Alignment and tagging

Lars Borin

Department of Linguistics, Uppsala University

Abstract

It is sometimes said that part of speech (POS) tags are likely to be the same for translation equivalent words. If this is correct, we could formulate the following hypothesis: It should be possible to use POS tagging for one language in combination with a word alignment system, in order to obtain a (partial) POS tagging for another language. This hypothesis is investigated both empirically—an experiment is described where POS tags were transferred from a POS tagged German text to a parallel Swedish text by automatic word alignment—and theoretically, in the form of a review of relevant linguistic work on the typology of POS systems. The conclusions are that the hypothesis seems to hold at least for closely related languages, that the findings of typological research do not contradict it (or a slightly modified form of it), but that further empirical research is needed.

1. Introduction

Is it a reasonable assumption, as made, e.g., by Melamed (1995:7) "that word pairs that are good translations of each other are likely to be the same parts of speech in their respective languages"? Sågvall Hein (p.c.) has made a similar observation based on the investigation of one-word sentence (fragment) alignments in one of the ETAP and PLUG project subcorpora, the Scania corpus (see Sågvall Hein this volume).[1]

If this assumption is correct about the relationship of part of speech (POS) labels, or tags, between the source language (SL) and target language (TL) texts, it could be used to advantage in parallel corpus linguistics, since in the case that we are in the possession of

1. a POS tagger for one language (the SL),
2. a set of parallel SL–TL texts, i.e., a parallel SL–TL corpus, and
3. an alignment algorithm for SL–TL word alignment (for this particular SL–TL pair or for general word alignment of any two languages),

we could formulate the following hypothesis: It should be possible to use the SL POS tagger in combination with the word alignment algorithm in order to obtain a (partial) POS tagging of the TL. The main advantage accruing from this would be the possibility of achieving an initial word class tagging of a text in a language for which no POS taggers are available. This initial POS tagging could then be

refined using methods which have been suggested in the literature (e.g., Màrquez *et al.* 1998; Borin to appear)

From a purely linguistic standpoint, there is reason to doubt that this assumption holds for the general case of any language compared with any other language, and for any part of speech. We will return to this question in section 3, where we review the linguistic literature on parts of speech in a cross-linguistic perspective.

Even though not universally valid, one might entertain the hypothesis that the assumption is more likely to hold for languages which either are closely related genetically—like Swedish and English—or have been in contact for a long time—as in the case of Swedish and Finnish. In order to test this hypothesis, we performed an experiment with the language pair Swedish–German. This experiment is described in section 2.

But if the languages are not close in the sense just mentioned, and even if they are, it is conceivable that not all parts of speech are equally likely to remain invariant when translating from one language to the other. If we could determine under what circumstances this is likely to be the case—or, alternatively, could formulate rules for how parts of speech are translated in those cases when they are not preserved, which would amount to a weaker, but no less useful, version of the initial hypothesis— we would still be able to transfer POS tags from the SL to the TL via links established by a word alignment algorithm. We will look into this matter more closely in section 3 below.

In order to test these hypotheses, one should test them with many language pairs, correlating the results with the degree of relatedness among the languages and the various parts of speech. Here, we make a start in this direction by investigating the language pair Swedish–German.

2. An experiment with POS tagging by word alignment

We made an experiment with POS tagging by word alignment on the language pair Swedish–German, as follows.

First, a Swedish–German parallel text was word aligned with a word alignment tool developed in our department (Tiedemann 1998, this volume, to appear) in the PLUG project (Sågvall Hein this volume). The text was one the ETAP and PLUG Swedish Government Policy Declarations (SGP) text pairs (see Sågvall Hein this volume). The alignment system first performs a sentence alignment with the method described by Gale and Church (1993), and then carries out word (and phrase) alignment within each sentence alignment unit, using a variety of linguistic and statistical information sources. The recall and precision of the word alignment were calculated by the use of a standard produced with the PLUG Link Annotator (Merkel *et al*. this volume), and were found to be: recall 39.76%

(46.39%, if we include partly correct alignments, i.e. part of a multi-word unit has been aligned, but not all of it), precision 77.95% (90.94% including partly correct alignments).

We see that comparatively few words are aligned; 40% is much below what a typical sentence alignment algorithm is capable of achieving, which is close to 100%, at least for this language pair (see Borin this volume). This is a partly due to the fact that word alignment is a much harder problem than sentence alignment, but partly also reflects a cautious approach to word alignment built into the word alignment program used (see Tiedemann to appear). The reward for this cautiousness is high alignment precision. Thus, most of the aligned SL words are correctly linked to their equivalents on the TL side.

The German text was POS tagged with Morphy, a freely available German morphological analyser and POS tagger (Lezius et al. 1998).[2]

For every German word–tag combination, if there was a word alignment with a Swedish word, that word was manually assigned the SUC tag (Ejerhed and Källgren 1997) most closely corresponding to the POS tag of the German word.

In Table 1, the resulting word alignments and their POS tags are shown for two sentence alignment units.

Table 1: Some Swedish–German word alignments in the ETAP SGP subcorpus, and their corresponding part-of-speech (POS) tags (a '*' marks bad tag correspondences).[3]

sentence alignment unit ID			
SUC POS	**Swedish token**	**German token**	**Morphy POS**
svdeprf83			
NN SIN	Industrins	Industrie	Industrie SUB GEN SIN FEM
NN SIN	anpassning	Anpassung	Anpassung SUB NOM SIN FEM
NN *SIN/PLU	krav	Anforderungen	Anforderung SUB AKK PLU FEM
KN	och	und	und KON NEB
NN PLU	processer	Prozesse	Prozeß SUB NOM PLU MAS
NN PLU	produkter	Produkte	Produkt SUB DAT SIN NEU
JJ	renare	reiner	rein ADJ ADV
VB	skall	sollen	sollen VER MOD 3 PLU
svdeprf102			
NN SIN	Livsmedelskontrollen	Nahrungsmittelkontrolle	Nahrungsmittelkontrolle SUB NOM SIN FEM
*VB	skärps	verschärft	verschärfen VER PA2

The accuracy of the Swedish POS tags assigned in the previous step was assessed manually in a subset of the aligned sentences (10 randomly selected sentence alignment units, containing 16 SL sentences). The results are shown in Table 2:

Table 2: Accuracy of Swedish POS tags assigned by word alignment

Sentences	*Aligned units (excl. punctuation)*			
16	78			
alignments	*correct*		*incorrect*	
	64 (82%)		14 (18%)	
	same	*different*	*same*	*different*
main category	61 (95%)	3 (5%)	1 (8%)	13 (92%)
NN subcategory number	27 (93%)	2 (7%)		

It turned out that only the major POS category (Noun, Verb, Adjective, etc.) was relevant for the comparison, since subcategories (Number, Case, Person, etc.) were generally not applicable even across such a comparatively short cross-lingual distance as that between German and Swedish. Hence, the table shows major category correspondences, with one exception, namely the NN (Morphy: SUB) subcategory *number* (7 PLU, 22 SIN in the text), where, contrary to what we just said, it turned out to be meaningful to compare the values, and where the German value turned out to be correct for the Swedish correspondence 27 times out of 29.

We see that for the correct alignments, the German tag is generally the correct one for the Swedish correspondence (in 95% of the cases), while the proportions are reversed for the incorrect alignments. This means that—at least for this language pair and this text type—POS tagging of the SL and word alignment can be used to accomplish a partial POS tagging of the TL, but also adds support to Melamed's (1995) claim that a "POS filter" is a good method for weeding out bad word alignment candidates, i.e. if we perform a word alignment on a parallel text where *both* language versions have been POS tagged, we should disfavour those alignment candidates whose POS tags do not coincide.

3. Results and discussion

We may suspect that the fairly promising results presented in the previous section are mainly due to the circumstance that Swedish and German are closely related languages, and that the situation would change if the languages involved were more dissimilar.[4]

This suspicion is strengthened if we look at some other language pairs in the ETAP corpus material. In examples 1–6 below, we give some translation

equivalents picked more or less at random in the parallel five-language ETAP IVT1 corpus (see Borin this volume). The intended correspondences are underlined in the examples, and their part of speech and other morphosyntactic information are provided at the end of each example.

(1) SE: Att flytta ut tunga myndigheter till Rinkeby, Tensta och Skärholmen är en idé som ligger i tiden. [VB INF + PRL]

PL: Przeprowadzka głównych urzędów do Rinkeby, Tensta i Skärholmen to pomysł na czasie. [NN FEM NOM SIN]

EN: Moving important public agencies to places like Rinkeby, Tensta and Skärholmen is an idea that is currently gaining ground. [VB GR]

(2) SE: Det är en följd av att Sverige skrivit under Schengen-avtalet om passamarbete mellan flera europeiska länder. [VB SUP + PRL]

PL: Takie jest następstwo podpisania przez Szwecję układu z Schengen o współpracy paszportowej między wieloma krajami europejskimi. [NVL NEU GEN SIN]

EN: This is one result of Sweden signing the Schengen Agreement on passport collaboration between several European countries. [VB GR]

(3) SE: Experterna tror på ökad tillväxt, fortsatt låga räntor och mer köpkraft för löntagarna. [PN]

FI: Asiantuntijat uskovat kasvun lisääntyvän, korkojen pysyvän alhaisina ja palkansaajien ostovoiman lisääntyvän. [VB ACT PR PTC GEN SIN]

EN: The experts are forecasting increased growth, low interest rates and greater purchasing power for wage-earners. [JJ]

(4) SE: För att locka resenärer sänker SJ biljettpriserna under våren. [VB INF]

FI: Matkustajien houkuttelemiseksi SJ alentaa lippujen hintoja kevään aikana. [NVL TRV SIN]

EN: To attract passengers, Swedish Rail will be reducing ticket prices in the spring. [VB INF]

(5) SE: De ska öva sig att tala svenska i studiecirklar [VB INF]

FI: He saavat harjoitella ruotsin puhumista opintopiireissä [NVL PTV SIN]

EN: They will practise speaking Swedish in study circles [VB GR]

(6) SE: Allt för många lämnar skolan utan att vara godkända. [VB PR ACT]

PL: Coraz więcej uczniów ryzykuje ukonczenie szkoły bez oceny dostatecznej. [NVL NEU ACC SIN]

EN: Far too many students face leaving school without pass grades. [VB GR]

We see that there seems to be less agreement in POS tags among these languages, which are still fairly similar as seen against the linguistic diversity in the world at large; all but one are Indo-European, and as we have already mentioned, that one—Finnish—has a long history of contact with Indo-European languages, which are known to have exerted profound influence on its vocabulary and structure (Hakulinen 1979).

Even if there are less *direct* POS correspondences—in the sense of a verb in language A always corresponding to a verb in language B, and the same for other parts of speech—between these and other languages, it is still conceivable that there may be *regular* correspondences, so that it would be possible to formulate linguistically motivated POS correspondence rules for a particular language pair.

In principle, such correspondence rules may be of two kinds:

1. universal rules (or universal tendencies), holding for all language pairs (or more likely: for all language pairs of a certain type, definable in linguistic terms);
2. those holding for a particular language pair only.

At least the second kind of rules can be found only by empirical investigation of a number of language pairs in a fashion similar to that described in section 2.

For the first kind of POS correspondence rules, we will now turn to the literature on language universals and linguistic typology as the place where we might find some research results bearing upon the issue of their existence and form.

The traditional part of speech inventory, a more fine-grained version of which makes up most POS tagsets, as well as the pre-terminal vocabulary of typical context-free phrase structure grammars, ultimately traces its heritage back to the Greek and Latin grammatical traditions (Jespersen 1924:58f; Vonen 1997, ch. 2). Even modern, heavily formalised grammatical frameworks, such as Generalized Phrase Structure Grammar (GPSG: Gazdar et al. 1985) and Head-Driven Phrase Structure Grammar (HPSG: Pollard and Sag 1994), and less formal, but still characterisable as formalistic, frameworks such as the successive versions of Generative Grammar (e.g., Radford 1988) tend to take this traditional part of speech inventory as primitive (i.e., given) categories of grammar, probably partly because the interest of the linguists developing these formalisms have lain elsewhere (in teasing out intricate problems of syntax), but possibly partly also simply because this inventory has stood the test of time and still represents “the most useful approach to linguistic categories” (Ramat 1999:173). The only real innovation in this area seems to have been Chomsky’s (1970) proposal that the parts of speech of the open word classes (or “lexical categories”, somewhat arbitrarily defined as Noun, Verb, Adjective and Preposition/Postposition; see Vonen 1997, ch. 2) be seen as complex categories, feature structures made up of the binary features ±N and ±V. Describing part of speech systems by feature

structures holds the potential, at least, for stating correspondence rules in a more general fashion than if word classes are treated as atomic entities, but to be useful in this regard, the feature structures should probably contain more information than the two features ±N and ±V.

While formalist grammatical traditions thus take the classical part of speech inventory for granted, functionally and cognitively oriented linguists aspire towards universally valid characterisation—or 'explanation'—of parts of speech as functionally or cognitively determined prototypes. Thus, Hopper and Thompson (1984) characterise prototypical verbs and nouns in discourse-functional terms, and Thompson (1988) goes on to define the cross-linguistic prototype 'adjective' in the same fashion (see also Givón 1984).

Still, there is scope for language-specific manifestations of these universal prototypes. Even though they represent distinctions that all languages are inclined to make, no language actually needs to make all of them always. As frequently happens in language description, we are dealing with tendencies, rather than absolutes. The actual part of speech inventory recognised for a particular language depends on many factors, including whims of history, and, consequently, universally valid generalisations regarding parts of speech have been hard to make. It has long been held that nouns and verbs are the only universal parts of speech, in the sense that they are found in all human languages (by necessity, some would say; cf. above and Sapir 1921:119), while other parts of speech appear only in some languages, but not in others. Even this fundamental division has been questioned, however, in that some languages have been described as having only verbs (e.g., Cayuga, see Ramat 1999), while other languages represent the opposite extreme, using no more than a handful of simple verbs (e.g., Kalam, see Pawley 1993).[5]

There is a growing interest among typologists in the properties of part of speech systems (see Anward *et al*. 1997 for a good overview of recent research in this area), but as far as I have been able to acertain, there have been no investigations of part of speech correspondences in translation.[6] This means that in a trivial sense, Melamed's conjecture "that word pairs that are good translations of each other are likely to be the same parts of speech in their respective languages" (1995:7), is necessarily false, because any word translated from, say, German into Cayuga (see above), would have to be translated into a verb, regardless of its original part of speech. At the same time, it means that we simply do not know whether there are universal correspondence rules, or tendencies, holding for parts of speech in translation, and which could make a modified version of the conjecture hold water, namely that there are systematic part of speech correspondences in translations. Asking whether there are such systematic correspondences is tantamount to asking whether there are interesting universal regularities holding for the mappings between different linguistic systems.[6] Thus, it seems that investigations of the kind presented here, if extended to more and to

more diverse languages, could make a contribution both to computational corpus linguistics and to linguistic typological research.

4. Conclusion

In brief, the conclusions tentatively to be drawn from the experiment described here is that the idea of using word alignment as a stand-in for, or as a complement to, POS tagging is viable and worth exploring further. However, it seems that certain prerequisites have to be fulfilled for it to work:

- The languages in question should be genetically or typologically close, at least pending more detailed research on correspondences between part of speech systems;
- A high word alignment precision is needed (high recall is good too, but if the precision is low, the results are too uncertain);
- Only coarse-grained POS tagging is possible with this approach.

Finally, it seems that investigations of the kind presented here are needed—although they must be extended to take into account many other languages, of various types—and could make a valuable contribution both to computational corpus linguistics and to linguistic typological research.

Notes

1 The research reported here was carried out within the ETAP project (see Borin this volume, for a description of this project), supported by the Bank of Sweden Tercentenary Foundation as part of the research programme *Translation and Interpreting—a Meeting between Languages and Cultures*. See http://www.translation.su.se/

2 In a comparison we made of two freely available German taggers, Morphy and TreeTagger (Schiller *et al.* 1995), Morphy actually came out in second place (Borin to appear). We still chose it for this experiment, however, because its larger and more fine-grained tag set corresponded better to the Swedish tag set used (the larger SUC tag set; see Ejerhed & Källgren 1997).

3 The abbreviations used in these and later examples are the following.

ACC: Accusative	ACT: Active	ADJ: Adjective
ADV: Adverb	AKK: Accusative	DAT: Dative
EN: English	FEM: Feminine	FI: Finnish
GEN: Genitive	GR: Gerund	INF: Infinitive
JJ: Adjective	KN: conjunction	KON: conjunction
MOD: Modal	NEB: Coordinating	NEU: Neuter
NN: Noun	NOM: Nominative	

NVL: (regular) Verbal Noun		PA2: Past Participle
PL: Polish	PLU: Plural	PN: Pronoun
PR: Present	PRL: Particle	PTC: Participle
PTV: Partitive	SE: Swedish	SIN: Singular
SUB: Noun	SUP: Supine	TRV: Translative
VB: Verb	VER: Verb	3: Third Person.

4 There is also the factor—always present—of translations tending to be more similar to their source language text, in all kinds of linguistic respects, than a comparable original target language text would be. We thus note that the fact that the translation is in 'translationese' may well in itself occasion an increase in the number of POS correspondences between the two texts, but we will not be able to delve deeper into this matter here (cf. Johansson this volume).

5 In the cited works, only the so-called open, or lexical word classes are considered, i.e. verbs, nouns, adjectives and adverbs. The existence or non-existence of parts of speech containing closed-class, or grammatical, or functional items is not under discussion.

6 Perhaps this is a special case of the general reluctance among linguists, noted by Salkie (this volume), to take on problems of translation.

7 In the same way that one could imagine that mappings between different colour systems obey certain general principles—e.g., if a language lacks a word for the colour 'violet', it uses a word which covers, i.a., 'brown' (this is only intended as an example; it is a fact about the history of Swedish colour terms, but I do not know if it is a valid generalisation about colour terms in languages in general)—it is conceivable that there are regularities (expressible in linguistic terms) in the mappings between different part of speech systems.

References

Anward, Jan, Edith Moravcsik and Leon Stassen (1996), 'Parts of speech: a challenge for typology', *Linguistic typology*, 1(2): 167–183.

Borin, Lars (this volume), '… and never the twain shall meet?'. 1–32.

Borin, Lars (to appear), 'Something borrowed, something blue: rule-based combination of POS taggers', in: *Proceedings of the 2nd international conference on language resources and evaluation (LREC2000)*, Athens, Greece.

Chomsky, Noam (1970), 'Remarks on nominalization', in: Roderick A. Jacobs and Peter S. Rosenbaum (eds.), *Readings in English transformational grammar*. Waltham, Mass.: Ginn. 184–221.

Ejerhed, Eva and Gunnel Källgren (1997), 'Stockholm Umeå Corpus version 1.0, SUC 1.0'. Department of Linguistics, Umeå University.

Gale, William A. and Kenneth W. Church (1993), 'A program for aligning sentences in bilingual corpora', *Computational linguistics*, 19(1): 75–102.

Gazdar, Gerald, Ewan Klein, Geoffrey Pullum and Ivan Sag (1985), *Generalized phrase structure grammar*. Oxford: Basil Blackwell.

Givón, Talmy (1984), *Syntax. A functional-typological introduction. Volume I*. Amsterdam: John Benjamins.

Hakulinen, Lauri (1979), *Suomen kielen rakenne ja kehitys*. Helsinki: Otava.

Hopper, Paul J. and Sandra A. Thompson (1984), 'The discourse basis for lexical categories in universal grammar', *Language*, 60(4): 703–752.

Jespersen, Otto (1924), *The philosophy of grammar*. London: George Allen & Unwin.

Johansson, Stig (this volume), 'Towards a multilingual corpus for contrastive analysis and translation studies'. 45–57.

Lezius, Wolfgang, Reinhard Rapp and Manfred Wettler (1998), 'A freely available morphological analyzer, disambiguator, and context sensitive lemmatizer for German', in: *COLING-ACL'98. 36th annual meeting of the Association for Computational Linguistics and 17th international conference on computational linguistics. Proceedings of the conference, Vol. I–II*. Montreal: Université de Montréal.

Màrquez, Lluís, Lluís Padró and Horacio Rodríguez (1998), 'Improving tagging accuracy by using voting taggers', in: *Proceedings of NLP + IA/TAL + AI '98*. Moncton, New Brunswick, Canada.

Melamed, Dan (1995), 'Automatic evaluation and uniform filter cascades for inducing *N*-best translation lexicons', in: *Proceedings of the third workshop on very large corpora*. Boston, Massachusetts. [Page references are to the version available through the author's Web Page: http://www.cis.upenn.edu/~melamed/]

Merkel, Magnus, Mikael Andersson and Lars Ahrenberg (this volume), 'The PLUG Link Annotator – interactive construction of data from parallel corpora'. 77–94.

Pawley, Andrew (1993), 'A language which defies description by ordinary means', in: W. A. Foley (ed.), *The role of theory in language description*. Berlin: Mouton de Gruyter. 87–129.

Pollard, Carl and Ivan A. Sag (1994), *Head-driven phrase structure grammar*. Chicago: The University of Chicago Press.

Radford, Andrew (1988), *Transformational grammar. A first course*. Cambridge: Cambridge University Press.

Ramat, Paolo (1999), 'Linguistic categories and linguists' categorizations', *Linguistics*, 37(1): 157–180.

Salkie, Raphael (this volume), 'How can linguists profit from parallel corpora?'.

Sapir, Edward (1921), *Language*. New York: Harcourt, Brace & World.

Sågvall Hein, Anna (this volume), 'The PLUG project: parallel corpora in Linköping, Uppsala, Göteborg: aims and achievements'. 59–75.

Schiller, Anne, Simone Teufel, Christine Stöckert and Christine Thielen (1995), 'Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS', Draft. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung / Universität Tübingen, Seminar für Sprachwissenschaft.

Thompson, Sandra A. (1988), ‘A discourse approach to the cross-linguistic category 'adjective'', in: John A. Hawkins (ed.), *Explaining language universals*. Oxford: Basil Blackwell. 167–185.

Tiedemann, Jörg (1998), ‘Extraction of translation equivalents from parallel corpora’, in: *Proceedings of the 11th Nordic conference on computational linguistics*, Copenhagen 28–29 January 1998 (NODALIDA’98), Center for Sprogteknologi, University of Copenhagen. 120–128.

Tiedemann, Jörg (this volume), ‘Uplug – a modular corpus tool for parallel corpora’. 107–122.

Tiedemann, Jörg (to appear), ‘Word alignment step by step’, in: *Proceedings of the 12th Nordic conference on computational linguistics*, Trondheim 9–10 December 1999 (NODALIDA’99).

Vonen, Arnfinn Muruvik (1997), *Parts of speech and linguistic typology. Open classes and conversion in Russian and Tokelau*, Oslo: Scandinavian University Press.