Punctuation and multilinguality:
Some reflections from a language engineering perspective

Diana Santos

Abstract

The aim of this paper is to question the notion of multilinguality, claiming that one cannot go above (or below) contrastive studies based on translation (which is a directional activity).

After defending this standpoint in abstract terms in the context of machine translation, I proceed by considering punctuation. This somewhat unusual subject was chosen because of its a priori easy definition (as opposed to e.g. tense and aspect, or part-of-speech) and, also, because of its special status in natural language processing, which will be commented upon more in detail below.

Having discussed some issues of punctuation in general, I turn to a brief description of differences between the use of punctuation marks in Portuguese, English and Norwegian.

As the empirical base, I analyse more closely the use of the colon and direct speech marking with the help of examples in the English Norwegian Parallel Corpus (ENPC). Based on the situations actually encountered in the language pairs English-Portuguese, English-Norwegian, and Norwegian-English, I discuss the possible multilingual options for handling these punctuation marks, concluding that the best way is to keep the several corresponding bilingual systems side to side.

To conclude, I express a warning against the concept of multilingual (aligned) corpora in general, presenting some additional arguments that make their constitution less appealing. The bottom line of the final section is thus: Carefully designed bilingual corpora are more appropriate for interlanguage comparison.

There is an upsurge of interest in multilinguality these days. Almost everyone is either creating multilingual systems from scratch or turning their previous systems into multilingual ones. So "multilinguality" has become a catchall, with a lot of people employing it totally unaware of the linguistic aspects which might be involved. This paper is an attempt to look into the assumptions underlying the concept of multilinguality, given that my previous work on machine translation (see e.g. Santos, 1993) and on the contrast of the tense and aspect systems of English and Portuguese (Santos, 1996) helped me forming definite opinions on contrastive matters. For the present paper, I chose as subject matter punctuation and as computational application the encoding of parallel corpora.

1. Some thoughts about "multilinguality"

The aim of the present paper is to question the notion of multilinguality, the "just add another language" idea. This excludes the straightforward cases of product localization, i.e., offering the same functionalities in another language. These products are still monolingual, but with a language parameter.

Outside this simple case, multilinguality is hard to define — it is even hard, I think, to see the need for it. Multilinguality is certainly not meant to replace monolinguality: one natural language per speaker is enough. In other words, and as far as I know, the aim is not to obtain a more expressive natural language.

A truly "multilingual" application (as opposed to parameterized monolinguality) is a system allowing language switching, that is, a system that is able to get input in more than one language and produce output in more than one language (and the output language(s) are different from the input language(s)), as sketched on the left handside of Figure 1 ($L_{1...n}$ stand for input languages, $L_{a...z}$ for output languages).

In the present paper, I will restrict my attention to natural language processing systems, i.e., I will disregard applications with canned output / fixed menus input — such as accessing Portuguese abstracts through a Norwegian menu. Now, if one requires an NLP system to analyse natural language input and produce NL output, however limited, one needs

- either translation into a language-independent representation (top-right of Fig. 1)
- or translation between language dependent representations (bottom right of Fig. 1)

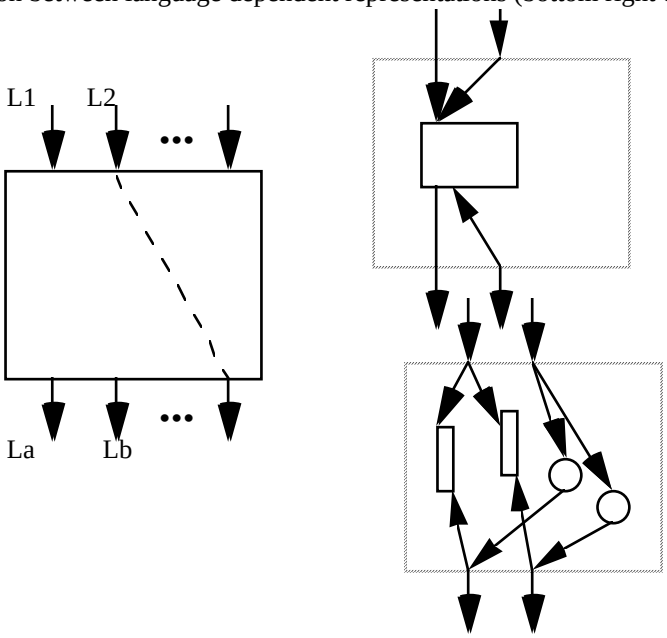


Figure 1

Everyone is aware of the advantage of the first possibility — which is not to say that it is possible in most domains. The hypothesis that it is possible to find it for general purpose systems has never been proved. In fact, if failed attempts could count as proofs of non-existence, it would have been ruled out long ago.

Even though in principle the inner box on the top right handside of Figure 1 could do much more than "mere" translation, the question of multilingual systems strikingly resembles the old question of machine translation (MT) architecture. And this is more so, because most knowledge that people might want to access is actually in natural language (and thus requires linguistic processing, and not other sorts of processing); and because machine translation itself has been repeatedly claimed to need all sorts of inference and non trivial processing in order to perform satisfactorily. So, I use MT to make my point about multilinguality.

There is a tradition in the machine translation literature that considers a sort of continuum between language dependency and depth of analysis, considering transfer as an engineering option. Such a view is illustrated in the left of Figure 2.

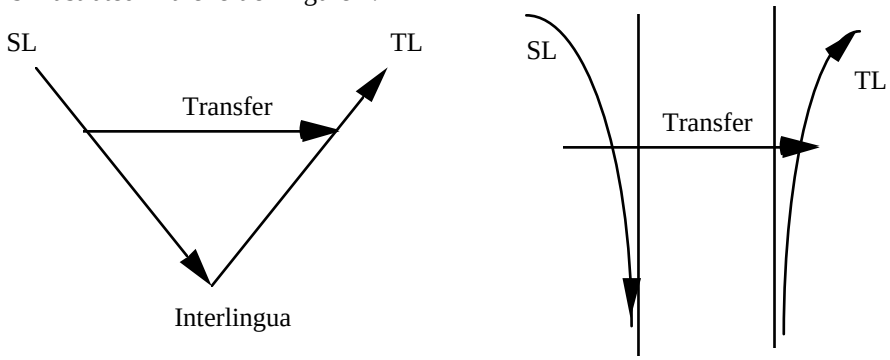


Figure 2

This was, however, already considered naive by Tsujii, one of the best-known MT developers, in his position paper at Coling'86, Tsujii (1986). Another possible model, displayed for example in Santos (1988:5), was the one on the right of Figure 2, where a residual fixed distance between languages could

not be overcome, no matter how deep an analysis, leading therefore to the conclusion that transfer is necessary. Such a residual fixed distance can also be related to Nitta's (1986) "idiosyncratic gap".

Now, I see no reason to expect that any depth of analysis of one language brings that language closer to another, and therefore the diagram that I believe to be right is the one of Figure 3.

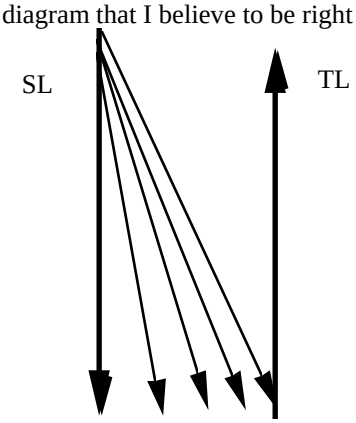


Figure 3

One can, of course, analyse any source-language text with the view to translate it into a target-language text. Such an analysis results in making the original text closer to the target language, in that one analyses the source text in terms of the categories of the target language. But, even if one claims to be using an interlingual representation, one is really seeing the source language through the eyes of the target language.

The intermediate representations are not language independent, they are language-pair dependent. My contention is that one always needs contrastive knowledge to be able to bridge the distance between two languages.

This is, by the way, also one of the conclusions reached by Tsujii in 1986: "the understanding results in such a system have to be specific to language pairs and not language universal" (1986:662), as displayed in Figure 4.

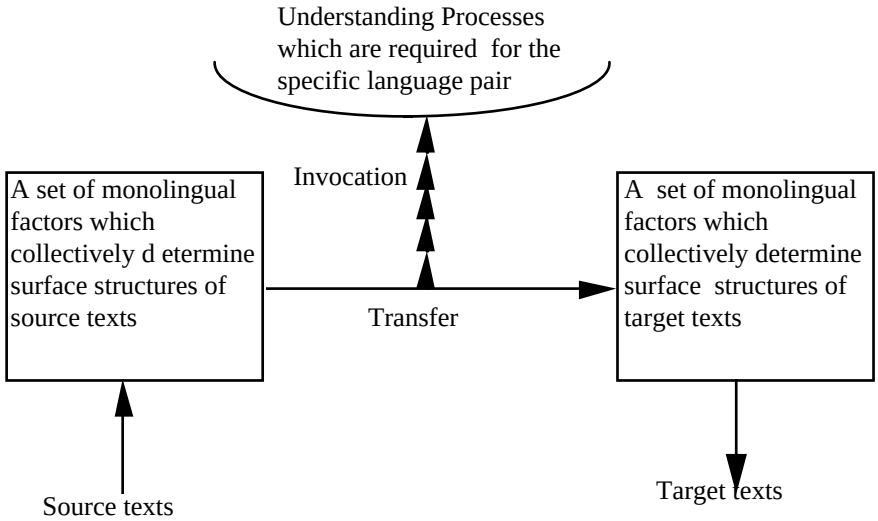
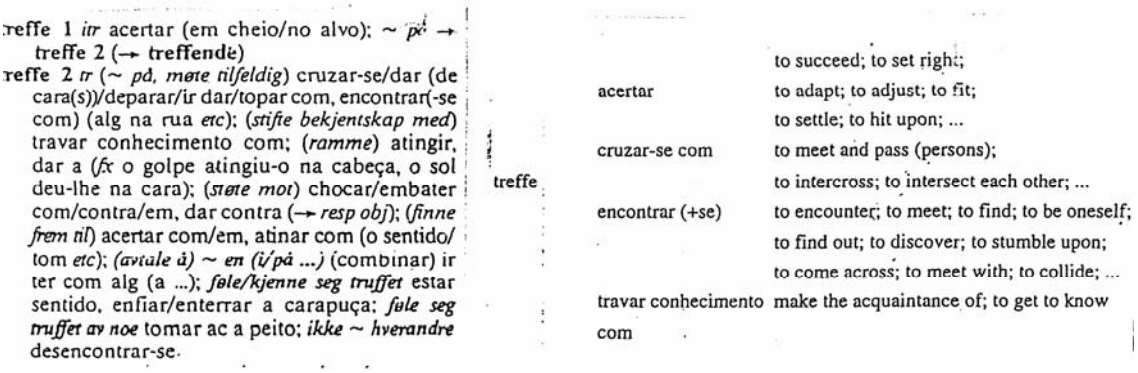


Figure 4 (adapted from Tsujii (1986:662, Fig 7))

Ultimately, the question of how to encode such contrastive knowledge is an engineering decision: contaminating the source language analysis with target language-relevant information or providing strict separation between contrastive and monolingual information are design criteria, not linguistically motivated options.

Still, one might hope to derive transitively contrastive information for other pairs. I.e., if one knew the contrast between A and B and B and C, one could automatically get the contrast between A and

C, without having to study it. This is, unfortunately, not possible due to the phenomenon of (contrastive) vagueness; see Santos (1997a). Figure 5 illustrates this fact by means of bilingual lexicographic material[1] from Norwegian to English, Norwegian to Portuguese, and Norwegian to English having Portuguese as intermediate pass.



treffe 1 *itr* acertar (em cheio/no alvo); ~ *på* → treffe 2 (→ **treffende**)
treffe 2 *tr* (~ *på, møte tilfeldig*) cruzar-se/dar (de cara(s))/deparar/ir dar/topar com, encontrar(-se com) (alg na rua *etc*); (*stifte bekjentskap med*) travar conhecimento com; (*ramme*) atingir, dar a (*fx* o golpe atingiu-o na cabeça, o sol deu-lhe na cara); (*støte mot*) chocar/embater com/contra/em, dar contra (→ *resp obj*); (*finne frem til*) acertar com/em, atinar com (o sentido/ tom *etc*); *(avtale å) ~ en (i/på ...)* (combinar) ir ter com alg (a ...); *føle/kjenne seg truffet* estar sentido, enfiar/enterrar a carapuça; *føle seg truffet av noe* tomar ac a peito; *ikke ~ hverandre* desencontrar-se.

treffe		to succeed; to set right;
	acertar	to adapt; to adjust; to fit;
		to settle; to hit upon; ...
	cruzar-se com	to meet and pass (persons);
		to intercross; to intersect each other; ...
	encontrar (+se)	to encounter; to meet; to find; to be oneself;
		to find out; to discover; to stumble upon;
		to come across; to meet with; to collide; ...
	travar conhecimento com	make the acquaintance of; to get to know

treffe hit; meet (with); come across; bump into

Figure 5

Compared with the right translations (at the bottom of Figure 5), one would have obtained as additional (and mistaken) translations for Norwegian *treffe* the English *discover, adapt, succeed,* etc.

In sum, this discussion amounts to the following points:

1. A multilingual system of a non-trivial kind requires translation across more than one language pair, i.e., it requires contrastive information for at least two different language pairs
2. Contrastive information is essentially language-pair dependent
3. It is not possible to derive contrastive information transitively for other pairs, due to the pervasiveness of contrastive vagueness

On the basis of this, I conclude that one cannot escape the need to study translation for each language pair (more precisely, for each direction of translation).

In other words, multilinguality does not exist over and above bilinguality, i.e., over and above translation from one language into another. Or, no matter how a multilingual system is designed, it will have to incorporate bilingual systems based on (directional) translation.

In order to show that it makes no sense to generalize, or create a higher-order notion of multilinguality which is qualitatively different from the study of two languages in contrast, I will discuss the concrete question of punctuation.

2. Some points on punctuation

Punctuation is pertinent to a discussion of multilinguality because it seems "the same" across (at least Western) languages. Showing the multiple problems it actually involves is thus an effective way to illustrate my point. After some general remarks, I will discuss the possible "multilingual" solutions to handle the colon and direct speech, in a system comprising Norwegian, Portuguese and English.

2.1 Punctuation as a linguistic system

Punctuation pertains to the written language, supposedly the counterpart of intonation and prosody. This has, however, been challenged by Nunberg (1990), who claims that it should be studied as a linguistic system in its own right:

"punctuation is usually regarded as a (highly imperfect and limited) device for transcribing certain

[1] Extracted from Kåre Nilsson's *Norsk-portugisisk ordbok*, Oslo: Universitetsforlaget, 1994, *Dicionário de Português-Inglês*, Porto: Porto Editora, 1989, and W.A. Kirkeby's *Norsk-Engelsk ordbok*, Oslo: Kunnskapsforlaget, 1986.

> of the prosodic and pausal features of speech. (...) the picture is both empirically unwarranted and theoretically incoherent. (...) It may turn out that punctuation and intonation are sometimes used to convey the same kinds of information, just as the English passive and the Romance reflexive do; but that should not be a starting point for the analysis of either system"

Formally, punctuation consists of a small vocabulary which gives clues as to how to interpret (and consequently read aloud) a written text. Closely connected with it are graphical conventions like paragraph, table of contents, footnotes, etc.

Traditionally, punctuation is perhaps the module of language competence most difficult to master in every language, probably because it pertains only to the written mode. In this respect, Franckel's (1989:25, my translation) remark is particularly relevant:

> "that some rules are violatable by native speakers means that the system has some degrees of freedom [...]; the fact itself that they are taught witnesses an immediate consciousness of their existence, if not of their organization. We can therefore, simplifying a little, state that the degree of epilinguistic conscience of the rules is inversely proportional to their necessity."

2.2 Punctuation as a style norm

Punctuation "rules" are extensively discussed in style manuals (see e.g. the references cited in note 13), as well as being a recurring theme in the newspaper's popular advice columns on "good language use".

It is probably undeniable that to know how to write well in one language requires one to internalize what is considered a good distribution of punctuation marks. See for instance the wide-spread Norwegian slogan "It is not a shame to write a dot".[2]

In a parallel fashion, style checkers do spend a lot of effort on this matter. In addition, they sometimes provide readibility indexes based on formulae where sentence length figures prominently.

It is true, however, that there seems to be two kinds of punctuation: the reliable one (dots, exclamation and question marks) and the unreliable one (all the rest). The latter are generally misused, and only specialists can confidently be trusted as far as their proper use is concerned. Commas are the most obvious example of such a chaotic situation.

2.3 Punctuation in natural language processing

It is interesting to note that "reliable" punctuation has acquired an important status in natural language processing – where, as Engh (forthcoming) correctly points out, there is no way to isolate subsystems and ignore the rest of the language as uninteresting:

Almost all NL analysers of running text (for Western languages) have a sentence separator module (pre-processor, tokenizer, "normalizer", etc.) that divides input based on punctuation marks. This gives punctuation a key role in parsing. In fact, I know of no system that can recover from a deficient sentence partitioning.

Also, almost all parallel text analysers use punctuation (again, usually the "reliable" punctuation) for their first processing: sentence alignment. ("Sentence" being defined precisely as what lies between two major punctuation marks.) As will be discussed below, this is not without problems either.

Perhaps, this is also the right moment to ask: What is the significance of the universally used measure "rate of one-to-one versus other sentence pairings"? It may reflect the complexity of the alignment process itself, but can it be taken to represent a fair guess of freeness of translation? Not if the language pair in question has obligatory translation of two clauses linked by a particular connective X into two sentences... Can it be taken as a good indicator of differences between languages? What if the only difference was punctuation itself, as ":" translates into "."? It is true that this is one difference among others, but should it be taken as more important than any other difference? Irrespective of the

[2] The first rule in the set of ten commandments for the proper use of Norwegian, hanging in the Nordic Department of the University of Oslo.

answer one gives to the above question, it is undeniable that punctuation has a special status in the alignment world.

Interestingly, it seems that there is a boom on computational studies of punctuation right now. While this is not the proper place to review the claims and achievements of those works, to some extent they point to conclusions that add to my points here: For example, Grefenstette & Tapanainen (1994), looking at the Brown corpus, show that it is not as easy as one might expect to distinguish sentence-final ("reliable"?) punctuation from the one used in abbreviations and other idiosyncracies. Briscoe (1994), and Briscoe and Carroll (1996) apply a text grammar based on Nunberg (1990) which covers 99.8% of the sentences but which yields thousands of analyses in 0.1% of the cases: "a text sentence containing eight commas (and no other punctuation) will have 3170 analyses. The multiple uses of commas cannot be resolved without access to (at least) the syntactic context of occurrence." (Briscoe & Carroll, 1996:140). They claim that "To our knowledge these are the first experiments which objectively demonstrate the utility of punctuation for resolution of syntactic ambiguity and improvement of parser coverage." (ibidem:147).

Jones (1994; 1996), not surprisingly, reaches the same conclusion, although in a different task: the insertion of punctuation marks in transcribed speech.

In any case, and as soon as one begins to look at the actual (messy) performance data, there is no doubt that punctuation hangs together with the rest of the linguistic signs, no matter whether one emphasizes their interrelationship (like in the quote of Briscoe & Carroll above) or rather contrast their utility, as opposed to lexical clues, as is done by Doran (1996:9): "the direct/indirect split is not the correct one, and [...] the choice of the verb and the other punctuation marks involved are [sic] more informative than the quotation marks".

2.4 Punctuation in contrastive studies

I have already mentioned that many alignment systems relied on punctuation. But also studies on the actual texts aligned have been directed or influenced by punctuation matters.

For example, Wikberg's (1996) study on questions in English and Norwegian, by selecting punctuation marks as reliable indicators of a linguistically interesting category, also demonstrates the importance of punctuation in corpus-based studies, cf.: "The '?' is a strong punctuation mark in the sense that it is a fairly reliable graphical Q-marker." (Wikberg, 1996:18) It is therefore specially interesting to note that in the rest of Wikberg's paper a fair number of cases is presented where the use of '?' is not consistent across translation.

Baker (1996) cites Johansson on the increase of sentence length in translated text in both directions English to Norwegian and Norwegian to English as evidence for the "universal" feature of explicitation; and Johansson and Ebeling (1996) present a first exploration of the ENPC relying heavily on punctuation. Baker also claims that changes of punctuation in translated text reveal what she calls standardization; while, as I pointed out in Santos (1997b), punctuation translationese is an overwhelming phenomenon in children's literature translated into Portuguese.

3. The use of punctuation is highly language dependent

Even though most Western languages use the same (punctuation) vocabulary, it is also well-known that there are significant differences between them, i.e., each language has its own conventions and preferences.[3]

[3] This is actually only part of the truth: Spanish inverted question marks, or Portuguese *travessão* (the long dash) had no place in lower ASCII, while the apostrophe is central in English but very marginal in Portuguese, and *reticências* (one proper character in typing machines and for all purposes in school in Portugal) is called in English "tree dots" and actually implemented as such in most character sets.

3.1 Some random examples

Let me cite a few by way of illustration:

- "It is not a shame to write a dot" vs. "It is not a shame to have a paragraph longer than a page." This is a somewhat anecdotical description of the differences between Norwegian, on the one hand, and German – and possibly most Romance languages – on the other. Fabricius-Hansen (forthcoming) uses the concept of information density to explain this contrastive fact.
- A comma is obligatory before an integrating clause in German, forbidden in Portuguese or English.
- A comma is obligatory after a restrictive relative clause in Norwegian, forbidden in Portuguese and English.
- A comma obligatorily surrounds rethorical sentential adverbs in Portuguese, but it is forbidden in Norwegian.
- A comma is required after a sentence initial adverb or prepositional phrase in English, but forbidden in Norwegian[4] and optional in Portuguese.

Even "reliable" punctuation marks do not behave the same way:

- In English, one can have ? or ! followed by a semi-colon or even by a comma, but that is not allowed in Portuguese, nor in Norwegian. Examples from Nunberg (1990:26;66-67):

  A quorum?; not on your life.
  What we want to know is when?, but he has told us only why.
  I am puzzled (who is he?).
  I am annoyed (he is a fool!).
- Neither is it considered good style to have a coordination with *e* ('and') of an affirmative and an interrogative or exclamative clause in Portuguese, something which does not seem to be a problem for writers of English, as these strange-looking coordinations in translated Portuguese show: [5]

  Acompanhou-os até à toca dos coelhos e ali estavam todos os carros!
  'He accompanied them to the rabbitt hole and there were all the cars!'

3.2 Direct speech in the three languages

Direct speech conventions in English, Norwegian, and Portuguese differ considerably. In fact, direct speech has been signalled as one of the major sources for misalignment for the English-Portuguese pair in Santos (1994); likewise, Johansson and Ebeling (1996:8) note that "differences in paragraphing seem to be mainly due to slightly different ways of representing dialogue".

It seems thus appropriate to look closer at the conventions in the three languages.

In English, direct speech (or at least its first "chunk") is part of the clause where a speech verb is expressed or left implicit, and is signalled by double quotes. Sometimes the author uses colon, other times a comma, sometimes nothing.

In Portuguese, direct speech is marked with paragraph and an initial long dash, possibly interrupted by a sentence expressing the speech verb. If the clause containing the speech verb precedes the direct speech, it ends in colon. If, instead of direct speech, one describes thought, then it is expressed inside "quotes" just like in English.

The following real examples from the ENPC help to show the variety of both languages. On purpose, they were all extracted from the same book(s) (*Strong Medicine* by Arthur Hailey and its

---

On the other hand, and if punctuation is meant to somehow reflect intonation (?), again one must remember that intonation itself is language specific: A question does not sound the same in English and Finnish, even if they, graphically, use the same question mark.

[4] Unless the phrase is uncommonly long, so that a clear pause might be imagined while reading the sentence.

[5] From *Viva o Nodi*, translation by Maria da Graça Moctezuma of Enid Blyton's *Hurrah for little Noddy!*. The English gloss is mine.

translation into Brazilian Portuguese, *Remédio Amargo*[6]).[7]

E: <s>He told her, deadpan, "Love can happen to the elderly, too."</s>
P: <s>Ele comentou, com uma expressão impassível: </s>
<s> — O amor também pode acontecer aos velhos. </s>

E: <s> Then, with a quiet , "Good night, John," to the husband, he went out. </s>
P: <s> Depois, em tom mais suave, ele acrescentou para o marido, ao se retirar: </s>
<s> — Boa noite, John. </s>

E: <s> Reliving, mentally, the events of three days earlier, Andrew said "You'll have to make allowance for my having been a little dazed at the time."</s>
P: <s> Reconstituindo mentalmente os acontecimentos dos últimos três dias, Andrew comentou: </s>
<s> — Espero que me dê o devido desconto por me encontrar um tanto atordoado na ocasião. </s>

E: <s>She would ask, "Can I help you?" to which the reply was usually, "When will *he* be free?" </s>
P: <s> E indagava: </s>
<s> — Em que posso servi-lo?</s>
<s> A resposta era quase que invariável: </s>
<s> — Quando *ele* estará livre? </s>

E: And a moment later: "I was about to make coffee, Miss de Grey. </s>
P: <s> Um momento depois, acrescentara:</s>
<s> — Eu ia fazer café quando chegou, Srta. de Grey. </s>

E: <s>"There'd be one for us," his wife explained patiently, "then one each for the children, and after they're born we'll want live-in help, so that's one more." </s>
P: <s> Ela explicou pacientemente:</s>
<s> — Um quarto para nós, dois para as crianças e mais um para a empregada que precisaremos depois que nascerem. </s>

Note that when the speech verb <u>follows</u> the direct speech, or is <u>embedded</u> in it, structural differences between English and Portuguese are much smaller. One could actually talk of a simple replacement of English quotes by Portuguese long dashes (but there is no endquote in Portuguese):[8]

E: <s>"That's better," he said as they lay down where the clothes had been. </s>
P: <s> — Assim é melhor — disse ele, quando se deitaram sobre o lugar em que antes se encontravam as roupas. </s>

E: <s> Just the same," Sam Hawthorne put in quietly, "Mrs. Jordan has a point."</s>
P: <s> — Mesmo assim, — interveio Sam Hawthorne, suavemente, — a Sra. Jordan tem um argumento forte. </s>

It should also be said that, sometimes, translators into Portuguese do copy down the English punctuation marks as they are, as the following example shows:

---

[6] This was the first English to Portuguese translation whose alignment I revised, and it certainly made an impression on me that, out of 128 non 1-1 alignment cases, 70 were due to the systematic differences in direct speech coding between the two languages.

[7] To make the examples more readable, all SGML tags encoding information not relevant for the point at hand were replaced by the actual original features they were meant to encode. Namely, long dashes, italics, and non-English characters are displayed, not SGML-encoded. <s> and tags were however left to make it clear which sentence and paragraph boundaries were involved.

[8] An alternative way of describing this difference is to say that quotation is signalled by one-way delimiters in Portuguese and two-way delimiters in English. It is also interesting to note that in English the main clause seems to be the one of the speech verb, while in Portuguese the speech clause is a parenthetical one, and thus a subordinate clause.

E: <s>Celia started to say, "Do we have time?" but was unable to finish because Andrew was kissing her.</s>

P: <s>Celia começou a indagar "Temos tempo?", mas foi incapaz de terminar porque Andrew a estava beijando.</s>

One might classify this example as translationese; on the other hand, it is undoubtedly the case that the English punctuation allows for a more complex embedding than Portuguese.

In Norwegian, on the other hand, the situation is not so clear: it seems that the Portuguese way (long dash and paragraph) is used in books, but the English way (with colon obligatory) is used in ordinary language. In other words, there is a sharp separation between typesetters' rules and the layman's knowledge. I present some examples to show the diversity; the first set coming from Norwegian original text, the second from translated Norwegian:[9]

> En utålmodig stemme hadde sagt: – Nei, du må vri.
> Da jeg så annonsen idag, tenkte jeg: Endelig!
> Malvin Paulsen sa:
> – Vi kunne kanskje ta et glass sammen og minnes Hans Georg.
> Han strøk hånden over haken: "Det kan vi leve foruten," sa han.
> Professor Castberg presiserer:
> "Det ligger i sakens natur at det i et demokratisk monarki ...
> Så hørte hun at hun selv sa: – Jeg vil ikke dra.
> – Det begynner å bli latterlig, sa Martin, og stemmesteg og ble rasende etter som han snakket: – Et som lite sted som dette er, ...
> Hun holdt fram en bolle varm geitemelk: – Drikk dette, sa hun.
>
> Nå slengte hun kåpen over en stol og sa langsomt, med vekt på hvert ord:
> "Ammoniakk.
> Alice Mair sa rolig: "God dag, Ryan," men uten å vente noe svar, lot det til.
> Og når en mann sier: "Gud er mannlig," kjenner mange kvinner det som om de har mistet retten til å be.
> Aila åpnet døren og i ansiktet hennes (...) leste han en så umiddelbar forståelse og frykt at han, som aldri hadde kommet hjem til læreren før, glapp ut med noe uhyre tåpelig: Jeg gikk tilfeldigvis forbi...

In any case, this description of direct speech conventions highlights some differences in the use of the colon in the three languages, so I looked into the colon occurrences in a little more detail.

3.3 The use of the colon in the three languages

Very briefly, one might describe the main functions of the English colon as explicitation (i.e., paraphrasable by "namely"), explanation and extended direct speech, as illustrated respectively in (a), (b) and (c).[10]

(a) All families behave alike: on this planet, "The King gives the key into the keeping of the Queen".

(b) No giant is going to come along and suck out all the water for you: that magic stuff is not going to help.

(c) Before we finish this discussion of the wound and the genius, we must ask the question: What gender might the water be?

As usual, the classifications are NOT meant to be mutually exclusive: best uses of whatever linguistic device are those in which it is able to convey more than one of its functions – in other words, be vague –, as I argued for in detail in Santos (1997a).

In Portuguese, one could say that the colon serves direct speech rendering (previously illustrated),

---

[9] Due to space limits and to the fact that the majority of readers will understand Norwegian, I do not present the English rendering of these examples.

[10] The examples are drawn from Robert Bly's *Iron John. A Book about Men* (American English).

explanation, and the "namely" purpose, as examples (d) and (e) show:[11]

(d) Pior, perigoso: fios eléctricos arrancados da parede e pendentes, com as pontas descarnadas.
(e) Alice esteve prestes a descrever o estado da casa: cimento nas retretes, fios eléctricos soltos, tudo; mas reteve-se por instinto.

In Norwegian, finally, I was able to find the use of colon associated mainly with direct spech (also previously illustrated), extended speech and the "namely" function, cf. respectively (f) and (g):[12]

(f) På den sto det med Ingrids skrift: 1 mark kaffe 1 kg ...
(g) Hun skjønte at mamma så dem: Henrik og henne.

These short descriptions are not meant to be exhaustive of usage and/or meaning,[13] but simply reflect a superficial browsing of the ENPC source texts and my own knowledge of Portuguese. Even though the corpus does not encompass all kinds of written texts, the fact that it was compiled with the objective to provide a comparable set of texts in English and Norwegian makes the results of my comparison more significant.

It should thus be emphasized that languages often appear to have the same devices and/or functions, but differences show up in their markedness and/or their frequency, as well as in the contexts where they are grammatical. Tobin (1994) is an excellent demonstration of such a claim, to which I also arrived in my own work on the contrast of the English and Portuguese tense and aspect systems in Santos (1996).

3.4 A quantitative wrapping up

Text	:	:<p>	: "	: –	: A	: a
Ah1 (BrE)	11	1	2	0	5	3
PoB	96	79	2	7	4	4
BM	68	3	60	0	5	0
Dl2 (BrE)	23	0	1	0	9	13
Po	136	84	17	19	3	13
BM	107	0	73	0	27	7
Jb1 (BrE)	30	0	0	0	9	21
PoB	37	0	8	1	5	22
BM	28	1	9	0	13	5
Ng1 (SAE)	28	0	0	2	4	22
PoB	43	0	3	10	1	29
BM	41	0	6	0	5	30
Pdj3 (BrE)	60	0	57	0	1	2
Po	55	42	2	7	1	3
BM	64	1	55	0	7	1
Wb1(BrE)	31	0	1	0	9	21
PoB	28	0	2	1	4	21
BM	13	0	1	0	3	9
Total E	188	1	61	2	37	82
Total P	395	205	34	45	18	92
Total N	321	5	204	0	60	52

Table 1

In order to give some quantitative view of these two phenomena, Table 1 presents, for the six English texts of ENPC translated into both Norwegian and Portuguese, the distribution and number of coccurrences of the colon, the colon followed by a new line or by a double quote or by a long dash

[11] Examples from the Brazilian Portuguese translation of Doris Lessing's *The good terrorist*.

[12] Examples from Herbjørg Wassmo's *Huset med den blinde glassveranda*.

[13] For that purpose the reader is invited to look at, for example, Gundersen et al. (1995) for Norwegian Bokmål, Estrela (1987) for European Portuguese, and McCaskill (s.d.) for (American) English. One should also note that, according to the latter, "Authorities disagree on usage of the colon and capitalization after a colon".

(irrespective of blank spaces), or by a capital or non-capital in the texts. (BM = Bokmål; Po = European Portuguese; PoB = Brazilian Portuguese; BrE = British English; SAE = South African English)

Table 2, on the other hand, presents the same data concerning all English and Norwegian texts available, also discriminated between original and translated text. The interpretation of such numbers is left to the reader, who is however advised to take into consideration the diverse uses of these punctuation marks and not only direct speech.

Language	:	:<p>	: "	: –	: A	: a
English	2730	435	577	3	466	1204
original	1408	221	241	2	157	770
translated	1322	214	336	1	309	434
Norwegian	3250	562	989	165	851	629
original	1565	362	264	131	546	229
translated	1685	200	725	34	305	400

Table 2

4. How to handle the colon in a multilingual system including E, N and P

I have argued that multilingual systems do not exist over and above bilingual ones; that punctuation is language dependent; that much natural language processing is based on punctuation. From this, it should follow that it is not a trivial matter to decide how to handle a punctuation mark whose use happens to be different in the three languages. I shall try to show this with the case of the colon.

In theory, there are at least three ways to go about the definitio of sentence or paragraph:

1. One can follow the conventions of each language, and make monolingual decisions based solely on traditional wisdom concerning each language. (This is not that easy, either, because "traditional wisdom" has not been primarily concerned with this kind of matter).

As a security measure, one should not use the same markup tags (why should the "sentence" markup tag have the same meaning in the three languages?), and have Ep, Pp and Np, Es, Ps, Ns ("Ep" standing for English paragraph, "Ps" for Portuguese sentence, etc.). Systematic parallelism would be an empirical question.

Example:

- *The colon in English would be a sentence separator if immediately followed by a capital letter; not otherwise.*
- *The colon in Portuguese is a sentence separator if followed by a capital letter, or by paragraph.*
- *The colon in Norwegian would be a sentence separator if immediately followed by a capital letter (or by a quote followed by a capital letter); not otherwise.*

Consequences of this approach (assuming that comma is not considered a sentence separator in English) for our particular corpus are a) a large number of 1 to 2 Es-Ps, b) a large number of 1 to 2 Es-Ns, c) a few 2 to 1 Ns-Es.

2. One can apply the same conventions to every language, i.e., define the meaning of ":", newline, etc. once and for all as far as tagging units are concerned. (A natural choice, in a project like the ENPC, would be to adopt the conventions for English, since English is the "pivot" language.[14]) If the conventions for English differed considerably from those of some other language, however, that would result in a strange encoding from a monolingual point of view, damaging compatibility with independent

[14] But even projects where such a choice might be more far-fetched, as the German-Norwegian Parallel Corpus of the Department of Germanic Languages of the University of Oslo, and the VerbMobil project (German-Japanese), a nation-wide German project, have used English tags and English analyses as intermediary steps.

encoding of texts in the language in question.

Another choice would be to adapt the minimal set of commonalities among the languages of the project. Such a stance, in addition to not complying exactly with encoding in any of the languages, would easily run into problems if a new language were added. In addition, it would most probably result in too little clues for the tags, i.e., much of the information provided by each language would not be taken into account.

Yet another possibility would be to choose the encoding that maximized the number of 1-1 alignments. Unfortunately, this might not work for all languages involved, in addition to not being consistent with general encoding practice in any of the languages.

Examples:

* *The colon is a sentence separator only if followed by paragraph.*
* *The colon is a sentence separator except if it is immediately followed by quotes or by noncapital letters.*

Disadvantage: strange encoding of the languages other than English, which would conflict with independent encoding of the languages in question.

3. One can decide to rely on semantic criteria with a view to make the annotation spirit of the several languages similar, but with the details different from traditional wisdom in any language. In other words, use the meaning of the tags as invariant, but allow the form of punctuation or typesetting to be unconstrained.

Example:

* *In addition to s or p units one could define tags like <dsq> "direct speech quotation", <dsi> "direct speech introduction", <dsc> "direct speech comment".*

The three following texts would thus consist of a <dsi>-unit followed by a <dsq>-unit, irrespective of quotes, paragraphs and/or colons.

N: Han sa bare: “Jeg vil hjemme”.

E: He said only, “I want to go home.”

P: Ele disse apenas:
— Quero ir para casa.

Even though this would apparently be the best way to proceed, it is also the most complex: there is no a-priori off-the-shelf procedure for any language; no way for a quasi-automated SGML encoding. The alignment might become more complex, and, worst of all, there is no guarantee of real semantic equivalence.

Lots of questions remain unclear: What about, for example, free indirect speech? Should it be encoded in yet a different manner? Should thought and speech receive the same or a different treatment? What about "extended direct speech"? How to count sentence alignment? For a look at some of these questions in a corpus, see Short et al. (1996), who make it clear that it is not a trivial matter to characterize patterns of speech and thought presentation in English. As for contrastive matters, see Salkie (forthcoming), comparing speaker presence and absence in English and German. Salkie amasses substantial evidence for the claim that the questions of reported speech are complex and deserve much investigation.

In any case, I hope to have shown that it is not a trivial matter to handle punctuation when dealing with more than two languages (while for two languages one could have both a source oriented approach or a contrastive approach).

As the three following examples concisely show, there is no single choice for the interpretation of colon and of colon followed by newline that would yield in all cases one-to-one mapping for the two target languages at stake.

E: “I was also told to ask you: Will you write up a case report, including your use of Lotromycin, for publication?”
P: — Pediram-me que lhe perguntasse: Escreverá um relatório sobre o caso, incluindo o uso de Lotromycina, para publicação em alguma revista médica?
N: “Han ba meg spørre om du ville skrive en rapport om dette tilfellet, for et medisinsk tidsskrift.”

E: And a moment later: "I was about to make coffee, Miss de Grey.
P: Um momento depois, acrescentara:
— Eu ia fazer café quando chegou, Srta. de Grey.
N: Og litt etter: "Jeg skulle akkurat til å lage litt kaffe.

E: The saleswoman — he observed again that she was young, probably no more than twenty-four — tossed the raincoat onto a chair. She spoke slowly and carefully.
"Ammonia, doctor.
P: A promotora — Andrew observou novamente que ela era jovem, não devia ter mais que 24 anos — largou a capa numa cadeira. E falou devagar, incisivamente:
— Amoníaco, Doutor.
N: Nå slengte hun kåpen over en stol og sa langsomt, med vekt på hvert ord:
"Ammoniakk.

5. A multilingual corpus ?

It is thus time to reflect on the question of a multilingual parallel corpus, of the sort of the one(s) receiving currently much attention (and funding) from European quarters.

If one considers how to build one, the problems presented above show that, in order to optimize alignment (putting into correspondence) two languages, one has to make decisions that either damage alignment with other languages or damage search on the monolingual side.

But this is not the only troubling issue about a multilingual parallel corpus. Issues of its constitution should also be considered: First, while one can follow a set of particular criteria for one particular language, one has to take into account the availability of translations into other languages in order to build a parallel corpus; second, in order to obtain a representative set of translators and translations one must allow a relatively wide time frame (it is not frequent to have several alternative translations of the same work at the same time); third, the same English original, for example, translated into Norwegian and Portuguese may reflect very different standards; last but not least, the number of translations in one direction can be several orders of magnitude larger than in the opposite direction.

Proportionality between genres and modes can, again, be very much language dependent: Even though I have no data as far as this matter is concerned, I suspect that the amount of Internet publishing, the proportion of publication of newspapers vs. magazines, and the ratio poetry:prosa:drama is different in the three languages.

Moreover, even if one is able to achieve a multilingual corpus which has, say, a considerable number of books in one language translated into two other languages; i.e., even when availability is not at stake, what should a comparison of translations into the two languages show? To some extent, such a comparison might elicit a possible similarity between the two target languages — but to identify it, a much better study would be to directly contrast the two languages in question. Because they have distinct systems, direct comparison between the translations could only[15] result in statements such as: Regarding the English phenomenon of X, Portuguese is closer to English than Norwegian. Without questioning the interest that such conclusions may have for translation studies in general, I should note that very few purely linguistic conclusions could be elicited this way.

[15] "Conclusions" of the sort Norwegian Y gets translated by Portuguese Z are obviously unwarranted, for the reasons expounded on Section 1 above.

In fact, even if the goal is of a practical sort, typically foreign language teaching, where such comparisons would seek to find difficulties specific to source language learning, i.e., general difficulties for learning/teaching English, it is hard to believe that it will be easier (or more difficult) for a Portuguese to learn English if he knows that a Norwegian has the same (or does not have the same) problem. It seems to me that, no matter the problem is general to all learners of English or just those with the native language X, teaching of English must be made relevant to the native language X. It won't help to invoke a language which the speaker does not know.

In a nutshell, I question the utility of amassing multilingual parallel corpora, as opposed to a set of bilingual corpora. Would it bring us more than studying each bilingual parallel corpus in detail? My answer is a clear no.

Acknowledgements

I am grateful to Stig Johansson for recommending me for the position of visiting researcher at the Department of British and American Studies, which in turn allowed me the invaluable access to the ENPC corpus. I am also grateful for having been able to participate in the building of the Portuguese part of the corpus, and for being allowed to participate in the international meetings on Languages in Contrast organized by him and his group in Oslo. The present paper actually originates from a presentation at the Fourth Nordic Symposium on Text-based Contrastive Studies, Oslo, 25 - 27 April 1997.

I also thank Jan Engh and Dag Gundersen for information on Norwegian punctuation and style and Jan for comments on the paper.

The studies described here were conducted in the Spring of 1997 on the ENPC corpus in the form in which it was available then. A list of the sources can be found by consulting the web page of the project, http://www.hd.uib.no/enpc.html.

References

Mona Baker. "Corpus-based translation studies: The challenges that lie ahead", in Harold Somers (ed.), *Terminology, LSP and Translation: Studies in language engineering in honour of Juan C. Sager*, Amsterdam / Philadelphia: John Benjamins Publishing Company, pp.175-86.

Ted Briscoe. "Parsing (with) Punctuation etc." Rank Xerox Reseach Centre, Grenoble, MLTT-TR-007, 1994.

Ted Briscoe & John Carroll. "A probabilistic LR parser of part-of-speech and punctuation labels", Jenny Thomas & Mick Short (eds.), *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*, Longman, 1996, pp.135-66.

Christine Doran. "Punctuation in Quoted Speech", in Bernard Jones (ed.), *Proceedings of the ACL Sigparse International Meeting on Punctuation in Computational Linguistics* (Santa Cruz, September 1996), HCRC/WP2, University of Edinburgh, pp.9-18.

Jan Engh. "Normer, grammatikk og databehandling", in Ruth Vatvedt Fjeld & Boye Wangensteen (eds.), *Festskrift til Dag Gundersen*, Oslo: Universitetsforlaget, forthcoming.

Edite Estrela. *Dúvidas do Falar Português: Consultório da Língua Portuguesa*, Editorial Notícias, 1987.

Cathrine Fabricius-Hansen. "Information density and translation, with special reference to German-Norwegian-English", in Stig Johansson & Signe Oksefjell (eds.), *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*, Rodopi, forthcoming.

Jean-Jacques Franckel. *Etude de quelques marqueurs aspectuels du français*, Librairie Droz, 1989.

Gregory Grefenstette & Pasi Tapanainen. "What is a word, What is a sentence? Problems of Tokenization", Rank Xerox Research Centre, Grenoble Laboratory, April 22, 1994.

Dag Gundersen, Jan Engh & Ruth Vatvedt Fjeld. *Håndbok i norsk: skriveregler, grammatikk og språklige råd fra a til å*, Oslo: Kunnskapsforlaget, 1995.

Stig Johansson & Jarle Ebeling. "Exploring the English-Norwegian Parallel Corpus", in Carol E. Percy, Charles F. Meyer & Ian Lancashire (eds.), *Synchronic Corpus Linguistics: Papers from the 16th International Conference on English Language Research on Computerized Corpora*, Amsterdam and Atlanta, Georgia: Rodopi, 1996, pp.41-56.

Bernard E. M. Jones. "Exploring the role of punctuation in parsing natural text", *Proceedings of COLING'94* (Kyoto, August 5-9, 1994), pp.421-5.

Bernard Jones. "What's The Point? A (Computational) Theory of Punctuation", PhD thesis, University of Edinburgh, 1996.

Mary K. McCaskill. *Grammar, Punctuation, and Capitalization: A Handbook for Technical Writers and Editors*, NASA SP-7084, available from http://sti.larc.nasa.gov/html/Chapt3/.

Yoshihiko Nitta. "Idiosyncratic Gap: A Tough Problem to Structure-bound Machine Translation", *Proceedings of COLING'86* (Bonn, 25-29 August 1986), 1986, pp.107-11.

Geoffrey Nunberg. *The linguistics of punctuation*, CSLI Lecture Notes, Number 18, 1990.

Raphael Salkie. "Not mentioning the speaker in English and German", in Stig Johansson & Signe Oksefjell (eds.), *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*, Rodopi, forthcoming.

Diana Maria de Sousa Marques Pinto dos Santos. "A fase de transferência de um sistema de tradução automática do inglês para o português" ('The transfer phase of an English to Portuguese machine translation system'), MSc dissertation, Instituto Superior Técnico, Universidade Técnica de Lisboa, October 1988.

Diana Santos. "Broad-coverage machine translation", in Karen Jensen, George E. Heidorn & Stephen D. Richardson, *Natural Language Processing: The PLNLP Approach*, Boston, Dordrecht, London: Kluwer Academic Press, 1993, pp.101-18.

Diana Santos. "Bilingual alignment and tense", *Proceedings of the Second Annual Workshop on Very Large Corpora* (Kyoto, August 4th, 1994), extended version as INESC Report AR/10-94.

Diana Maria de Sousa Marques Pinto dos Santos. "Tense and aspect in English and Portuguese: a contrastive semantical study", PhD dissertation, Instituto Superior Técnico, Universidade Técnica de Lisboa, June 1996.

Diana Santos. "The importance of vagueness in translation: Examples from English into Portuguese", *Romansk Forum* 5, Juni 1997, pp.43-69.

Diana Santos. "O tradutês na literatura infantil traduzida em Portugal", *Actas do XII Encontro da Associação Portuguesa de Linguística* (Lisboa, 1-3 de Outubro de 1997).

Mick Short, Elena Semino & Jonathan Culpeper. "Using a corpus for stylistics reserach: speech and thought presentation", in Jenny Thomas & Mick Short (eds.), *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*, Longman, 1996, pp.110-131.

Yishai Tobin. *Invariance, Markedness and Distinctive Feature Analysis: A Contrastive Study of Sign Systems in English and Hebrew*, John Benjamins Publishing Company, 1994.

Jun-ichi Tsujii. "Future Directions of Machine Translation", *Proceedings of COLING'86* (Bonn, 25-29 August 1986), 1986, pp.655-68.

Jun-ichi Tsujii. "What Is a Cross-Linguistically Valid Interpretation of Discourse?", in Dan Maxwell, Klaus Schubert & A.P.M. Witkam (eds.), *New Directions in Machine Translation*, Conference Proceedings, Budapest 18-19/8/88, Foris Publishers, 1988, pp.157-66.

Kai Wikberg. "Questions in English and Norwegian: Evidence from the English Norwegian Parallel Corpus", in Carol E. Percy, Charles F. Meyer & Ian Lancashire (eds.), *Synchronic Corpus Linguistics: Papers from the 16th International Conference on English Language Research on Computerized Corpora*, Amsterdam and Atlanta, Georgia: Rodopi, 1996, pp.17-28.