these particles need not be restricted to the focus of the sentence (as the term previously used for them may suggest) but that they may also occur in the topic; this observation has led to the introduction of the notion of contrastive topic and to the distinction between the focus of the sentence as a whole (global focus) and the local focus of the focalizer.

While Part I of the case studies (Section 3) contains analyses of phenomena that belong to grammar, Part II covers a domain that traditionally might be relegated to the domain of pragmatics. However, as the arguments presented in numerous writings on topic–focus articulation quoted in Section 4.1 and supporting the semantic relevance of this phenomenon, a description of the information structure of the sentence is an indispensable part of any functionally conceived grammatical theory. On the other hand, coreference relations (with the exception of grammatical coreference) and discourse relations do go beyond the sentence structure and therefore they were not analyzed in detail in the theoretical framework the PDT annotation is based on. In this sense, the analysis presented in Sections 4.2 and 4.3 brings observations that have not yet been included in a systematic way in any description of Czech.

An irreplaceable role in the process of recognition and implementation of the improvements to the theory is played by the human annotators themselves; though the manual annotation is rather expensive, its effect is doubtless: the annotators have to work consistently applying the existing guidelines and they supply many observations that uncover linguistic details hitherto not registered. The usefulness, abundance and originality of these observations is best documented by the publication of the modern scientific syntax of Czech based on PDT (Panevová et al., 2014).

6.2. Outlook

It is indisputable, however, that some particular phenomena require further analysis. Many empirical problems are connected with coordinated constructions. The studies of elliptic coordinations are planned for the detection of the formal criteria for the possibility of restoration their underlying representation in contrast to the pragmatic conditions for their application belonging to the domain of text structure and discourse relations. Another domain of further work relates to the reflection of the results achieved in our analysis in the dictionary build-up. Selected data extracted from PDT 3.0 will be incorporated into the valency dictionary: e.g. completion of the list of words with the ability of control and the proposal of the description of the interplay between morphological meanings of verbs and the realization of their valency frames in the sentence.

In the particular case of the Prague Dependency Treebank, there is one feature that distinguishes it from annotation schemes worked out for other languages, namely the fact that annotation on all layers together with the annotation of discourse relations, coreference, and associative relations is applied to the same collection of full texts (and partly also on parallel English texts). This makes it possible to look for an interplay of these layers and to try and use the complex annotation for some particular projects. For instance, we have started a research in the interplay of syntactic structure, information structure, and coreference relations based on the notion of the activation hierarchy of elements of the stock of knowledge as proposed by Hajičová and Vrbová (1982) and elaborated further e.g., in Hajičová (1993, 2003, 2012) and Hajičová and Vidová-Hladká (2008). The underlying hypothesis for our analysis of discourse structure was formulated as follows: A finite mechanism exists that enables the addressee to identify the referents on the basis of a partial ordering of the elements in the stock of knowledge shared by the speaker and the addressees (according to the speaker's assumption), based on the degrees of activation (salience) of referents. The following three basic heuristics (a) through (c) based on the position of the items in question in the topic or in the focus of the sentence, on the means of expression (noun, pronoun) and on the previous state of activation have been formulated to determine the degrees of salience of the elements of the stock of shared knowledge:

(a) In the flow of communication, a discourse referent enters the discourse, in the prototypical case, first as contextually non-bound, thus getting a high degree of salience. A further occurrence of the referent is contextually bound, the item still has a relatively high degree of salience, but lower than an element referred to in the focus (as contextually non-bound) in the given sentence.
(b) If an item is not referred to in the given sentence, the degree of salience is lowered; the fading is slower with a referent that had in the previous co-text occurred as contextually bound; this heuristics is based on the assumption that a contextually bound item has been 'standing in the foreground' for some time (as a rule, it was introduced in the focus, then used as contextually bound, maybe even several times) and thus its salience is reinforced; it disappears from the set of the highly activated elements of the stock of shared knowledge in a slower pace that an item which has been introduced in the focus but then dropped out, not rementioned. If the referent has faded too far away it has to be re-introduced in the focus of the sentence.
(c) If the difference in the degree of salience of two or more items is very small, then the identification of reference can be done only on the basis of inferencing.

These three basic heuristics served as a basis for our formulation of several rules for the assignment of the degrees of salience, which have been applied to numerous text segments to check how the determination of these degrees may help reference resolution. Thanks to the richly annotated corpus of PDT, we basically have at our disposal

all of the information we need for an application of our rules for activation assignment: the underlying sentence representation with restored (superficial) deletions as well as with part-of-speech information, the Topic-Focus assignment (via the TFA attribute with values contextually-bound and contextually non-bound) and coreferential chains for nominal and pronominal realization of referential expressions. The activation algorithm has already been implemented and applied to (selected but full) documents, the 'activation' diagrams have been visualized and the task now is to test the hypotheses our approach is based on and the possibilities the approach offers for text analysis and generation on a larger portion of the PDT collection.

Acknowledgements

The authors gratefully acknowledge the support from the Grant Agency of the Czech Republic (project No. P406/12/0658) and from the Ministry of Education, Youth and Sports (projects LH14011 and LM2015071). The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

Bibliography

Bejček, Eduard and Pavel Straňák. Annotation of Multiword Expressions in the Prague Dependency Treebank. *Language Resources and Evaluation*, 44(1–2):7–21, 2010.

Bejček, Eduard, Jarmila Panevová, Jan Popelka, Lenka Smejkalová, Pavel Straňák, Magda Ševčíková, Jan Štěpánek, Josef Toman, Zdeněk Žabokrtský, and Jan Hajič. Prague Dependency Treebank 2.5. Data/software, 2011.

Bejček, Eduard, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. Prague Dependency Treebank 3.0. Data/software, 2013.

Bémová, Alevtina, Jan Hajič, Barbora Vidová Hladká, and Jarmila Panevová. Morphological and Syntactic Tagging of the Prague Dependency Treebank. In *Journées ATALA – Corpus annotés pour la syntaxe; ATALA Workshop – Treebanks*, pages 21–29, Paris, 1999. Université Paris.

Bhaskararao, Peri and Karumuri Venkata Subbarao. *Non-nominative subjects*, volume 1. John Benjamins Publishing, 2004.

Böhmová, Alena, Jan Hajič, Eva Hajičová, and Barbora Hladká. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In *Treebanks: Building and Using Syntactically Annotated Corpora*, chapter 7, pages 103–128. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The TIGER Treebank. In Hinrichs, E. and K. Simov, editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pages 24–41, 2002.

Brants, Thorsten. Inter-Annotator Agreement for a German Newspaper Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, 2000. European Language Resources Association.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, pages 969–974, 2006.

Camacho, J. A. *Null Subjects*. Cambridge University Press, 2013.

Cinková, Silvie, Josef Toman, Jan Hajič, Kristýna Čermáková, Václav Klimeš, Lucie Mladová, Jana Šindlerová, Kristýna Tomšů, and Zdeněk Žabokrtský. Tectogrammatical Annotation of the Wall Street Journal. *The Prague Bulletin of Mathematical Linguistics*, (92):85–104, 2009.

Clancy, Steven J. *The chain of being and having in Slavic*, volume 122. John Benjamins Publishing, 2010.

Cohen, Jacob. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Daneš, František, Helena Běličová, Mirek Čejka, Emil Dvořák, Miroslav Grepl, Karel Hausenblas, Zdeněk Hlavsa, Jana Hoffmannová, Josef Hrbáček, Jan Chloupek, Petr Karlík, Eva Macháčková, Olga Müllerová, Bohumil Palek, Jiří Nekvapil, Jiří Novotný, Petr Piťha, Hana Prouzová, Milena Rulfová, Blažena Rulíková, Otakar Šoltys, Ludmila Uhlířová, and Stanislav Žaža. *Mluvnice češtiny. 3. Skladba [Grammar of Czech. 3. Syntax]*. Academia, Prague, 1987.

Fillmore, Charles J. The Case for Case Reopened. *Syntax and Semantics*, 8(1977):59–82, 1977.

Fillmore, Charles J. *Form and Meaning in Language. Volume 1. Papers on Semantic Roles*. CSLI Publications, Stanford University Press, 2003.

Firbas, Jan. K otázce nezákladových podmětů v současné angličtině. Příspěvek k teorii aktuálního členění větného. *Časopis pro moderní filologii*, 39:22–42; 165–173, 1957. (An abbreviated and modified English version of this contribution was published as Non-thematic subjects in Contemporary English, TLP 2, Prague: Academia, 239–256.).

Giger, Markus. *Resultativa im modernen Tschechischen: unter Berücksichtigung der Sprachgeschichte und der übrigen slavischen Sprachen*, volume 69. Peter Lang, Bern – Berlin – Bruxelles – Frankfurt a.M. – New York – Oxford – Wien, 2003.

Hajič, Jan. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová (ed. Eva Hajičová)*. Karolinum, Charles University Press, Prague, 1998.

Hajič, Jan. Complex Corpus Annotation: The Prague Dependency Treebank. In Šimková, Mária, editor, *Insight into the Slovak and Czech Corpus Linguistics*, pages 54–73. Veda, Bratislava, 2005.

Hajič, Jan and Václav Honetschläger. Annotation Lexicons: Using the Valency Lexicon for Tectogrammatical Annotation. *The Prague Bulletin of Mathematical Linguistics*, (79–80):61–86, 2003.

Hajič, Jan and Zdeňka Urešová. Linguistic Annotation: from Links to Cross-Layer Lexicons. In Nivre, Joakim and Erhard Hinrichs, editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 69–80, Vaxjo, Sweden, 2003. Vaxjo University Press.

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In Nivre, Joakim and Erhard Hinrichs, editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68, Vaxjo, Sweden, 2003. Vaxjo University Press.

Hajič, Jan, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razímová, and Zdeňka Urešová. Prague Dependency Treebank 2.0. Data/software, 2006.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Prague Czech–English Dependency Treebank 2.0, 2011.

Hajič, Jan, Eva Hajičová, Marie Mikulová, Jiří Mírovský, Jarmila Panevová, and Daniel Zeman. Deletions and node reconstructions in a dependency-based mutlilevel annotation scheme. In Gelbukh, Alexander, editor, *16th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 9041 of *Lecture Notes in Computer Science*, pages 17–31, Berlin / Heidelberg, 2015. Springer.

Hajičová, Eva. *Issues of Sentence Structure and Discourse Patterns*. Charles University Press, Prague, 1993.

Hajičová, Eva. Aspects of discourse structure. In Vertan, Christina, editor, *Natural language processing between linguistic inquiry and system engineering*, pages 47–54, Iasi, 2003. Editura Universitatii Alexandru Ioan Cuza.

Hajičová, Eva. On scalarity in information structure. *Linguistica Pragensia*, XXII(2):60–78, 2012.

Hajičová, Eva and Barbora Vidová-Hladká. What Does Sentence Annotation Say about Discourse? In *18th International Congress of Linguists, Abstracts*, pages 125–126, Seoul, Korea, 2008. The Linguistic Society of Korea.

Hajičová, Eva, Petr Pajas, and Kateřina Veselá. Corpus Annotation on the Tectogrammatical Layer: Summarizing the First Stages of Evaluations. *The Prague Bulletin of Mathematical Linguistics*, 77:5–18, 2002.

Hajičová, Eva, Jiří Mírovský, and Katja Brankatschk. A Contrastive Look at Information Structure: A Corpus Probe. In *Proceedings of the 6th Congres de la Societe Linguistique Slave*, pages 47–51, Aix-en-Provence, 2011. Univ. de Provence.

Hajičová, Eva, Marie Mikulová, and Jarmila Panevová. Reconstruction of Deletions in a Dependency-based Description of Czech: Selected Issues. In Hajičová, Eva and Joakim Nivre, editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 131–140, Uppsala, Sweden, 2015. Uppsala University.

Hajičová, Eva and Jarka Vrbová. On the Role of the Hierarchy of Activation in the Process of Natural Language Understanding. In Horecký, Ján, editor, *Proceedings of the 9th Conference on Computational Linguistics*, pages 107–113, Prague, 1982. Academia.

Hajičová, Eva, Barbara Partee, and Petr Sgall. *Topic–Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer Academic Publishers, Dordrecht, 1998.

Hajič, Jan, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, and Alevtina Bémová. A manual for analytic layer tagging of the Prague Dependency Treebank. Technical Report TR-1997-03, 1997.

Hajič, Jan, Eva Hajičová, Marie Mikulová, and Jiří Mírovský. Prague Dependency Treebank. To be published in Handbook on Linguistic Annotation, eds. N. Ide and J. Pustejovsky. Berlin / Heidelberg: Springer, 2015.

Hajičová, Eva. Postavení rematizátorů v aktuálním členění věty [Position of Rhematizers in the Topic–Focus Articulation]. *Slovo a slovesnost*, 56(4):241–251, 1995.

Hausenblas, Karel. Slovesná kategorie výsledného stavu v dnešní češtině. *Naše řeč*, 46:13–28, 1963.

Jínová, Pavlína, Jiří Mírovský, and Lucie Poláková. Semi-Automatic Annotation of Intra-Sentential Discourse Relations in PDT. In Hajičová, Eva, Lucie Poláková, and Jiří Mírovský, editors, *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA) at Coling 2012*, pages 43–58, Bombay, 2012.

Kingsbury, Paul and Martha Palmer. From TreeBank to PropBank. In *Proceedings of LREC 2002*, pages 1989–1993, Las Palmas, Canary Islands, Spain, 2002.

Koktová, Eva. *Word-order based grammar*, volume 121. Walter de Gruyter, 1999.

Komárek, Miroslav, Jan Petr, Jan Kořenský, Anna Jirsová, Naďa Svozilová, Karel Hausenblas, Jan Balhar, Emil Dvořák, Milena Rulfová, Zdeňka Hrušková, Jarmila Panevová, Eva Buráňová, Libuše Kroupová, and Oldřich Uličný. *Mluvnice češtiny. 2. Tvarosloví [Grammar of Czech. 2. Morphology]*. Academia, Prague, 1986.

Lopatková, Markéta, Zdeněk Žabokrtský, and Václava Kettnerová. *Valenční slovník českých sloves [Valency Dictionary of Czech Verbs]*. Nakladatelství Karolinum, Praha, 2008.

Marcus, Mitchell, Beatrice Santorini, and Marcinkiewicz Mary Ann. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330, 1993.

Mathesius, Vilém. Slovesné časy typu perfektního v hovorové češtině [Verbal tenses of the perfective type in colloquial Czech]. *Naše řeč*, 9(7):200–202, 1925.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. Annotating Noun Argument Structure for NomBank. In *Proceedings of LREC 2004*, pages 803–806, Lisbon, Portugal, 2004.

Mikulová, Marie and Jan Štěpánek. Ways of Evaluation of the Annotators in Building the Prague Czech-English Dependency Treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1836–1839, Valletta, 2010. European Language Resources Association.

Mírovský, Jiří, Lucie Mladová, and Šárka Zikánová. Connective-Based Measuring of the Inter-Annotator Agreement in the Annotation of Discourse in PDT. In Huang, Chu-Ren and Dan Jurafsky, editors, *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 775–781, Beijing, 2010. Tsinghua University Press.

Mírovský, Jiří, Pavlína Jínová, and Lucie Poláková. Does Tectogrammatics Help the Annotation of Discourse? In Kay, Martin and Christian Boitet, editors, *Proceedings of the 24th International Conference on Computational Linguistics*, pages 853–862, Bombay, 2012.

Nedoluzhko, Anna. *Rozšířená textová koreference a asociační anafora (Koncepce anotace českých dat v Pražském závislostním korpusu) [Textual coreference and associative anaphora: The conception of annotation of Czech data in the Prague Dependency Treebank]*. Charles University in Prague, Institute of Formal and Applied Linguistics, Prague, 2011.

Nedoluzhko, Anna and Jiří Mírovský. How Dependency Trees and Tectogrammatics Help Annotating Coreference and Bridging Relations in Prague Dependency Treebank. In Hajičová, Eva, Kim Gerdes, and Leo Wanner, editors, *Proceedings of the Second International Conference on Dependency Linguistics, Depling 2013*, pages 244–251, Praha, Czechia, 2013. Univerzita Karlova v Praze, Matfyzpress.

Pajas, Petr and Jan Štěpánek. System for Querying Syntactically Annotated Corpora. In Lee, Gary and Sabine Schulte im Walde, editors, *Proceedings of the ACL–IJCNLP 2009 Software Demonstrations*, pages 33–36, Suntec, 2009. Association for Computational Linguistics.

Panevová, Jarmila. On Verbal Frames in Functional Generative Description, Parts I, II. *The Prague Bulletin of Mathematical Linguistics*, 22, 23:3–40, 17–52, 1974–75.

Panevová, Jarmila. Verbal Frames Revisited. *The Prague Bulletin of Mathematical Linguistics*, 28: 55–72, 1977.

Panevová, Jarmila. Valency Frames and the Meaning of the Sentence. In Luelsdorff, Ph. L., editor, *The Prague School of Structural and Functional Linguistics*, pages 223–243. Benjamins Publ. House, Amsterdam-Philadelphia, 1994.

Panevová, Jarmila and Magda Ševčíková. The Role of Grammatical Constraints in Lexical Component in Functional Generative Description. In Apresjan, Valentina, Boris Iomdin, and Ekaterina Ageeva, editors, *Proceedings of the 6th International Conference on Meaning-Text Theory*, pages 134–143, Praha, Czechia, 2013. Univerzita Karlova v Praze.

Panevová, Jarmila, Eva Hajičová, Václava Kettnerová, Markéta Lopatková, Marie Mikulová, and Magda Ševčíková. *Mluvnice současné češtiny. 2 [Grammar of Modern Czech. 2]*. Karolinum, Prague, 2014.

Poláková, Lucie, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, 2013. Asian Federation of Natural Language Processing.

Poláková, Lucie, Pavlína Jínová, and Jiří Mírovský. Genres in the Prague Discourse Treebank. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1320–1326, Reykjavik, 2014. European Language Resources Association.

Prasad, Rashmi, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report IRCS-08-01, Philadelphia, 2007.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2961–2968, Marrakech, 2008. European Language Resources Association.

Preinhaelterová, Ludmila. Systemic Ordering of Complementations in English as Tested with Native Speakers of British English. *Linguistica Pragensia*, 7(97):12–25, 1997.

Putnam, Hilary. Some Issues in the Theory of Grammar. In Jakobson, Roman, editor, *The Structure of Language and Its Mathematical Aspects, Proceedings of Symposia in Applied Mathematics*, pages 25–42, Providence, 1961. American Mathematical Society.

Rooth, Mats. *Association with Focus*. PhD thesis, GLSA, Dept. of Linguistics, University of Massachusetts, Amherst, 1985.

Rysová, Kateřina. *O slovosledu z komunikačního pohledu [On Word Order from the Communicative Point of View]*. Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, 2014a.

Rysová, Kateřina, Jiří Mírovský, and Eva Hajičová. On an apparent freedom of Czech word order. A case study. In *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 93–105, Warszawa, Poland, 2015a. IPIPAN.

Rysová, Kateřina, Magdaléna Rysová, and Eva Hajičová. Topic–Focus Articulation in English Texts on the Basis of Functional Generative Description. Technical Report TR 2015-59, Prague, Czechia, 2015b.

Rysová, Magdaléna. Alternative Lexicalizations of Discourse Connectives in Czech. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2800–2807, Istanbul, 2012.

Rysová, Magdaléna. Verbs of Saying with a Textual Connecting Function in the Prague Discourse Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 930–935, Reykjavik, 2014b.

Ševčíková, Magda and Jiří Mírovský. Sentence Modality Assignment in the Prague Dependency Treebank. In Sojka, Petr, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue: 15th International Conference, TSD 2012*, pages 56–63, Berlin/Heidelberg, 2012. Springer.

Sgall, Petr. Functional Sentence Perspective in a Generative Description of Language. *Prague Studies in Mathematical Linguistics*, 2:203–225, 1967.

Sgall, Petr. Towards a Definition of Focus and Topic. *Prague Bulletin of Mathematical Linguistics*, 31:3–25, 1979.

Sgall, Petr, Ladislav Nebeský, Alla Goralčíková, and Eva Hajičová. *A Functional Approach to Syntax in Generative Description of Language*. American Elsevier Publishing Company, New York, 1969.

Sgall, Petr, Eva Hajičová, and Eva Benešová. *Topic, Focus and Generative Semantics*. Scriptor, Kronberg/Taunus, 1973.

Sgall, Petr, Eva Hajičová, and Eva Buráňová. *Aktuální členění věty v češtině [Topic–Focus Articulation in Czech]*. Academia, Prague, 1980.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel Publishing Company, Dordrecht, 1986.

Šmilauer, Vladimír. *Novočeská skladba [Syntax of Modern Czech]*. Ing. Mikuta, Prague, Czechia, 1947.

Štěpánková, Barbora. *K funkci výrazů částicové povahy ve výstavbě textu, zejména k jejich roli v aktuálním členění. [On the function of particles in the structure of text, especially on their role in topic-focus articulation]*. PhD thesis, Charles University, Prague, 2013.

Tesnière, Lucien. *Eléments de syntaxe structurale*. Librairie C. Klincksieck, Paris, 1959.

Urešová, Zdeňka. *Valence sloves v Pražském závislostním korpusu [Valency of Verbs in the Prague Dependency Treebank]*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, 2011a.

Urešová, Zdeňka. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex) [Valency Dictionary of the Prague Dependency Treebank (PDT-Vallex)]*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, 2011b.

Veselá, Kateřina, Jiří Havelka, and Eva Hajičová. Annotators' Agreement: The Case of Topic–Focus Articulation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 2191–2194, Lisbon, 2004.

Zikánová, Šárka. What do the Data in Prague Dependency Treebank Say about Systemic Ordering in Czech? *The Prague Bulletin of Mathematical Linguistics*, 86:39–46, 2006.

Zikánová, Šárka. Problematické syntaktické struktury: k rozborům aktuálního členění v Pražském závislostním korpusu [Probematic syntactic structures: on topic-focus articulation analysis in the Prague Dependency Treebank]. In Polách, Vladimír, editor, *Svět za slovy a jejich tvary, svět za spojením slov*, pages 233–240. Univerzita Palackého, Olomouc, 2008.

Zikánová, Šárka, Miroslav Týnovský, and Jiří Havelka. Identification of Topic and Focus in Czech: Evaluation of Manual Parallel Annotations. *The Prague Bulletin of Mathematical Linguistics*, 87:61–70, 2007.

Address for correspondence:
Eva Hajičová
hajicova@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics, Charles University
Malostranské nám. 25
118 00 Prague 1
Czech Republic

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 125–146

Efficient Word Alignment with Markov Chain Monte Carlo

Robert Östling, Jörg Tiedemann

Department of Modern Languages, University of Helsinki

Abstract

We present EFMARAL, a new system for efficient and accurate word alignment using a Bayesian model with Markov Chain Monte Carlo (MCMC) inference. Through careful selection of data structures and model architecture we are able to surpass the FAST_ALIGN system, commonly used for performance-critical word alignment, both in computational efficiency and alignment accuracy. Our evaluation shows that a phrase-based statistical machine translation (SMT) system produces translations of higher quality when using word alignments from EFMARAL than from FAST_ALIGN, and that translation quality is on par with what is obtained using GIZA++, a tool requiring orders of magnitude more processing time. More generally we hope to convince the reader that Monte Carlo sampling, rather than being viewed as a slow method of last resort, should actually be the method of choice for the SMT practitioner and others interested in word alignment.

1. Introduction

Word alignment is an essential step in several applications, perhaps most prominently phrase-based statistical machine translation (Koehn et al., 2003) and annotation transfer (e.g. Yarowsky et al., 2001). The problem is this: given a pair of translationally equivalent sentences, identify which word(s) in one language corresponds to which word(s) in the other language. A number of off-the-shelf tools exist to solve this problem, but they tend to be slow, inaccurate, or both. We introduce EFMARAL, a new open-source tool[1] for word alignment based on partially collapsed Gibbs sampling in a Bayesian model.

[1]The source code and documentation can be found at `https://github.com/robertostling/efmaral`

 Corresponding author: `robert.ostling@helsinki.fi`
Cite as: Robert Östling, Jörg Tiedemann. Efficient Word Alignment with Markov Chain Monte Carlo. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 125–146. doi: 10.1515/pralin-2016-0013.

2. Background

In order to understand the present work, we first need to formalize the problem and introduce the family of models used (Section 2.1), describe their Bayesian extension (Section 2.2), the Markov Chain Monte Carlo algorithm used for inference (Section 2.3) and its particular application to our problem (Section 2.4).

2.1. The IBM models

The IBM models (Brown et al., 1993) are asymmetric generative models that describe how a *source language* sentence generates a *target language* sentence though a set of latent alignment variables. Since the task at hand is to align the words in the source and target language sentences, the words in both sentences are given, and we are left with inferring the values of the alignment variables.

Formally, we denote the k:th sentence pair $\langle \mathbf{s}^{(k)}, \mathbf{t}^{(k)} \rangle$ with the source sentence $\mathbf{s}^{(k)}$ containing words $s_i^{(k)}$ (for each word index $i \in 1 \dots I^{(k)}$) and the target sentence $\mathbf{t}^{(k)}$ containing words $t_j^{(k)}$ (for $j \in 1 \dots J^{(k)}$).

Each sentence pair $\langle \mathbf{s}^{(k)}, \mathbf{t}^{(k)} \rangle$ is associated with an alignment variable $\mathbf{a}^{(k)}$, where $a_j^{(k)} = i$ indicates that target word $t_j^{(k)}$ was generated by source word $s_i^{(k)}$. This implies an n-to-1 mapping between source and target words, since each target word is aligned to exactly one source word, while each source word can be aligned to zero or more target words.

Sentences are assumed to be generated independently, so the probability of generating a set of parallel sentences $\langle \mathbf{s}, \mathbf{t} \rangle$ is

$$P(\mathbf{t}|\mathbf{s}, \mathbf{a}) = \prod_{k=1}^{K} P(\mathbf{t}^{(k)}|\mathbf{s}^{(k)}, \mathbf{a}^{(k)}) \tag{1}$$

For simplicity of notation, we will drop the sentence index (k) in the following discussion and let $\langle \mathbf{s}, \mathbf{t} \rangle$ instead denote a single sentence pair, without loss of generality due to the independence assumption between sentences.

A source word type e is associated with a *lexical distribution*, modeled by a categorical distribution with parameter vector θ_e. In the simplest of the IBM models (model 1), the probability of generating a target sentence $\mathbf{t}$ is defined as the probability of independently generating each of the J target words independently from the lexical distributions of their respective aligned source words.

$$P(\mathbf{t}|\mathbf{s}, \mathbf{a}) \propto \prod_{j=1}^{J} \theta_{s_{a_j}, t_j} \tag{2}$$

IBM model 1 assumes a uniform distribution for $P(\mathbf{a})$, which effectively means that the word order of the sentences are considered irrelevant. This is clearly not true

in real translated sentences, and in fact a_j and a_{j+1} tend to be strongly correlated. Most research on word alignment has assumed some version of a *word order model* to capture this dependency. Perhaps the simplest version is used in IBM model 2 and the FAST_ALIGN model (Dyer et al., 2013), which are based on the observation that $j/J \approx a_j/I$, in other words that sentences tend to have the same order of words in both languages. This is however a very rough approximation, and Vogel et al. (1996) instead proposed to directly model $P(a_{j+1} - a_j = x|I)$, which describes the length x of the "jump" in the source sentence when moving one word forward in the target sentence, conditioned on the source sentence length I.

Although the IBM models allow n-to-1 alignments, not all values of n are equally likely. In general, high values of n are unlikely, and a large proportion of translations are in fact 1-to-1. The value of n depends both on the particular languages involved (a highly synthetic language like Finnish translated into English would yield higher values than a French to English translation) and on the specific word type. For instance, the German *Katze* 'cat' would typically be translated into a single English word, whereas *Unabhängigkeitserklärung* would normally be translated into two (*independence declaration*) or three words (*declaration of independence*). This can be modeled by defining the *fertility* $\phi(i) = \sum_{j=1}^{J} \delta_{a_j=i}$ of a source token s_i, and introducing a distribution for $P(\phi(i) = n|s_i = e)$ for each source word type e.

A large number of models based on the same general assumptions have been explored (Brown et al., 1993; Toutanova et al., 2002; Och and Ney, 2003), and the interested reader may want to consult Tiedemann (2011) for a more thorough review than we are able to provide in this work.

2.2. Bayesian IBM models

The IBM models make no a priori assumptions about the categorical distributions that define the model, and most authors have used maximum-likelihood estimation through the Expectation-Maximization algorithm (Dempster et al., 1977) or some approximation to it. However, when translating natural languages the lexical distributions should be very sparse, reflecting the fact that a given source word tends to have a rather small number of target words as allowable translations, while the vast majority of target words are unimaginable as translations.

These constraints have recently been modeled with sparse and symmetric Dirichlet priors (Mermer and Saraçlar, 2011; Mermer et al., 2013; Riley and Gildea, 2012) which, beyond capturing the range of lexical distributions we consider likely, also turn out to be mathematically very convenient as the Dirichlet distribution is a conjugate prior to the categorical distribution. The d-dimensional Dirichlet distribution is defined over the space of d-dimensional categorical distributions, and is parameterized by the d-dimensional vector $\boldsymbol{\alpha} > 0$. If $\mathbf{X} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, the probability density function of $\mathbf{X}$ is

given by

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \prod_{i=1}^{d} x_i^{\alpha_i - 1} \tag{3}$$

where the normalization constant Z is given by the multinomial beta function

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{d} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{d} \alpha_i)} \tag{4}$$

A symmetric Dirichlet distribution has $\alpha_i = \alpha_j$ for all i, j, with the interpretation in our case that no particular translation is preferred a priori for any source word type, as this has to be estimated from the data. By also setting $\alpha \ll 1$ we favor sparse lexical distributions where most probabilities are close to zero.

While it is possible to treat $\boldsymbol{\alpha}$ as a latent variable to be inferred, good results can be obtained by using a fixed value roughly in the range of 10^{-6} to 10^{-2} (Riley and Gildea, 2012). Another direction of research has explored hierarchical distributions such as the Pitman-Yor process (Pitman and Yor, 1997) instead of the Dirichlet distribution for the translation distribution priors (Gal and Blunsom, 2013; Östling, 2015). Such distributions offer even greater flexibility in specifying prior constraints on the categorical distributions, but at the cost of less efficient inference. Since the gain in accuracy has turned out to be limited and computational efficiency is an important concern to us, we will not further consider hierarchical priors in this work.

2.3. Markov Chain Monte Carlo

Several different methods have been used for inference in IBM alignment models. Starting with Brown et al. (1993), maximum-likelihood estimation through the Expectation-Maximization (EM) algorithm has been a popular choice. This method is generally efficient for simple models without word order or fertility distributions, but computing the expectations becomes intractable for more complex models such as IBM model 4 so approximative hill-climbing methods are used instead.

Another disadvantage of using plain EM inference with the IBM models is that it is unable to incorporate priors on the model parameters, and as was pointed out in the previous section this deprives us of a powerful tool to steer the model towards more realistic solutions. Riley and Gildea (2012) presented a method to extend the EM algorithm to IBM models with Dirichlet priors, through Variational Bayes inference. Unfortunately, their method inherits the complexity issues of earlier EM approaches.

The inference approach chosen by most authors working on Bayesian IBM models (Mermer and Saraçlar, 2011; Gal and Blunsom, 2013; Östling, 2015) is Gibbs sampling (Gelfand and Smith, 1991), a special case of the Markov Chain Monte Carlo (MCMC) method which we will briefly summarize here.

Given a probability function $p_M(\mathbf{x})$ of some model M on parameter vector $\mathbf{x}$, MCMC provides us with the means to draw samples from p_M. This is done by constructing a Markov chain with values of $\mathbf{x}$ as states, such that its stationary distribution is identical to p_M. In practice, this means deriving expressions for the transition probabilities $P(\mathbf{x}'|\mathbf{x})$ of going from state $\mathbf{x}$ to state $\mathbf{x}'$. Since the number of states is enormous or infinite in typical applications, it is essential that there is some way of sampling efficiently from $P(\mathbf{x}'|\mathbf{x})$. With Gibbs sampling, this is done by sampling one variable from the parameter vector $\mathbf{x}$ at a time, conditioned on all other variables: $P(x_i|x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m)$ which we will write as $P(x_i|\mathbf{x}^{(-i)})$ to indicate conditioning on all elements of x except at index i. All positions i are then sampled in some arbitrary but fixed order. By choosing suitable distributions for the model, the goal in designing a Gibbs sampler is to make sure that this distribution is easy to sample from.

2.4. Gibbs sampling for Bayesian IBM models

The Bayesian version of IBM model 1 defines the following probability over the parameter vector, which consists of the alignment vector $\mathbf{a}$ and the lexical distribution vectors θ_e for each e in the source target vocabulary:

$$P(\mathbf{a}, \theta) = P(\mathbf{s}, \mathbf{t}, \mathbf{a}, \theta, \boldsymbol{\alpha}) \propto \left(\prod_{k=1}^{K} \prod_{j=1}^{J^{(k)}} \theta_{s^{(k)}_{a^{(k)}_j}, t^{(k)}_j} \right) \cdot \left(\prod_{e=1}^{E} \prod_{f=1}^{F} \theta_{e,f}^{\alpha_f - 1} \right) \tag{5}$$

since $\mathbf{s}$, $\mathbf{t}$ and $\boldsymbol{\alpha}$ are constant.

A straightforward Gibbs sampler can be derived by observing that

$$P(x_i|\mathbf{x}^{(-i)}) = \frac{P(\mathbf{x})}{P(\mathbf{x}^{(-i)})} = \frac{P(\mathbf{x}^{(-i)}, x_i)}{P(\mathbf{x}^{(-i)})}$$

which means that

$$P(a_j = i|\mathbf{a}^{(-j)}, \theta) = \frac{P(\mathbf{a}^{(-j)}, a_j = i, \theta)}{P(\mathbf{a}^{(-j)}, \theta)} \propto \theta_{s_{a_j}, t_j} \tag{6}$$

and

$$P(\theta_e = x|\mathbf{a}, \theta^{(-e)}) = \frac{P(\theta^{(-e)}, \theta_e = x|\mathbf{a})}{P(\theta^{(-e)}|\mathbf{a})} = \frac{\prod_{f=1}^{F} x_f^{\alpha_f + c_{e,f} - 1}}{B(\boldsymbol{\alpha}_e + \mathbf{c}_e)} \tag{7}$$

where $c_{e,f}$ is the number of times that word e is aligned to word f given $\mathbf{a}$, $\mathbf{s}$ and $\mathbf{t}$. Equation (7) is a consequence of the fact that the Dirichlet distribution is a conjugate prior to the categorical distribution, so that if

$$\mathbf{x} \sim \mathrm{Dir}(\boldsymbol{\alpha})$$
$$z \sim \mathrm{Cat}(\mathbf{x})$$

then given a sequence z of $|z|$ samples from $\mathrm{Cat}(x)$ we have

$$x|z \sim \mathrm{Dir}(\alpha + c(z)) \tag{8}$$

where

$$c(z)_m = \sum_{i=1}^{|z|} \delta_{z_i=m}$$

is the number of samples in z that are equal to m. This can be easily shown from the definition of the Dirichlet distribution using Bayes' theorem:

$$P(x|\alpha, z) \propto P(z|\alpha, x)P(\alpha, x) \tag{9}$$

$$\propto \prod_{i=1}^{d} x^{\alpha_i - 1} \prod_{i=1}^{|z|} x_{z_i} \tag{10}$$

$$= \prod_{i=1}^{d} x^{\alpha_i - 1} \prod_{m} x^{c(z)_m} \tag{11}$$

$$= \prod_{i=1}^{d} x^{\alpha_i + c(z) - 1} \tag{12}$$

which is the (unnormalized) Dirichlet distribution with parameter $\alpha + c(z)$.

Equation (6) and Equation (7) can be used for sampling with standard algorithms for categorical and Dirichlet distributions, respectively, and together they define an *explicit* Gibbs sampler for the Bayesian IBM model 1. While simple, this sampler suffers from poor mixing (Östling, 2015, section 3.3) and is not a competitive algorithm for word alignment. However, much better performance can be achieved by using a *collapsed* sampler where the parameters θ_e are integrated out so that we only have to derive a sampling equation for the alignment variables $P(a_j = i|a^{(-j)})$.

First we use Equation (5) to derive an expression for $P(a|s, t, \alpha)$, from which the final sampler can be computed as

$$P(a_j = i|a^{(-j)}, s, t, \alpha) = \frac{P(a^{(-j)}, a_j = i|s, t, \alpha)}{P(a^{(-j)}|s, t, \alpha)} \tag{13}$$

Since the elements of a are exchangeable, a sufficient statistic for a is the count vector $c(\cdot)$ where each element

$$c(a, e, f)_{e,f} = \sum_{k=1}^{K} \sum_{j=1}^{J^{(k)}} \delta_{s^{(k)}_{a^{(k)}_j} = e \wedge t^{(k)}_j = f} \tag{14}$$

represents the number of times that source word type e is aligned to target word type f under the alignment $\mathbf{a}$. Next, we marginalize over each of the lexical distributions θ_e.

$$P(\mathbf{a}|\mathbf{s},\mathbf{t},\boldsymbol{\alpha}) = \prod_{e=1}^{E} \int_{\Delta} P\left(\mathbf{a}_{\{j|s_{a_j}=e\}}|\theta_e,\mathbf{s},\mathbf{t},\boldsymbol{\alpha}\right) P\left(\theta_e|\mathbf{s},\mathbf{t},\boldsymbol{\alpha}\right) d\theta_e \tag{15}$$

Substituting from Equation (5) into the integral we have

$$P(\mathbf{a}|\mathbf{s},\mathbf{t},\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{e=1}^{E} \int_{\Delta} \prod_{f=1}^{F} \theta_{e,f}^{c(\mathbf{a},\mathbf{s},\mathbf{t})_{e,f}+\alpha_f-1} \, d\theta_e \tag{16}$$

where the innermost product can be recognized as an unnormalized $\mathrm{Dir}\left(\boldsymbol{\alpha}+\mathbf{c}(\mathbf{a},\mathbf{s},\mathbf{t})\right)$ distribution which has normalization factor $B\left(\boldsymbol{\alpha}+\mathbf{c}(\mathbf{a},\mathbf{s},\mathbf{t})\right)$, so that the final expression becomes

$$P(\mathbf{a}|\mathbf{s},\mathbf{t},\boldsymbol{\alpha}) = \prod_{e=1}^{E} \frac{B\left(\boldsymbol{\alpha}+\mathbf{c}(\mathbf{a},\mathbf{s},\mathbf{t})\right)}{B(\boldsymbol{\alpha})} \tag{17}$$

$$= \prod_{e=1}^{E} \frac{\Gamma\left(\sum_{f=1}^{F} \alpha_f\right) \prod_f \Gamma\left(\alpha_f + c(\mathbf{a},\mathbf{s},\mathbf{t})_{e,f}\right)}{\Gamma\left(\sum_{f=1}^{F} (\alpha_f + c(\mathbf{a},\mathbf{s},\mathbf{t})_{e,f})\right) \prod_f \Gamma(\alpha_f)} \tag{18}$$

Combining Equation (13) with Equation (18) gives us an expression where almost all of the terms are cancelled out, except when $s_i = e$ and $t_j = f$ for which $c(\mathbf{a},\mathbf{s},\mathbf{t})_{e,f}$ and $c(\mathbf{a}^{(-j)},\mathbf{s},\mathbf{t})_{e,f}$ differ by 1. We are left with a remarkably simple sampling distribution:

$$P\left(a_j = i|\mathbf{a}^{(-j)},\mathbf{s},\mathbf{t},\boldsymbol{\alpha}\right) = \frac{\alpha_{t_j} + c(\mathbf{a}^{(-j)},\mathbf{s},\mathbf{t})_{s_i,t_j}}{\sum_{f=1}^{F} \left(\alpha_f + c(\mathbf{a}^{(-j)},\mathbf{s},\mathbf{t})_{s_i,f}\right)} \tag{19}$$

By repeatedly sampling each a_j in turn from Equation (19) we are guaranteed to, in the limit, obtain an unbiased sample from $P(\mathbf{a})$ under the model. What we are really interested in, however, is to estimate the marginal distributions $P(a_j = i)$ as closely as possible while using as little computation as possible, given a sequence of correlated samples $\mathbf{a}^{(t)}$ for time $t \in 1 \ldots T$. Given a sequence of samples $\mathbf{a}^{(t)}$ we can then approximate the marginal distributions

$$P(a_j = i) = \mathop{\mathbb{E}}_{P(\mathbf{a})}[\delta_{a_j=i}] = \sum_{t=1}^{\infty} \delta_{a_j^{(t)}=i} \approx \frac{1}{T} \sum_{t=1}^{T} \delta_{a_j^{(t)}=i} \tag{20}$$

In practice $\mathbf{a}^{(0)}$ will be initialized either from a uniform distribution or by using the output of a simpler model, and the samples will gradually become more independent of $\mathbf{a}^{(0)}$ as t increases. Since $\mathbf{a}^{(0)}$ is likely to lie in a low-probability region of the model,

so do the initial samples, and it is common to use a *burn-in* period and disregard all $\mathbf{a}^{(t)}$ for $t < t_0$. To further ameliorate the problem of initialization bias, it is possible to run several independently initialized samplers and average their results. Combining these methods the marginal distribution approximation becomes

$$P(a_j = i) \approx \frac{1}{N(T - t_0 + 1)} \sum_{n=1}^{N} \sum_{t=t_0}^{T} \delta_{a_j^{(n,t)}=i} \tag{21}$$

where N is the number of independent samplers and t_0 is the length of the burn-in period. Finally, a better estimate can be obtained by applying the Rao-Blackwell theorem (Blackwell, 1947; Gelfand and Smith, 1991), which allows us to re-use the computations of $P(a_j = i|\mathbf{a}^{(-j)})$ during sampling and averaging these distributions rather than $\delta_{a_j^{(t)}=i}$. The final approximation then becomes

$$P(a_j = i) \approx \frac{1}{N(T - t_0 + 1)} \sum_{n=1}^{N} \sum_{t=t_0}^{T} P(a_j^{(n,t)} = i|\mathbf{a}^{(n,t)(-j)}) \tag{22}$$

3. Methods

We now turn to the particular models and algorithms implemented in EFMARAL, presenting our Bayesian HMM model with fertility, the Gibbs sampler used as well as the details on how to make it computationally efficient.

3.1. Alignment model

Our goal in this work is to find a word alignment algorithm that is both accurate and efficient. Previous studies have shown that good word order and fertility models are essential to high accuracy (Brown et al., 1993; Och and Ney, 2003), along with reasonable priors on the parameters (Mermer and Saraçlar, 2011; Östling, 2015). As was discussed in Section 2.3, MCMC algorithms and in particular collapsed Gibbs sampling are particularly suitable for inference in this class of models, as long as the convergence of the Markov chain are sufficiently fast. Even within this class of algorithms there are some trade-offs between accuracy and computational efficiency. In particular, hierarchical priors have been shown to somewhat improve accuracy (Östling, 2015, p. 65), but in spite of improved sampling algorithms (Blunsom et al., 2009) it is still considerably more costly to sample from models with hierarchical priors than with Dirichlet priors.

For these reasons, we use a HMM model for word order based on Vogel et al. (1996) as well as a simple fertility model, and the complete probability of an alignment is essentially the same as Equation (5) with extra factors added for the word order and

fertility model:

$$\begin{aligned}
&P(\mathbf{s}, \mathbf{t}, \mathbf{a}, \theta, \psi, \pi, \alpha, \beta, \gamma) \\
&\quad \propto \left(\prod_{k=1}^{K} \prod_{j=1}^{J^{(k)}} \theta_{s_{a_j^{(k)}}, t_j^{(k)}} \right) \cdot \left(\prod_{e=1}^{E} \prod_{f=1}^{F} \theta_{e,f}^{\alpha_f - 1} \right) \\
&\quad \cdot \left(\prod_{k=1}^{K} \prod_{j=1}^{J^{(k)}+1} \psi_{a_j^{(k)} - a_{j-1}^{(k)}} \right) \cdot \left(\prod_{m=m_{min}}^{m_{max}} \psi_m^{\beta_m - 1} \right) \\
&\quad \cdot \left(\prod_{k=1}^{K} \prod_{i=1}^{I^{(k)}} \pi_{s_i^{(k)}, \phi(i, \mathbf{a}^{(k)})} \right) \cdot \left(\prod_{e=1}^{E} \prod_{n=0}^{n_{max}} \pi_{e,n}^{\gamma_n - 1} \right)
\end{aligned} \tag{23}$$

where $\psi \sim \mathrm{Dir}(\beta)$ are the categorical distribution parameters for the word order model $P(a_j - a_{j-1} = m)$, and $\pi_e \sim \mathrm{Dir}(\gamma)$ for the fertility model $P(\phi(i, \mathbf{a}) | s_i = e)$. In our experiments we fix $\alpha = 0.001$, $\psi = 0.5$ and $\gamma = 1$, but these parameters are not very critical as long as $0 < \alpha \ll 1$.

The IBM models naturally allow unaligned source language words, but in order to also allow target words to not be aligned we use the extension of Och and Ney (2003) to the HMM alignment model, where each source word s_i (from sentence $\mathbf{s}$ of length I) is assumed to have a special NULL word s_{i+I}. The NULL word generates lexical items from the distribution θ_{NULL}, and the word order model is modified so that

$$P(a_j = i + I | a_{j-1} = i') = p_{\text{NULL}} \delta_{i=i'} \tag{24}$$

$$P(a_j = i + I | a_{j-1} = i' + I) = p_{\text{NULL}} \delta_{i=i'} \tag{25}$$

$$P(a_j = i | a_{j-1} = i' + I) = \psi_{i-i'} \tag{26}$$

where p_{NULL} is the prior probability of a NULL word alignment (fixed to 0.2 in our experiments).

We collapse the sampler over θ and ψ in the same manner as was shown in Section 2.4 and obtain the following approximate[2] sampling distribution:

$$\begin{aligned} P(a_j = i|\mathbf{a}^{(-j)}, \mathbf{s}, \mathbf{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto & \frac{\alpha_{t_j} + c(\mathbf{a}^{(-j)}, \mathbf{s}, \mathbf{t})_{s_i, t_j}}{\sum_{f=1}^{F} \left(\alpha_f + c(\mathbf{a}^{(-j)}, \mathbf{s}, \mathbf{t})_{s_i, f}\right)} \\ & \cdot \frac{\beta_{i - a_{j-1}} + c'(\mathbf{a}^{(-j)})_{i - a_{j-1}}}{\sum_{m=m_{min}}^{m_{max}} \left(\beta_m + c'(\mathbf{a}^{(-j)})_m\right)} \\ & \cdot \frac{\beta_{a_{j+1} - i} + c'(\mathbf{a}^{(-j)})_{a_{j-1} - i}}{\sum_{m=m_{min}}^{m_{max}} \left(\beta_m + c'(\mathbf{a}^{(-j)})_m\right)} \\ & \cdot \frac{\pi_{s_i, \phi(i, \mathbf{a}^{(-j)})+1}}{\pi_{s_i, \phi(i, \mathbf{a}^{(-j)})}} \end{aligned} \tag{27}$$

While collapsing over the θ is essential for acceptable mixing in the Markov chain, this is not the case for $\boldsymbol{\pi}$. Instead, we alternate between sampling from Equation (27) and

$$\boldsymbol{\pi}_e \sim \mathrm{Dir}\left(\boldsymbol{\gamma} + c''(\mathbf{a})_e\right) \tag{28}$$

where $c''(\mathbf{a})_e$ is the count vector over the fertility distribution for source word e given alignments $\mathbf{a}$. The advantage of this is that the last product of Equation (27) can be precomputed, saving computation in the inner loop in exchange for the (relatively minor) expense of also sampling from Equation (28).

3.2. Computational efficiency

From Equation (27) it is clear that the computational complexity of sampling sentence k is $O(I^{(k)}J^{(k)})$, since every alignment variable $a_j^{(k)}$ for each $j \in 1 \ldots J^{(k)}$ needs to evaluate the expression in 27 once for each $i \in 1 \ldots I^{(k)}$, and each evaluation requires constant time assuming that the sums are cached. Since sentence lengths are approximately proportional across languages, $I^{(k)} \approx \lambda J^{(k)}$ for some constant λ, this gives a total complexity of $O(\sum I^2)$ per iteration of sampling $\mathbf{a}$. Note that the complexity does not change as we go from Equation (19) for the simple IBM model 1 to Equation (27) for the more complex model with word order and fertility.

In contrast, the corresponding Expectation-Maximization (EM) algorithm for IBM alignment models has $O(\sum I^2)$ complexity in the E-step only for models with simple or no word order model. The HMM-based model of Vogel et al. (1996) can still be implemented relatively efficiently using dynamic programming, but complexity increases to $O(\sum I^3)$. For models with fertility computing the expectations instead becomes intractable, and previous authors have solved this by using approximative

[2]The approximation consists of ignoring the dependence between the two draws from the word order jump distribution (second and third factors).

greedy optimization techniques (Brown et al., 1993) or local Gibbs sampling (Zhao and Gildea, 2010). The main advantage of EM over a collapsed Gibbs sampler is that the former is trivial to parallelize, which makes well-implemented parallel EM-based implementations of simple alignment models with $O(\sum I^2)$ complexity, such as FAST_ALIGN (Dyer et al., 2013), a strong baseline performance-wise.

Algorithm 1 Inner loop of our sampler for IBM model 1

function SAMPLE($a_j^{(k)(-j)}$)
  ▷ Initialize cumulative probability
  $s \leftarrow 0$
  for all $i \in 1 \dots I^{(k)}$ **do**
    ▷ Load denominator reciprocal (small array random access)
    $D^{-1} \leftarrow d_{k,i}$
    ▷ Load numerator index (sequential access)
    $L \leftarrow l_{k,i,j}$
    ▷ Load numerator (large array random access)
    $N \leftarrow u_L$
    ▷ Compute unnormalized probability (one multiplication)
    $\hat{p} \leftarrow D^{-1}U$
    ▷ Accumulate probabilities (one addition)
    $s \leftarrow s + \hat{p}$
    ▷ Store cumulative probability (sequential access)
    $p_i \leftarrow s$
  end for
  ▷ Sample from a uniform distribution on the unit interval
  $r \sim \text{Uniform}(0, 1,)$
  $r \leftarrow r \cdot p_I$
  ▷ Find the lowest i such that $p_i > r$
  $i \leftarrow 1$
  while $p_i \leq r$ **do**
    $i \leftarrow i + 1$
  end while
  $a_j^{(k)} \leftarrow i$
end function

If a collapsed Gibbs sampler is to be a viable option for performance-critical applications, we must pay attention to details. In particular, we propose utilizing the fixed order of computations in order to avoid expensive lookups. Recall that variables $a_j^{(k)}$ are sampled in order, for $k = 1 \dots K$, $j = 1 \dots J^{(k)}$. Now, for each pair $\langle k, j \rangle$ we need

to compute

$$\frac{\alpha_{t_j} + c(\mathbf{a}^{(-j)}, \mathbf{s}, \mathbf{t})_{s_i,t_j}}{\sum_{f=1}^{F} \left(\alpha_f + c(\mathbf{a}^{(-j)}, \mathbf{s}, \mathbf{t})_{s_i,f}\right)}$$

which, if the numerator sum and the reciprocal of the denominator sum are cached in memory, involves two table lookups and one multiplication. Since multiplication is fast and the denominator reciprocal is stored in a relatively small dense array, most attention has to be paid to the numerator lookup, which apart from the constant α_{t_j} is a sparse matrix with non-zero counts $c(\mathbf{a}^{(-j)}, \mathbf{s}, \mathbf{t})_{s_i,t_j}$ only where s_i and t_j are aligned. The standard solution would therefore be to use a hash table with $\langle s_i, t_j \rangle$ as keys to ensure memory efficiency and constant-time lookup. However, most counts are in fact guaranteed to always be zero, as only words from the same parallel sentence pair can be aligned. We are therefore able to construct a count vector $\mathbf{u}$ and an index table $\mathbf{l}$ such that $u_{l_{k,i,j}} = c(\mathbf{a}^{(-j)}, \mathbf{s}, \mathbf{t})_{s_i,t_j} + \alpha_{t_j}$. At the expense of some extra memory usage we are able to achieve the lookup with only two operations, one of which is a cache-efficient sequential memory access. With this method, the inner loop of the sampler for IBM model 1 thus contains only six operations, outlined in algorithm 1. Adding the HMM word order model, two more sequential memory loads and two multiplications are needed, and adding the fertility model requires one more memory load and a multiplication.

4. Related work

In this section we relate our work mainly to the literature on Bayesian models of word alignment, as well as computationally efficient methods for this problem. A comprehensive survey of word alignment methods is beyond the scope of this article, for this we refer the reader to Tiedemann (2011).

Much of research into word alignment has been based on the pioneering work of Brown et al. (1993), and we have already introduced part of their family of IBM alignment models in Section 2.1. Their most advanced models still perform competitively after nearly two decades, but due to their complexity (with exact inference being intractable) many have suggested simpler alternatives, typically by keeping the lexical translation model intact and introducing computationally convenient word order and fertility models so that inference with the Expectation-Maximization (EM) algorithm remains tractable. Notable examples include the simple HMM-based model of Vogel et al. (1996) and the even simpler reparametrized IBM model 2 of Dyer et al. (2013). Neither of these include a model for word fertility, but Toutanova et al. (2002) showed that a simplified fertility model (which only counts alignments from consecutive target words) can be added to the HMM model without increasing the complexity of inference, and more recently this has also been achieved for a general fertility model (Quirk, 2013).

The EM algorithm requires computing the expected values of the alignments, $\mathbb{E}\left[\delta_{a_j=i}\right]$, given the current values of the model parameters. The authors cited above all dealt with this fact by analytically deriving expressions for exact computation of these expectations in their models. Zhao and Gildea (2010) instead chose to use Gibbs sampling to approximate these expectations, which allowed them to perform efficient inference with EM for a HMM model with fertility. Riley and Gildea (2012) later showed how Variational Bayesian techniques can be used to incorporate priors on the parameters of the IBM models, with only minor modifications to the expressions for the alignment expectations.

Recently, several authors have disposed with EM altogether, relying entirely on Gibbs sampling for inference in IBM-based models with Bayesian priors of varying complexity (Mermer and Saraçlar, 2011; Mermer et al., 2013; Gal and Blunsom, 2013; Östling, 2015). Of these, Gal and Blunsom (2013) and to some extent Östling (2015) prioritize maximizing alignment accuracy, which is obtained by using complex hierarchical models. Mermer et al. (2013) use Dirichlet priors with IBM models 1 and 2 to obtain efficient samplers, which they implement in an approximate fashion (where dependencies between variables are ignored during sampling) in order to facilitate parallelization. This article follows previous work by the first author (Östling, 2015), which however was focused on alignment of short parallel text for applications in language typology and transfer learning, rather than efficient large-scale alignment for use with statistical machine translation systems.

5. Results

In this section we first investigate the effect of different parameter settings in EFMARAL, then we proceed with a comparison to two other influential word alignment systems with respect to the performance of statistical machine translation (SMT) systems using the alignments. Since computational efficiency is an important objective with EFMARAL, we report runtime for all experiments.

The following three systems are used in our comparison:

GIZA++: The standard pipeline of IBM models with standard settings of 5 iterations of IBM 1, 5 iterations of the HMM model, and 5 iterations of IBM model 3 and 4 with Viterbi alignments of the final model (Och and Ney, 2003). Class dependencies in the final distortion model use automatically created word clusters using the MKCLS tool, 50 per language.

FAST_ALIGN: An log-linear reparameterization of IBM model 2 using efficient inference procedures and parameter estimations (Dyer et al., 2013). We use the options that favor monotonic alignment points including the optimization procedures that estimate how close they should be to the monotonic diagonal.

EFMARAL: Our implementation of the MCMC alignment approach proposed in this article.

Since these tools all use asymmetric models, we ran each aligner in both directions and applied the GROW-DIAG-FINAL-AND (Section 5.1) or GROW-DIAG-FINAL (Section 5.2) symmetrization heuristic (Och and Ney, 2003, p. 33). This method assumes a set of binary alignments, so for EFMARAL we produce these by choosing the single most probable value for each a_j: $\arg\max_i P(a_j = i)$. In this way the results are more easily comparable to other systems, although some information is lost before the symmetrization step and methods have been explored that avoid this (Matusov et al., 2004; Östling, 2015, pp. 46–47).

5.1. Alignment quality experiments

As discussed in Section 2.4, there are two ways of trading off computing time for approximation accuracy: increasing the number of independent samplers, and increasing the number of sampling iterations. Here we explore the effects of these trade-offs on alignment accuracy.

Following Och and Ney (2003), most subsequent research has compared the results of automatic word alignment to hand-annotated data consisting of two sets of links: S, containing *sure* tuples $\langle i, j\rangle$ where the human judgment is that s_i and t_j must be aligned, and $S \subseteq P$, containing *possible* tuples $\langle i, j\rangle$ where s_i and t_j may be linked. Given a set A of alignments to be evaluated, they define the measures precision (p), recall (r), and alignment error rate (AER) as follows:

$$p = \frac{|A \cap P|}{|A|} \tag{29}$$

$$r = \frac{|A \cap S|}{|P|} \tag{30}$$

$$\text{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \tag{31}$$

While popular, the AER measure is biased towards precision rather than recall and correlates poorly with machine translation performance. Fraser and Marcu (2007) instead suggest to use the F-measure, which favors a balance between precision and recall as defined in Equation (29) and Equation (30):

$$F_\alpha = \left(\frac{\alpha}{p} + \frac{1-\alpha}{r}\right)^{-1} \tag{32}$$

In our experiments, we report both AER and $F_{0.5}$.

In order to evaluate alignment quality we are limited to language pairs with annotated alignment data. For this reason, we use the corpora and test sets from the WPT 2003 and 2005 shared tasks (Mihalcea and Pedersen, 2003; Martin et al., 2005). In addition, we also use the Swedish-English part of the Europarl corpus version 7

Table 1. Data sets used for our alignment quality experiments. The total number of sentences in the respective corpora are given along with the number of sentences and gold-standard (S)ure and (P)ossible alignment links in the corresponding test set.

Corpus	Sentences Training	Sentences Test	|S|	|P|
English-French	1,130,588	447	4,038	17,438
English-Romanian	48,641	200	5,034	5,034
English-Inuktitut	333,185	75	293	1,972
English-Hindi	3,556	90	1,409	1,409
English-Swedish	692,662	192	3,340	4,577

(Koehn, 2005) with test set from Holmqvist and Ahrenberg (2011). The data sets are presented in Table 1, where it can be noted they differ both in size and in annotation style. In particular, the English-Romanian and English-Hindi data only have one set of gold-standard links, so that $S = P$, the English-French and English-Inuktitut data have $|S| \ll |P|$, while the English-Swedish data lies somewhere in between.

Table 2: Results of our alignment quality experiments. All timing and accuracy figures use means from five independently initialized runs. Note that lower is better for AER, higher is better for $F_{0.5}$. All experiments are run on a system with two Intel Xeon E5645 CPUs running at 2.4 GHz, in total 12 physical (24 virtual) cores.

Configuration	Quality AER	Quality $F_{0.5}$	Time (seconds) CPU	Time (seconds) Wall
English-French				
FAST_ALIGN	15.3	86.2	4,124	243
1x iterations, 2 samplers	8.2	92.3	741	270
4x iterations, 2 samplers	8.1	92.2	2,700	809
16x iterations, 2 samplers	8.1	92.1	10,557	2,945
1x iterations, 1 samplers	9.1	91.4	470	248
1x iterations, 4 samplers	7.8	92.6	1,324	298
1x iterations, 8 samplers	**7.6**	**92.9**	2,456	330

Continued on next page

Configuration	AER	$F_{0.5}$	CPU	Wall
English-Hindi				
FAST_ALIGN	67.3	32.7	27	2
1x iterations, 2 samplers	48.3	51.7	107	12
4x iterations, 2 samplers	49.0	51.0	416	46
16x iterations, 2 samplers	51.0	49.0	1,664	183
1x iterations, 1 samplers	49.4	50.6	81	10
1x iterations, 4 samplers	47.5	52.5	146	13
1x iterations, 8 samplers	**46.7**	**53.3**	238	17
English-Inuktitut				
FAST_ALIGN	28.7	78.1	752	48
1x iterations, 2 samplers	22.3	81.5	160	62
4x iterations, 2 samplers	19.7	83.7	560	199
16x iterations, 2 samplers	**17.3**	**86.0**	2,176	747
1x iterations, 1 samplers	23.8	80.1	98	56
1x iterations, 4 samplers	19.6	84.1	259	64
1x iterations, 8 samplers	18.4	85.3	515	72
English-Romanian				
FAST_ALIGN	32.5	67.5	266	17
1x iterations, 2 samplers	28.7	71.3	167	47
4x iterations, 2 samplers	29.0	71.0	648	173
16x iterations, 2 samplers	29.5	70.5	2,580	682
1x iterations, 1 samplers	29.8	70.2	97	43
1x iterations, 4 samplers	28.2	71.8	320	53
1x iterations, 8 samplers	**27.9**	**72.1**	656	59
English-Swedish				
FAST_ALIGN	20.5	79.8	12,298	671
1x iterations, 2 samplers	13.1	87.0	1,606	589
4x iterations, 2 samplers	11.4	88.6	5,989	1,830
16x iterations, 2 samplers	**10.6**	**89.4**	23,099	6,519
1x iterations, 1 samplers	13.8	86.3	1,005	538
1x iterations, 4 samplers	13.2	86.8	2,681	626
1x iterations, 8 samplers	11.7	88.3	6,147	839

Table 2 shows the result of varying the number of samplers and iterations for all the language pairs under consideration. As a baseline for each language pair, we use FAST_ALIGN as well as the default EFMARAL configuration of two independent samplers, running $x = \lfloor 100/\sqrt{K} \rfloor$ sampling iterations where K is the number of parallel sentences in the data (with the additional constraint that $4 \leq x \leq 250$). Following

the practice set by Brown et al. (1993), each model is initialized with the output of a simpler model. For the full HMM+fertility model, we run $\lfloor x/4 \rfloor$ sampling iterations of IBM model 1 initialized with uniformly random alignments, use the last sample to initialize the fertility-less HMM model that we also run for $\lfloor x/4 \rfloor$ iterations. Finally, x samples are drawn from the full model and the final alignments are estimated from these using Equation (22).

The experiments described in Table 2 were carried out on a system with dual Intel Xeon E5645 CPUs, with a total of 24 virtual cores available. Even though this setup strongly favors FAST_ALIGN's parallel implementation, EFMARAL is faster for the largest corpus (where speed matters most) in terms of both wall time and CPU time, and for all but the smallest corpora in CPU time. This trend will also be seen in Section 5.2, where even larger parallel corpora are used for our machine translation experiments.

As expected, increasing the number of independently initialized samplers consistently results in better alignments, in line with research on model averaging for a wide range of machine learning models. When it comes to increasing the number of sampling iterations the result is less clear: for some pairs this seems even more important than the number of independent samplers, whereas for other pairs the quality metrics actually change for the worse. Recall that the samplers are initialized with a sample from the fertility-less HMM model, and that the correlation to this sample decreases as the number of samples from the HMM model with fertility increases. Decreasing quality therefore indicates that for that particular language pair and annotation style, the fertility model performs worse than the mix between the fertility and fertility-less models obtained by using a small number of samples. When interpreting these results, it is also important to keep in mind that the quality metrics are computed using discretized and symmetrized alignments, which are related in a quite complex way to the probability estimates of the underlying model.

From a practical point of view, one should also consider that additional independent samplers can be run in parallel, unlike additional sampling iterations which have a serial dependence. For this reason and because of the consistent improvements demonstrated in Table 2, increasing the number of samplers should be the preferred method for improving alignment quality at the cost of memory and CPU time.

5.2. Machine translation experiments

In order to test the effect of word alignment in a downstream task, we conducted some experiments with generic phrase-based machine translation. Our models are based on the Moses pipeline (Koehn et al., 2007) with data coming from the Workshop on Statistical Machine Translation. In our setup we use the news translation task from 2013 with translation models for English to Czech, German, Spanish, French and Russian and vice versa. Parallel training data comes from Europarl version 7 (Koehn, 2005) (for all language pairs except Russian-English) and the News Commentary corpus version 11. For language modeling, we use the monolingual data sets from Eu-

Table 3. Data used for training SMT models (all counts in millions). Parallel data sets refer to the bitexts aligned to English and their token counts include both languages.

	Monolingual		Parallel	
Language	Sentences	Tokens	Sentences	Tokens
Czech	8.4	145	0.8	41
German	23.1	425	2.1	114
English	17.3	411	–	–
Spanish	6.5	190	2.0	109
French	6.4	173	2.0	114
Russian	10.0	178	0.2	10

roparl and News Commentary as well as the shuffled news texts from 2012. We did not use any of the larger news data sets from more recent years to avoid possible overlaps with the 2013 test set. We apply a pipeline of pre-processing tools from the Moses package to prepare all data sets including punctuation normalization, tokenization, lowercasing and corpus cleaning (for parallel corpora). Statistics of the final data sets are listed in Table 3.

All language models use order five with modified Kneser-Ney smoothing and are estimated using KenLM (Heafield et al., 2013). Word alignments are symmetrized using the GROW-DIAG-FINAL heuristics and we use standard settings to extract phrases and to estimate translation probabilities and lexical weights. For reordering we use the default distance-based distortion penalty and parameters are tuned using MERT (Och, 2003) with 200-best lists.

Table 4 shows the performance of our SMT models given alignments from the different word alignment systems. The left-hand part of the table contains results when using full word forms for the word alignment systems, whereas the results in the right-hand part were obtained by removing any letters after the four first from each word, as a form of approximate stemming since all the languages in our evaluation are predominantly suffixing. Though seemingly very drastic, this method improves accuracy in most cases since data sparsity is a major problem for word alignment.

Next we turn to the computational cost of the experiments just described, these are found in Table 5. In almost all cases, EFMARAL runs faster by a comfortable margin. The only exception is for the smallest dataset, Russian-English, where FAST_ALIGN uses slightly less wall time (but still much more CPU time). This trend is also present in the alignment quality experiments in Section 5.1 with mostly smaller corpora, where EFMARAL is only faster for the largest corpus.[3]

[3]Due to different computing environments, only four CPU cores were available per aligner in the SMT experiments, versus 24 cores in the alignment quality experiments.

Table 4. Results from our SMT evaluation. The BLEU scores are the maximum over the Moses parameters explored for the given word alignment configuration.

Translation pair	BLEU score					
	No stemming			4-prefix stemming		
	EFMARAL	GIZA++	FAST_ALIGN	EFMARAL	GIZA++	FAST_ALIGN
Czech-English	**23.43**	23.29	22.77	**23.58**	23.57	23.44
English-Czech	**16.22**	15.97	15.69	**16.11**	15.96	15.88
German-English	23.60	**23.86**	22.84	23.54	**23.80**	23.08
English-German	**17.83**	17.69	17.50	**17.77**	17.70	17.65
Spanish-English	**28.50**	28.43	28.25	28.57	**28.69**	28.20
English-Spanish	27.39	**27.51**	27.08	**27.49**	**27.49**	27.08
French-English	**28.50**	28.45	28.06	**28.69**	28.67	28.33
English-French	**27.73**	27.57	27.22	27.66	**27.71**	27.16
Russian-English	**20.74**	20.14	19.55	**20.96**	20.65	20.38
English-Russian	**15.89**	15.55	15.07	**16.17**	16.13	15.77

Table 5. Timings from the word alignments for our SMT evaluation. The values are averaged over both alignment directions. For these experiments we used systems with 8-core Intel E5-2670 processors running at 2.6 GHz.

		Time (seconds)					
		Wall	CPU	Wall	CPU	Wall	CPU
Translation pair	Stem	EFMARAL		GIZA++		FAST_ALIGN	
Czech-English	no	**303**	**462**	13,089	13,083	465	1,759
Czech-English	yes	**233**	**361**	12,035	12,033	311	1,200
German-English	no	**511**	**766**	42,077	41,754	1,151	4,407
German-English	yes	**377**	**558**	43,048	43,023	813	3,115
Spanish-English	no	**500**	**782**	39,047	39,003	1,034	3,940
Spanish-English	yes	**346**	**525**	38,896	38,866	758	2,911
French-English	no	**696**	**1,088**	41,698	41,646	1,681	6,423
French-English	yes	**383**	**583**	40,986	40,907	805	3,101
Russian-English	no	122	**206**	3583	3581	**107**	382
Russian-English	yes	87	**151**	3148	3143	**78**	292

6. Concluding remarks

We hope that the reader at this point is convinced that Bayesian alignment models with Markov Chain Monte Carlo inference should be the method of choice for researchers who need to align large parallel corpora. To facilitate a practical shift towards this direction, we have released the EFMARAL tool which the evaluations in this article show to be both accurate, computationally efficient, and useful as a component of practical machine translation systems.

Acknowledgments

Computational resources for this project were provided by CSC, the Finnish IT Center for Science.[4]

Bibliography

Blackwell, David. Conditional Expectation and Unbiased Sequential Estimation. *The Annals of Mathematical Statistics*, 18(1):105–110, 03 1947. doi: 10.1214/aoms/1177730497. URL http://dx.doi.org/10.1214/aoms/1177730497.

Blunsom, Phil, Trevor Cohn, Sharon Goldwater, and Mark Johnson. A Note on the Implementation of Hierarchical Dirichlet Processes. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 337–340, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1667583.1667688.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, June 1993. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=972470.972474.

Dempster, A. P., N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N13-1073.

Fraser, Alexander and Daniel Marcu. Measuring Word Alignment Quality for Statistical Machine Translation. *Computational Linguistics*, 33(3):293–303, Sept. 2007. ISSN 0891-2017. doi: 10.1162/coli.2007.33.3.293. URL http://dx.doi.org/10.1162/coli.2007.33.3.293.

Gal, Yarin and Phil Blunsom. A Systematic Bayesian Treatment of the IBM Alignment Models. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA, 2013. Association for Computational Linguistics.

[4] https://www.csc.fi/

Gelfand, Alan E. and Adrian F. M. Smith. Gibbs Sampling for Marginal Posterior Expectations. Technical report, Department of Statistics, Stanford University, 1991.

Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of ACL*, pages 690–696, 2013.

Holmqvist, Maria and Lars Ahrenberg. A Gold Standard for English-Swedish Word Alignment. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, number 11 in NEALT Proceedings Series, pages 106–113, 2011.

Koehn, Philipp. Europarl: A Parallel Corpus for Statistical Machine Translation. In *The Tenth Machine Translation Summit.*, Phuket, Thailand, 2005.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1073445.1073462. URL http://dx.doi.org/10.3115/1073445.1073462.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL*, pages 177–180, 2007.

Martin, Joel, Rada Mihalcea, and Ted Pedersen. Word Alignment for Languages with Scarce Resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05, pages 65–74, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1654449.1654460.

Matusov, Evgeny, Richard Zens, and Hermann Ney. Symmetric Word Alignments for Statistical Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/1220355.1220387.

Mermer, Coşkun and Murat Saraçlar. Bayesian Word Alignment for Statistical Machine Translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 182–187, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-88-6. URL http://dl.acm.org/citation.cfm?id=2002736.2002775.

Mermer, Coşkun, Murat Saraçlar, and Ruhi Sarikaya. Improving Statistical Machine Translation Using Bayesian Word Alignment and Gibbs Sampling. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):1090–1101, May 2013. ISSN 1558-7916. doi: 10.1109/TASL.2013.2244087.

Mihalcea, Rada and Ted Pedersen. An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond - Volume 3*, HLT-NAACL-PARALLEL '03, pages 1–10, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1118905.1118906. URL http://dx.doi.org/10.3115/1118905.1118906.

Och, Franz Josef. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*, pages 160–167, 2003.

Och, Franz Josef and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, Mar. 2003. ISSN 0891-2017. doi: 10.1162/089120103321337421. URL http://dx.doi.org/10.1162/089120103321337421.

Östling, Robert. *Bayesian Models for Multilingual Word Alignment*. PhD thesis, Stockholm University, 2015. URL http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-115541. ISBN 978-91-7649-151-5.

Pitman, Jim and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997. doi: 10.1214/aop/1024404422.

Quirk, Chris. Exact Maximum Inference for the Fertility Hidden Markov Model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 7–11, 2013. URL http://aclweb.org/anthology/P/P13/P13-2002.pdf.

Riley, Darcey and Daniel Gildea. Improving the IBM Alignment Models Using Variational Bayes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 306–310, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390665.2390736.

Tiedemann, Jörg. *Bitext Alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2011.

Toutanova, Kristina, H. Tolga Ilhan, and Christopher Manning. Extensions to HMM-based Statistical Word Alignment Models. In *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 87–94, 2002. URL http://ilpubs.stanford.edu:8090/557/.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann. HMM-Based Word Alignment in Statistical Translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 836–841, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics. doi: 10.3115/993268.993313. URL http://dx.doi.org/10.3115/993268.993313.

Yarowsky, David, Grace Ngai, and Richard Wicentowski. Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics. doi: 10.3115/1072133.1072187. URL http://dx.doi.org/10.3115/1072133.1072187.

Zhao, Shaojun and Daniel Gildea. A Fast Fertility Hidden Markov Model for Word Alignment Using MCMC. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 596–605, Cambridge, MA, USA, October 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D10-1058.

Address for correspondence:
Robert Östling
robert.ostling@helsinki.fi
PL 24 (Unionsgatan 40, A316)
00014 Helsingfors universitet, Finland

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 147–158

Qualitative: Python Tool for MT Quality Estimation Supporting Server Mode and Hybrid MT

Eleftherios Avramidis

German Research Center for Artificial Intelligence (DFKI) Berlin, Germany

Abstract

We are presenting the development contributions of the last two years to our Python open-source Quality Estimation tool, a tool that can function in both experiment-mode and online web-service mode. The latest version provides a new MT interface, which communicates with SMT and rule-based translation engines and supports on-the-fly sentence selection. Additionally, we present an improved Machine Learning interface allowing more efficient communication with several state-of-the-art toolkits. Additions also include a more informative training process, a Python re-implementation of QuEst baseline features, a new LM toolkit integration, an additional PCFG parser and alignments of syntactic nodes.

1. Introduction

After almost a decade of research, evaluating Machine Translation (MT) remains an active topic. Through the years, a multitude of methods have been researched, in a continuous effort to assess whether the MT output adheres to the users expectations. For a significant amount of time, ranking has been the dominant approach for MT equality, since it seems a relatively robust way to circumvent the subjectivity of perceiving quality (Callison-Burch et al., 2007, 2008; Bojar et al., 2015, etc.).

Many automatic metrics have been developed in order to measure MT quality by comparing it to the reference translations (e.g. Papineni et al., 2002), facing the limitation that the reference represents usually only one of the possible translations. As a more recent development, Quality Estimation (QE) has shifted the focus from the reference translations towards the translations themselves, by identifying qualitative features that can be indications of a good translation.

 Corresponding author: eleftherios.avramidis@dfki.de
Cite as: Eleftherios Avramidis. Qualitative: Python Tool for MT Quality Estimation Supporting Server Mode and Hybrid MT. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 147–158.
doi: 10.1515/pralin-2016-0014.

The work presented here is a programming effort to ease research in both aspects sketched above. We present the latest version of Qualitative (Avramidis et al., 2014), a QE tool that processes translations as a ranking, in an attempt to learn better the human preferences. At the same time, extensive engineering takes place to devise new features by enabling various natural language processing (NLP) methods.

The version of the toolkit presented here is a result of more than two years of development and offers a data processing unit, a powerful feature generation engine with feature selection, a machine learning interface and a collection of evaluation metrics. Additionally, it can perform hybrid machine translation by communicating with several MT engines and combining their output on a sentence level. The development takes place in GitHub [1] and further contributions are welcome.

2. Related work

Few pieces of software on QE have been released with an open source. QuEst (Specia et al., 2013), previously also known as Harvest, is the most established one, as it has been used as a baseline for the yearly WMT Shared Tasks on QE (e.g. Callison-Burch et al., 2012). The main difference with our approach is that it uses two different pieces of software for feature generation and machine learning, where the former is written in Java and the latter in Python. Additionally, many parts of it operate only in batch mode. For these two reasons, in contrast to our software, operating in a real-usage scenario (e.g. server mode) with sentence-level requests is non-trivial. Its latest version, QuEst++ (Specia et al., 2015), additionally supports word-level and document-level QE.

Some most recent software focuses on an another level of granularity, namely word-level QE. WceLig (Servan et al., 2015) is a tool which introduces support for various target languages, handles glass-box, lexical, syntactic and semantic features for estimating confidence at word-level. Marmot (Logacheva et al., 2016), focuses on word-level and phrase-level QE and is written in Python. It offers a modular architecture, users can easily add or implement new parsers, data representations and features that fit their particular use cases, whereas it can be easily plugged into a standard experiment workflow.

In contrast to most of the above software, the approach of the software presented here focuses on a double-usage scenario for both scientific experimentation and real-usage. Feature generators and machine learning support both batch mode and sentence-level mode, whereas the functionality can be easily plugged into web-services and other software that requires QE functionality. Furthermore, it offers a dynamic pipeline architecture, including wrappers for NLP tools written in several programming languages.

[1] The source code, along with basic documentation for installation, execution and further development can be found at `https://github.com/lefterav/qualitative`

3. Design

The software has been developed based on a multilateral design that serves the operational requirements sketched above. This section includes the main architecture of the pipeline and the interoperability features with embedded software.

3.1. Architecture

The software consists of:

- a **data processing unit** able to read XML and text-based input,
- a **pre-processing stage** that performs the necessary string normalisation process for the languages at hand,
- a **machine translation** module, which communicates with external MT systems and handles sentence-level system combination,
- a **feature generation engine** that produces hundreds of dynamic black-box and glass-box features, by harvesting the output of embedded open-source NLP tools,
- a **machine learning** interface that embeds widely-used ML libraries, including data conversion to their internal structures. Additionally there are pairwise wrappers that allow the usage of binary classifiers for ranking and
- an **evaluation package** that includes several metrics for measuring ranking and translation performance.

3.2. Interoperability

A detailed diagram of the architecture can be seen in Figure 1. Additionally to the core architecture which is seen in the horizontal axis, the system integrates external components developed in various programming languages. These 25 components are integrated using 9 different approaches, including native Python libraries, sockets to the Java Virtual Machine (JVM), wrappers, system pipes and remote service APIs (e.g. JSON, REST).

The majority of these tools are seemlessly integrated and available as callable Python objects throughout the entire pipeline. For instance, Truecasing (Wang, Wei and Knight, Kevin and Marcu, 2006) is done with the original Moses scripts via Perl pipes, features from PCFG parsing are collected through a JVM socket from Berkeley Parser (Petrov et al., 2006), whereas Machine Translation is fetched from Moses (Koehn et al., 2006) via XML-RPC. More details on the interoperability interfaces can be found in Avramidis (2016).

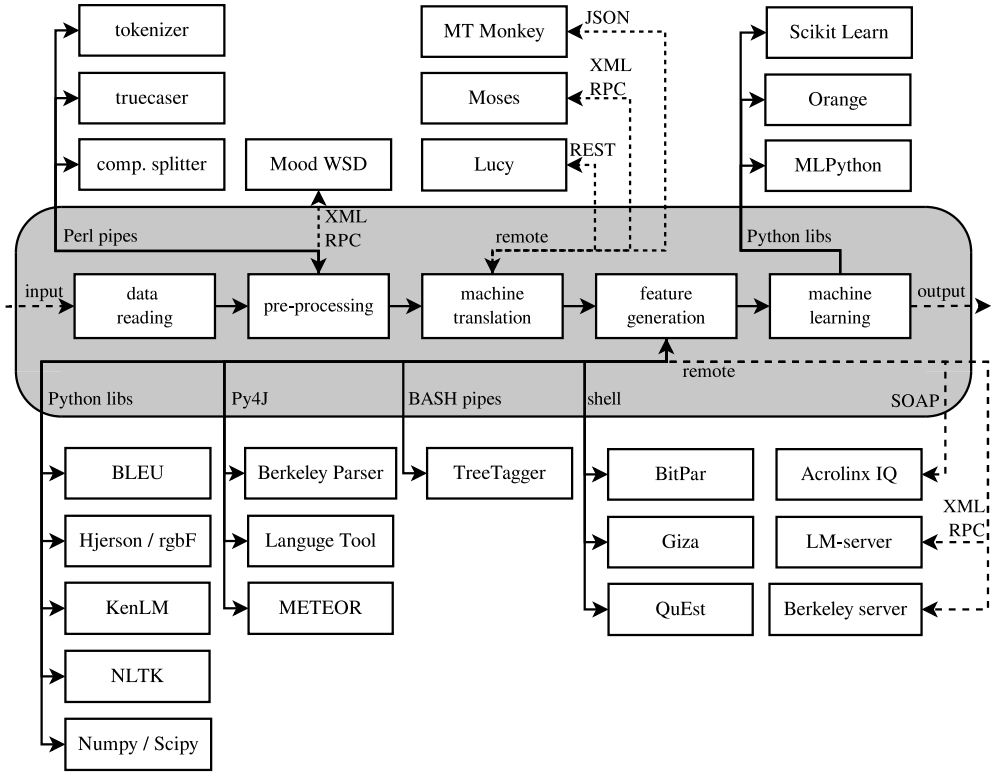


Figure 1. Full diagram of the components that have been integrated into the application. Source: (Avramidis, 2016)

```
<?xml version="1.0" encoding="utf-8"?> <jcml>
  <judgedsentence langsrc="en" id="11" langtgt="de">
    <src>Go to System Preferences</src>
    <tgt system="pilot_0" berkley-loglikelihood="-84.9089431054"
         berkeley-n="19" rank="2">Gehen Sie zu Systemeinstellungen</tgt>
    <tgt system="pilot_2" berkley-loglikelihood="-74.6569913407"
         berkeley-n="5" rank="3">Sprung zu Systemeinstellungen</tgt>
    <ref berkley-loglikelihood="-83.3551531463"
         berkeley-n="18" >Gehen Sie zu Systemeinstellungen</ref>
  </judgedsentence>
</jcml>
```

Figure 2. Sample JCML file, containing a source sentence, the reference and two translations with Berkeley Parser scores and human ranks

4. New functionality

For the generic functionality, including instructions on how to run the tool, the reader is referred to (Avramidis et al., 2014) and for the underlying methodology to Avramidis (2012a). Here, we outline the functionality introduced in the latest version.

4.1. Data Processing

The majority for the read/write processing of the software is done in a special XML format, the JCML format, which stands for *Judged Corpus Markup Language*. It is a simple XML format that has been devised so that it allows dynamic feature lists but at the same time it can be inspected manually. The latest improvements include incremental reading and writing, a feature which solved many memory-related issues, given the big volume of some data sets. There are also several scripts that allow the conversion from and to other common formats. A sample of such a file can be seen in Figure 2.

4.2. Machine translation

One of the basic applications of the automatic ranking is the possibility to combine differen systems on the sentence level. Such a method is often referred to as a case of *hybrid* MT when it combines different types of systems (e.g. statistical and rule-based). This version offers a new package that handles the communication with translation engines by connecting to remote APIs. It currently supports Moses (Koehn et al., 2006), Lucy (Alonso and Thurmair, 2003), as well as MT-Monkey (Tamchyna et

al., 2013) for accessing deployed server installations. The communication with the engines allows fetching translations and glass-box features (translation probability, unknown words etc.), when these are made available by the engine.

Additionally, some techniques of hybridisation are included, such as serial post-editing of a rule-based system with a statistical system (SMT) (as implemented for Avramidis et al., 2015), partial translation of terminology from the rule-based system with SMT, and SMT including an alternative decoding path from WSD disambiguation (Avramidis et al., 2016).

The machine translation interface, apart from being a step of the QE pipeline, it can also act as a standalone application, or as a web-service pluggable via XML-RPC.

4.3. Feature Generation

The modular interface of the feature generation pipeline allows easy plugging of new Feature Generators. These are classes which process the text of the sentences and return numerical values that describe some aspects of quality. The existing functionality, presented in the past, includes usage of language models, PCFG parsing, cross-target BLEU and METEOR (Banerjee and Lavie, 2005), language correction, IBM-1 probabilities, as well as token counts.

The new version offers additionally:

- **word alignment** based on the IBM-1 model (Brown et al., 1993), allowing to derive the count of aligned PCFG tree spans, nodes and leaves between the source and the target sentence. Whereas this generates hundreds of sparse features, the most prominent of them are expected to help isolate systems that fail to translate grammatically important chunks of the source,
- relative and absolute **position** of every PCFG tag within the sentence, with the goal to capture wrong positioning of grammatical chunks in languages where this is important (e.g. German),
- a re-implementation of the **WMT baseline features** (Callison-Burch et al., 2012) in Python, including the average number of translations per source word in the segment as given by IBM-1 model with probabilities thresholded in different ways, and the average number of occurrences of the target word within the target segment,
- integration of **KenLM** (Heafield, 2011) via its Python library, which allows efficient of loading compiled language models, removing the previous requirement for an external LM server,
- a wrapper for the PCFG parser **BitPar** (Schmid, 2004), as an alternative for Berkeley Parser (integration based on van Cranenburgh, 2010; van Cranenburgh et al., 2010),
- a wrapper for the **TreeTagger** (Schmid, 1994), which acquires the necessary POS tags for Hjerson (Popović, 2011)

- a connector to the XML-RPC of MoodWSD (Weissenborn et al., 2015), an external word sense disambiguator

4.4. Machine Learning

A new more transparent and modular internal interface allows for integration of several external machine learning (ML) toolkits. The integration of every ML toolkit should extend an abstract class named `Ranker`. This should implement some basic functions, such as training on a batch of sentences, or producing the ranking given one source sentence and its translations. The implementation of every ML toolkit is also responsible of converting the given sentence data and its features to the data structures understood by the toolkit. Binary classifiers, where available, are wrapped to provide a ranker's functionality.

Currently the following toolkits and learning methods are supported:

- Orange (Demšar et al., 2004) with k-Nearest Neighbours, Logistic Regression with Stepwise Feature Set Selection or L2-regularisation and C45 trees,
- SciKit Learn (Pedregosa et al., 2011) with Support Vector Machines with Grid parameter optimisation over cross-validation, Decision Trees, Gaussian Naïve Bayes, Linear and Quadratic Discriminant Analysis, Bagging, Adaptive Boosting and Gradient Boosting and feature selection methods such as Recursive Feature Elimination with Cross-Validation
- MLpython[2] with listwise ranking methods, such as ListNet (Cao et al., 2007).

4.5. Evaluation

The evaluation phase is the last part of the experiment process, as the trained models are tested against gold-sets and need to be evaluated accordingly. The evaluation phase offers a wide range of ranking metrics, with latest additions the inverse-weighed Kendall's τ and its theoretical p-values and confidence intervals. Finally, the evaluation phase includes automatic metric scores (BLEU, METEOR, TER, WER, Hjerson) for the performance of the system combination and its components against the reference translation.

4.6. Experimental management

Similarly to the previous version, experimenting over the training of new models is organised by using the ExpSuite (Rückstieß and Schmidhuber, 2011). This allows the exhaustive exploration of several experimental settings in parallel. We have extended the functionality to provide out-of-the-box parellelised **cross-validation** for any given dataset. Additionally, the split training and test-sets of the cross-validation are **cached**

[2]MLpython is described at `http://www.dmi.usherb.ca/~larocheh/mlpython/`

in a common directory, so that they can be re-used for different experimental settings which require the same original dataset. The experiments can be resumed from the step they were left, in case of any unexpected interruption.

The experiment pipeline keeps a structured log of every step of the experiment, which may include the results of the evaluation, but also details about the machine learning process (e.g. the beta coefficients of a log-linear model, or weights of a linear model). The trained models are also dumped in external files, so that they can be re-used later. After all iterations and cross-validation folds of the experiment are concluded, a script allows for creating a comma-separated table that compares all experimental settings against a desired set of metrics.

5. Further work

There are often upgrades to the integrated external software that fix issues or provide additional functionality. Although sticking to older tested versions usually suffices, further work may include adaptations for newer versions of this software. In this direction, adjusting the current Python 2.7 code to support Python 3 would be useful.

Whereas the interface for machine learning over ranking has been re-developed as outlined above, most parts of the pipeline have been used for other types of quality estimation, such as quality score prediction for single outputs (Avramidis, 2012b) and error prediction (Avramidis and Popovic, 2014). Small extensions to provide abstract classification and regression interfaces for all ML toolkits would be desirable.

We are also aware that the glass-box feature integration requires extensions to support most MT-engines, although this faces the barrier that not all glass-box features are easily available.

Finally, big-amounts of data, despite the huge potential for machine learning, create bottlenecks in case they must be analyzed or processed selectively. We plan to support more effective data types (e.g. JSON). A possible solution would include the implementation of smart databases and other data-effective techniques.

Acknowledgements

This work has received support by the EC's FP7 (FP7/2007-2013) under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches". Early stages have been developed with the support of the projects TaraXÚ and QT-Launchpad. Many thanks to: Slav Petrov for modifying the Berkeley Parser in order to allow modification of parsing parameters; Andreas van Cranenburgh, Hieu Hoang, Philipp Koehn, Nitin Madnani, Laurent Pointal, Maja Popović, Josh Schroeder and David Vilar as parts of their open source code have been included in some of our scripts.

Bibliography

Alonso, Juan A and Gregor Thurmair. The Comprendium Translator system. In *Proceedings of the Ninth Machine Translation Summit*. International Association for Machine Translation (IAMT), 2003.

Avramidis, Eleftherios. Comparative Quality Estimation: Automatic Sentence-Level Ranking of Multiple Machine Translation Outputs. In *Proceedings of 24th International Conference on Computational Linguistics*, pages 115–132, Mumbai, India, dec 2012a. The COLING 2012 Organizing Committee. URL http://www.aclweb.org/anthology/C12-1008.

Avramidis, Eleftherios. Quality estimation for Machine Translation output using linguistic analysis and decoding features. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 84–90, Montréal, Canada, jun 2012b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W12-3108.

Avramidis, Eleftherios. Interoperability in MT Quality Estimation or wrapping useful stuff in various ways. In *Proceedings of the LREC 2016 Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 1–6. Language Science Press, 2016. URL http://www.cracking-the-language-barrier.eu/wp-content/uploads/Avramidis.pdf.

Avramidis, Eleftherios and Maja Popovic. Correlating decoding events with errors in Statistical Machine Translation. In Sangal, Rajeev, Jyoti Pawar, and Dipti Misra Sharma, editors, *Proceedings of the 11th International Conference on Natural Language Processing. International Conference on Natural Language Processing (ICON-2014), 11th International Conference on Natural Language Processing, December 18-21, Goa, India*. International Institute of Information Technology, Natural Language Processing Association, India, 2014. URL https://www.dfki.de/lt/publication{_}show.php?id=7605.

Avramidis, Eleftherios, Lukas Poustka, and Sven Schmeier. Qualitative: Open source Python tool for Quality Estimation over multiple Machine Translation outputs. *The Prague Bulletin of Mathematical Linguistics*, 102:5–16, 2014. URL http://ufal.mff.cuni.cz/pbml/102/art-avramidis-poustka-schmeier.pdf.

Avramidis, Eleftherios, Maja Popovic, and Aljoscha Burchardt. DFKI's experimental hybrid MT system for WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation. Workshop on Statistical Machine Translation (WMT-2015), 10th, September 17-18, Lisbon, Portugal*, pages 66–73. Association for Computational Linguistics, 2015. URL http://aclweb.org/anthology/W15-3004.

Avramidis, Eleftherios, Burchardt, Aljoscha, Vivien Macketanz, and Ankit Srivastava. DFKI's system for WMT16 IT-domain task, including analysis of systematic errors. In *Proceedings of the First Conference on Machine Translation*, pages 415–422, Berlin, Germany, aug 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W16/W16-2329.

Banerjee, Somnath and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, 2005.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, sep 2015. Association for Computational Linguistics. URL http://aclweb.org/anthology/W15-3001.

Brown, Peter F, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2): 263–311, 1993. ISSN 0891-2017.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation (StatMT'07)*, pages 136–158, Prague, Czech Republic, jun 2007. Association for Computational Linguistics. doi: 10.3115/1626355.1626373. URL http://www.statmt.org/wmt07/pdf/WMT18.pdf.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, jun 2008. Association for Computational Linguistics. URL www.aclweb.org/anthology/W08-0309.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, jun 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W12-3102.

Cao, Zhe, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM, 2007.

Demšar, Janez, Blaž Zupan, Gregor Leban, and Tomaz Curk. Orange: From Experimental Machine Learning to Interactive Data Mining. In *Principles of Data Mining and Knowledge Discovery*, pages 537–539, 2004. doi: 10.1007/b100704.

Heafield, Kenneth. KenLM : Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, number 2009, pages 187–197, Edinburgh, Scotland, jul 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W11-2123$\delimiter"056E30F$nhttp://kheafield.com/code/kenlm.

Koehn, Philipp, Wade Shen, Marcello Federico, Nicola Bertoldi, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Ondrej Bojar, Richard Zens, Alexandra Constantin, Evan Herbst, and Christine Moran. Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180, Prague, Czech Republic, jun 2006.

Logacheva, Varvara, Chris Hokamp, and Lucia Specia. MARMOT: A Toolkit for Translation Quality Estimation at the Word Level. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, jul 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135.

Pedregosa, F, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Petrov, Slav, Leon Barrett, Romain Thibaux, and Dan Klein. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia, jul 2006. Association for Computational Linguistics.

Popović, Maja. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96(-1):59–68, 2011. doi: 10.2478/v10108-011-0011-4.PBML. URL `www.mt-archive.info/PBML-2011-Popovic.pdf`.

Rückstieß, Thomas and Jürgen Schmidhuber. A Python Experiment Suite. *The Python Papers*, 6 (1):2, 2011.

Schmid, Helmut. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Schmid, Helmut. Efficient Parsing of Highly Ambiguous Context-free Grammars with Bit Vectors. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. doi: 10.3115/1220355.1220379. URL `http://dx.doi.org/10.3115/1220355.1220379`.

Servan, Christophe, Ngoc-Tien Le, Ngoc Quang Luong, Benjamin Lecouteux, and Laurent Besacier. An Open Source Toolkit for Word-level Confidence Estimation in Machine Translation. In *The 12th International Workshop on Spoken Language Translation (IWSLT'15)*, Da Nang, Vietnam, dec 2015. URL `https://hal.archives-ouvertes.fr/hal-01244477`.

Specia, Lucia, Kashif Shah, José Guilherme Camargo de Souza, and Trevor Cohn. QuEst - A translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria, aug 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P13-4014`.

Specia, Lucia, Gustavo Paetzold, and Carolina Scarton. Multi-level Translation Quality Prediction with QuEst++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China, jul 2015. Association for Computational Linguistics and The Asian Federation of Natural Language Processing. URL `http://www.aclweb.org/anthology/P15-4020`.

Tamchyna, Aleš, Ondřej Dušek, Rudolf Rosa, and Pavel Pecina. MTMonkey: A Scalable Infrastructure for a Machine Translation Web Service. *The Prague Bulletin of Mathematical Linguistics*, 100:31–40, oct 2013. ISSN 0032-6585. URL `http://ufal.mff.cuni.cz/pbml/100/art-tamchyna-dusek-rosa-pecina.pdf`.

van Cranenburgh, Andreas. Enriching Data-Oriented Parsing by blending morphology and syntax. Technical report, University of Amsterdam, Amsterdam, 2010. URL https://unstable.nl/andreas/ai/coglang/report.pdf.

van Cranenburgh, Andreas, Galit W Sassoon, and Raquel Fernández. Invented antonyms: Esperanto as a semantic lab. In *Proceedings of the 26th Annual Meeting of the Israel Association for Theoretical Linguistics (IATL 26).*, volume 26, 2010.

Wang, Wei and Knight, Kevin and Marcu, Daniel. Capitalizing machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conferenc*, pages 1–8, New York, 2006. URL http://dl.acm.org/citation.cfm?id=1220836.

Weissenborn, Dirk, Leonhard Hennig, Feiyu Xu, and Hans Uszkoreit. Multi-Objective Optimization for the Joint Disambiguation of Nouns and Named Entities. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 596–605, Beijing, China, 2015. Association for Computer Linguistics. ISBN 978-1-941643-72-3. URL http://aclweb.org/anthology/P/P15/P15-1058.pdf.

Address for correspondence:
Eleftherios Avramidis
eleftherios.avramidis@dfki.de
German Research Center for Artificial Intelligence (DFKI GmbH)
Language Technology Lab
Alt Moabit 91c
10559 Berlin, Germany

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 159–168

Otedama: Fast Rule-Based Pre-Ordering for Machine Translation

Julian Hitschler[a], Laura Jehl[a], Sariya Karimova[ac], Mayumi Ohta[a], Benjamin Körner[a], Stefan Riezler[ab]

[a] Computational Linguistics, Heidelberg University, Germany
[b] IWR, Heidelberg University, Germany
[c] Kazan Federal University, Russia

Abstract

We present Otedama,[1] a fast, open-source tool for rule-based syntactic pre-ordering, a well established technique in statistical machine translation. Otedama implements both a learner for pre-ordering rules, as well as a component for applying these rules to parsed sentences. Our system is compatible with several external parsers and capable of accommodating many source and all target languages in any machine translation paradigm which uses parallel training data. We demonstrate improvements on a patent translation task over a state-of-the-art English-Japanese hierarchical phrase-based machine translation system. We compare Otedama with an existing syntax-based pre-ordering system, showing comparable translation performance at a runtime speedup of a factor of 4.5–10.

1. Introduction

Syntactic pre-ordering is a commonly used pre-processing technique in state-of-the-art machine translation systems. It attempts to adjust the syntax of the source language to that of the target language by changing the word order of a source sentence prior to translation. Originally, this technology was devised to overcome a weakness of classical phrase-based translation systems (Koehn et al., 2003), which usually penalize moving target phrases far away from their source positions. This is a major

[1] An open-source version is available at `https://github.com/StatNLP/otedama`. Otedama is named after a Japanese juggling game.

 Corresponding author: `riezler@cl.uni-heidelberg.de`
Cite as: Julian Hitschler, Laura Jehl, Sariya Karimova, Mayumi Ohta, Benjamin Körner, Stefan Riezler. Otedama: Fast Rule-Based Pre-Ordering for Machine Translation. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 159–168. doi: 10.1515/pralin-2016-0015.

source of errors when translating between languages with heterogeneous and dissimilar sentence structures. Hierarchical phrase-based translation systems do not place a similar prior penalty on phrase reordering during decoding, however, such systems have been shown to profit from syntactic pre-ordering as well (de Gispert et al., 2015).

Otedama implements a variant of the approach of Genzel (2010), which learns cascading lists of rules for syntactic transformation. Unlike early works on pre-ordering which rely on hand-written rules (Collins et al., 2005), Otedama automatically extracts rules from parse trees. While recent work has applied other learning algorithms to the problem of learning pre-ordering models (Lerner and Petrov, 2013; Jehl et al., 2014; de Gispert et al., 2015; Nakagawa, 2015), automatic rule-learning is a popular and suitable choice for pre-ordering systems because syntactic features are discrete and relatively dense and the resulting models allow for very fast application to system input. In particular, the rule-based approach deals well with the combinatorial explosion that is incurred when attempting to train in-domain pre-ordering models on data with a high prevalence of long sentences, as we demonstrate in our system evaluation. Furthermore, our approach is compatible with nearly any external parsing framework, in difference to approaches where preordering and parsing need to be tightly connected for improved performance (Nakagawa, 2015). Despite the fact that pre-ordering continues to be an important technique in high-quality machine translation, so far, there is a lack of an open-source implementation of learners and online application systems for pre-ordering that are convenient to use and fast enough to be suitable for rapid prototyping and meaningful comparison of new approaches to existing baselines. We created Otedama to address this lack. We compare Otedama to two variants of an open-source pre-orderer by Neubig et al. (2012) which induces a bracketing transduction grammar for producing a re-ordered string. Our system yields comparable improvements in translation quality at a runtime speedup of a factor of 4.5–10. Our tool is available as open-source code and compatible with nearly any external parser.

2. Specifications

2.1. Model Formalism

Our model formalism follows the work of Genzel (2010). The model is trained based on syntactic parse trees of the source sentences in a parallel corpus, in addition to a bilingual word alignment. Parse trees are obtained from an automatic dependency parser. By introducing head nodes, non-projective dependency graphs are converted to a tree format (see Figure 2a). In order to obtain good results, the parser should produce labeled dependencies. Otedama provides bindings to the Stanford Parser and is fully compatible with any parser that produces POS tags and dependency labels in the `CoNLL` output format,[2] such as, for example, the Parzu parser for

[2] `http://ilk.uvt.nl/conll/`

```
function ExampleRule(Node N)
    if N.tag = _VBD AND
      N.dep = root AND
      N.parent.tag = ROOT AND
      N.children[1].tag = VBD AND
      N.children[1].dep = head AND
      N.children[2].tag = _NN AND
      N.children[2].dep = dobj then
        swap(N.children[1], N.children[2])
    end if
end function
```

Figure 1: Example pre-ordering rule. This rule swaps a past tense verb (VBD) with a noun phrase (_NN).

German (Sennrich et al., 2009).[3] Thus, Otedama is able to process a wide variety of source languages.

The rules learned by Otedama comprise a matching context for nodes in a parse tree, expressed in terms of the POS tags and dependency labels of a node, its parent node, and a sequence of neighboring children of the node. Furthermore, a reordering operation is defined on the sequence of neighboring children, which permutes the positions of the children of the node, thereby making syntactic changes to the source language corpus with the goal of approximating the target language syntax. An example rule is given in Figure 1. Rules are learned iteratively, meaning that a rule which is found useful (i.e. that increases the alignment monotonicity of the training corpus) is first applied to the entire corpus before further rules are tested and applied. This results in a cascading list of rules, which can be applied to source language parse trees before translation.

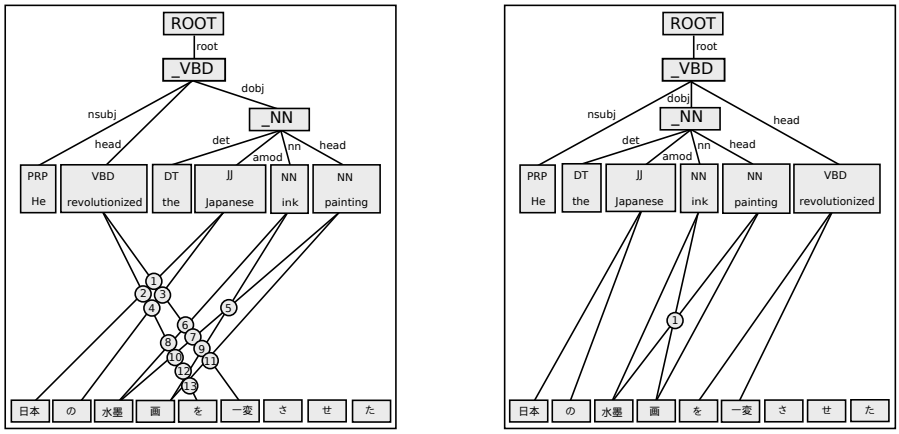
In order to avoid combinatorial explosion, the permutations are restricted to a sliding window, the size of which can be specified to be either 2, 3, or 4 (henceforth referred to as parameter l). For a window size of four, child node reordering is therefore restricted to children that are at most three nodes apart from each other for any one rule.

2.2. Training Procedure

We follow the basic procedure delineated by Genzel (2010) for learning pre-ordering rules, with some modifications. We first describe the basic training procedure.

This objective of the training procedure is to minimize the number of alignment crossings (Genzel, 2010) in a parallel training corpus. For example, the sentence pair shown in Figure 2a has 13 alignment crossings. Applying the example rule from Figure 1 reduces the number of alignment crossings to 1, as shown in Figure 2b. This

[3] https://github.com/rsennrich/ParZu



(a) Before rule application

(b) After rule application

Figure 2: Example parse tree with POS tags, dependency labels and alignment crossings (shown as numbered dots). The left side shows the original tree, the right side shows the same tree after applying the rule from Figure 1. This rule swaps the second and third child node of the "`_VBD`"-node, resulting in a reduction of alignment crossings from 13 to 1. Some alignment links were omitted for clarity.

metric is trivial to compute for any word alignment and can be used as a proxy measure for evaluating pre-ordering models.

Training proceeds iteratively until a convergence criterion is reached or a fixed number of iterations or amount of runtime have elapsed. Each iteration carries out a two-step procedure: In the first step, all possible candidate rules from a small, random subset of the training corpus are generated. This is done by extracting all rule contexts (consisting of POS-tags and dependency labels of the current node, its parent, and a span of its child nodes up to the window size), and pairing them with all possible permutations of the given nodes for all dependency trees in the subset. Only such candidate rules which locally increase alignment monotonicity are generated. In the second step, each candidate rule is evaluated by applying it to the entire training corpus (or, in Genzel's case, a second, larger subset thereof) and measuring the reduction of alignment crossings. The candidate rules which increase the alignment monotonicity of the training corpus and fulfill certain variance constraints are added to the final rule set and applied to all training data. The partially reordered training set is then used as input to the next iteration in which a new random subset is sampled for rule candidate generation. This iterative procedure has two advantages: First, it is likely to extract the most general rules, i.e. the rules which have the largest effect on all

training examples in the beginning, and then subsequently proceed to more specific rules, because sentences with frequent syntactic constructions are more likely to be sampled than those with infrequent ones. This means that if training time is limited, the most important rules will still be extracted. Second, even though the rule is restricted to a window of a small number of nodes, the iterative procedure allows for rules to be applied sequentially to the child nodes of the same parent, thus achieving long-range permutations.

The following paragraphs describe our modifications of Genzel's training recipe.

Scalability Genzel's approach was designed for a MapReduce architecture, since the rule extraction and evaluation at each training iteration can be easily parallelized. Training on a large training set, however, requires a large parallel infrastructure running MapReduce. Our implementation aims to eliminate the dependency on MapReduce, allowing the user to run Otedama on a single machine, and to provide a flexible solution which can be adjusted for the available resources. We achieve these goals by dynamically adjusting the size of the sampled subset of the training instances which is used to generate candidate rules at each iteration. Since the candidate generation step is much more expensive than the rule evaluation, we can still calculate crossing alignment reduction on the entire training set, allowing for good generalization. The initial sample size (henceforth referred to as parameter m) can be set by the user. If the number of rules learned in the last iteration is below a minimum value of 20 or above a maximum value of 1000, the sample size is adjusted in the following iteration by doubling or halfing it. The candidate rule generation and scoring then follows Genzel's recipe. Our implementation supports multi-threading on a single machine.

Variance Constraints Depending on the quality of the parser and the syntactic complexity of the training data, feedback on the effect of rules from reduction in alignment crossings can be quite noisy. Otedama provides the option of specifying a variance constraint for rule candidates in the form of a minimum ratio between sentences in the training data on which alignment crossing is reduced, compared to those on which it increases (henceforth referred to as parameter v). For example, setting $v = 2$ means that only such rules that increase the alignment monotonicity of at least two times the number of sentences than the number of sentences for which alignment monotonicity decreases should be retained during training.

Rule Application Otedama provides two options for making rule application more flexible: At training time, it can be specified whether rules with feature sets that are subsets of the matching context should also be extracted as candidates. For a rule extracted from a node with two children, where the original matching context comprises eight features (the POS-tag and dependency labels of the parent, the node, and the two children, respectively), this means that all 254 nonempty subsets of the origi-

nal matching context are extracted and evaluated as candidate rules. After candidate rules are created from a node, they are ordered increasingly by the size of their feature sets. This means that Otedama first evaluates the most general rules with the fewest required features and, if no rule is found to increase alignment monotonicity and fulfill the specified variance constraints, proceeds to try the more specific rules. In addition, we allow fuzzy matching of rules by specifying the maximum number of features to be matched globally, at both training and test times.

3. Evaluation

As demonstrated by de Gispert et al. (2015), it is potentially beneficial to train pre-ordering models on in-domain data rather than to deploy a model trained on a different domain. For our evaluation we selected the challenging domain of patents. Patents contain very long, convoluted sentences, specialized vocabulary and idioms. They are thus a poorer fit for a general purpose parser than other data. However, we will show that pre-ordering improves translation quality on translation from English into Japanese (EN-JA) and from German into English (DE-EN).

Baseline SMT system All our systems use the hierarchical phrase-based paradigm (Chiang, 2007) as implemented by the `cdec` decoder (Dyer et al., 2010). Our English-Japanese system was trained on the NTCIR7 patent MT data set by Utiyama and Isahara (2007) (1.8M sentence pairs). We used the official development sets dev1 and dev2 for tuning, while reserving dev3 for evaluation. The Japanese side was segmented using MeCab[4]. The English side was tokenized and true-cased using scripts from the Moses toolkit[5]. Our German-English system was trained on 750K sentence pairs from the PatTr corpus (Wäschle and Riezler, 2012). Both sides of the data set were tokenized and true-cased. Both systems used MGIZA++[6] for word alignment. The alignments were produced by the training script provided in the Moses toolkit with the default number of IBM Model 1–4 and HMM iterations (5 iterations of IBM Model 1 and HMM Model, 3 iterations IBM Model 3 and 4). Alignments were symmetrized using the grow-diag-final-and heuristic. We trained a 5-gram target side language model using `lmplz` (Heafield et al., 2013). Rule extraction was performed using `cdec`'s default parameters (maximum rule span = 15, maximum number of symbols per rule = 5). Our system was tuned with the pairwise ranking optimizer `dtrain` (Simianer et al., 2012). Tuning was run for 15 iterations, using a k-best size of 100 and a constant learning rate of 0.00001. Final weights were obtained by averaging over all

[4] `http://taku910.github.io/mecab/`

[5] `https://github.com/moses-smt/mosesdecoder`

[6] `http://www.cs.cmu.edu/~qing/giza/`

system	# crossing alignments	% of baseline
Baseline	1840536	100
Otedama	1465120	79.60
Lader class	1447312	78.64
Lader full	1364459	74.13

Table 1: EN-JP crossing scores

iterations. Both tuning and decoding used a cube pruning pop-limit of 200. We report BLEU scores we calculated using `MultEval` (Dyer et al., 2011) on tokenized output.

Pre-orderer training Pre-ordering models were trained on 100,000 parallel sentence pairs from our parallel training data. [7] English source data was parsed with the Stanford Parser (Socher et al., 2013), German source data with the Parzu parser (Sennrich et al., 2009). In contrast to Genzel (2010), we used IBM Model 4 alignments instead of IBM Model 1 alignments in order to reduce noise on our smaller data set. We re-used the symmetrized word alignments that were created during baseline training, as described in the previous paragraph. We tried out various configurations of Otedama's hyper-parameters window size $l \in \{3, 4\}$ and variance constraints $v \in \{0, 2, 5, 10\}$. For both language pairs, $l = 3$ and $v = 2$, without fuzzy rule matching or feature subsets performed best. The number of matching features was therefore set to the maximum value of 10.[8]The initial subsample size for rule candidate generation was kept constant at $m = 10$ throughout our experiments. Training was stopped after exhaustion of a fixed runtime budget. The rules extracted by Otedama were applied to our MT training and testing data, and word alignments of the training data were re-calculated using GIZA++ with the same settings as above. Hierarchical rule extraction and tuning were then carried out in the standard way, using the same settings as our baseline system.

Instructions for training and applying Otedama, along with a small example script and recommended parameter values, are provided on the GitHub page.

Comparative pre-ordering system We compare our system to `Lader` (Neubig et al., 2012).[9] `Lader` is an open-source reordering tool for machine translation. It performs a

[7]The size of the data set was chosen due to the runtime constraints of our comparison system in order to ensure meaningful comparison. Due to its parallelized architecture, Otedama could process significantly larger data sets during learning with only negligible overhead costs.

[8]This is the maximum number for a window size of 3. With a window of 4, there would be at most 12 features.

[9]`https://github.com/neubig/lader`

large-margin training treating the parser's derivation tree as a latent variable. `Lader` allows to define various features. One possibility is to train a pre-ordering model only from the parallel text and word classes (*Lader class*). Another option is to enhance the model with additional linguistically informed features (POS tags and parse trees), if available (*Lader full*). In both cases, the phrase table was also used as a feature to keep frequent contiguous phrases. In our experiments we have replicated the standard feature set from Neubig et al. (2012) using feature templates. To calculate classes we utilized word classes computed by GIZA++. To obtain POS tags and parse trees the Stanford tagger and the Stanford lexicalized PCFG parser were used, for both English and German as source languages. We trained our models with the default learner Pegasos. Due to the time constraints training was stopped after 10–15 iterations and the model with the best alignment crossing score selected. We did not observe improvements in alignment monotonicity on the training data past 5 iterations. Training times were comparable for Otedama and Lader models, depending on the exact Lader model specifications.

Intrinsic Evaluation by Crossing Score The crossing score counts the number of crossing alignments in a heldout set of 10K sentence pairs. For EN-JA Otedama and `Lader` achieved crossing score reductions of over 20 points. However, `Lader` performed slightly better than Otedama under the intrinsic metric. Our implementation includes a script, `crossing-score.py`, for evaluating crossing score on heldout data. This measure is useful for quickly evaluating different configurations of Otedama, as well as comparing it to other pre-ordering approaches.

Evaluation by MT Performance Table 2 shows BLEU scores for Otedama and `Lader` experiments. For English-Japanese translation, Otedama performed on par with the *Lader class* model. Both models significantly outperformed the baseline by 0.7 and 0.8 BLEU. The best model, *Lader full*, outperformed Otedama by 1.2 BLEU points. However, the ranking changes for the German-English experiment. Here, Otedama and *Lader full* were indistinguishable, and both systems significantly improved over the baseline. *Lader class* did not produce a significant improvement, showing the importance of syntactic information for this language pair.

While the `Lader` models were equally good or better in terms of BLEU, this came at the cost of speed. Table 2 lists running times for the different systems. The benchmark was conducted on 100 randomly selected sentences, running 10 threads in parallel. We include parsing time for *Otedama* and *Lader full* (*Lader class* does not require parsing). Otedama ran 4.5–10 times faster than `Lader`, making it more practical to use, especially on large data.

4. Conclusion

We have presented a fast, flexible open-source implementation of automatic rule-learning for source-side pre-ordering from dependency-annotated aligned parallel

system	BLEU		Seconds/sentence	
	EN-JA	DE-EN	EN-JA	DE-EN
Baseline	31.6	38.1		
Otedama	32.3*	38.8*	0.64	0.35
Lader class	32.4*	38.4	2.89 ($\times$4.5)	2.38 ($\times$6.9)
Lader full	33.5*+	38.6*	4.58 ($\times$7.1)	3.72 ($\times$10.7)

Table 2: BLEU scores and run times (including parsing, where necessary) for different pre-ordering tools. * indicates a statistically significant improvement over the baseline. + indicates a statistically significant improvement over Otedama.

training data. Our tool supports several external parser formats, and has shown promising results on the difficult task of patent translation. Compared to another open-source pre-ordering tool, we achieved a speedup of 4.5–10 while maintaining translation performance.

Bibliography

Chiang, David. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, 2007.

Collins, Michael, Philipp Koehn, and Ivona Kučerová. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, Stroudsburg, PA, USA, 2005.

de Gispert, Adrià, Gonzalo Iglesias, and Bill Byrne. Fast and Accurate Preordering for SMT using Neural Networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, 2015. Association for Computational Linguistics.

Dyer, Chris, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. cdec: A Decoder, Alignment, and Learning Framework for Finite-State and Context-Free Translation Models. In *Proceedings of the ACL 2010 System Demonstrations*, Uppsala, Sweden, 2010.

Dyer, Chris, Jonathan H. Clark, Alon Lavie, and Noah A. Smith. Unsupervised Word Alignment with Arbitrary Features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, 2011.

Genzel, Dmitriy. Automatically Learning Source-side Reordering Rules for Large Scale Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, August 2010.

Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 2013.

Jehl, Laura, Adrià de Gispert, Mark Hopkins, and Bill Byrne. Source-side Preordering for Translation using Logistic Regression and Depth-first Branch-and-Bound Search. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics*, Edmonton, Alberta, Canada, 2003.

Lerner, Uri and Slav Petrov. Source-Side Classifier Preordering for Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, October 2013.

Nakagawa, Tetsuji. Efficient Top-Down BTG Parsing for Machine Translation Preordering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, 2015.

Neubig, Graham, Taro Watanabe, and Shinsuke Mori. Inducing a Discriminative Parser to Optimize Machine Translation Reordering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, Stroudsburg, PA, USA, 2012.

Sennrich, Rico, Gerold Schneider, Martin Volk, and Martin Warin. A New Hybrid Dependency Parser for German. In *Proceedings of the GSCL-Conference*, Potsdam, Germany, 2009.

Simianer, Patrick, Stefan Riezler, and Chris Dyer. Joint feature selection in distributed stochastic learning for large-scale discriminative training in SMT. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, South Korea, 2012.

Socher, Richard, John Bauer, Christopher D. Manning, and Ng Andrew Y. Parsing with Compositional Vector Grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

Utiyama, Masao and Hitoshi Isahara. A Japanese-English patent parallel corpus. In *Proceedings of MT summit XI*, Copenhagen, Denmark, 2007.

Wäschle, Katharina and Stefan Riezler. Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus. *Multidisciplinary Information Retrieval*, pages 12–27, 2012.

Address for correspondence:
Stefan Riezler
riezler@cl.uni-heidelberg.de
Computational Linguistics
Heidelberg University
Im Neuenheimer Feld 325
69120 Heidelberg, Germany

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 169–179

FaDA: Fast Document Aligner using Word Embedding

Pintu Lohar, Debasis Ganguly, Haithem Afli, Andy Way, Gareth J.F. Jones

ADAPT Centre
School of computing
Dublin City University
Dublin, Ireland

Abstract

FaDA[1] is a free/open-source tool for aligning multilingual documents. It employs a novel crosslingual information retrieval (CLIR)-based document-alignment algorithm involving the distances between embedded word vectors in combination with the word overlap between the source-language and the target-language documents. In this approach, we initially construct a pseudo-query from a source-language document. We then represent the target-language documents and the pseudo-query as word vectors to find the average similarity measure between them. This word vector-based similarity measure is then combined with the term overlap-based similarity. Our initial experiments show that s standard Statistical Machine Translation (SMT)-based approach is outperformed by our CLIR-based approach in finding the correct alignment pairs. In addition to this, subsequent experiments with the word vector-based method show further improvements in the performance of the system.

1. Introduction

A crosslingual document alignment system aims at efficiently extracting likely candidates of aligned documents from a comparable corpus in two or more different languages. Such a system needs to be effectively applied to a large collection of documents. As an alternative approach, a state-of-the-art machine translation (MT) system (such as Moses, Koehn et al., (2007)) can be used for this purpose by translateing every source-language document with an aim of representing all the documents in the

[1] Available at `https://github.com/gdebasis/cldocalign/`

 Corresponding author: haithem.afli@adaptcentre.ie
Cite as: Pintu Lohar, Debasis Ganguly, Haithem Afli, Andy Way, Gareth J.F. Jones. FaDA: Fast Document Aligner using Word Embedding. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 169–179.
doi: 10.1515/pralin-2016-0016.

same vocabulary space. This in turn facilitates the computation of the text similarity between the source-language and the target-language documents. However, this approach is rather impractical when applied to a large collection of bilingual documents, because of the computational overhead of translating the whole collection of source-language documents into the target language.

To overcome this problem, we propose to apply an inverted index-based cross-language information retrieval (CLIR) method which does not require the translation of documents. As such, the CLIR apporach results in much reduction computation compared to the MT-based method. Hence we refer to our tool using the CLIR approach as the *Fast document aligner (FaDA)*. Our FaDA system works as follows. Firstly, a pseudo-query is constructed from a source-language document and is then translated with the help of a dictionary (obtained with the help of a standard word-alignment algorithm (Brown et al., 1993) using a parallel corpus). The pseudo-query is comprised of the representative terms of the source-language document. Secondly, the resulting translated query is used to extract a ranked list of documents from the target-language collection. The document with the highest similarity score is considered as the most likely candidate alignment with the source-language document.

In addition to adopted a standard CLIR query-document comparison, the FaDA systems explores the use of a word-vector embedding approach with the aim of building a semantic matching model in seeks to improve the performance of the alignment system. The word-vector embedding comparison method is based on the relative distance between the embedded word vectors that can be estimated by a method such as 'word2vec' (Mikolov et al., 2013). This is learned by a recurrent neural network (RNN)-based approach on a large volume of text. It is observed that the inner product between the vector representation of two words u and v is high if v is likely to occur in the context of u, and low otherwise. For example, the vectors of the words 'child' and 'childhood' appear in similar contexts and so are considered to be close to each other. FaDA combines a standard text-based measure of the vocabulary overlap between document pairs, with the distances between the constituent word vectors of the candidate document pairs in our CLIR-based system.

The remainder of the paper is organized as follows. In Section 2, we provide a literature survey of the problem of crosslingual document alignment. In Section 3, the overall system architecture of *FaDA* is described. In Section 4, we describe our experimental investigation. The evaluation of the system is explained in Section 5. Finally, we conclude and suggest possible future work in Section 6.

2. Related Work

There is a plethora of existing research on discovering similar sentences from comparable corpora in order to augment parallel data collections. Additionally, there is also existing work using the Web as a comparable corpus in document alignment. For example, Zhao and Vogel (2002) mine parallel sentences from a bilingual compa-

rable news collection collected from the Web, while Resnik and Smith (2003) propose a web-mining-based system, called STRAND, and show that their approach is able to find large numbers of similar document pairs. Bitextor[2] and ILSPFC[3] follow similar web-based methods to extract monolingual/multilingual comparable documents from multilingual websites.

Yang and Li (2003) present an alignment method at different levels (title, word and character) based on dynamic programming (DP) to identify document pairs in an English-Chinese corpus collected from the Web, by applying the longest common subsequence to find the most reliable Chinese translation of an English word. Utiyama and Isahara (2003) use CLIR and DP to extract sentences from an English-Japanese comparable corpus. They identify similar article pairs, consider them as parallel texts, and then align the sentences using a sentence-pair similarity score and use DP to find the least-cost alignment over the document pair.

Munteanu and Marcu (2005) use a bilingual lexicon to translate the words of a source-language sentence to query a database in order to find the matching translations. The work proposed in Afli et al. (2016) shows that it is possible to extract only 20% of the true parallel data from a collection of sentences with 1.9M tokens by employing an automated approach.

The most similar work to our approach is described in Roy et al. (2016). In this documents and queries are represented as sets of word vectors, similarity measure between these sets calculated, and then combine with IR-based similarities for document ranking.

3. System architecture of FaDA

The overall architecture of FaDA comprises two components; (i) the CLIR-based system, and (ii) the word-vector embedding system.

3.1. CLIR-based system

The system diagram of our CLIR-based system is shown in Figure (1). The source-language and the target-language documents are first indexed, then each of the indexed source-language documents is used to construct a pseudo-query. However, we do not use all the terms from a source-language document to construct the pseudo-query because very long results in a very slow retrieval process. Moreover, it is more likely that a long query will contain many 'outlier' terms which are not related to the core topic of the document, thus reducing the retrieval effectiveness. Therefore, we use only a fraction of the constituent terms to construct the pseudo-query, which are considered to be suitably representative of the document.

[2] http://bitextor.sourceforge.net/

[3] http://nlp.ilsp.gr/redmine/projects/

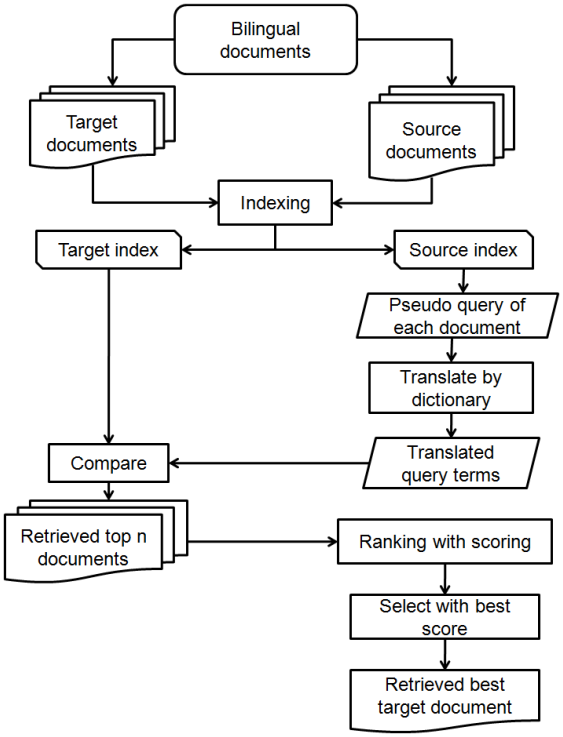


Figure 1. Architecture of the CLIR-based system

To select the terms to include in the pseudo-query we use the score shown in Equation (1), where $\mathrm{tf}(t, d)$ denotes the term frequency of a term t in document d, $\mathrm{len}(d)$ denotes the length of d, and N and $\mathrm{df}(t)$ denote the total number of documents and the number of documents in which t occurs, respectively. Furthermore, $\tau(t, d)$ represents the term-selection score and is a linear combination of the normalized term frequency of a term t in document d, and the inverse document frequency (idf) of the term.

$$\tau(t, d) = \lambda \frac{\mathrm{tf}(t, d)}{\mathrm{len}(d)} + (1 - \lambda) \log\left(\frac{N}{\mathrm{df}(t)}\right) \quad (1)$$

It is obvious that in Equation (1) the terms that are frequent in a document d and the terms that are relatively less frequent in the collection are prioritized. The parameter λ controls the relative importance of the tf and the idf factors. Using this function, each term in d is associated with a score. This list of terms is sorted in de-

creasing order of this score. Finally, a fraction σ (between 0 and 1) is selected from this sorted list to construct the pseudo-query from d. Subsequently, the query terms are translated by a source-to-target dictionary, and the translated query terms are then compared with the indexed target-language documents. After comparison, the top-n documents are extracted and ranked using the scoring method in Equation (3), which is explained in Section 3.2.1. Finally, to select the best candidate for the alignment, we choose the target-language document with the highest score.

3.2. Word-vector embedding-based system

In addition to the CLIR framework described in Section 3.1, we also use the vector embedding of words and incorporate them with the CLIR-based approach in order to estimate the semantic similarity between the source-language and the target-language documents. This word-embedding approach facilitates the formation of "bag-of-vectors" (BoV) which helps to express a document as a set of words with one or more clusters of words where each cluster represents a topic of the document.

Let the BoW representation of a document d be $W_d = \{w_i\}_{i=1}^{|d|}$, where $|d|$ is the number of unique words in d and w_i is the i^{th} word. The BoV representation of d is the set $V_d = \{x_i\}_{i=1}^{|d|}$, where $x_i \in \mathbb{R}^p$ is the vector representation of the word w_i. Let each vector representation x_i be associated with a latent variable z_i, which denotes the topic or concept of a term and is an integer between 1 and K, where the parameter K is the total number of topics or the number of Gaussians in the mixture distribution. These latent variables, z_is, can be estimated by an EM-based clustering algorithm such as K-means, where after the convergence of K-means on the set V_d, each z_i represents the cluster id of each constituent vector x_i. Let the points $C_d = \{\mu_k\}_{k=1}^{K}$ represent the K cluster centres as obtained by the K-means algorithm. The posterior likelihood of the query to be sampled from the K Gaussian mixture model of a document d^T, centred around the μ_k centroids, can be estimated by the average distance of the observed query points from the centroids of the clusters, as shown in Equation (2).

$$P_{WVEC}(d^T|q^S) = \frac{1}{K|q|} \sum_i \sum_k \sum_j P(q_j^T|q_i^S) q_j^T \cdot \mu_k \tag{2}$$

In Equation (2), $q_j^T \cdot \mu_k$ denotes the inner product between the query word vector q_j^T and the k^{th} centroid vector μ_k. Its weight is assigned with the values of $P(q_j^T|q_i^S)$ which denote the probability of translating a source word q_i^S into the target-language word q_j^T. It is worth noting that a standard CLIR-based system is only capable of using the term overlap between the documents and the translated queries, and cannot employ the semantic distances between the terms to score the documents. In contrast, the set-based similarity, shown in Equation 2, is capable of using the semantic distances and therefore can be used to try to improve the performance of the alignment system.

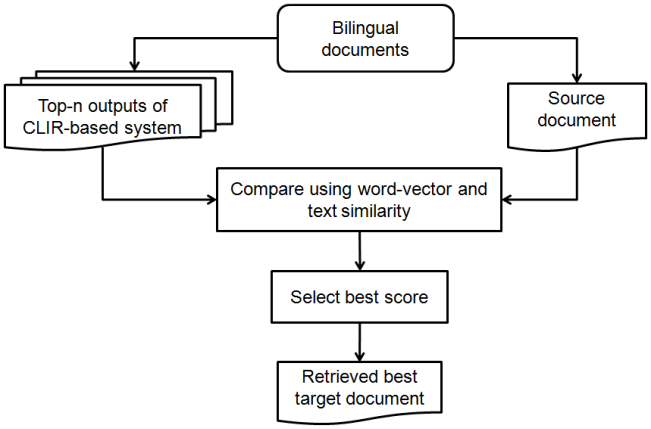


Figure 2. Architecture of the word vector embedding-based system

3.2.1. Combination with Text Similarity

Although the value of $P(d^T|q^S)$ is usually computed with the BoW representation model using language modeling (LM) (Ponte, 1998; Hiemstra, 2000) for CLIR (Berger and Lafferty, 1999), in our case we compute it with a different approach as shown in Equation (2). From a document d^T, the prior probability of generating a query q^S is given by a multinomial sampling probability of obtaining a term q_j^T from d^T. Then the term q_j^T is transformed with the term q_i^S in the source language. The priority belief (a parameter for LM) of this event is denoted by λ. As a complementary event to this, the term q_j^T is also sampled from the collection and then transformed into q_i^S, with the prior belief $(1-\lambda)$. Let us consider that $P_{LM}(d^T|q^S)$ denotes this probability which is shown in Equation (3).

$$P_{LM}(d^T|q^S) = \prod_j \sum_i \lambda P(q_i^S|q_j^T)P(q_j^T|d^T) + (1-\lambda)P(q_i^S|q_j^T)P_{coll}(q_j^T) \quad (3)$$

In the next step, we introduce an indicator binary random variable to combine the individual contributions of the text-based and word vector-based similarity. Let us consider that this indicator is denoted by α. We can then construct a mixture model of the two query likelihoods as shown in Equation (2) and Equation (3) for the word vector-based and the text-based methods, respectively. This combination is shown in Equation (4):

$$P(d^T|q^S) = \alpha P_{LM}(d^T|q^S) + (1-\alpha)P_{WVEC}(d^T|q^S) \quad (4)$$

3.2.2. Construction of Index

The K-means clustering algorithm is run for the whole vocabulary of the words which can cluster the words into distinct semantic classes. These semantic classes are different from each other and each of them discusses a global topic (i.e., the cluster id of a term) of the whole collection. As a result of this, semantically related words are embedded in close proximity to each other.

While indexing each document, the cluster id of each constituent term is retrieved using a table look-up, so as to obtain the per-document topics from the global topic classes. The words of a document are stored in different groups based on their cluster-id values. Then the the cluster centroid of each cluster id is computed by calculating the average of the word vectors in that group. Consequently, we obtain a new representation of a document d as shown in Equation (5).

$$\mu_k = \frac{1}{|C_k|} \sum_{x \in C_k} x, C_k = \{x_i : c(w_i) = k\}, i = 1, \ldots, |d| \quad (5)$$

In the final step, the information about the cluster centroids is stored in the index. This helps to compute the average similarity between the query points and the centroid vectors during the retrieval process. The overall architecture of the word vector embedding-based approach is shown in Figure 2. It can be observed that this approach is combined with the text similarity method and makes use of the top-n outputs from the CLIR-based system to compare with the source document for which we intend to discover the alignment. In contrast, a system which is solely based on CLIR methodology simply re-ranks the top-n retrieved documents and selects the best one (as seen in Figure 1). Therefore, this extended version of our system facilitates the comparison of the document pair in terms of both the text and word-vector similarity as a continuation of our previous work (Lohar et al., 2016).

4. Experiments

4.1. Data

In all our experiments, we consider French as the source-language and English as the target language. Our experiments are conducted on two different sets of data, namely (i) *Euronews*[4] data extracted from the *Euronews* website[5] and (ii) the WMT '16[6] test dataset. The statistics of the English and French documents in the Euronews and the WMT-16 test datasets are shown in Table 1. The baseline system we use is based on

[4] https://github.com/gdebasis/cldocalign/tree/master/euronews-data

[5] http://www.euronews.com

[6] http://www.statmt.org/wmt16/

dataset	English	French
Euronews	40,419	39,662
WMT-16 test dataset	681,611	522,631

Table 1. Statistics of the dataset.

the Jaccard similarity coefficient[7] (JSC) to calculate the alignment scores between the document pair in comparison. This method focuses on the term overlap between the text pair and solves two purposes: (i) NE matches are extracted, and (ii) the common words are also taken into consideration.

In our initial experiments it was found that the Jaccard similarity alone produced better results than when combined with the cosine-similarity method or when only the cosine-similarity method was used. Therefore we decided to use only the former as the baseline system. We begin by using this method without employing any MT system and denote this baseline as 'JaccardSim'. Furthermore, we combine JaccardSim with the MT-output of the source-language documents to form our second baseline which is called 'JaccardSim-MT'.

4.2. Resource

The dictionary we use for the CLIR-based method is constructed using the EM algorithm in the IBM-4 word alignment (Brown et al., 1993) approach using the Giza++ toolkit (Och and Ney, 2003), which is trained on the English-French parallel dataset of Europarl corpus (Koehn, 2005). To translate the source language documents, we use Moses which we train on the English-French parallel data of Europarl corpus. We tuned our system on Euronews data and apply the optimal parameters on WMT test data.

5. Results

In the tuning phase, we compute the optimal values for the (empirically determined) parameters as follows; (i) $\lambda = 0.9$, (ii) $M = 7$, that is when we use 7 translation terms, and (iii) 60% of the terms from the document in order to construct the pseudo-query. The results on the Euronews data with the tuned parameters are shown in Table 2, where we can observe that the baseline approach (JaccardSim) has a quadratic time complexity (since all combinations of comparison are considered) and takes more than 8 hours to complete. In addition to this, the runtime exceeds 36 hours when combined with the MT system. In contrast, the CLIR-based approach takes only 5 minutes

[7] https://en.wikipedia.org/wiki/Jaccard_index

Method	Parameters		Evaluation Metrics			Run-time
	τ	M	Precision	Recall	F-score	(hh:mm)
JaccardSim	N/A	N/A	0.0433	0.0466	0.0448	08:30
JaccardSim-MT	N/A	N/A	0.4677	0.5034	0.4848	36:20
CLIR ($\lambda = 0.9$)	0.6	7	**0.5379**	**0.5789**	**0.5576**	**00:05**

Table 2. Results on the development set (EuroNews dataset).

Method	Parameters					Recall	Run-time
	λ	τ	M	K	α		(hhh:mm)
JaccardSim	N/A	N/A	N/A	N/A	N/A	0.4950	130:00
CLIR	0.9	0.6	7	N/A	N/A	0.6586	007:35
CLIR-WVEC	0.9	0.6	7	20	0.9	0.6574	023:42
CLIR-WVEC	0.9	0.6	7	50	0.9	**0.6619**	024:18
CLIR-WVEC	0.9	0.6	7	100	0.9	0.6593	025:27

Table 3. Results on the WMT test dataset.

to produce the results. Moreover, the "JaccardSim" method has a very low effectiveness and can only lead to a considerable improvement when combined with MT. The CLIR-based approach produces the best results both in terms of precision and recall.

Table 3 shows the results on the WMT test dataset in which the official evaluation metric was only the recall measure to estimate the effectiveness of the document-alignment methods. However, we do not use "JacardSim-MT" system for the WMT dataset since it is impractical to translate a large collection of documents as it requires an unrealistically large amount of time.

We can draw the following observations from Table 3: (i) due to having a quadratic time complexity, the JaccardSim method has a high runtime of 130 hours. In contrast, the CLIR-based system is much faster and consumes only 7 hours. Additionally, it produces much higher recall than the JaccardSim method; (ii) the word-vector similarity method helps to further increase the recall produced by the CLIR-based approach, and (iii) a cluster value of 50 results in the highest value of recall among all values tested.

6. Conclusion and Future Work

In this paper we presented a new open-source multilingual document alignment tool based on a novel CLIR-based method. We proposed to use the measurement of the distances between the embedded word vectors in addition to using the term

overlap between the source and the target-language documents. For both the Euronews and WMT data, this approach produces a noticeable improvement over the Jaccard similarity-based baseline system. Moreover, an advantage of using the inverted index-based approach in CLIR is that it has a linear time complexity and can be efficiently applied to very large collections of documents. Most importantly, the performance is further enhanced by the application of the word vector embedding-based similarity measurements. We would like to apply our approach to other language pairs in future.

Acknowledgements

This research is supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (`www.adaptcentre.ie`) at Dublin City University.

Bibliography

Afli, Haithem, Loïc Barrault, and Holger Schwenk. Building and using multimodal comparable corpora for machine translation. *Natural Language Engineering*, 22(4):603 – 625, 2016.

Berger, Adam and John Lafferty. Information Retrieval As Statistical Translation. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 222–229, New York, NY, USA, 1999. ACM. ISBN 1-58113-096-1. doi: 10.1145/312624.312681.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19: 263–311, June 1993. ISSN 0891-2017.

Hiemstra, Djoerd. *Using Language Models for Information Retrieval*. PhD thesis, Center of Telematics and Information Technology, AE Enschede, The Netherlands, 2000.

Koehn, Philipp. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, volume 5, pages 79–86, Phuket, Thailand, 2005.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Prague, Czech Republic, 2007.

Lohar, Pintu, Haithem Afli, Chao-Hong Liu, and Andy Way. The adapt bilingual document alignment system at wmt16. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, 2016.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of NIPS '13*, pages 3111–3119, Lake Tahoe, USA, 2013.

Munteanu, Dragos Stefan and Daniel Marcu. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504, 2005. ISSN 08912017.

Och, Franz Josef and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29:19–51, March 2003. ISSN 0891-2017.

Ponte, Jay Michael. *A language modeling approach to information retrieval*. PhD thesis, University of Massachusetts, MA, United States, 1998.

Resnik, Philip and Noah A. Smith. The Web as a parallel corpus. *Comput. Linguist.*, 29:349–380, September 2003. ISSN 0891-2017.

Roy, Dwaipayan, Debasis Ganguly, Mandar Mitra, and Gareth J. F. Jones. Representing Documents and Queries as Sets of Word Embedded Vectors for Information Retrieval. *CoRR*, abs/1606.07869, 2016.

Utiyama, Masao and Hitoshi Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 72–79, Sapporo, Japan, 2003.

Yang, Christopher and Kar Wing Li. Automatic construction of English/Chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, 54:730–742, June 2003. ISSN 1532-2882. doi: 10.1002/asi.10261.

Zhao, Bing and Stephan Vogel. Adaptive Parallel Sentences Mining from Web Bilingual News Collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 745–748, Washington, DC, USA, 2002. IEEE Computer Society. ISBN 0-7695-1754-4.

Address for correspondence:
Haithem Afli
haithem.afli@adaptcentre.ie
School of Computing, Dublin City University,
Dublin 9, Ireland

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 181–192

Language Adaptation for Extending Post-Editing Estimates for Closely Related Languages

Miguel Rios, Serge Sharoff

Centre for Translation Studies, University of Leeds

Abstract

This paper presents an open-source toolkit for predicting human post-editing efforts for closely related languages. At the moment, training resources for the Quality Estimation task are available for very few language directions and domains. Available resources can be expanded on the assumption that MT errors and the amount of post-editing required to correct them are comparable across related languages, even if the feature frequencies differ. In this paper we report a toolkit for achieving language adaptation, which is based on learning new feature representation using transfer learning methods. In particular, we report performance of a method based on Self-Taught Learning which adapts the English-Spanish pair to produce Quality Estimation models for translation from English into Portuguese, Italian and other Romance languages using the publicly available Autodesk dataset.

1. Introduction

A common problem with automatic metrics for Machine Translation (MT) evaluation, such as BLEU (Papineni et al., 2002), is the need to have reference human translations (Specia et al., 2010). Also such metrics work best on a corpus of segments, while they are not informative for evaluation of individual segments. The aim of Quality Estimation (QE) is to predict a quality score for a segment output by MT without its reference translation, for example, to predict Translation Edit Rate (TER), i.e., the distance between the raw MT output and its revised human output (Snover et al., 2006).

From the implementation viewpoint, the QE task can be framed as a regression problem aimed at predicting the amount of human TER, without the reference translations available. This helps in deciding whether an MT sentence can be a suitable

 Corresponding author: s.sharoff@leeds.ac.uk
Cite as: Miguel Rios, Serge Sharoff. Language Adaptation for Extending Post-Editing Estimates for Closely Related Languages. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 181–192.
doi: 10.1515/pralin-2016-0017.

basis for human Post-Editing (PE) or it would be better to translate this sentence from scratch. The QE methods mostly rely on supervised Machine Learning (ML) algorithms aimed at computing similarity scores between a source sentence and its machine translations using a variety of sources of information, which are used as features to train a supervised ML algorithm to predict QE scores. Specia et al. (2013) developed QuEst, a baseline QE framework, which uses simple features quantifying the complexity of the source segment and its match to the machine translation output.

However, currently existing training datasets are only available for a limited number of languages. For example, in the WTM'15 QE task the available pairs were en-es and en-de,[1] which have been evaluated on the same domain (news). The end users of MT need a wider variety of language pairs and domains for evaluation. So far there has been little research to deal with this problem. Turchi and Negri (2014) proposed an automatic approach to produce training data for QE in order to tackle the problem of scarce training resources. Specia et al. (2010) used baseline QE framework across different domains and languages (i.e. en-es to en-dk). In our earlier work (Rios and Sharoff, 2015) we proposed using Transfer Learning (TL) for a training dataset from the WMT'14 QE task to predict PE labels, i.e., 'Perfect' vs 'Near miss' vs 'Low quality'.

In this paper, we describe the implementation of a transfer-based QE workflow to produce a large number of QE models for predicting the TER score by utilising the notion of relatedness between languages. More specifically, we use TL to learn better feature representations across related languages. Our intuition is that sentences with similar quality scores are near-neighbours in terms of QE features across related languages. In other words, good or bad quality sentences translated into Spanish (i.e., available training data) show similar characteristics to sentences translated into Portuguese (i.e., unlabelled data). This makes it possible to train a prediction algorithm by sharing information from the available labelled dataset with unlabelled datasets for related languages. However, to achieve reasonable prediction rate we need to adapt the feature representation for the dataset for the unlabelled language pair.

In this paper, we will present the Self-Taught Learning (STL) approach (Section 2), discuss the experimental setup (Section 3) and the implementation details of our toolkit (Section 3.3). We will also describe the dataset and analyse the results (Section 4).

2. Transfer Learning Methods

Transfer Learning aims to transfer information learned in one or more source tasks, i.e., using labelled datasets, to improve learning in a related target task without new annotations, i.e., using unlabelled datasets (Pan and Yang, 2010).

From the viewpoint of notation, the transfer models start with l labelled training examples $\{(x_1, y_1), (x_2, y_2)..., (x_l, y_l)\}$ and u unlabelled training examples $\{z_1, z_2, ..., z_u\}$.

[1] Throughout the paper we will be using the two-letter ISO codes to indicate the languages

The labels belong to a set of labels $\mathcal{Y}$ for the classification task or they are real-valued numbers for the regression task.

2.1. Self-Taught Learning

Raina et al. (2007) propose a semi-supervised transfer learning method, which does not assume that the unlabelled dataset is drawn from the same distribution as the labelled one. The unlabelled data is used to learn a lower dimensional representation of the input features. With this representation the labelled data can be used to learn a prediction model in the lower dimensional space, which better fits the unlabelled dataset.

Formally, the steps to perform STL are defined as:

1 Learn the dimensionality reduction for the unlabelled set z_i.
2 Compute a new representation for the labelled training dataset x_i.
3 Use standard classification/prediction methods with the new training dataset $f(\hat{x}_i) = y_i$.

The dimensionality reduction in our case is based on autoencoders. The autoencoder uses backpropagation to learn mapping the inputs to their own values via a hidden layer. The method learns an approximation function $h_{W,b}(z) \approx z$ similar the identity function, where W are the weights and b the bias. In Step 2, the labelled training data x_i is transformed by using the same parameters from the autoencoded unlabelled data $\hat{z}_i$. The new representation of the training data $\hat{x}_i$ is used to learn a prediction model in Step 3. The size of the lower-dimensional space is given by the number of units in the hidden layer. The STL model can be expanded to take into account several unknown signals, such as language pairs or domains.

Stacked autoencoders can perform a series of transformations of a labelled dataset given different autoencoders learned on several unlabelled datasets. In other words, each autoencoder is a layer L_n where the output of one autoencoder L_1 becomes the input of the following autoencoder L_2. For example, a two layer model has parameters $(W, b) = (W^1, b^1, W^2, b^2)$ for two stacked autoencoders.

3. Methodology

In this section, we describe the QE features and the transfer learning setup.

3.1. Features

The QE features come from information about the source sentence, its MT output and information about relations between them. The QuEst framework (Specia et al., 2013) implements 17 language-independent features classified into three types:

Complexity Indicators Features related to the difficulty in translating the source sentence, such as, the number of tokens of the source sentence, its language model and average number of translations in the phrase tables.

Fluency Indicators Features related to how fluent the MT output is, such as the language model of the target sentence.

Adequacy Indicators Features related to how much meaning is preserved in the MT output, such as, ratios of tokens between the source and target, ratio of punctuation and syntactic similarity. The QuEst framework also uses features related to a specific decoding process when available, such as, global score of the system and number of hypotheses in the n-best list.

In addition, we use a second set of features based on bilingual embeddings (Hermann and Blunsom, 2013), i.e., words and sentences from the source and target languages are positioned in a shared multidimensional representation, which assumes that words and sentences from one language are neighbours with words and sentences with similar meanings from another language. The motivation for introducing embeddings is to expand the range of Adequacy indicators using simple resources. The bilingual embeddings are induced from parallel data from the target domain. We build each sentence as an additive composition of individual word vectors. The final vector is a concatenation of vectors from the source sentence and its machine translation. The final embedding vector for the experiments consists of 200 features.

3.2. Implementation Details

Texts in related languages are treated as unlabelled data. For example, the available en-es labelled dataset is used to transfer information into the unlabelled en-pt sentences to predict their QE scores. We compare the transfer-based QE workflow that uses Self-Taught Learning (STL) against the Baseline with no transfer. There developed workflows can tackle both QE scenarios: the prediction of HTER and the classification of post-editing effort.

For all the HTER prediction workflows we use the Support Vector Regression (SVR) algorithm with the RBF kernel from scikit-learn (Pedregosa et al., 2011). The hyperparameters C and ϵ have been determined analytically following (Cherkassky and Ma, 2004): $\epsilon = 3\sigma(y)\sqrt{\ln(n)/n}$ and $C = mean(y)+3\sigma(y)$ where y is HTER in the training set, σ is the standard deviation, n is the number of observations.

For STL we modified the autoencoder implementation from Theano (Bergstra et al., 2010). The STL model first finds the weights W, b from the unlabelled z_i dataset by training a sparse autoencoder. Second, the model produces a modified training dataset by using the unlabelled weights on an second autoencoder. The modified training dataset is a lower-dimensional representation of the input features. A new test dataset can be predicted by using the weights W, b to represent the data points into the same lower-dimensional space. However, we do not have access to any development datasets for tuning the z_i autoencoder for our unlabelled language pairs. For the parameter selection of the unlabelled autoencoder, as suggested in Bergstra and Bengio (2012), we run a random search over a split of the modified training dataset (90% training, 10% validation) in order to find: the size of the hidden dimension, the

desired average activation sparsity parameter (ρ), the weight decay parameter (λ) and the sparsity penalty (β).

The stacked STL setup can be used for language pairs where the source is different form the available training dataset. For example, the training dataset is en-es and the objective test is fr-es. The first autoencoder is trained with en-fr and the second autoencoder with fr-es, which projects training en-es first into the space of en-fr, and then into fr-es.

In addition to STL, we also experimented with other TL strategies, namely multi-view learning and Transductive SVM. The multi-view learning framework tries to jointly optimise different views of the same input (Xu et al., 2013). Spectral methods such as Canonical Correlation Analysis (CCA) can be used to learn a subspace shared by label and unlabelled data. The Spectral method is straightforward to apply to two-view data. In our case, the first view is the labelled data x_i and the second view is the unlabelled data z_i. CCA learns two projections $A_m \epsilon \mathbb{R}^{l \times m}$ where l are the labelled instances and m the number of features, and $B_m \epsilon \mathbb{R}^{u \times m}$. We use A_m to project each instance of the test dataset into $\hat{x}_i$. For the Spectral Learning setup, we use the CCA implementation from MATLAB[2] and the same SVR setup as for STL. For example, the available dataset is en-es and the test objective en-pt. We use CCA to learn the projections of en-es and en-pt. The en-es test is projected into the same lower space with A_i, and then, we use the projected datasets for training and testing respectively.

The methods described above can be used in different QE scenarios by changing from SVR to SVM. In particular for the classification of post-editing effort, Transductive Support Vector Machine (TSVM) takes into consideration a particular test dataset and tries to minimise errors only on those particular instances (Vapnik, 1995). The TSVM model learns a large margin hyperplane classifier using labelled training data, but at the same time it forces that hyperplane to be far from the unlabelled data, and the method transfers the information from labelled instances to the unlabelled. We use SVMlin[3] for training the TSVM. TSVM uses an Linear kernel with no hyper-parameter optimisation. We select the heuristic Multi-switch TSVM. Each instance in the unlabelled dataset is added to the training dataset. For classification, we implement the one-against-one strategy, and the final decision is given by voting.

The standard QE baseline measures HTER prediction without any adaptation, i.e., the en-es QE prediction model is applied to en-pt data.

For the regression scenario, we report the Mean Absolute Error (MAE), Root Mean Squared Error (RSME) and Pearson correlation. Our main evaluation metric is the Pearson correlation as suggested in Graham (2015).

[2] http://uk.mathworks.com/help/stats/canoncorr.html

[3] http://vikas.sindhwani.org/

3.3. QE Transfer Command Usage

In this section, we show the different implemented methods for transfer-based QE. The repository contains the scripts for extracting the features and implementations of transfer-based QE methods, where each transfer workflow uses the same input parameters. The first step of the transfer-based workflow is to extract features for: the labelled dataset, the unlabelled and test datasets. For the feature representation, we have available two feature extractor scripts. The QuEst feature extractor that depends on QuEst[4] and Moses. The bilingual embeddings feature extractor that depends on BICVM[5].

The next step is to train and predict the test dataset. We developed different QE adaptation tools based on transfer-learning such as: STL, stacked STL, CCA all for regression and classification, and TSVMy for classification. The final and optional step is to measure the predicted HTER against a gold-standard annotation of the test dataset.

In addition, we implemented the analytical method to estimate the parameters ϵ and C of the SVR, where the input is the training examples features.

Preliminary results show that STL outperforms other transfer learning methods over both regression and classification. We show the use of the STL transfer method given the QuEst baseline features. The input parameters of the adapted QE based on STL with SVR for HTER prediction are: (1) training examples features, (2) training labels, (3) unlabelled training examples features, (4) test features, (5) output, (6) epsilon parameter for SVR, (7) C parameter for SVR and (8) size of hidden layer for the autoencoder. Parameters (6)-(8) will be determined as discussed above if not provided explicitly. An example of the command is as follows:

```
python stlSVR.py \
--training-examples autodesk.training.en-es.feat \
--training-labels autodesk.training.en-es.hter \
--unlabelled-examples autodesk.training.en-pt.feat \
--test autodesk.test.en-pt.feat \
--output autodesk.en-pt.pred \
--epsilon 41.06 \
--c 0.232 \
--hidden-layer 50
```

The default parameters for the autoencoder have been selected via random search over a split on the labelled language dataset. It is worth noticing that we do not constraint the number of hidden units during the learning of the autoencoder. Thus, we

[4] http://www.quest.dcs.shef.ac.uk/

[5] https://github.com/karlmoritz/bicvm

set a bound for random search from 0 to 100 units, and for our example the optimum number of units has been detected as *50*.

4. Experiments

In this section, we describe the datasets used to train and evaluate our transfer learning model for pairs of related languages. We show the results of the STL transfer-based QE and we also discuss the predicted scores for different language pairs.

4.1. Dataset

In this paper, we experimented with the Autodesk PE dataset (Zhechev, 2012).[6] The Autodesk corpus contains the source, MT and PE segments for several languages. The corpus consist of user manuals, marketing and educational material for the Autodesk applications, such as AutoCAD, REVIT, Inventor. We use the following language pairs showed in Table 1, with a 70/30% split for the training/test data.

Language Pair	Training Labelled	Training Unlabelled	Test
en-es	24,073	-	8,025
en-pt	-	28,886	9,629
en-it	-	30,311	10,104
en-fr	-	38,469	12,824
en-ru	30,905	-	10,302
en-cs	-	20,997	7,000
en-pl	-	24,853	8,285
fr-es	-	10,000	1,000
it-es	-	10,000	1,000
pt-es	-	10,000	1,000

Table 1. Autodesk training and test number of segments used in this study.

We use as labelled training data *en-es* for the Romance family and *en-ru* for the Slavonic family. The remaining language pairs were used as unlabelled and test data for each family. Given that the Czech dataset is much smaller, it has been only used for tuning/testing. The unlabelled set has been produced by running the remaining English segments **not included** in the en-cs set through Google MT.

The QE score (HTER) is the minimum number of edit operations (TER) between the MT output and PE. We use Tercom (Snover et al., 2006) to compute the HTER scores between the post-edited and MT segments.

[6] https://autodesk.app.box.com/v/autodesk-postediting

We produce the pt-es, it-es and fr-es language pairs by intersecting the English segments. For example, the same English segments present in en-pt and en-es produces the pt-es alignment for both MT and PE. For extracting the QuEst features, we use Moses (Koehn et al., 2007) and KenLM (Heafield, 2011) with a 3-gram language model (LM).

4.2. Results

In this section, we show the results of our proposed STL QE workflow against the standard QE Baseline. We built the transfer-based QE and baseline models for the language directions in Table 2.

Training labelled	Test unlabelled
en-es	en-pt, en-it, en-fr
	pt-es, it-es, fr-es
en-ru	en-cs, en-pl

Table 2. Language directions workflows.

The upper bound for our TL methods is the standard QE setup in which the same feature set is used for training and testing on the same language pair, en-es and en-ru in our case (Table 3).

Training en-es		en-es
Upper baseline	MAE	0.14
	RSME	0.18
	Pearson	**0.53**
Training en-ru		en-ru
Upper baseline	MAE	0.18
	RSME	0.27
	Pearson	**0.47**

Table 3. Upper-bound baseline for labelled language pairs.

Table 4 shows the transfer results for the workflows. Over the Romance pair we can see consistent and considerable improvement over the baseline with no adaptation, e.g., $0.35 \rightarrow 0.52$ for correlation in the case of en-es→en-pt TL, which approaches the

Training en-es		en-pt	en-it	en-fr
STL	MAE	0.14	0.16	0.17
	RMSE	0.17	0.21	0.22
	Pearson	**0.52**	**0.40**	**0.30**
Baseline	MAE	0.16	0.18	0.18
	RMSE	0.20	0.23	0.23
	Pearson	0.35	0.26	0.24

Training en-ru		en-cs	en-pl
STL	MAE	0.19	0.19
	RMSE	0.25	0.25
	Pearson	**0.41**	**0.46**
Baseline	MAE	0.20	0.21
	RMSE	0.26	0.27
	Pearson	0.32	0.33

Table 4. Transfer learning results.

upper baseline of 0.53 for training and testing on the same language pair (en-es). For the Slavonic language pairs we also reach the upper baseline for the en-pl pair.

Training en-es		pt-es	it-es	fr-es
STL	MAE	0.18	0.18	0.18
	RSME	0.23	0.22	0.22
	Pearson	**0.19**	**0.23**	**0.21**
Stacked STL	MAE	0.20	0.58	0.24
	RMSE	0.25	0.62	0.30
	Pearson	0.07	0.06	0.02
Baseline	MAE	0.19	0.19	0.18
	RSME	0.23	0.24	0.22
	Pearson	0.14	0.17	0.10

Table 5. Transfer learning results with en-es training into test: pt-es, it-es and fr-es.

Table 5 shows the transfer results **across** the Romance language pairs. The training is en-es and we adapt to pt-es, it-es and fr-es.

Table 6 shows the transfer-based results and the Baseline for comparison between more distant languages. As expected, the performance of TL is much lower between non related languages, i.e., no useful adaptation is taking place.

Training en-es		en-cs	en-pl
STL	MAE	0.22	0.25
	RMSE	0.29	0.32
	Pearson	*0.08*	*0.11*
Baseline	MAE	0.23	0.22
	RSME	0.31	0.29
	Pearson	*0.11*	*0.09*

Table 6. Transfer learning results with en-es training into test: en-cs and en-pl.

The features of the source and target directions affect the results of the transfer methods, i.e. complexity, fluency and adequacy indicators. For example, in the case of STL adaptation from en-es to pt-es, there is no agreement between the features of the source languages (en vs pt, complexity indicators) given they are not closely related, but the target languages are closely related. However, when the source languages are the same (en-es → en-pt) and the target languages are closely related, i.e. the fluency indicators can be transformed, the overall performance improves nearly up to the level of the labelled (upper-bound) pair baseline.

In addition to RMSE and correlation scores, there is a danger that adaptation can produce a narrow range of predicted values in comparison to the test set. We analyse the results of transfer by presenting the range of HTER predicted scores at (10%, 90%) quantiles, i.e. by trimming 10% of the most extreme values, which are likely to contain the outliers.

The (10%, 90%) quantile range for en-es → en-pt is as follows: Gold (0.0, 0.53), STL (0.23, 0.46) and Baseline(0.16, 0.47). The spread of the STL predicted values is slightly less than the baseline. For the stacked STL a possible reason for the low performance is related to over-fitting. The range for en-es → pt-es is: Gold (0.0, 0.57) and Stacked STL (0.50, 0.50). A better configuration of en-es → pt-es with the the stacked STL can be: en-es (training), es-pt (first layer) and pt-es (second layer). The motivation is to find an agreement between the source and the target features with the addition of more closely languages in terms of the induced lower dimensional space.

5. Conclusions and Future Work

We present an open-source toolkit[7] for transferring QE features from a single training dataset to closely related languages via Self-Taught Learning. We also developed other transfer learning methods for the task of QE prediction. It has been found successful in prediction the PE operations on the Autodesk dataset for the Romance and Slavonic families. For the reasons of testing the method the language pairs involved in

[7] https://github.com/mriosb08/palodiem-QE

the experiment do have suitable training resources. However, such sizeable datasets are rare. Even the Autodesk set only covers three Slavonic languages, while only German is available for the Germanic languages in this set.

One possibility for further research concerns the expansion of the available labelled resources with adaptation to different *domains* in addition to the language families, for example, by transferring predictions from the original domain of the Autodesk PE set to other domains with only unlabelled data available.

Acknowledgements

This study was funded as a research grant by Innovate UK and ZOO Digital.

Bibliography

Bergstra, James and Yoshua Bengio. Random Search for Hyper-parameter Optimization. *J. Mach. Learn. Res.*, 13:281–305, Feb. 2012. ISSN 1532-4435.

Bergstra, James, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a CPU and GPU Math Expression Compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation.

Cherkassky, Vladimir and Yunqian Ma. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks*, 17(1):113–126, 2004.

Graham, Yvette. Improving Evaluation of Machine Translation Quality Estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing*, 2015.

Heafield, Kenneth. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July 2011.

Hermann, Karl Moritz and Phil Blunsom. A Simple Model for Learning Multilingual Compositional Semantics. *CoRR*, abs/1312.6173, 2013.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007.

Pan, Sinno Jialin and Qiang Yang. A Survey on Transfer Learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, Oct. 2010.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Raina, Rajat, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. Self-taught Learning: Transfer Learning from Unlabeled Data. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 759–766, New York, NY, USA, 2007. ISBN 978-1-59593-793-3.

Rios, Miguel and Serge Sharoff. Large Scale Translation Quality Estimation. In *The Proceedings of the 1st Deep Machine Translation Workshop*, Praha, Czech Republic, 2015.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas,*, 2006.

Specia, Lucia, Dhwaj Raj, and Marco Turchi. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50, 2010.

Specia, Lucia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. QuEst - A translation quality estimation framework. In *51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL, pages 79–84, Sofia, Bulgaria, 2013.

Turchi, Marco and Matteo Negri. Automatic Annotation of Machine Translation Datasets with Binary Quality Judgements. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014.

Vapnik, Vladimir N. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.

Xu, Chang, Dacheng Tao, and Chao Xu. A Survey on Multi-view Learning. *CoRR*, abs/1304.5634, 2013. URL `http://arxiv.org/abs/1304.5634`.

Zhechev, Ventsislav. Machine Translation Infrastructure and Post-editing Performance at Autodesk. In *Proc AMTA*, San Diego, CA, 2012.

Address for correspondence:
Serge Sharoff
`s.sharoff@leeds.ac.uk`
Centre for Translation Studies,
Parkinson Bldg, University of Leeds
LS2 9JT, UK

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 193–204

RuLearn: an Open-source Toolkit for the Automatic Inference of Shallow-transfer Rules for Machine Translation

Víctor M. Sánchez-Cartagena[a], Juan Antonio Pérez-Ortiz[b], Felipe Sánchez-Martínez[b]

[a] Prompsit Language Engineering, Spain
[b] Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain

Abstract

This paper presents ruLearn, an open-source toolkit for the automatic inference of rules for shallow-transfer machine translation from scarce parallel corpora and morphological dictionaries. ruLearn will make rule-based machine translation a very appealing alternative for under-resourced language pairs because it avoids the need for human experts to handcraft transfer rules and requires, in contrast to statistical machine translation, a small amount of parallel corpora (a few hundred parallel sentences proved to be sufficient). The inference algorithm implemented by ruLearn has been recently published by the same authors in *Computer Speech & Language* (volume 32). It is able to produce rules whose translation quality is similar to that obtained by using hand-crafted rules. ruLearn generates rules that are ready for their use in the Apertium platform, although they can be easily adapted to other platforms. When the rules produced by ruLearn are used together with a hybridisation strategy for integrating linguistic resources from shallow-transfer rule-based machine translation into phrase-based statistical machine translation (published by the same authors in *Journal of Artificial Intelligence Research*, volume 55), they help to mitigate data sparseness. This paper also shows how to use ruLearn and describes its implementation.

1. Introduction

Although statistical machine translation (SMT) has been the leading MT paradigm during the last decade, its application may be limited by the availability of parallel corpora. When parallel corpora sufficiently big to build a competitive SMT system are not available, rule-based machine translation (RBMT) is an appealing option. How-

 Corresponding author: vmsanchez@prompsit.com
Cite as: Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez. RuLearn: an Open-source Toolkit for the Automatic Inference of Shallow-transfer Rules for Machine Translation. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 193–204. doi: 10.1515/pralin-2016-0018.

ever, if the RBMT system has to be developed from scratch, the cost in terms of time spent by trained linguists can be prohibitively high.

In this paper, we present ruLearn, an open-source toolkit with which to automatically infer shallow-transfer RBMT rules from very small parallel corpora and existing RBMT dictionaries. The underlying methodology has been described in depth elsewhere (Sánchez-Cartagena et al., 2015): multiple rules with different generalisation levels are generated from bilingual phrases extracted from the parallel corpus and the minimum set of rules that correctly reproduces the bilingual phrases is selected. In this way, conflicts between rules are effectively solved at a global level. The rules produced by ruLearn are encoded in the format used by the Apertium shallow-transfer RBMT platform (Forcada et al., 2011) but they can be adapted to other platforms. They can be easily modified by human experts and can co-exist with hand-crafted rules.

Transfer rules are the linguistic resource in Apertium that requires the deepest linguistic knowledge in order to be created. Apertium translates by analysing the source-language (SL) text into an SL intermediate representation (IR), transferring it into a TL IR, and generating the final translation from the TL IR. The transfer step makes use of transfer rules and bilingual dictionaries while the analysis and generation steps require monolingual morphological dictionaries. Transfer rules encode the operations to be carried out in order to deal with the grammatical divergences between the languages. Thus, ruLearn reduces the difficulty of creating Apertium-based RBMT systems for new language pairs. ruLearn has been successfully used in the development of Apertium-based RBMT systems for Chinese→Spanish (Costa-Jussà and Centelles, 2015) and Serbian↔Croatian (Klubička et al., 2016)

The rules obtained with ruLearn can also be integrated into a phrase-based SMT system by means of the hybridisation strategy we developed (Sánchez-Cartagena et al., 2016) and released as an open-source toolkit (Sánchez-Cartagena et al., 2012). When shallow-transfer rules extracted from the same training corpus are integrated into a phrase-based SMT system, the translation knowledge contained in the parallel corpus is generalised to sequences of words that have not been observed in the corpus, thus helping to mitigate data sparseness.

The rest of the paper is organised as follows: next section presents the most prominent related rule inference approaches in literature. Section 3 describes the rule inference algorithm implemented by ruLearn. A summary of the most relevant results is presented in Section 4. Implementation details and usage instructions are provided in Section 5. The paper ends with some concluding remarks.

2. Related work

There have been other attempts to automatically infer transfer rules for RBMT. ruLearn is greatly inspired by the work of Sánchez-Martínez and Forcada (2009). It overcomes the most relevant limitations of their work: the low expressiveness of their formalism, which is not able to encode rules that are applied regardless of the morphological inflection attributes of the words they match and hence limits the generali-

sation power of their approach;[1] and the fact that their algorithm generates rules that usually prevent the application of other, more convenient rules, when they are used in the Apertium RBMT platform. ruLearn explicitly takes into account the interaction between rules when the RBMT engine chooses which rules to apply and avoids the generation of rules that harm translation quality.

Probst (2005) developed a method with which to learn transfer rules from a small set of bilingual segments obtained by asking bilingual annotators to translate a controlled, parsed corpus. The main differences between her approach and ruLearn are the following: first, her method learns hierarchical syntactic rules that are integrated in a statistical decoder (thus the system can mitigate the impact of errors introduced by the rules) whereas ruLearn produces flat, shallow-transfer rules that are used by a pure RBMT system; and, second, her approach solves conflicts between rules in a greedy fashion rather than choosing the most appropriate ones according to a global minimisation function. Varga and Yokoyama (2009) also developed a rule inference method addressed to small parallel corpora. The differences with ruLearn are similar to those that have just been described: the rules inferred by their approach are also hierarchical syntactic rules that must be used in a system with a statistical decoder.

Finally, Caseli et al. (2006) present a method in which shallow-transfer rules and bilingual dictionaries are learnt from a parallel corpus. It mainly differs from ruLearn in the way in which bilingual phrases are generalised to obtain rules. Unlike ruLearn, their approach does not generalise the rules to unseen values of morphological inflection attributes and deals with conflicts between rules in a greedy manner.

Among the rule inference approaches listed in this section, only those by Sánchez-Martínez and Forcada (2009) and Caseli et al. (2006) have been released as open-source toolkits.[2] We expect ruLearn to be a useful alternative to these tools thanks to its strong generalisation power and its ability to effectively solve rule conflicts.

3. Automatic inference of shallow-transfer rules

3.1. Generalised alignment template formalism

Instead of directly inferring shallow-transfer rules, ruLearn infers simpler units called generalised alignment templates (GATs) which are converted into Apertium shallow-transfer rules at the end of the whole process. GATs are easier to obtain from parallel corpora than Apertium shallow-transfer rules. The SL and TL IRs in Apertium consist of sequences of *lexical forms*. A lexical form, e.g. *car* `N-gen:ϵ.num:sg`, consists of a lemma (*car*), a lexical category (`N` = noun) and a set of morphological inflection attributes and their values (`gen:ϵ.num:sg` = empty gender and singular num-

[1] For instance, four different rules are needed by the approach of Sánchez-Martínez and Forcada (2009) in order to swap a noun followed by an adjective when translating from Spanish to English: one for each possible combination of gender and number.

[2] Available at `https://sourceforge.net/projects/apertium/files/apertium-transfer-tools/` and `https://sourceforge.net/projects/retratos/` respectively.

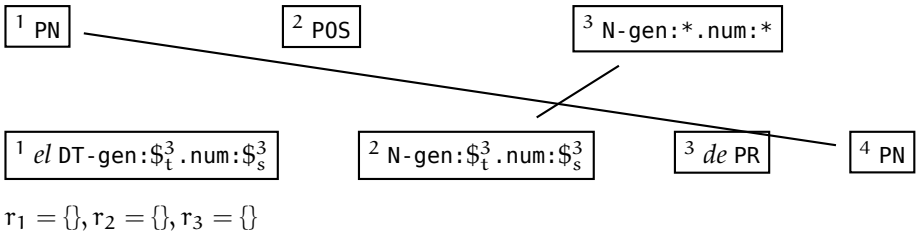


Figure 1. GAT for the translation of the English Saxon genitive construction into Spanish. It will produce the Spanish translation "el gato de Juan" from English "John's cat".

ber). A GAT processes a sequence of SL lexical forms together with their translation according to the bilingual dictionary of the RBMT system and performs the required operations to ensure that the output is grammatically correct in the TL.

Figure 1 shows a GAT that encodes the translation of the English Saxon genitive construction into Spanish. It matches a sequence of 3 English lexical forms defined by the SL word classes depicted at the top of the figure: a proper noun (PN) with any lemma followed by the possessive ending (POS) and a (common) noun with any lemma, any gender and any number. The wildcard value ($*$) for an SL morphological inflection attribute means that any value is allowed. Our formalism also permits defining the lemmas that a sequence of lexical forms must have in order to match a GAT. The GAT in Figure 1 generates a sequence of 4 TL lexical forms defined by the TL word classes: a determiner (DT) whose lemma is *el*, a noun whose lemma is obtained after looking up the SL noun that matched the GAT in the bilingual dictionary (there is an alignment link between them), a preposition (PR) whose lemma is *de* and a proper noun whose lemma is obtained after looking up the SL proper noun in the bilingual dictionary. The genders of the TL determiner and noun are copied from the TL lexical form obtained after looking up the SL noun in the bilingual dictionary ($\$_t^3$; the SL noun is the third matching SL lexical form), while the number is copied from the same SL lexical form without dictionary look-up ($\$_s^3$). Attributes $\$_t^3$ and $\$_s^3$ are *reference attributes* because their values depend on the SL lexical forms that match the GAT. Finally, restrictions (r_i) define the values of morphological inflection attributes the matching SL lexical forms must have after being looked up in the bilingual dictionary in order to match the GAT. In the running example, no restrictions are imposed. See the publication by Sánchez-Cartagena et al. (2015, Sec. 3) for more details.

3.2. Rule inference algorithm

In the first step of the rule inference algorithm implemented by ruLearn (all the steps are summarised in Figure 2), bilingual phrases are obtained from the parallel corpus following a strategy similar to that usually followed for obtaining bilingual phrases during SMT training (Koehn, 2010). From each bilingual phrase, many dif-

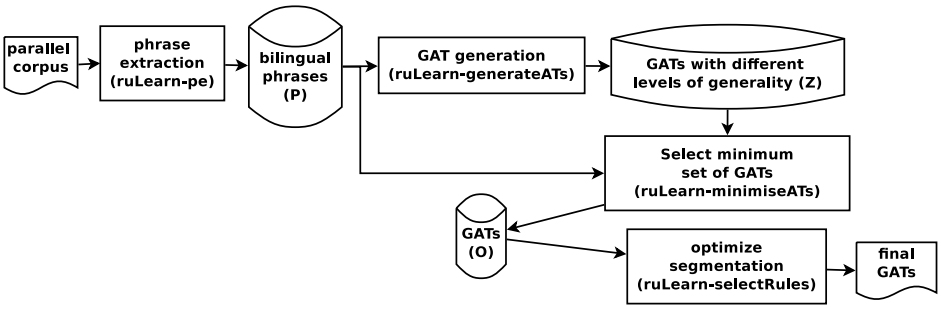


Figure 2. Steps followed to obtain a set of shallow-transfer rules from a parallel corpus.

ferent GATs that correctly *reproduce* it, that is, when applied to the SL phrase, the corresponding TL phrase is obtained, are generated. GATs with different levels of generalisation are obtained by using different sets of wildcard and reference attributes, and also with different lexicalised SL word classes (SL word classes that only match specific lemmas). Only a subset of these GATs will be part of the output of ruLearn.

In order to produce high-quality rules, a set of GATs that correctly translates at least all the bilingual phrases extracted from the parallel corpus should be chosen. These GATs should be as general as possible in order to extend the linguistic knowledge from the corpus to unseen SL segments. This is achieved by selecting the minimum set of GATs needed to correctly reproduce all the bilingual phrases. In addition, just before selecting them, those GATs that correctly reproduce a low proportion of the bilingual phrases they match (the proportion is controlled by a threshold δ) are removed.

The minimisation problem is formalised as follows. Let P be the set of bilingual phrases, Z the set of GATs, $\mathcal{G}(z)$ the set of bilingual phrases correctly reproduced by the GAT $z \in Z$ and $\mathcal{B}(z)$ the set of bilingual phrases matched but not correctly reproduced by z (i.e. when z is applied to the SL side of the bilingual phrase, the TL side is not obtained). The relation of specificity between GATs is defined by the function more_specific(z_i, z_j), whose value is `true` if z_i is more specific than z_j, that is, if z_i contains more lexicalised words or less wildcard and reference attributes than z_j. This function is only defined for GATs with the same sequence of SL lexical categories, as explained later in this section. The minimum set of GATs $O \subseteq Z$ is chosen subject to the following constraints:

1. Each bilingual phrase pair has to be correctly reproduced by at least one GAT that is part of the solution:

$$\bigcup_{z_i \in O} \mathcal{G}(z_i) = P$$

2. If a GAT z_i that is part of the solution incorrectly reproduces a bilingual phrase pair $p \in P$, there is another GAT z_j that is part of the solution, is more specific

than z_i and correctly reproduces p:

$$\forall z_i \in O, \forall p \in \mathcal{B}(z_i), \exists z_j \in O : \text{more\_specific}(z_j, z_i) \ \wedge \ p \in \mathcal{G}(z_j)$$

The solution generally looks like a hierarchy with a mix of general rules and more specific rules fixing the cases not correctly translated with the general ones. The problem can be solved in a reasonable amount of time when the quantity of bilingual phrases and GATs is relatively small (a common situation when the amount of training parallel corpora is scarce) by splitting it into one independent subproblem for each different sequence of SL lexical categories. Each subproblem is formulated as an integer linear programming problem (Garfinkel and Nemhauser, 1972) and solved using the state-of-the-art *branch and cut* algorithm (Xu et al., 2009).

After solving the minimisation subproblems, GATs with certain sequences of SL lexical categories are discarded. This is necessary because, in Apertium, the segmentation of the input SL sentences into chunks (sequences of SL lexical forms that are processed together by a rule) is done by the rules to be applied, which are chosen by the engine in a greedy, left-to-right, longest match fashion. It is necessary to avoid that lexical forms that should be processed together (because they are involved in the same linguistic phenomenon) are assigned to different chunks. The minimum set of SL text segments (*key segments*) in the SL side of the training corpus which need to be translated by a rule to obtain the highest similarity with the TL side is first identified. Afterwards, the set of sequences of SL categories that ensure that the maximum number of key segments get translated properly are selected and those GATs with a sequece of SL lexical categories not found in that set are discarded. Finally, those GATs which produce the same translations that a sequence of shorter GATs would produce are removed and the remaining GATs are encoded as Apertium shallow-transfer rules. More details can be found in the paper by Sánchez-Cartagena et al. (2015, Sec. 4).

4. Evaluation of the tool

The rule inference algorithm implemented by ruLearn was exhaustively evaluated in the paper by Sánchez-Cartagena et al. (2015). Experiments comprised 5 different language pairs. For each of them, shallow-transfer rules were inferred from parallel corpora of different sizes (from 100 to 25 000 parallel sentences) and the resulting rules were integrated in Apertium and automatically evaluated using a test parallel corpus.

The evaluation showed that ruLearn clearly outperforms the approach by Sánchez-Martínez and Forcada (2009). Furthermore, the number of inferred rules is significatly smaller. When the languages involved are closely-related, a few hundred parallel sentences proved to be sufficient to obtain a set of competitive transfer rules, since the addition of more parallel sentences did not result in great translation quality improvements. For instance, Figure 3 shows the results of the automatic evaluation of the Spanish→Catalan rules produced by ruLearn from fragments of differenct sizes

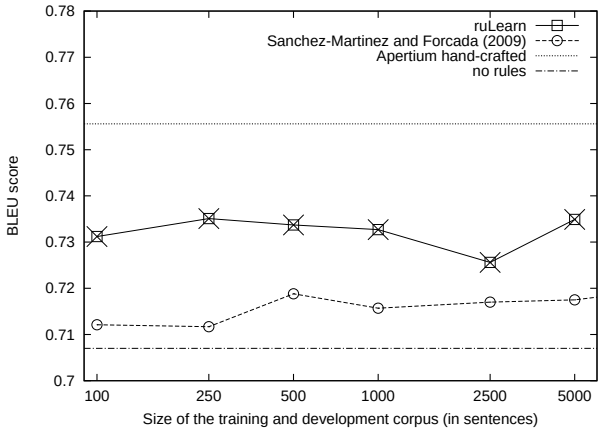


Figure 3. Translation quality (measured using BLEU) achieved by the Spanish→Catalan shallow-transfer rules produced by ruLearn, the rules produced by the approach by Sánchez-Martínez and Forcada (2009), the hand-crafted rules included in Apertium and an empty set of rules. If the difference between the rules obtained with the two rule inference approaches is statistically significant according to paired bootstrap resampling (Koehn, 2004) with $p \leq 0.05$ *and* $1\,000$ *iterations, a diagonal cross is placed on top of the points that represent the results of the approach that performs best.*

of a parallel corpus extracted from the newspaper *El Periodico de Catalunya*.[3] The test corpus was built by randomly selecting sentences from the parallel corpus *Revista Consumer Eroski* (Alcázar, 2005), which contains product reviews. The evaluation metric used was BLEU (Papineni et al., 2002). More details about the evaluation can be found in the paper by Sánchez-Cartagena et al. (2015).

The high complexity of the minimisation problem, which is caused by the generalisation of morphological inflection attributes (with wildcard and reference attributes), made very difficult the evaluation of the inference algorithm with training corpora bigger than $5\,000$ sentences. Disabling that generalisation allowed ruLearn to scale to bigger corpora and reach, and in some cases surpass, the translation quality of the Apertium hand-crafted rules. For instance, Figure 4 shows the results of the automatic evaluation of the Spanish→English rules produced by ruLearn from fragments of different sizes of the *Europarl* (Koehn, 2005) parallel corpus (minutes from the European Parliament). The test corpus was *newstest2013*[4], which contains pieces of news. Note that ruLearn outperforms the hand-crafted rules for the biggest training corpus.

[3] `http://www.elperiodico.com/`

[4] `http://statmt.org/wmt13/translation-task.html`

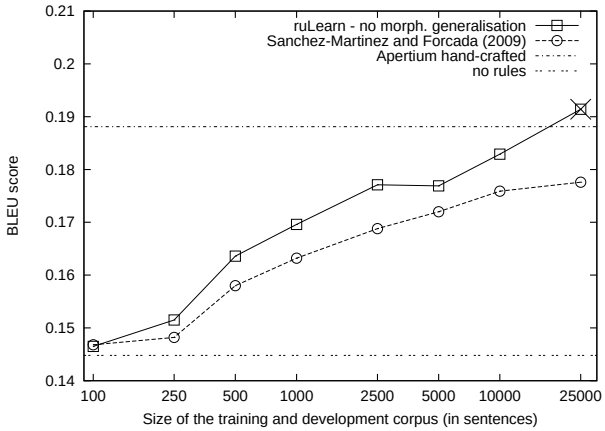


Figure 4. Translation quality (measured using BLEU) achieved by the Spanish→English shallow-transfer rules produced by ruLearn after disabling generalisation of morphological inflection attributes, the rules produced by the approach by Sánchez-Martínez and Forcada (2009), the hand-crafted rules included in Apertium and an empty set of rules. A diagonal cross over a square point indicates that ruLearn outperforms the hand-crafted rules by a statistically significant margin according to paired bootstrap resampling (Koehn, 2004) with $p \leq 0.05$ *and* $1\,000$ *iterations.*

Moreover, we proved that ruLearn can be successfully combined with a hybridisation approach (Sánchez-Cartagena et al., 2016) in order to allow an SMT system enhanced with linguistic information from RBMT to be built using dictionaries as the only hand-crafted linguistic resource. According to our evaluation, a hybrid system with automatically inferred rules is able to attain the translation quality achieved by a hybrid system with hand-crafted rules and, even when it does not, it often performs better than a pure SMT system and a hybrid system that only uses RBMT dictionaries.

5. Technical details and usage instructions

5.1. Getting ruLearn

ruLearn source code can be downloaded from the Apertium Subversion repository at `https://svn.code.sf.net/p/apertium/svn/trunk/ruLearn`. It is licensed under GNU GPL v3. and distributed as a GNU Autotools[5] package. It currently can only be compiled and executed under GNU/Linux.

[5] `http://www.gnu.org/software/autoconf/` and `http://www.gnu.org/software/automake/`

```
<np> <pos> <n><*gen><*num> |
el<det><def><)3gen><(3num> <n><)3gen><(3num> de<pr>
<np> | 0:3 1:0 2:1 |
```

Figure 5. GAT in Figure 1 encoded in one the intermediate files generated by ruLearn. Fields are separated by |. The first field represents SL word classes, the second field contains TL word classes, the third one contains word alignments and the last one, restrictions.)3 represents the reference attribute $\$_t^3$ in Figure 1, while (3 represents $\$_s^3$.

5.2. Program design

ruLearn is written in `Bash` and `Python`. There is an independent command-line program for each step of the algorithm (their names are depicted in Figure 2) and a wrapper program that executes all the steps. Given the huge amount of bilingual phrases and GATs that need to be processed, communication between the different modules is done by writing and reading intermediate files. The results of each step of the algorithm are written in a different subdirectory. This allows users to understand the different steps of the algorithm and even to customise the algorithm by adding new steps. It also makes easier the reuse of some of the steps from previous executions of the algorithm. `Bash` is used to manage and check the availability of all the intermediate files while the core algorithm is implemented in `Python`.

GATs for each sequence of SL lexical categories are stored in a different file (Figure 5 shows how the GAT in Figure 1 is encoded in an intermediate file) and bilingual phrases are organised in a similar way. This way of storing the data eases the parallelisation of the different minimisation subproblems and increases the simplicity of the core algorithm code. By default, all the available CPUs of the machine are used to solve the minimisation subproblems thanks to the use of the `parallel` tool.[6] In order to increase the parallelisation degree and hence speed up the process, the minimisation subproblems can be scattered across different machines.

5.3. Usage instructions

Compilation and installation of ruLearn can be performed by means of the commands depicted below. The `configure` program checks whether all the software dependencies are met. The most important ones are a recent version of Apertium and the `PuLP`[7] Python module, which contains the linear programming solver.

```
$ ./autogen.sh
$ ./configure && make && make install
```

[6] `https://joeyh.name/code/moreutils/`

[7] `http://pypi.python.org/pypi/PuLP`

```
[tag groups]
gender:m,f,mf,GD,nt
number:sg,pl,sp,ND
...
[tag sequences]
n:gender,number
...
```

Figure 6. Fragment of a linguistic configuration file. The `[tag groups]` section defines the values the morphological inflection attributes can take. The `[tag sequences]` section defines that all the nouns must contain a gender and a number.

In order to produce rules, ruLearn needs a training parallel corpus, a development corpus (used to determine the best value for the threshold δ described in Section 3.2), the path to the source code of the Apertium linguistic package of the language pair for which rules will be inferred (because Apertium linguistic resources are needed in order to analyse the training corpus) and a linguistic configuration file, which contains a set of Apertium-specific and language-pair-dependent configuration parameters. The most important ones are *tag groups* and *tag sequences*. The former define the allowed values for each type of morphological inflection attribute while the latter define the sequence of attributes for each lexical category (Figure 6 shows an example). They are needed in order to map lexical forms as they encoded in the Apertium dictionaries to a representation compatible with the GAT formalism, in which the type of each morphological inflection attribute is explicitly defined. For instance, a feminine singular noun is represented in Apertium as `<n><f><sg>` (`f` stands for *feminine* and `sg` stands for *singular*). The fact that `f` represents a gender and the set of possible values a gender can take is not explitly encoded anywhere in Apertium, but this information is needed by the rule inference algorithm in order to be able to introduce wildcard and reference attributes. Examples of linguistic configuration files for different language pairs are shipped with ruLearn.

The following command runs the rule inference algorithm:

```
$ ruLearn --source_language SOURCE_LANGUAGE_CODE --target_language
TARGET_LANGUAGE_CODE  --corpus TRAINING_CORPUS --dev_corpus
DEVELOPMENT_CORPUS --data_dir SOURCE_OF_APERTIUM_LANGUAGE_PAIR
--work_dir OUTPUT_DIRECTORY --config LINGUISTIC_CONFIG_FILE
```

Results are written into the directory `OUTPUT_DIRECTORY`. When the inference process finishes, ruLearn prints the best value of δ and the path to the file with the resulting set of rules. If a test corpus is defined with the `--test_corpus` option, ruLearn translates it with the automatically inferred rules and prints the BLEU and TER scores obtained.

6. Concluding remarks

We have presented ruLearn: an open-source toolkit for the automatic inference of shallow-transfer rules from scarce parallel corpora and morphological dictionaries. ruLearn produces rules that can be used in the Apertium platform without further modification and are able to reach the translation quality of hand-crafted rules. The software architecture of the toolkit allows it to deal with the complexity of the rule inference algorithm by introducing a high degree of parallelisation.

Concerning future research lines, the rule formalism could be extended with a new type of GAT in order to further improve the generalisation power and the translation quality achieved between languages that are not closely related. These new GATs would operate on sequences of chunks instead of sequences of words and would be encoded as Apertium *interchunk* rules (Forcada et al., 2011, Sec. 2.1). ruLearn could also be used when a parallel corpus is not available if a crowdsourcing (Wang et al., 2013) approach is followed. Finally, we plan to integrate our open-source tool for hybridisation (Sánchez-Cartagena et al., 2012) into ruLearn in order to ease the use of automatically inferred rules in a phrase-based SMT system.

Acknowledgements

Research funded by the Spanish Ministry of Economy and Competitiveness through projects TIN2009-14009-C02-01 and TIN2012-32615, by Generalitat Valenciana through grant ACIF/2010/174, and by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran).

Bibliography

Alcázar, A. Consumer Corpus: Towards linguistically searchable text. In *Proceedings of BIDE (Bilbao-Deusto) Summer School of Linguistics 2005*, Bilbao, Spain, 2005.

Caseli, H. M., M. G. V. Nunes, and M. L. Forcada. Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation*, 20(4):227–245, 2006.

Costa-Jussà, M. R. and J. Centelles. Description of the Chinese-to-Spanish Rule-Based Machine Translation System Developed Using a Hybrid Combination of Human Annotation and Statistical Techniques. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 15(1), 2015.

Forcada, M. L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, 2011. Special Issue: Free/Open-Source Machine Translation.

Garfinkel, R. S. and G. L. Nemhauser. *Integer programming*, volume 4. Wiley New York, 1972.

Klubička, F., G. Ramírez-Sánchez, and N. Ljubešić. Collaborative development of a rule-based machine translator between Croatian and Serbian. *Baltic Journal of Modern Computing*, 4(2), 2016.

Koehn, P. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 4, pages 388–395, Barcelona, Spain, 2004.

Koehn, P. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit X*, pages 12–16, Phuket, Thailand, September 2005.

Koehn, P. *Statistical Machine Translation*. Cambridge University Press, 2010.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. doi: 10.3115/1073083.1073135. URL `http://www.aclweb.org/anthology/P02-1040`.

Probst, K. *Automatically Induced Syntactic Transfer Rules for Machine Translation under a Very Limited Data Scenario*. PhD thesis, Carnegie Mellon University, 2005.

Sánchez-Cartagena, V. M., F. Sánchez-Martínez, and J. A. Pérez-Ortiz. An open-source toolkit for integrating shallow-transfer rules into phrase-based satistical machine translation. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 41–54, Gothenburg, Sweden, June 2012.

Sánchez-Cartagena, V. M., J. A. Pérez-Ortiz, and F. Sánchez-Martínez. A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. *Computer Speech & Language*, 32(1):46–90, 2015. Hybrid Machine Translation: integration of linguistics and statistics.

Sánchez-Cartagena, V. M., J. A. Pérez-Ortiz, and F. Sánchez-Martínez. Integrating rules and dictionaries from shallow-transfer machine translation into phrase-based statistical machine translation. *Journal of Artificial Intelligence Research*, 55:17–61, 2016.

Sánchez-Martínez, F. and M. L. Forcada. Inferring shallow-transfer machine translation rules from small parallel corpora. *Journal of Artificial Intelligence Research*, 34(1):605–635, 2009.

Varga, I. and S. Yokoyama. Transfer rule generation for a Japanese-Hungarian machine translation system. In *Proceedings of the Machine Translation Summit XII*, Ottawa, Canada, 2009.

Wang, A., C. Hoang, and M.Y. Kan. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47(1):9–31, 2013.

Xu, Y., T. K. Ralphs, L. Ladányi, and M. J. Saltzman. Computational experience with a software framework for parallel integer programming. *INFORMS Journal on Computing*, 21(3), 2009.

Address for correspondence:
Víctor M. Sánchez-Cartagena
`vmsanchez@prompsit.com`
Prompsit Language Engineering
Av. Universitat s/n. Edifici Quorum III. E-03202 Elx, Spain

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 205–213

Lexicographic Tools to Build New Encyclopaedia of the Czech Language

Aleš Horák, Adam Rambousek

Natural Language Processing Centre
Faculty of Informatics, Masaryk University

Abstract

The first edition of the Encyclopaedia of the Czech Language was published in 2002 and since that time it has established as one of the basic reference books for the study of the Czech language and related linguistic disciplines. However, many new concepts and even new research areas have emerged since that publication. That is why a preparation of a complete new edition of the encyclopaedia started in 2011, rather than just re-printing the previous version with supplements. The new edition covers current research status in all concepts connected with the linguistic studies of (prevalently, but not solely) the Czech language. The project proceeded for five years and it has finished at the end of 2015, the printed edition is currently in preparation. An important innovation of the new encyclopaedia lies in the decision that the new edition will be published both as a printed book and as an electronic on-line encyclopaedia, utilizing the many advantages of electronic dictionaries.

In this paper, we describe the lexicographic platform used for the Encyclopaedia preparation and the process behind the work flow consisting of more than 3,000 pages written by nearly 200 authors from all over the world. The paper covers the process of managing entry submissions, the development of tools to convert word processor files to an XML database, tools to cross-check and connect bibliography references from free text to structured bibliography entries, and the preparation of data for the printed publication.

1. Introduction

The first edition of the Encyclopaedia of the Czech Language (Bachmannová et al., 2002) was published in 2002. Since that time it has been adopted as one of the basic reference books for the study of the Czech language and related linguistic disciplines

 Corresponding author: rambousek@fi.muni.cz
Cite as: Aleš Horák, Adam Rambousek. Lexicographic Tools to Build New Encyclopaedia of the Czech Language. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 205–213. doi: 10.1515/pralin-2016-0019.

not only in the Czech Republic, but by Bohemists all over the world. However, many new concepts and even new research areas have emerged since that publication. The Encyclopaedia editorial board (led by Petr Karlík) decided to prepare a complete new edition of the encyclopaedia, rather than just a re-print with supplements. The new edition covers current research as well all the concepts of linguistic studies connected with the Czech language. The project is coordinated by a team at the Faculty of Arts, Masaryk University, it started in 2011 and has finished at the end of 2015. Currently (June 2016), the encyclopaedia data undergoes the final proofreading and the final processing phase before publishing. The final version of the New Encyclopaedia contains 1569 entries, spanning over 3,000 pages, written by 191 authors.

As an important innovation of the original encyclopaedia from 2002, the new edition is primarily organized as an electronic encyclopaedia, utilizing the advantages of the electronic dictionaries. The printed version will be published by a well-known Czech publishing house Nakladatelství Lidové noviny based on the preprocessed data exported from the electronic edition. This move to electronic publishing is in line with recent trends in dictionary publishing (Tarp, 2012; Verlinde and Peeters, 2012). The DEB platform was selected as the dictionary writing system for the preparation of the new edition.

2. The DEB Platform Overview

Based on the experience with several tens of dictionary projects, the team at the NLP Centre FI MU has designed and implemented a universal dictionary writing system that can be exploited in various lexicographic applications to build large lexical databases. The system has been named Dictionary Editor and Browser (Horák and Rambousek, 2007), shortly DEB,[1] and has been used in more than twenty lexicographic projects since 2005, e.g. the development of the Czech Lexical Database (Rangelova and Králík, 2007), or currently running projects of the Pattern Dictionary of English Verbs (Hanks, 2004), Family names in UK (Hanks et al., 2011), and highly multimedial Dictionary of Czech Sign Language (Rambousek and Horák, 2015).

The DEB platform is based on the client-server architecture, which brings along a lot of benefits. All the data are stored on the server side and a considerable part of the client-side functionality is also implemented on the server, thus the client application can be very lightweight. The DEB platform approach provides very good tools for team cooperation: all data modifications are immediately seen by all involved users. The server also provides well arranged authentication and authorization functions. Unlike other dictionary writing systems (both commercial, and open-source), the DEB platform is not limited to one type of language or knowledge resources. DEB supports requirements of many frequently used resource types, while most of the applications

[1] http://deb.fi.muni.cz

specialize only on one type of data. The DEB platform and related projects are covered in detail in Rambousek (2015).

3. The Encyclopaedia Editing Process

3.1. The Editing Team Management

The encyclopaedia text preparation team consists of 191 authors, supported by 15 other team members. Here, the DEB platform support for complex access rights is utilized – all the users are hierarchically organized as *entry authors*, *entry referees*, *area administrators*, *editing assistants*, and *encyclopaedia coordinators* with various levels of access to the dictionary data. For example, the entry authors may only edit the entries assigned to them by the area administrator. The system can also limit access to all entries (or a selected subset) for some users during various editing phases, eg. for batch update. The editors can compare several versions of each entry – the original document provided by the author(s), the XML file stored in the database, the HTML preview for checking, and the final electronic version.

The management system also provide reporting tools to track progress of individual entries and overall encyclopaedia, such as:

- the current *editing phase* of an entry (posted by an author, converted to XML database, proofread, electronic version verified, confirmed by area administrator, etc.),
- the *number of work-in-progress* and *finished entries*,
- the *number of entries* and *"normalized" pages*[2] written by each of the authors,
- the option to see and compare the *editing history* of each entry.

Apart from the full history of document changes, the system also provides daily backups of the database.

3.2. Entry Editing and Conversion

Since the source materials of the original edition were prepared as a set of word processing documents, and mainly because some of the authors could not use the online editing tools, it was decided by the editorial board that in the first stage of the editing process, the entries will be written in the form of a word processing documents. To allow the use of the new features for the electronic encyclopaedia, special markup tags inside the standard document text were introduced for the project:

- the entry headword and its variants,
- the entry category classification,
- the entry author(s),
- splitting the entry text to two parts – a standard part for public users, and an advanced part for researchers and experienced scholars,

[2] 1,800 characters of text per page

- the list of bibliographic references,
- a definition (form of a sub-entry) in the text,
- multimedia attachment files (images, audio recordings, video files), either included inline in the text, or on separate pages,
- cross-references to other entries, sub-entries or bibliographic references.

At the first step, documents provided by the authors in several word processing formats are unified by automatic conversion to the Open Document format (Brauer et al., 2011).

In the next step, the ODF documents are converted to an internal XML format. The word processor instructions and special markup tags are converted to semantic tags of the encyclopaedia XML format. Wherever possible, the included images are converted to vector graphic formats to provide the best image quality both in the electronic and the printed edition. During this phase, varying text formatting is unified to the common layout of the New Encyclopaedia.

All of the 1569 entries were regularly processed and converted during the Encyclopaedia editing. As more documents were provided by the authors, the conversion tools were continuously updated to handle various input document formats.

After the upload and conversion, the documents are stored in the DEB XML database and edited via the online tools. It is also possible to download the entries for offline editing and upload the updated version later.

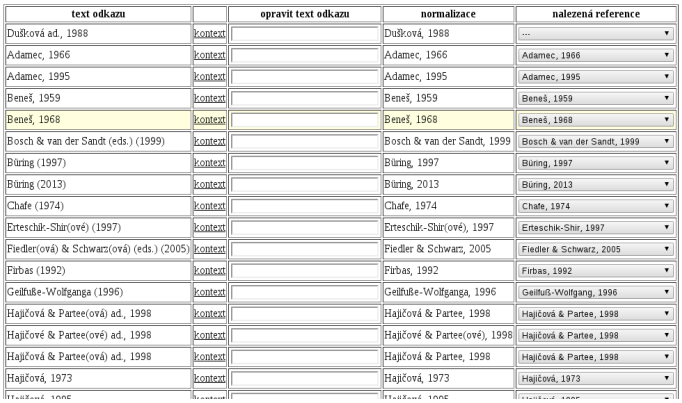
3.3. Cross-references Checks

Although the subject areas and entry lists were planned beforehand, many changes were introduced during the writing process. Sometimes, completely new entries emerged to describe the current state of linguistic research. In other cases, entries were split or merged for the best presentation of the concepts and spreading the length of entries more evenly. However, such changes could cause various errors in entries cross-referencing.

In the final phase of the Encyclopaedia preparation, all cross-references between entries were checked and verified. The lexicographic system tools scanned all the entries and their connections, reporting any inconsistencies or links to missing entries. The editing assistants then browsed through and fixed each of the errors, either with updating the cross-reference to another entry, creating new variant headword, or deleting the link. During this process, several entries were identified that were omitted during the writing phase and needed to be added.

3.4. Bibliographic References Processing

After the final form of all entries was delivered, the bibliography lists and all bibliographic references in the text were verified. Since the encyclopaedia texts are written in Czech, the bibliographic references within the entry texts may come in different



text odkazu		opravit text odkazu	normalizace	nalezená reference
Dušková ad., 1988	kontext		Dušková, 1988	---
Adamec, 1966	kontext		Adamec, 1966	Adamec, 1966
Adamec, 1995	kontext		Adamec, 1995	Adamec, 1995
Beneš, 1959	kontext		Beneš, 1959	Beneš, 1959
Beneš, 1968	kontext		Beneš, 1968	Beneš, 1968
Bosch & van der Sandt (eds.) (1999)	kontext		Bosch & van der Sandt, 1999	Bosch & van der Sandt, 1999
Büring (1997)	kontext		Büring, 1997	Büring, 1997
Büring (2013)	kontext		Büring, 2013	Büring, 2013
Chafe (1974)	kontext		Chafe, 1974	Chafe, 1974
Erteschik-Shir(ové) (1997)	kontext		Erteschik-Shir(ové), 1997	Erteschik-Shir, 1997
Fiedler(ová) & Schwarz(ová) (eds.) (2005)	kontext		Fiedler & Schwarz, 2005	Fiedler & Schwarz, 2005
Firbas (1992)	kontext		Firbas, 1992	Firbas, 1992
Geilfuße-Wolfganga (1996)	kontext		Geilfuße-Wolfganga, 1996	Geilfuß-Wolfgang, 1996
Hajičová & Partee(ová) ad., 1998	kontext		Hajičová & Partee, 1998	Hajičová & Partee, 1998
Hajičové & Partee(ové) ad., 1998	kontext		Hajičové & Partee(ové), 1998	Hajičová & Partee, 1998
Hajičová & Partee(ová) ad., 1998	kontext		Hajičová & Partee, 1998	Hajičová & Partee, 1998
Hajičová, 1973	kontext		Hajičová, 1973	Hajičová, 1973

Figure 1. Verification and matching of bibliographic references.

inflected forms (grammar cases, masculine/feminine name endings, etc.). As a first step, a uniform and unique representation of each item in the bibliography list was created. Although the authors followed the CSN ISO 690-2 standard (CSN690, 2011) for references, many items contained some sort of spelling or typing errors. All inconsistencies to the standard were reported and fixed.

In the next step, all references in the entry text were transformed to the same unified form and matched against the entry bibliography list. From the total of 16,252 bibliography links, 95 % were correctly matched using the uniform representation. See Figure 1 for an example of the bibliography checking form to verify and interlink the bibliographic references. The remaining cases consisted of following issues that were handled by the editing assistants:

- an unusual form or misspelled name, year or other part of the bibliography reference,
- a bibliography entry not following the standard form,
- a choice of two (or more) publications by the same author in the same year,
- a missing bibliographic entry,
- a misplaced reference in the entry text.

3.5. Final Proofreading by Authors

When the conversion, verification and finalization processes were successfully carried out, all the authors were asked to proofread their own entries before submitting the final data to the publisher.

For this purpose, an entry representation similar to the printed edition was created in the PDF format and prepared for download on personalized author checking web

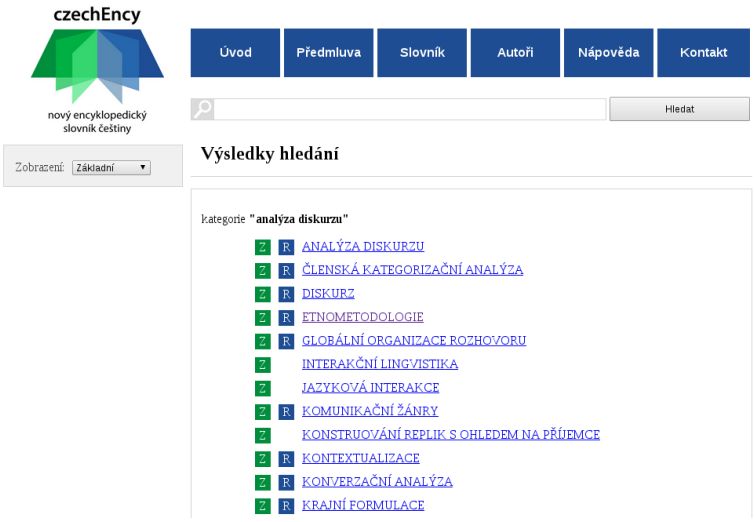


Figure 2. Search results displaying entries from a selected category. The green and blue squares indicate the proficiency level (standard/advanced) of the entry parts available.

pages. Subsequently, the authors were able to enter the proofreading comments and requests into a special web-based form. All identified issues[3] were then transferred to the database by the team of editing assistants. During the review, 110 entries written in an early stage of the project were (usually briefly) updated to reflect the current research. Because of the changes, it was required to verify some of the cross-references and bibliographic references again.

3.6. The Printed Edition

The printed edition of the Encyclopaedia is going to be published by one of the largest Czech publishing houses, Nakladatelství Lidové noviny. The lexicographic system contains tools to prepare the data in the format used by the publisher for typesetting:

- each entry is saved in a separate XML file,
- the metadata are updated (e.g. author abbreviation is changed to full name),
- the cross-references are updated to link correctly to the saved XML files,
- all included multimedia files are downloaded,
- all images are saved in all available formats to provide the best quality for typesetting.

[3]There were about 1,700, mostly small, corrections reported.

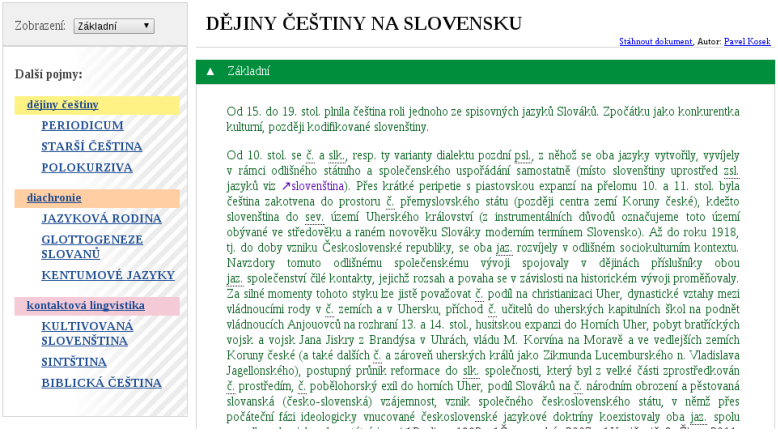


Figure 3. Preview of an entry, with links to more information in the same category.

Due to file conversions, cross-references checks, and various document updates, preparation of all 1569 entries for the publisher takes one hour. Without additional features, the complete export of the database takes less then 5 minutes.

3.7. The Electronic Edition

The New Encyclopaedia edition takes the advantage of the electronic and multimedia dictionary features to help users with navigation in the encyclopaedia, to obtain extra relevant information, and to better understand the concepts. See Figure 2 for an example of search results, and Figure 3 for an example of an entry preview. The DEB platform tools take care of properly encoding and providing all the following information:

- cross-references to other entries or entry parts,
- links to external websites,
- references to the bibliography, with the possibility to query external library resources,
- images, charts, diagrams etc.,
- sound recordings (e.g. samples of various dialects, see Figure 4),
- animations and video recordings (e.g. video recordings showing sign language gestures),
- explanations of the abbreviations used in the text or bibliography lists.

To make the encyclopaedia content comprehensible and useful for different reader groups, the entries can be described in two different levels of proficiency, i.e. the entry text can contain a standard part and an advanced part. Out of the total of 1,569 entries

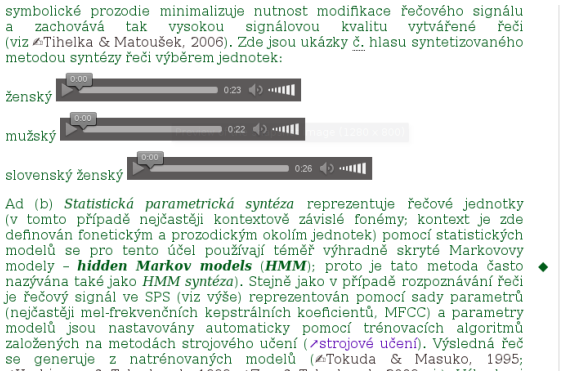


Figure 4. Example of an entry with inline sound recordings.

in the encyclopaedia, 1,093 entries contain just the standard part, 193 entries contain only the advanced part, and 283 entries have both descriptive parts.

On the encyclopaedia website, readers may choose their preference of the default description level. For example, readers may hide the advanced information and when they search for an entry, only the standard entries or descriptions are provided.

The system tracks the most often visited entries in each category and provides hints to extra information in related categories for readers interested in certain topics.

4. Conclusions

We have described the tools and processes utilized to build the New Encyclopaedia of Czech, the largest electronic encyclopaedia devoted to the Czech Language and related linguistic studies. The presented lexicographic tools successfully supported the team of more than 200 authors and assistants during creation of both printed and electronic version of one of the most important resource for the study of the Czech language and many connected areas of linguistic research.

Acknowledgements

This paper describes the Encyclopaedia created as a result of the Czech Science Foundation project P406/11/0294. This work was also partially supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071 and by the national COST-CZ project LD15066. The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047.

Bibliography

Bachmannová, Jarmila, Petr Karlík, Marek Nekula, and Jana Pleskalová. *Encyklopedický slovník češtiny*. Lidové noviny, Praha, 2002.

Brauer, Michael, Patrick Durusau, Gary Edwards, David Faure, Tom Magliery, and Daniel Vogelheim. ISO/IEC 26300:2006: Open Document Format for Office Applications 1.2, 2011.

CSN690, 2011. *ČSN ISO 690 (01 0197) Informace a dokumentace - Pravidla pro bibliografické odkazy a citace informačních zdrojů*. Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, Praha, 3rd edition, 2011.

Hanks, Patrick. Corpus Pattern Analysis. In *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France, 2004. Universite de Bretagne-Sud.

Hanks, Patrick, Richard Coates, and Peter McClure. Methods for Studying the Origins and History of Family Names in Britain. In *Facts and Findings on Personal Names: Some European Examples*, pages 37–58, Uppsala, 2011. Acta Academiae Regiae Scientiarum Upsaliensis.

Horák, Aleš and Adam Rambousek. DEB Platform Deployment – Current Applications. In *RASLAN 2007: Recent Advances in Slavonic Natural Language Processing*, pages 3–11, Brno, Czech Republic, 2007. Masaryk University.

Rambousek, Adam. *Creation and Management of Structured Language Resources*. PhD thesis, Faculty of Informatics, Masaryk University, 2015.

Rambousek, Adam and Aleš Horák. Management and Publishing of Multimedia Dictionary of the Czech Sign Language. In *Natural Language Processing and Information Systems - 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015, Passau, Germany, June 17-19, 2015. Proceedings*, Lecture Notes in Computer Science. Springer, 2015.

Rangelova, Albena and Jan Králík. Wider Framework of the Research Plan Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century. In *Proceedings of the Computer Treatment of Slavic and East European Languages 2007*, pages 209–217, Bratislava, Slovakia, 2007.

Tarp, Sven. Theoretical challenges in the transition from lexicographical p-works to e-tools. In Granger, Sylviane and Magali Paquot, editors, *Electronic Lexicography*, pages 107–118. Oxford University Press, Oxford, 2012. ISBN 978-0-19-965486-4. doi: 10.1093/acprof:oso/9780199654864.001.0001.

Verlinde, Serge and Geert Peeters. Data access revisited: The Interactive Language Toolbox. In Granger, Sylviane and Magali Paquot, editors, *Electronic Lexicography*, pages 147–162. Oxford University Press, Oxford, 2012. ISBN 978-0-19-965486-4. doi: 10.1093/acprof:oso/9780199654864.001.0001.

Address for correspondence:
Adam Rambousek
rambousek@fi.muni.cz
Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, Brno, Czech Republic

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016

INSTRUCTIONS FOR AUTHORS

Manuscripts are welcome provided that they have not yet been published elsewhere and that they bring some interesting and new insights contributing to the broad field of computational linguistics in any of its aspects, or of linguistic theory. The submitted articles may be:

- long articles with completed, wide-impact research results both theoretical and practical, and/or new formalisms for linguistic analysis and their implementation and application on linguistic data sets, or
- short or long articles that are abstracts or extracts of Master's and PhD thesis, with the most interesting and/or promising results described. Also
- short or long articles looking forward that base their views on proper and deep analysis of the current situation in various subjects within the field are invited, as well as
- short articles about current advanced research of both theoretical and applied nature, with very specific (and perhaps narrow, but well-defined) target goal in all areas of language and speech processing, to give the opportunity to junior researchers to publish as soon as possible;
- short articles that contain contraversing, polemic or otherwise unusual views, supported by some experimental evidence but not necessarily evaluated in the usual sense are also welcome.

The recommended length of long article is 12–30 pages and of short paper is 6–15 pages.

The copyright of papers accepted for publication remains with the author. The editors reserve the right to make editorial revisions but these revisions and changes have to be approved by the author(s). Book reviews and short book notices are also appreciated.

The manuscripts are reviewed by 2 independent reviewers, at least one of them being a member of the international Editorial Board.

Authors receive a printed copy of the relevant issue of the PBML together with the original pdf files.

The guidelines for the technical shape of the contributions are found on the web site `http://ufal.mff.cuni.cz/pbml`. If there are any technical problems, please contact the editorial staff at `pbml@ufal.mff.cuni.cz`.