(German) Corpus Representativity, Bigrams, and PoS-Tagging Quality

Karel OLIVA
Austrian Research Institute
for Artificial Intelligence (ÖFAI)
Schottengasse 3
Wien, Austria, A-1010
kareloliva@hotmail.com

Pavel KVĚTOŇ
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25
Praha, Czech Republic, CZ-118 00
pavel.kveton@seznam.cz

Abstract

After some theoretical discussion on the issue of (different meanings of the term) *representativity of a corpus*, this paper takes this issue into practice and shows how a representative (in one of the meanings) corpus of German can be achieved. The approach is based on the idea of application of "invalid bigrams", i.e. of abolishing pairs of adjacent tags which constitute an incorrect configuration in a text of German (e.g., the bigram [ARTICLE,FINITE VERB]). On this spot, the paper puts forward a list of such bigrams for the STTS tagset (widely used for PoS-tagging German corpora). The power of the approach is illustrated on the results achieved on the NEGRA corpus. Finally, some general implications for tagging and taggers are mentioned.

1 The Representativity Issue

"Representativity" is a noun which bears two theta-roles: representativity is necessarily a *representativity of a* (representing) *agent wrt. a patient* (being – or not being, as it might happen – represented). A typical example from the area of corpus linguistics is the representativity of a corpus wrt. some language phenomenon.

In this paper, we intend to scrutinize the issue of representativity of a part-of-speech (PoS) tagged corpus in some detail. In order to stay on a non-trivial but still easily understandable level, let us consider the case of representativity of the NEGRA corpus of German wrt. to bigrams contained[1]. In this case, the phenomena whose presence and relative frequency are at stake are:

- bigrams, i.e. pairs [First,Second] of tags of words occurring in the corpus adjacently and in this order
- unigrams, i.e. the individual tags.

We shall define the *qualitative representativity* wrt. bigrams as the kind of representativity meeting the following two complementary requirements:

- the representativity wrt. *the presence of all valid bigrams* of the language in the corpus, which means that if any bigram [First,Second] is a bigram in a correct sentence of the language, then such a bigram occurs also in the corpus - this requirement might be called *positive representativity*
- the representativity wrt. *the absence of all invalid bigrams* of the language in the corpus, which means that if any bigram [First,Second] is a bigram which cannot occur in a correct (i.e. grammatical) sentence of the language, then such a bigram does not occur in the corpus - this requirement might be called *negative representativity*.

If a corpus is both positively and negatively representative, then indeed it can be said to be a qualitatively representative corpus. In our particular example this means that a bigram occurs in a qualitatively representative (wrt. bigrams) corpus if and only if it is a possible bigram in the language (and from this it already follows that any unigram occurs in such a corpus if and only if it is a possible unigram[2]). From this formulation, it is also clear that the

[1] The case of a trigrams, used more usual in tagging practice, would be almost identical but require more lengthy explanations. For the conciseness of argument, we thus limit the discussion to bigrams.

[2] This assertion holds only on condition that each sentence of the language is of length two (measured in words) or longer. Analogical limitation holds in the case of trigrams, etc.

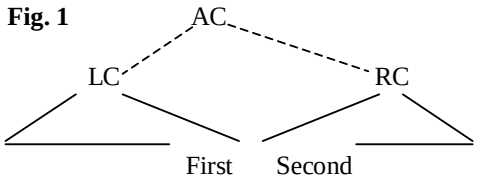
qualitative representativity depends on the notion of grammaticality, that is, on the "language competence" – on the ability of distinguishing between a grammatical and an ungrammatical sentence.

The *quantitative representativity* of a corpus wrt. bigrams can then be approximated as the requirement that the frequency of any bigram and any unigram occurring in the corpus be in the proportion "as in the language performance" to the frequency of occurrence of all other bigrams or unigrams, respectively[3]. However, even when its basic idea is quite intuitive and natural, it is not entirely clear whether the quantitative representativity can be formalized rigorously. At stake is measuring the frequency of a bigram (and of a unigram) within the "complete language performance", understood as set of utterances of a language. This set, however, is infinite if considered theoretically (i.e. as set of all possible utterances in the language) and finite but practically unattainable if considered as a set of utterances realized within a certain time span. For this reason, we refrain from quantitative representativity in the following and we deal with qualitative representativity only.

2 The Bigram Issue or Towards Negative Representativity wrt. German Bigrams

From a linguistic viewpoint, the pair of tags [First,Second] is a *valid bigram* in a certain natural language if and only if there exists a sentence (at least one) in this language which contains two adjacent words bearing the tags First and Second, respectively. Such a sentence then can be assigned its structure, and hence a valid bigram [First,Second] comes into being via a structural configuration where there occur two adjacent constituents LC (for "Left Constituent") and RC (for "Right Constituent"), such that LC immediately precedes RC and the last (rightmost) element of the terminal yield of LC is First and the first (leftmost) element of the terminal yield of RC is Second, cf. Fig. 1, where also the common ancestor (not necessarily the mother) of LC and RC is depicted (as AC, "Ancestor Constituent").

Fig. 1



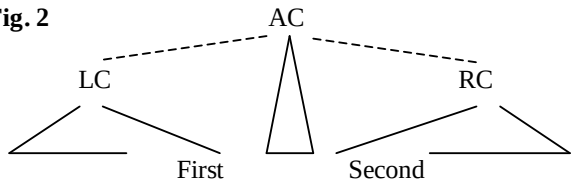
Accordingly, the pair of tags [First,Second] is a *linguistically invalid bigram* in a certain natural language if and only if there exists no grammatically correct sentence in this language which contains two adjacent words bearing the tags First and Second, respectively. Seen from a structural perspective, [First,Second] is an invalid bigram if one or more of the following obtains:

- the configuration from Fig. 1 is impossible because in all constituents LC, First must necessarily be followed by some other lexical material (example: the bigram [ARTICLE, FINITE VERB] is impossible in German since in any LC - NPs, PPs, Ss etc. - article must be followed by (at least) a noun/adjective/ numeral before an RC (in this case a VP or S) can start)
- the configuration from Fig. 1 is impossible because in all constituents RC, Second must necessarily be preceded by some other lexical material (example: the bigram [SEPARABLE VERB PREFIX, POSTPOSITION] is impossible in German since in any RC - NPs, PPs, Ss, etc. - a postposition must combine with some preceding lexical material displaying (morphological) case before such a constituent can be combined with any other material into a higher unit
- the configuration from Fig. 1 is impossible because LC and RC can never occur as adjacent sisters standing in this order - cf. Fig. 2 (example: the bigram [FINITE VERB, FINITE VERB] is impossible in German since according to the rules of German orthography any two finite verbs / verb phrases must be separated from each other by at least a conjunction (coordinating or subordinating) and/or by a comma[4]).

[3] From this it easily follows that any quantitatively representative corpus is also a qualitatively representative corpus.

[4] The categorization of a particular invalid bigram into the class (i), (ii) or (iii) depends obviously on the shape of constituent structure adopted. However, different categorization cannot change the fact of the invalidity of the particular bigram.

Fig. 2



For a particular language with a particular tagset, the set of invalid bigrams can be obtained by a reasonable combination of (i) simple empirical methods leaning on the language performance that can be gained from a corpus with (ii) a careful competence-based ("linguistic") analysis of the language facts.

In our case, we used the German NEGRA corpus hand-tagged with the STTS tagset. Put very simply, we created a set of all bigrams which occurred in this corpus five or less times (including no occurrences) and then checked this set manually, since the presence of a bigram in a corpus still does not guarantee that the bigram is valid (the bigram or the source text might be erroneous - the corpus is not necessarily negatively representative) and likewise its absence does not automatically imply that the bigram is an invalid one (the corpus need not be positively representative). For the STTS tagset consisting of 54 tags, the size of the set of invalid bigrams thus obtained went into tens (its substantial part is discussed in more detail below). For larger tagsets, e.g., the tagset for Czech described in Hajič and Hladká (1998), we conjecture that the cardinality of this set will reach thousands, forcing some factorisation (e.g., by PoS and subPoS) for reasons of practical manageability. Tedious as such manual checking is, it is certainly less demanding (measured in hours of manpower) than hand-tagging of a reasonably sized learning corpus, and it is also a very rewarding as to results, since the set of invalid bigrams is a powerful tool for error detection in corpora already tagged and for avoiding errors in tagging raw texts, since:

- the presence of an invalid bigram in a tagged corpus signals an error in this corpus
- an invalid bigram should never be used in - and hence never come into being as a result of - tagging a raw corpus.

The preceding, however, holds only if the following non-trivial presuppositions are met:

- first, all words in the text are to be used in their primary function. In particular, metalinguistic usage is not taken into consideration, otherwise counterexamples (i.e. correct usage of bigrams marked as invalid) can be found easily, cf. the sentence *Das Wort die ist ein Artikel* where the otherwise invalid bigram [ARTICLE, FINITE VERB] (cf. above) is to be found
- second, all sentences in the corpus are correct wrt. the language of the corpus; in existing large corpora, however, this condition is as a rule not met.

Taking the this into account, we have to conclude that:

- the presence of an invalid bigram in a tagged corpus signals either an error in tagging or a error in the source text or a metagrammatical usage of some word(s) in the text
- the impossibility of assigning other than an invalid bigram in tagging (typically because the morphological analysis did not provide any other options for the tagger to choose from) might have the following reasons: (i) a genuine error in the source text or (ii) an incorrect/incomplete morphological analysis (typical with unknown words) or (iii) metalinguistic usage of some word(s).

From this it follows that if we wish to achieve a correctly tagged corpus, then, in the case of a corpus already tagged, any detected occurrence of an inva -checked and corrected when appropriate (i.e. at least in the cases where a tagging error was detected). Mind that hand-checking is necessary since the decision whether the origin of the invalid bigram is a tagging error, a source text error or a metagrammatical usage, can be performed solely on the basis of linguistic competence. In addition, in the particular case of a corpus which is to be used as a training corpus for statistical taggers, it is even advisable to correct also the errors in the source text, since otherwise the learning corpus will not be (qualitatively) representative. With sentences containing metalinguistic expressions, we would tentatively argue that they should be marked as such and excluded from the learning process. As for what to do in the case of a corpus which is yet to be tagged (i.e. in the case of active tagging), we shall discuss the issue briefly in the Conclusions.

In order to be able to apply the approach using the invalid bigrams in practice, it is necessary to have a list of the invalid bigrams. In the following, we put forward a list of (German) invalid bigrams consisting of the tags of the STTS tagset. The overview is organized in such a way that each its item starts with the respective bigram, which consists either of two genuine tags or it may contain a "variable" X which is then specified more closely in the description following the bigram proper. If two tags behave similarly in the bigram, they have been packed together onto one position and their disjunction is marked off by a slash. A reasonable knowledge of the STTS tagset is needed for understanding the descriptions, for this cf. (Schiller et al. 1999). The tags FM, ITJ, XY and $(are excluded from the following overview, unless specifically mentioned.

- *[X,PRELS]*: PRELS introduces the relative clause, i.e. it must stand very close to its beginning, preceded by a clause separator (typically a comma or coordinating conjunction), inbetween the two only a preposition can intervene; since a relative pronoun has to follow its antecedent, it cannot stand at the very beginning of a sentence (it cannot be preceded by beginning of sentence - BOS). Hence, the bigram [X,PRELS] is incorrect for all X ≠ $, , $(, KON , APPR. Exception to this rule is attested once in NEGRA, in the sentence 6870 where the relative pronoun *die* starts a stand-alone relative sentence: *(Oder beispielsweise Leute, die an ihre Idee glaubten.) Die/PRELS gegen großen Widerstand, gegen die gesamte etablierte Wissenschaft gekämpft haben...*
- *[X,PRELAT]*: this kind of relative pronoun displays the same properties as PRELS plus it can stand on the position of a genitive attribute; this means that it can be preceded (only) by any material mentioned for PRELS and in addition by a noun; i.e. the bigram [X,PRELAT] is incorrect for X ≠ $, , $(, KON, APPR, NN, NE
- *[PRELAT,X]*: PRELAT must necessarily be followed by an NP (or at least by a remnant of an NP), so that X must be a tag marking a word which possibly can start an NP, hence tags APPO, APZR, KOUS, PTKVZ, VVFIN, VVIMP, VVINF, VAFIN, VAIMP, VAINF, VMFIN, VMINF are ruled out, and further impossible are also the following ones: (i) $. (the sentence cannot end immediately after the attributive relative pronoun), (ii) PWS (the NP following the PRELAT cannot be a *wh*-NP, and any of the pronouns *wer, was* cannot even occur at its beginning), (iii) KON (the NP to follow PRELAT cannot start by a coordinating conjunction, even not of the type *weder* (in *weder-noch*), *entweder* (in *entweder-oder*) etc.). Further ruled out are bigrams [PRELAT,PRELAT] and [PRELAT, PRELS]. In the real performance, many more bigrams are in fact ruled out, since, e.g., constructions like *das Schiff, dessen aufzubrechen/VVIZU wollende Mannschaft ...* are indeed possible in the competence but not attested in the performance
- *[X,APPO/APZR]*: APPO/APZR must be immediately preceded by some nominal material (typically by NN, NE, PPER, PDS, PRELS, PWS; possible but without empirical evidence from NEGRA are elliptical constructions where ADJA, PPOSAT, CARD stand in front of APPO/APZR) or by a comma; it is impossible, however, for any other material to immediately precede APZR or APPO, hence the bigram [X,APPO/APZR] is incorrect for all X ≠ $, , $(, NN, NE, PPER, PDS, PRELS, PWS, ADJA, PPOSAT, CARD
- *[X,KOUS]:* a subordinating conjunction has to stand at the beginning of the respective subordinate clause, preceded by a clause separator (typically a comma or coordinating conjunction) or directly by the beginning of sentences (BOS); inbetween the clause separator and the subordinating conjunction, only a preposition or a "short" adverb can intervene (e.g., *ohne dass er wusste, erst wenn ...*), i.e. the bigram [X,KOUS] is incorrect for X ≠ BOS, $, , $(, KON, APPR, ADV. If another configuration occurs, e.g., NN KOUS, it signals either a tagging error or a syntactic problem (e.g., NEGRA sentence No. 11818 *Einen Tag/NN nachdem/KOUS der ASC Darmstadt und der Ausrüster die Verträge kündigten....* is KOUS really the appropriate part-of-speech for *nachdem* in this sentence, and how comes there is a subordinated sentence which does not start (and maybe even contain) a subordinating conjunction ?) or there occurs a genuine

ungrammaticality in the source text (e.g., NEGRA sentence 11684 *Das Ethos des preußischen Berufsbeamtentums genoß einen hohen Stellenwert, FR-Porträt/NN als/KOUS er der Chef im Rathaus war.*)

- *[ART/APPRART/APPR,X]*: nothing verbal incl. separable prefix but excl. the *zu* particle (since this stands also with verbal adjectives - *die zu renovierende Wohnung*), no relative pronoun (cf. above, pronoun on the second position of the bigram), no KOUI, no APPO and no APZR can stand immediately after an article or a preposition (or their aggregate); two articles or prepositions are however allowed, and in fact in German even examples like *eine Tonnage von/APPR bis/APPR zu/APPR über/APPR 200.000 BRT* (unattested, but easily constructible) are possible ...
- *[PTKA,X]*: the PTKA particles (*zu, allzu, am*) stand regularly with adjectives ADJA, ADJD or adverbs ADV (occasionally also VVPP) and rarely with PIS/PIAT (*zu wenig essen, zu wenige Besucher*); any other combinations are ruled out, hence this bigram is incorrect if X ≠ ADJA, ADJD, VVPP, ADV, PIS, PIAT
- *[PTKZU,X]*: the typical position of the verbal particle *zu* is in front of an infinitive verb form, alternatively it may occur also in front of an attributively used verbal adjective (*die zu renovierende Wohnung*), and this even in case this adjective is modified by an adverb (*die ganz nötig zu renovierende Wohnung*), and of course it can stand in front of inverted commas; i.e. the bigram [PTKZU,X] is incorrect whenever X ≠ VVINF, VMINF, VAINF, ADJA, $(
- *[PTKVZ,X]*: a separable verbal prefix occurs most typically in the position of the Rechte Satzklammer, that is, it can be followed either by the interpunction marking off the end of the sentence/clause or by material standing extraposed in the Nachfeld; on rare occasions, it can stand as the single element of the Vorfeld of a Verb-second clause (ex.: *Aus/PTKVZ schaltet/VVFIN man es mit diesem Knopf*), being thus followed by a finite form of a main verb (not by an auxiliary[5], not by a modal). Hence, the set of invalid bigrams depends crucially on the material allowed to occur in the Nachfeld, which most typically can be a prepositional phrase (started by a preposition), or an adverb, or a heavy infinitive phrase (which never starts by an infinitive verb, more likely by a KOUI like *um* or *ohne*), or a relative clause (which has to be separated by a comma, however) and which never can be an auxiliary or modal. The definition of invalidity of this bigram thus depends on the grammatical tolerance towards material in the Nachfeld, but in any case this bigram is incorrect if X = VMFIN, VMINF, VAINF, VAIMP, VVINF, VVIMP. Interesting is the case of X = PTKVZ, i.e. the case of two separable prefixes following immediately each other, which, according to standard grammatical wisdom, should be impossible; however, examples like *Er handelte den Vertrag mit aus* cast serious doubts on such statements
- *[X,VVIMP/VAIMP]:* Imperative[6] must be generally clause initial, and can be preceded only by a very restricted set of expressions: *Ich weiss, dass du es machen kannst, doch/PTKANT mache/VVIMP es nicht; Bitte/PTKANT warten Sie; Wenn du es nicht selbst machen kannst, dann/ADV lass deine Freunde es machen* and of course it is possible that an imperative, exactly because it is clause initial, can be preceded by a comma (or by some other interpunction sign, for that matter) or by a coordinating conjunction. However, any other material is ruled out in standard German, i.e. this bigram is incorrect if X ≠ ADV, PTKANT, KON, $, , $(
- *[KOUI,X]*: KOUI is a conjunction introducing an infinitive VP, hence X cannot be from the set {VAFIN, VMFIN, VVFIN, VAIMP, VVIMP, PTKVZ} of finite verb forms (joined by a separable prefix)
- *no two finite verb forms can follow each other immediately:*
  Carthesian product {VAFIN,VMFIN,VVFIN, VAIMP,VMINP,VVIMP} x {VAFIN, VMFIN,VVFIN,VAIMP,VMINP,VVIMP} is impossible (it is an invalid bigram)

---

5 Note, however, that also copular and existential *sein/werden*, all kinds of *haben* (in particular the *haben* of possession) and all their derivatives are tagged as auxiliaries in STTS. ☹

6 STTS contains no tag for an imperative of a modal verb - hence only VVIMP/VAIMP is mentioned.

- *two interpunction signs following each other:* the configuration where two interpunction signs, both different from a fullstop, follow each other and both are different from inverted commas or both are the same kind of inverted commas or both are fullstops constitute an invalid bigram: e.g., two fullstops, two commas, colon and comma, ...
- *[VMFIN,PTKVZ]:* since a modal verb never takes a separable prefix, its finite form cannot be immediately followed by it
- *[KOKOM,PTKVZ/VAIMP/VVIMP]:* any of the two comparative particles (*als, wie*) can be followed by neither a separable prefix nor an imperative form of any verb.

Of practical importance are also the following invalid bigrams where one element of the pair is specified lexically (not by a tag):

- *[ART/APPR/APPRART,man]:* an article, a preposition or their aggregate cannot be followed by the pronoun *man*, for the reason that *man* behaves as if it were a personal pronoun in nominative - and an article never forms an NP with a personal pronoun, and a preposition can never be followed by any nominative case form
- *[BOS,$.]:* this is an invalid bigram since no sentence can start with (or: consist only of) its final punctuation.

Some bigram configurations are open for (linguistic) discussion. Such a case is, for instance, the attributive elements (such as ADJA, PIAT, PIDAT, PPOSAT) which have to be generally followed by an NP, so that at least finite verb forms following them should be ruled out - however, since ellipses might occur, even though especially when following PIAT, PIDAT they are improbable (e.g., they are not attested in NEGRA), we do not include such bigrams among the invalid ones. Generally, also many other bigrams are possible theoretically, but are not attested in the competence.

Another point of discussion is of course the generalisation of the approach from invalid bigrams to invalid trigrams, invalid tetragrams, etc. A possible strategy of learning some - but not all - of the invalid *n*-grams (n>2) from a tagged corpus by the strategy of "loosening" the invalid bigrams (i.e., for a known invalid bigram [First,Second] by allowing some material to occur inbetween First and Second) is described in (Květoň and Oliva, to appear 2002).

As examples of invalid trigrams might serve:

- *[ART/APPRART,ADJD/ADV,X]:* since an article or article+preposition aggregate has to combine with some nominal (case-marked) material to its right before it can combine with anything verbal, the trigram is invalid for X from {VAFIN, VMFIN, VVFIN, VAIMP, VVIMP, VAINF, VMINF, VVINF, PTKVZ}
- *[ADJD/ADV,NN/NE/PPER/PDS/PIS/PPOSS/PRF,APZR]:* the configuration adverb + nominal (noun or pronoun) + right part of circumposition is impossible since an adverb can modify (i) neither a noun to its right (cf. *der Tisch links/ADV* vs. **der links Tisch*) (ii) nor an adjective to its left (*die gründlich/ADV renovierte Wohnung* vs. **die renovierte gründlich/ADV Wohnung*) and hence cannot stand on this position within a nominal construction which ends with the APZR and starts (somewhere to the left) with an APPR (this APPR has to be there, since it creates the left pendant to the APZR).

As an example of an invalid tetragram, we might put forward:

- *[ART,APPR,NN/NE,APPO]* which is invalid since APPR and APPO cannot occur both around a single noun – this were in such a configuration enforced by the presence of the ART (the trigram [APPR,NN/NE,APPO] is a valid trigram, however, cf. *der Nachricht von/APPR Reuters/NE nach/APPO* !).

3 The Quality Issue or Results of Practical Application

Employing the invalid bigrams (and some extensions to it) as an error-detection technique, we were up to now able to correct 3.773 errors in the NEGRA corpus, and we can guarantee that the corrected version of the corpus is now negatively representative wrt. bigrams. Since we aimed at achieving a truly correct corpus, suitable, e.g., for training statistical taggers, we corrected all kinds of errors. The prevailing part of the errors detected was that of incorrect tagging (only less than 8% were genuine ungrammaticalities in the source, about 26% were errors in segmentation). Based on this, we were able to confirm the expected fact that the quality (i.e. representativity) of the learning

corpus has a paramount importance for the quality of the tagger trained on this corpus. We made some experiments in this direction and figured out that for the trigram-based TnT tagger (Brants 2000), the result of training on the corrected NEGRA (negatively representative wrt. bigrams) brought a relative error improvement of slightly over 10% as compared to training on the original NEGRA.

This also shows the directions of future work: the extension from (negative) representativity wrt. bigrams to (negative) representativity wrt. trigrams, which might possibly help to discover more errors in the tagging of the NEGRA corpus. In particular, there exist invalid trigrams [First,Second,Third] which cannot be detected as such (i.e. as invalid) by the method (even with the "loosened" invalid bigrams) if [First,Second], [Second,Third] and [First,Third] are all possible bigrams[7]. Mind in this connection the fact that even if the set of all trigrams is much larger than the set of all bigrams, a very substantial subset of this set need not be searched through manually once the previous results concerning invalid bigrams are available, since:

- all those candidates [First,Second,Third] for invalid trigram which contain an invalid bigram [First,Second] or [Second,Third] can be discarded automatically from the search space (these are invalid as bigrams, hence certainly also invalid as trigrams)
- all those candidates [First,Second,Third] for invalid trigram which have been discovered as "valid extended bigrams" (discussed in Květoň and Oliva, to appear 2002) are to be eliminated automatically from the search space, too, since they are already known to be possible trigrams.

Finally, it should not remain neglected that in a tagged corpus, the method sketched above allows not for detecting errors only, but also for detecting inconsistencies in hand-tagging (i.e. differences in application of a given tagging scheme by different human annotators and/or in different time), and even inconsistencies in the tagging guidelines. An issue of its own is also the area of detecting and tagging idioms/ collocations, in the case when these take a form which makes them deviate from the rules of standard syntax. Thus, in the following we present a selection of collocations which were found during the work on NEGRA and which are in some way syntactically deviant (and hence we did not take them into consideration when defining the invalid bigrams)[8]:

ohne wenn und aber	*Augen zu und durch*	
mehr oder minder	*mit von der Partie*	
ab und zu	*nach und nach*	*nach wie vor*
drum herum	*nichts wie weg*	*durch und durch*
je nachdem	*darüber hinaus*	*vor sich hin*
ein paar	*ein wenig*	*ein bisschen*
ein für allemal	*jung und alt*	*angst und bange*
zu Recht	*dann und wann*	*von einst*
zu eigen machen	*dicht an dicht*	*von neuem*
hin und wieder	*Vorhang auf*	*oben ohne*

Special cases are constituted by the following collocation-like constructions:

- *<NP> Revue passieren lassen* where - on an approach disregarding collocations - the verb *passieren* would take two objects, *Revue* and the *NP*
- *was für ein <NP>*, where the nominal group *ein <NP>* can (in an appropriate context) occur also in nominative, in spite that it follows the "preposition" *für*
- *die <NUMERAL> Mann, alle Mann an Bord* - the word *Mann* serves here as a "measure word" (almost as in Chinese or Japanese ☺), takes no plural and is hardly a noun in the usual sense; cf. that this construction is impossible with any other noun *(*die 60 Frau, *die 60 Person)*
- the verbal collocations *wie folgt, d.h. (das heisst)* and *s. (siehe)*, which - in spite of having the form of a finite verb - can occur as elements of a clause, i.e. within a syntactic environment of another finite verb without being separated by a comma (cf., e.g., *die Liste sieht wie folgt aus*), the reason being that in this usage they do not give rise to a separate clause (e.g. the verb *folgt* in the above example does not take any subject, etc.).

[7] A possible example is the trigram [PREPOSITION, PREPOSITION, RELATIVE PRONOUN].

[8] For reasons of space, we provide no explanations to particular cases, however, even a moderate knowledge of German syntax makes it clear that in sentential contexts these collocations if tagged using just local morphological information give rise to non-standard syntactic constructions (and non-standard bigrams).

Of some interest might be also the following numbers: taking the 54 tags of STTS and enriching them with the tags BOS and EOS (for beginning and end of sentence, respectively), the complete bigram set has 56*56 = 3.136 bigrams. In the corrected version of the NEGRA corpus, only 947 bigrams of this set occur more than 5 times, and 457 bigrams have between one and five occurrences. The rest of 1.732 bigrams (i.e. considerably more than the half of the bigram set) do not occur at all (however, only a small part of them is genuinely invalid in the above sense !).

Conclusion

The main contribution of this paper lies in showing one possibility of combining the linguistic performance (as documented in corpora) with the linguistic competence (i.e. the expertise of a linguist) in order to achieve better corpora (better tagging results).

The primary practical outcome of this idea is that of correcting the NEGRA corpus, at least to an extent that it becomes negatively representative wrt. bigrams (i.e. that no invalid bigram occurs in the corrected version unless it is licensed by, e.g., a collocation; obviously we do not guarantee that the resulting corpus is positively representative wrt. bigrams - in fact we know it is not, cf. the numbers given in the final paragraph of Sect. 3 - and we do not know whether it is negatively representative wrt. trigrams even though we performed a limited search for a couple of invalid trigrams).

Moreover, there is another, more profound[9] or at least more general, result of the approach: the suggestion that avoiding errors (in tagging) is better than correcting them. In particular, we would like to argue that the idea of marrying performance with competence in the area of tagging forces the advent of interactive taggers. The experience gathered in our work shows that human intervention during the tagging process is unavoidable if errors are to be avoided (human correction of the errors committed being the only other option). The reason for this is that it is *only* the human linguistic knowledge (linguistic competence) together with understanding the text (semantics, pragmatics) which can decide what to do in cases where an invalid bigram (in the general case: *n*-gram) has no alternative. In other words, it is only the human language competence which can decide whether the occurrence of such configurations is due to a genuine error in the source text (and to decide whether such an error has to be corrected, and how) or due to other factors discussed above.

This holds for all kinds of taggers, statistical ones (*n*-gram and maximum entropy based) and rule-based ones (Brill-style and constraint grammar style) alike, and this is also the moral to be learnt for further developments, if the aim at achieving high-quality PoS-tagged corpora should become reality in the near future.

Acknowledgements

This work has been sponsored by the *Fonds zur Förderung der wissenschaftlichen Forschung (FWF)*, Grant No. P12920. The *Austrian Research Institute for Artificial Intelligence (ÖFAI)* is supported by the *Austrian Federal Ministry of Education, Science and Culture*.

References

Brants T. (2000) *TnT – A Statistical Part-of-Speech Tagger.* In "Proceedings of the 6th Applied Natural Language Processing Conference", Seattle

Hajič J. and B. Hladká (1998) *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset.* In "Proceedings of ACL/Coling '98", Montreal

NEGRA®: `http://www.coli.uni-sb.de/sfb378/NEGRA-corpus`

Květoň P. and K. Oliva (to appear 2002): *Achieving an Almost Correct PoS-Tagged Corpus*. In: "Proceeding of the 5th international conference Text, Speech and Dialogue TSD 2002", Lecture notes in artificial intelligence , Springer, Berlin

Schiller A., S. Teufel, C. Stöckert and C. Thielen (1999) *Guidelines für das Tagging deutscher Text-corpora*. University of Stuttgart / University of Tübingen

Skut W., B. Krenn, T. Brants and H. Uszkoreit (1997) *An Annotation Scheme for Free Word Order Languages.* In "Proceedings of the 3rd Applied Natural Language Processing Conference", Washington D.C.

[9] Even when sounding trivial.