Against multilinguality

Diana Santos
Linguateca, SINTEF

1. Introduction

An obvious assumption of the present workshop is that multilingual corpora are useful, and should be built and investigated. In the present paper, I would like to point out that this is far from straightforward and actually remains to be proved.

In addition, and in a more constructive vein, I want to present some examples that show that the right encoding depends crucially on what one wants to do with the corpus (using now bilingual corpora as the domain of experience), and that there is no such thing as the right general purpose (multilingual or bilingual) corpus as much as there was no way to get at a general-purpose problem solver, or a general-purpose dictionary.

The paper is structured as follows: I start with raising some problems with

a) the definition of multilinguality in general
b) what people call multilingual applications
c) the purpose of multilingual corpora
d) the encoding of multilingual corpora

And go on with the main message of the paper: there is no general purpose corpora; what you compile and encode depends on what you want to do. I try to give some concrete examples based on my work with monolingual and parallel corpora. I end the paper with the suggestion that, rather than start talking about standards, we should be looking at methodological issues and actual corpus use – hypotheses tested and actual advantages.

2. What is multilinguality?

I have argued at length in (Santos 1999) that there was no satisfactory definition of multilinguality over and above bilinguality. In the present section I repeat the main points of that paper, adding some new evidence in the subsections below.

In (Santos 1999), I contended that

- the main and only context where multilinguality made sense in itself was the one dealing with truly multilingual people, who use different languages in different spheres (allegedly a rare user community for corpus applications)
- "handling several languages well enough" is a too broad and little relevant definition (since it could be done by several monolingual agents)
- multilinguality, to be more than a mere sum of monolingual applications or behaviour, has always to involve some sort of translation (from one language to others) – if one excludes language identification or language ranking (basically "meta-natural language" applications)
- all multilingual systems are just piecewise bilingual, there is no multilinguality over and above bilinguality

This last point will be detailed in the next section, when multilingual applications are discussed. It should be born in mind, however, that while the goal of (Santos 1999) was to defend language-dependent – primarily monolingual – applications, the purpose of the present text is to propose concentrating on *bilingual* corpora (and applications) instead of multilingual ones.

2.1 What are multilingual applications?

(Santos 1999) analysed in detail several multilingual applications, such as generation and interlingual MT, to show that they were basically N-times bilingual, no matter the claims often advanced.

I will now add, in the present paper, further examples of attempts to go multilingual, which have not, in my opinion, managed to overcome the conceptual difficulty of having multilinguality adding value to what is bilinguality, or rather, involving translation between *two* languages.

Let us start with cross language information retrieval (CLIR), whose purpose is to get at information no matter in which language it is conveyed. Here again, one has a multilingual system, whose more natural architecture involves either translation of the query into several monolingual IR systems, or translation of the results into the language required by the user. The ranking of several results in different languages is then a "meta-natural language" application and, as far as I know, is being handled as a typical IR task, using no language clues.

I am, in any case, willing to accept the fact that multilingual corpora can allow a thorough evaluation of CLIR (as well as all other CL-endeavours, such as CL-summarization, CL-query answering, etc., which can be called “meta-NL” applications, since the language is itself a variable of the application). However, it has rarely been the case that the need for such applications has been motivated in the first place.

For example, to my knowledge it has never been demonstrated that there is enough different content on the Web on the same topic and in different languages in order for CLIR to be a relevant application: the CLEF experiment (Peters and Braschler 2002), although undoubtedly a major advancement in promoting evaluation campaigns in Europe and in languages other than English, may be rather artificial, since it comprises newspaper text in seven languages at the same time – where most of the “international news” would be repeated seven times and no need to find them in other languages except if one was writing a journalism research paper.[1] To measure what would be the user gain – in the line for example of evaluations such as the one performed by (Resnik 1997) for gisting – should be conceptually a first step. As we argued in (Aires and Santos 2002), to my knowledge there has been as yet no research on Web content per language (the few characterizations of Web *content* are always either in general (Broder 2002) or restricted to English (Spink et al 2002)).

2.2 What is the purpose of multilingual corpora?

One may roughly separate three main uses of corpora: research, education, and evaluation (obviously not mutually exclusive, in theory).

As to evaluation of multilingual *applications*, I have above argued that – except for CL-ones – they are non-existent (or rather, that they can be reduced to bilingual ones). *Education* in several languages at the same time is not, to my knowledge, usual, and I have never seen discussed or even proposed that it should be done simultaneously. Therefore, *research* in multilingual issues remains the only argument for creating multilingual corpora.

But the need for research pressupposes that there are truly multilingual questions that make sense to ask, and therefore try to answer. Since I myself cannot come up with any,[2] I will survey some published work on multilingual corpora, to see if they can be found.

Let us start with considering the work based on the Oslo Multilingual Corpus (OMC) (see e.g. Johansson 2002), as well as several recent books on parallel corpora (Johansson and Oksefjell 1998; Véronis 2000; Borin 2002), some of them touching the issues of multilingual corpora as well.

(Johansson 2002:47) states: “if we want to gain insight into language and translation generally, and at the same time highlight the characteristics of each language, it is desirable to extend the comparison beyond language pairs”. It is unfortunately neither evident how one is to gain this insight, using a multilingual corpus, nor how a comparison can be extended beyond language pairs. My impression is that the corpus linguistics literature in general has been too vague in explaining why and what for, even though a lot is being written about how [to compile / encode a corpus].

(Fabricius-Hansen 1998) is, to my knowledge, the first to actually use a multilingual corpus to substantiate her claims on information density, by comparing translations from German into both Norwegian and English, and trying to identify strategies of information splitting in the two language pairs. It is not clear whether the use of two different bilingual corpora (one German - Norwegian, the other German - English) would not allow her to draw the same conclusions, though.

In (Véronis 2000), there is only one paper about multilingual corpora or applications, that of Simard, discussed below; (Brown et al. 2000), although concerned with CLIR, only handles the bilingual case.

(Simard 2000) independently repeats my point that it is not obvious whether the existence of multilingual corpora can make new applications possible, and offers his paper on multilingual text alignment with the following justification: "what this chapter intends to show is that while trilingual or *multilingual text alignments may not be interesting in themselves,* any additional version of a translated text should be viewed as additional information that can and should be used to produce better bilingual alignments, and therefore a better knowledge of bilingual translational equivalences" (Simard, 2002:50, emphasis added). However, it remains to be convincingly shown that this is the case, especially since the only experiments reported concerned the Bible, which is not exactly the kind of text one would develop most applications for (or base most research on modern languages on).

As regards (Borin 2002), and though this book includes several papers featuring the fashionable multilingual label in their titles, there is not a single one which is actually concerned with more than

---

[1] German social science documents and French bibliographic references, which make up the scientific collection part, seem to me not to be multilingual (although CLIR) and thus their discussion lies outside the subject of this paper.

[2] Excluding diachronic studies comparing the evolution of languages of a same family.

parallel cases of bilingual corpora. In fact, when the authors describe multilingual corpora they simply show that their system works with several language pairs... never with language trios or quartets.

The only exception is the paper on the compilation of *bilingual* dictionaries using the Web (Grefenstette 2002), which uses a multilingual "corpus" (the Web) and might be extended to trilingual (or multilingual) dictionaries. Grefenstette shows that automatic compilation works for languages with wide Web presence (and for which there already exist bilingual dictionaries), and claims that his algorithm could be applied to language pairs for which there are as yet no bilingual dictionaries and which have little Web presence. He does not produce any evidence on this claim, though.

To me, it is hard to think of any application or human activity which could be enhanced by having a multilingual (instead of a bilingual) resource. Would a multilingual thesaurus explain better a concept to a human being? Would a multilingual dictionary help people solve their tasks in a better way? I have never seen any published (or other) evidence on that. But those, in my opinions, would be *the* valid arguments to build multilingual corpora.

If one looks at more linguistic or translation-theoretic approaches of using multilingual corpora, one is conspicuously still at the descriptive level. I am quite positive in general to exploratory studies, but, in this case, the question is – are descriptions of three or more languages in contrast (or in comparison) an asset for a language learner? a translator? a language engineer? even a linguist?[3]

2.3 Use of multilingual corpora

It is a fair point that, if one had the same text available in several languages, one might use the information more explicit in other languages to improve monolingual capabilities (as (Dagan and Itai 1994) claimed and at least for MT demonstrated, or as Simard argued as quoted above). However, in isolation this seems a weak argument to build multilingual corpora in the first place, instead of trying to improve the systems by less costly and less random ways.

That it is not automatically an advantage even for bilingual studies, or at least that we still have a long way to go until we know how to use this kind of corpora, is actually nicely illustrated by my own attempt at investigating the advantages of using a multilingual corpus for throwing some light on the differences between the prepositions *com* (Portuguese for *with*) and *med* (Norwegian for *with*), reported in (Santos, 2002).

In fact, the purpose of my study was strictly bilingual, but due to the unfortunate fact that there are no bilingual Norwegian-Portuguese corpora available, I was forced to investigate the use of trilingual (Portuguese-English-Norwegian and Norwegian-English-Portuguese) corpora to see whether they still could give me some valuable information. This incidentally fitted in nicely within a small Nordic project whose purpose was to investigate clever ways to improve CALL for the Nordic languages (Borin et al. 2001). For Portuguese-speaking learners of Norwegian as the user group, I wanted to see what could be gained through corpus-based education resources and/or research, with the help of a pivot language (unsurprisingly, English).

So, in addition to use (somewhat) comparable corpora in the two languages;[4] I turned to multilingual corpora to find out why the use of the prepositions *com* and *med* (both standardly corresponding to English *with*) was so different in Portuguese and Norwegian. First, I tried to use the OMC to look at the distribution patterns of *with* into *com* and *med* or not (given that this corpus has some English source texts translated into the two languages), to see whether there was agreement or some obvious differences between the translations into the two languages. This was, unfortunately, not possible, because, although the OMC is a multilingual corpus with translations into six languages, it did not offer a multilingual query capacity for the language trio I was interested with, and so I was forced to look independently at the translations involving *with* and/or *com* and *with* and/or *med* for each language pair.

So, I had to use a "fictive" trilingual corpus composed of the ENPC (Johansson et al. 1999) and COMPARA (Frankenberg-Garcia and Santos 2000), thus Norwegian - English and English - Portuguese, respectively, to compare the use of these two prepositions. The results,[5] displayed in Tables 1 and 2, show that there is considerable deviation between the use of English *with* and each of

---

[3] In my opinion, linguistics works that are based in a bunch of languages of which the linguist has reduced competence, or is dependent on informants, are a curse for linguistics as a scientific endeavour, but this is probably too contentious a point to be discussed here (not the right forum, anyway).

[4] The data and results can be found in Santos (2002).

[5] The data presented here concerns COMPARA in a version below 1.6 (April 2002). At the time of writing the present paper, COMPARA, in version 3.2, features 25 text pairs as opposed to the 16 of that time. COMPARA is a collaboration project with Ana Frankenberg's team. Access to COMPARA is available from http://www.linguateca.pt/COMPARA.

these prepositions, but could not shed any further light about the differences between the prepositions themselves.

Table 1. Relationship between *with* and *com* in COMPARA and ENPC (En-Po)

Pattern	COMPARA	ENPC (En-Po)	Total
Com in original text	757		
Com in translation from English	1657	2313	3970
With in original text	1063	1585	2648
With in translation from Portuguese	882		
Com translated by *with*	480 (63.4%)		
Com not translated by *with*	277 (36.6%)		
With translated by *com*	804 (75.6%)	1144 (72.2%)	1948 (73.5%)
With not translated by *com*	259 (24.0%)	441 (27.8%)	700 (26.4%)
Com in translation not corresponding to *with*	751 (45.3%)	1028 (44.4%)	1779 (44.8%)
With in translation not corresponding to *com*	387 (43.9%)		

Table 2. Relationship between *with* and *med* in ENPC (En-No-En) (Fiction and Non-Fiction)

Pattern	Fiction	Non Fiction	Total ENPC
Med in original text	4799	2431	7230
Med in translation from English	5319	2903	8222
With in original text	3258	1799	5057
With in translation from Norwegian	3557	1782	5339
Med translated by *with*	2678 (58.0%)	1280 (52.6%)	3958 (54.7%)
Med not translated by *with*	2014 (42.0%)	1151 (47.3%)	3165 (43.8%)
With translated by *med*	2338 (71.7%)	1280 (71.1%)	3618 (71.5%)
With not translated by *med*	920 (28.2%)	511 (28.4%)	1431 (28.2%)
With in translation not corresponding to *med*	879 (24.7%)	567 (31.8%)	1446 (27.0%)
Med in translation not corresponding to *with*	2761 (51.9%)	1461 (50.3%)	4222 (51.3%)

These raw data show that *med* is more frequent than both *with* and *com*, and that a sizeable number of occurrences of these two prepositions in translated text does not correspond to *with* (44-45% for *com* and 50-52% for *med*). If one looks more closely at what these mismatches correspond some results are displayed for the English-Portuguese pair in Tables 3, 4, 5 and 6. (Codes for tables 4-6, all concerning COMPARA, are explained in Table 3.)

Table 3. Kinds of mismatch when *com* and not *with*: CO - COMPARA

Kind / Code	translation from E->P (random in 5 pairs CO)		Translation from E->P (all in 2.5 pairs, ENPC)	Translation from P to E (random in 10 pairs CO)	Total	
Different preposition/P	61	30.5%	21	27	109	27%
Direct object vs. PP(com)/V	29	14.5%	16	13	58	14.5%
Adverb/Adv	29	14.5%	15	18	60	15.5%
Adjective/Adj	21	11.5%	10	3	34	8.5%
Verb vs verb+PP(com)/VV	13	6.6%	6	3	22	5.3%
Verb vs. PP(com)/vvv	4	2%	3	4	11	2.8%
Discourse/d	12	6%	4	7	23	5.8%
and vs. *com*/and	2	1%		4	6	1.5%
Reordering/R	29	14.5%	25	21	75	18.8%
» explicit args	2		6			
» head swithching	2		6	7		
» *parecido* com			2			
Total	200		100	100	400	

Table 4. Kinds of mismatch when *com* and not *with*, Portuguese translated into English

Pair	UA's	Words	*com*	P	V	Adv	Adj	VV	vvv	d	and	R	Total	% of *com*
PBJAT1	219	3470	17		1					2			3	18%
PBJAT1	219	3470	17					1		1		3	5	29%

PBMA1	841	12614	83	5	9	5	1			1		7	28	34%
PBMA2	491	11554	81	7	2	5	1	1		3		3	22	27%
PBMA3	661	11021	66	4	1	6	1	1		1	1	9	24	36%
PBPC1	772	10775	97	10	9	2	5		1	1		22	50	52%
PBPM1	1025	14036	100	16	8	2	1	1	1		6	5	40	40%
PPEQ1	303	6754	46	4	1	3	5		5	5		2	25	54%
PPMC1	1394	23499	217	17	8	17	2		1	4	2	12	63	29%
PPSC1	615	9440	33	6	3	1			1			5	16	48%
Total	6540		**757**	69	42	41	16	4	9	18	9	68	**276**	**36%**
% mism				25	15	15	5.8	1.4	3.3	6.5	3.3	25		

Table 5. Kinds of mismatch when *with* and not *com*, Portuguese translated into English

Pair	*with*	P	V	Adv	Adj	VV	vvv	D	and	R	*sem*	N	Total	% of *with*
PBJAT1	27	2	1					1		3			7	26%
PBJAT1	21	1	1		1			1		2			6	28%
PBMA1	97	19	5		2	1	1	1	3	5	2		39	40%
PBMA2	105	19	3	2		2				9	1	1	37	35%
PBMA3	97	17	3			2		7	1	13	2		45	46%
PBPC1	69	8	3						1	9	1		22	32%
PBPM1	89	9	2		4	2		5		3	1		27	30%
PPEQ1	59	15	2	1	1	1				6		1	27	46%
PPMC1	249	60	12	3	6		2	1	1	7	1	1	94	38%
PPSC1	69	25	4	1	5	1	1	1		10	1	1	50	72%
Total	**882**	175	36	6	19	9	4	17	6	67	9	4	**350**	**40%**
% mism.		50	10	1.7	5.4	2.5	1.1	4.9	1.7	19	2.5	1.1		

Table 6. Kinds of mismatch when *with* and not *com*, English translated into Portuguese

Pair	*with*	P	V	Adv	Adj	VV	vvv	D	*e (and)*	R	*sem*	N/pron	Total	% of *with*
EBDL1T1	301	40	6		2	6	1	1	3	19	2	2	82	27%
EBDL1T2	301	27	6		1	2		1	1	22	1	1	62	21%
EBJB1	130	17	4			2		1		3	1	1	29	22%
EBJT1	211	20	3	1	4	3				9	4		45	21%
ESNG1	120	6	5		2	2				1	1		17	14%
Total	**1063**	110	24	1	9	15	1	4	4	64	9	4	**235**	**22%**
% mism.		47	10	.42	3.8	6.3	.42	1.7	1.7	27	3.8	1.7		

Even though a considerable number of cases seems to be lexically determined, it is interesting to note that the correspondence pattern of presence/absence of the prepositions *com* and *with* seems to be rather common, irrespective of the direction of translation and, to a lesser degree, of actual preposition.

My conclusion, as far as a multilingual corpus (of this kind) is concerned, is that interesting bilingual findings can be read for each directly related/translated pair of languages,[6] but no generalization is possible to three languages (and thus to two not directly related/translated, which was my initial aim). In other words, the sum of two bilingual experiments and contrasts did not illuminate the particular issue I was preoccupied with (the missing bilingual question). That is: a trilingual corpus may be *less* than a bilingual corpus.

3. Encoding of multilingual corpora

But not only it is hard to find cases where it pays to deal simultaneously with more than two languages, it may be that such a requirement may pose unexpected challenges. In (Santos 1998), I described in considerable detail encoding questions related to direct speech (and, in particular, punctuation issues) in a multilingual corpus including Portuguese, English and Norwegian, pressing the conclusion that there was simply *no* way to solve the problem without letting one language dictate the conventions.

[6] Even though I have no opportunity to dwell on them in the present text.

In that paper, I claimed that there was no way to give an uniform treatment of the semi-colon and be faithful to the three languages (in the sense of using the language conventions of each for encoding) and then have a similar alignment procedure for the three pairs.

In other words, alignment (and, more generally, to put things in correspondence) requires having a theory of what is common, and there may be no consistent theory if one has three languages. One can always find one theory with two languages, for example by using translation practice, or forcing source (or target) interpretation to dictate the structure; but in the context of three languages there are simply cases where one cannot do that.

In COMPARA, which is bilingual, the choice taken was to let originals define sentence alignment, which is always one source sentence to (parts or numbers of) target sentences. We maintained a language-specific version of sentence separation in the two languages, as far as there are such things as a standard English sentence and a standard Portuguese sentence (see below), but we do not have a standard notion of sentence (independent of language), which means that the two parts E to P and P to E may differ subtly.

Before we leave the topic of encoding, a final note is in place. Although I claimed in (Santos 1998b) that, conceptually, the following three things should be kept separate: corpus proper, corpus encoding system and corpus interface (or use); for all practical purposes, if you cannot use a corpus with a multilingual query capability (as it was the case with the OMC for the trio English-Norwegian-Portuguese[7]) you might as well state that you were not conversing with a multilingual corpus, but simply with two bilingual ones.

So, what is a multilingual corpus in terms of minimum query capabilities might also be a pertinent question, not always made crystal clear.

4. What the corpus user wants the corpus for is determinant

In any case, one important thing – probably the single most relevant one – is what one wants the (multilingual) corpus for, and this is actually the main message conveyed in the present paper.

Do we want a multilingual corpus for evaluation of cross-language applications? Do we want a multilingual corpus to automatically create lexicographic multilingual resources? Do we want a multilingual corpus to study more than two languages at the same time?

Different wishes call for radically different encoding procedures and choices, as illustrated, for the treebank case, in (Santos 2003). Let me present very simple, one might even call trivial, examples related to monolingual or parallel corpora.

4.1 Words

First of all, for most of the units that could be devised in any of the languages, there is never a precise and unambiguous definition, even in one language alone. Let me recall (Grefenstette and Tapanainen 1994)'s "What is a word? What is a sentence?" questions dealing with English but equally pertinent if translated into any language that has words about basic units of language...

So, I can cite on this respect, with Portuguese as the object language, that (Santos and Bick 2000) measured as much as 14% tokenization differences among two different systems for Portuguese. And this was not an isolated property of a particular system, as was painfully illustrated in (Santos et al. 2003), where five different morphological analysers cut the tokenization pie in widely different ways.

Assuming, now, that morphological analysers have some understanding of the notion of word (or of what is the basic language unit that can be classified), this implies that parallel corpora compiled by each of these groups would feature grossly different characteristics (and we might as well expect different word alignment results, and perhaps significant differences in every lexical study performed).

4.2 Sentences

If the word is tricky to define and difficult to agree upon, the sentence does not fare any better. In fact, sentence separation programs (see our Web page at the AC/DC project site[8] for a detailed description of the basic algorithm) were revised differently for two different projects with different goals. This led to the (maybe surprising) situation that the same institution, Linguateca[9], displays and serves three different "sentence flavours", depending on the kind of application that the corpora are built for:

[7] This is no technical shortcoming of the OMC, since this possibility exists for other language trios. Simply at the time of the research reported here it was not implemented for Portuguese.
[8] http://acdc.linguateca.pt/acesso/atomizacao.html
[9] Linguateca is a publically funded virtual organization for developing free and publically available resources for the Portuguese language, with three main lines of action: maintaining a Web portal for

1. Namely, for the AC/DC project (Santos and Bick 2000), a sentence defines a linguistically meaningful monolingual concordance unit;

2. For the Floresta Sintá(c)tica treebank (Afonso *et al.* 2002), a sentence attempts to mirror something as close to a linguist's definition of sentence for purposes of syntactic structure as possible;

3. For the English-Portuguese parallel corpus COMPARA (Frankenberg-Garcia and Santos 2000), a Portuguese sentence was chosen as what gives the best results in terms of aligning with English. In fact, sometimes, in cases where there was no general consensus, we took some radical choices, such as not considering (in some cases) the colon as a sentence separator for Portuguese.

Summing up, the differences are related to the pertinent notion of sentence in each application, given that the notion of sentence plays a different role in each of the projects.

4.3 Alignment

The previous example can be multiplied at will for almost every thing one encodes in parallel corpora.

Let us take alignment issues next: how to evaluate an aligner (or the alignment result thereby produced) depends on what the user wants to do with that alignment: Should it reflect the extent of the differences among the two languages, or should it actually put into correspondence more or less the same "content" independent of form?

Note that this would give very different encoding strategies to the seemingly simple bi-texts you can query in COMPARA that involve the *reordering* of alignment units. One can separate the correspondence itself from the ordering in the text (as we did), but it is doubtful how much of this needs to be encoded or actually falls under the label "alignment" as generally understood.

4.4 Word correspondence

Word correspondence issues: Again, how to measure a word aligner depends on what its application is: Was it designed to extract correct translation pairs, to detect translationese, to find reliable clues for sentence alignment, or to be used in CLIR?

If your goal is to find correspondence in context, you should be able to "align" rather different words; if your purpose is to create a bilingual dictionary, you should beware of creative translation.

4.5 Clause correspondence

Clause correspondence, a task in between the two previous ones, concerning clauses, which are considered by some the most meaningful unit to work with, can also be shown to depend widely on what is intended.

In fact, identifying clauses in different languages is dependent on one's wish to e.g. a) identify simple translations for MT evaluation; b) create a bilingual valency dictionary; or c) study particular translation shifts and/or performance.

While the "chassé-croisé"[10] described in (Santos 1996) in a contrastive study requires a fine-grained notion of clause, it is of no consequence for the evaluation of translation systems (since it is always the union of the two clauses which should be translated by two new clauses).

And, while the existence of higher perception or saying verbs is irrelevant for the purpose of aligning the object clauses to create a bilingual dictionary, it exceedingly matters for discourse studies and the automatic verification of translations.

4.6 Semantic studies

Take now a rather different marking of parallel corpora, namely aspectual encoding, with the purpose of e.g. identify aspectual gaps (Santos 1998c). In order to perform such study, one would need to have explicit annotation of English and Portuguese aspectual classes (which, as I argued at length in (Santos 1996), are different).

If, on the other hand, one used the same labels for both languages, one would be, at most, able to (and interested in) measuring surface similarity in translation.

NLP of Portuguese, developing and improving Portuguese language resources, and organizing evaluation contests for Portuguese. It was created as a follow-up of the Computational Processing of Portuguese project, and has currently nodes in Oslo, Braga, Lisbon and Oporto. Cf. www.linguateca.pt.

[10] This term was coined in (Vinay and Darbelnet, 1977:105) to describe a switch of verbs in main and adjunct clauses very common in translation. See (Santos 1996:248ff) for ample illustration.

5. Concluding remarks

It seems too early to start discussing standards when there is not yet any given practice of using (and therefore being able to assess) multilingual corpora. In fact, I hope to have shown in this paper that, prior to how, the question of *why* build multilingual corpora should be addressed.

The question of what should be encoded (paid attention to, documented) should be crucial to every project, but different projects have different goals, which makes the idea of achieving a standard as useless as impractical. In fact, basic methodological issues -- how and what proceed to a given goal given a corpus, how to validate results and so on -- are far from settled in our discipline. We should devote more time to the semantics of what we do, rather than to its syntax (how to wrap it).

Another consideration that seems relevant when talking about standards is the existence of two kinds of corpora: those too big to be humanly revised and edited, but which are required for statistical processing, and those into which creation a considerable care is put. These two kinds of corpora are often complementary, and are typically used by different kinds of researchers, respectively language engineers, and linguists. It has rarely been the case that the same kind of encoding is used for both kinds of projects, even though superficially they may use the same labels. So, apparent use of same "standard" labels may in fact be even misleading.

Even in the cases where one needs multilingual corpora to evaluate multilingual applications (of the kind called above “meta-NL” applications), the NLP and linguistics community might be looking at Web IR instead of creating corpora from the scratch. Sampling the Web seems to be the most effective way of getting a truly multilingual parallel corpus, which in addition is realistic in the sense that reflects language use and a large field of application.

Using (Gaizauskas 1998)’s terminology, I suggest, furthermore, that one should perform user-visible evaluation of (multilingual) corpora instead of arguing and agreeing on user-transparent criteria. And the same for all application areas mentioned above (such as alignment, semantics or clause correspondence).

Concluding, this paper basically argues against creating multilingual corpora and encoding standards for them, when so many interesting and challenging unsolved questions, as well as useful applications, are to be found in bilingual (and monolingual) corpora. It uses two kinds of arguments:

1. that there is no need for multilingual corpora in the first place (there is no multilinguality above bilinguality)
2. that corpora have to be created with a purpose, and different purposes usually require different encoding (decisions)

A third practical argument would be that there are several more relevant kinds of parallel corpora that should be first taken into account, such as sound and text; multiple paraphrases; multiple translations (not necessarily or especially into different languages); multiple revisions of a "same" text; multiple corrections or correction suggestions of a non-native text; multiple versions of a same event (as described by different witnesses), etc. etc. All of these extend the standard parallel case (translations or comparable content, in two languages) in a realistic situation, without entering the "virtual" world of dealing with three or more languages at the same time.

My plea is to measure and evaluate the (application) contexts[11] where it would be useful to have (or process an existing) multilingual corpus, prior to creating corpora or trying to agree on standards for their encoding.

6. Acknowledgements

I am grateful to the two reviewers, to one for encouragement and to the other for a detailed critical discussion of a previous version of the paper. I thank Stig Johansson, and Jarle and Signe Ebeling, for granting me access to the ENPC and the OMC, and for critical comments on the preposition contrast at an early presentation, but most of all for having included Portuguese in the corpus in the first place. Some of the ideas in this paper have been previously presented at the University of Oslo, and I am grateful to Kjell Johann Sæbø for the invitation, and to Stig Johansson for feedback.

---

[11] One reviewer noted that I was mixing applications and research all over the paper. I do it on purpose, since I believe every research must have some goal. (The goal may be to improve an application, or answer some question(s) that are relevant for theory). And for all purposes discussed here, it was not relevant which kind of (meta)goal the corpus compiler has in mind.

References

Afonso Susana, Bick Eckhard, Haber Renato, Santos Diana 2002 "Floresta sintá(c)tica": a treebank for Portuguese. In *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain, pp 1698-1703.

Aires Rachel, Santos Diana 2002 Measuring the Web in Portuguese. In Matthews Brian, Hopgood Bob, Wilson Michael (eds), *Euroweb 2002 conference*, Oxford, UK, pp 198-9.

Borin Lars (ed.) 2002 *Parallel Corpora, Parallel Worlds*. Amsterdam and New York, Rodopi.

Borin Lars, Carlson Lauri, Santos Diana 2001 Corpus based language technology for computer-assisted learning of Nordic languages: Squirrel. Progress Report September 2001. In Holmboe Henrik (ed), *Nordisk sprogteknolog. Nordic language technology. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*, København: Museum Tusculanums Forlag, Københavns Universitet, pp 257-270.

Broder Andrei 2002 A Taxonomy of Web Search. *SIGIR Forum* **36**(2): 3-10.

Brown Ralf D, Carbonell Jaime G., Yang Yiming 2000 Automatic dictionary extraction for cross-language information retrieval. In Véronis Jean (ed), *Parallel Text Processing*, Dordrecht, Kluwer Academic Publishers, pp 275-98.

Dagan Ido, Itai Alon 1994 Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Computational Linguistics* **20** (4): 563-96.

Fabricius-Hansen Cathrine 1998 Information density and translation, with special reference to German-Norwegian-English. In Johansson Stig, Oksefjell Signe (eds), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*, Amsterdam & Atlanta, Rodopi, pp 197-234.

Frankenberg-Garcia Ana, Santos Diana 2000 Introducing COMPARA, the Portuguese-English parallel translation corpus. Presented at *CULT'2000*, to appear in Bernardini S., Zanettin F., Stweart D. (eds) *Corpora in translator education* (provisional title) Manchester, St. Jerome.

Gaizauskas Robert 1998 Evaluation in language and speech technology. *Computer Speech and Language* **12** (4): 249-262.

Grefenstette Gregory, Tapanainen Pasi 1994 What is a word, What is a sentence? Problems of Tokenization. In *Proceedings of the 3rd International Conference on Computational Lexicography (COMPLEX'94)*, Budapest, pp 79-87.

Grefenstette Gregory 2002 Multilingual corpus-based extraction and the Very Large Lexicon. In Borin Lars (ed) *Parallel Corpora, Parallel Worlds*. Amsterdam and New York, Rodopi, pp 137-49.

Johansson Stig, Oksefjell Signe (eds) 1998 *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*, Amsterdam & Atlanta, Rodopi.

Johansson Stig, Ebeling Jarle, Oksefjell Signe 1999 English-Norwegian Parallel Corpus: Manual. Oslo: Department of British and American Studies, University of Oslo, 1999/2002, available from http://www.hf.uio.no/iba/prosjekt/ENPCmanual.html [accessed 14 March 2003].

Johansson Stig 2002 Towards a multilingual corpus for contrastive analysis and translation studies. In Borin Lars (ed.) *Parallel Corpora, Parallel Worlds*. Amsterdam and New York, Rodopi, pp 47-59.

Peters Carol, Braschler Martin 2002 The Importance of Evaluation for Cross-Language System Development: the CLEF Experience. In *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain, Vol I, pp 122-7.

Resnik Philip 1997 Evaluating Multilingual Gisting of Web Pages. In *Natural Language Processing for the World Wide Web, Papers from the 1997 AAAI Spring Symposium* (Stanford, March 24-26, 1997), Menlo Park, California: AAAI Press, pp 129-135.

Santos Diana Maria de Sousa Marques Pinto dos 1996. *Tense and aspect in English and Portuguese: a contrastive semantical study*. Unpublished PhD dissertation, Instituto Superior Técnico, Lisbon, Portugal, June 1996.

Santos Diana 1998 Punctuation and multilinguality: Reflections from a language engineering perspective. In Ydstie Jo Terje, Wollebæk Anne C (eds), *Working Papers in Applied Linguistics* 4/98, Oslo, Department of Linguistics, Faculty of Arts, University of Oslo, pp 138-60.

Santos Diana 1998b Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts. In *Proceedings of The First International Conference on Language Resources and Evaluation*, Granada, 28-30 May 1998, Vol. 1, pp 475-81.

Santos Diana 1998c Perception verbs in English and Portuguese. In Johansson Stig, Oksefjell Signe (eds), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*, Amsterdam: Rodopi, pp 319-342.

Santos Diana 1999 Toward Language-specific Applications. *Machine Translation* **14** (2): 83-112.

Santos Diana, Bick Eckhard 2000 Providing Internet access to Portuguese corpora: the AC/DC project. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000*, Athens, pp 205-210.

Santos Diana 2002 *Med* e *com*: um estudo contrastivo português – norueguês. *Romansk Forum* **16** (2): 1029-1042.

Santos Diana 2003 Timber! Issues in treebank building and use. To appear in *Proceedings of PROPOR 2003* (Faro, 26-27 June 2003).

Santos Diana, Costa Luís, Rocha Paulo 2003 Cooperatively evaluating Portuguese morphology. To appear in *Proceedings of PROPOR 2003* (Faro, 26-27 June 2003).

Simard Michel 2000 Multilingual text alignment: Aligning three or more versions of a text. In Véronis Jean (ed), *Parallel Text Processing*, Dordrecht, Kluwer Academic Publishers, pp 49-67.

Spink Amanda, Ozmutlu Seda, Ozmutlu Huseyin C., Jansen Bernard J. 2002 U.S. versus European Web searching trends. *SIGIR Forum* **36**(2): 32-8.

Véronis Jean (ed) 2000 *Parallel Text Processing*, Dordrecht, Kluwer Academic Publishers.

Vinay, J.-P., Darbelnet J. 1977 *Stylistique Comparée du Français et de l'Anglais: Méthode de traduction*, Didier, Paris. Nouvelle édition révue et corrigée.