The Prague Bulletin of Mathematical Linguistics 78, 2002

Linguistically Motivated Bigrams in Part-of-Speech Tagging of Language Corpora

Karel Oliva and Pavel Květoň

Abstract

After some discussion concerning the issues of corpus representativity in the first paragraphs, this paper presents a simple yet in practice very efficient technique serving for automatic detection of those positions in a Part-of-Speech tagged corpus where an error is to be suspected. The approach is based on the idea of creating and then applying a set of "invalid bigrams", i.e. of pairs of adjacent Part-of-Speech tags which constitute an incorrect configuration in a tagged text of a particular language (in English, e.g., the bigram *[ARTICLE, FINITE VERB]*). Further, the paper describes the generalization of the "invalid bigrams" into a certain set of "invalid *n*-grams", for any natural *n*, which indeed provides a powerful tool for error detection in a corpus. Some implementation issues are also presented, as well as evaluation of results of the approach when used for error detection in the NEGRA corpus of German. Finally, general implications for the quality of results of statistical taggers are discussed.

Illustrative examples in the text are taken mainly from German, and hence at least a basic command of this language would be helpful for their understanding – due to the complexity of the necessary accompanying explanation, the examples are neither glossed nor translated. However, the central ideas of the paper should be understandable also without any knowledge of German.

1. Errors in Part-of-Speech Tagged Corpora

The importance of correctness (absence of errors) of language resources in general and of tagged corpora in particular cannot probably be overestimated. However, the definition of what constitutes an error in a tagged corpus depends on the intended usage of this corpus.

If we consider a quite typical case of a Part-of-Speech (PoS) tagged corpus used for training statistical taggers, then an error is defined naturally as any deviation from the regularities which the system is expected to learn; in this particular case this means that the corpus should contain neither errors in assignment of PoS-tags nor ungrammatical constructions in the corpus body[1], since if any of the two cases (wrong tagging, ungrammatical input) is present in the corpus, then the training process necessarily:

- gets a confused view of probability distribution of configurations (e.g., trigrams) in a correct text and/or, even worse (and, alas, much more likely)
- gets positive evidence also about configurations (e.g., trigrams) which should not occur as the output of tagging linguistically correct texts, while simultaneously getting less evidence about correct configurations.

If we consider PoS-tagged corpora destinated for testing NLP systems, then obviously they should not contain any errors in tagging (since this would be detrimental to the validity of results of the testing) but on the other hand they should contain a certain amount of ungrammatical constructions, in order to test the behaviour of the system on a realistic input.

[1] In this paper we on purpose do not distinguish between "genuine" ungrammaticality, i.e. one which was present already in the source text, and ungrammaticality which came into being as a result of faulty conversion of the source into the corpus-internal format, e.g., incorrect tokenization, OCR-errors, etc.

Both these cases share the quiet presupposition that the tagset used is linguistically adequate, i.e. it is sufficient for unequivocal and consistent assignment of tags to the source text[2].

As for using annotated corpora for linguistic research, it seems that even inadequacies in the tagset are tolerable provided they are marked off properly – in fact, the spots in the corpus where the tagset proves linguistically inadequate might well be quite an important source of linguistic investigation since, more often than not, they constitute direct pointers to occurrences of linguistically "interesting" (or at least "difficult") constructions in the text.

This paper, hence, will be mainly concerned with the issues of errors in a PoS-tagged corpus, that is, with the theoretical basis and possibilities of application of methods which can be used for detecting errors in a standing PoS-tagged corpus and, in the final paragraphs, also with proposals of techniques serving for avoiding errors in PoS-tagging in case of corpora yet untagged.

2. Issues of Corpus Representativity

In corpus linguistics, the term *representativity* is used to express

- either the fact that the corpus is balanced wrt. the kinds (genres) of text from which the texts[3] constituting the corpus are taken
- or the fact that the corpus contains the full range of examples of a certain linguistic (e.g., syntactic) phenomenon or set of phenomena – such as agreement, subcategorization, word order, etc.

In the current paper, we shall ignore the first of the above readings and take into consideration only the second one, but even here we shall possibly diverge from what can be thought of as "standard linguistic intuitions".

The definition of a (general) phenomenon might vary considerably, and in particular, it need not be in accord with the standard linguistic approaches. Thus, in this paper, we intend to scrutinize the issue of representativity of a PoS-tagged corpus wrt. to *bigrams*[4]. In this case, the phenomena[5] 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 at least once also in the corpus – this 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 might be called *negative representativity*.

[2] This problem might be – in a very simplified form – illustrated on an example of a tagset introducing tags for *NOUN*s and *VERB*s only, and then trying to tag the sentence *John walks slowly* – whichever tag is assigned to the word *slowly*, it is obviously an incorrect one. Natural as this requirement of linguistic adequacy might seem, it is in fact not met fully satisfactorily in any tagset we are aware of.

[3] The term "text" is to be understood very broadly – in particular, not only as a written form of a language, since there of course exist also corpora of spoken language.

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

[5] In an indeed broadly understood sense of the word "phenomenon".

If a corpus is both positively and negatively representative, then indeed it can be said to be a qualitatively representative corpus[6]. 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[7]). From this formulation, it is also clear that the 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 occurrences of all other bigrams or unigrams, respectively[8]. However, even when its basic idea is quite intuitive and natural, it is not entirely clear whether quantitative representativity can be formalized rigorously. At stake is measuring the occurrences of a bigram (and of a unigram) within the "complete language performance", understood as the set of utterances of a language. This set, however, is infinite if considered theoretically (i.e. as the 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 (also, due to immanent language change, it is questionable whether the concept of set of utterances over a time span is a true performance of a single language). Notwithstanding these problems, the frequencies are used in practice (e.g., for the purpose of training statistical taggers), and hence it is useful to state openly what they really mean: in our example, it is the relative frequencies of the bigrams (and unigrams) in a particular (training or otherwise referential) corpus. For this reason, since we would not like to be bound to a particular corpus, we refrain from quantitative representativity in the following and we shall deal only with qualitative representativity.

3. Invalid Bigrams

Our starting point is the search for "invalid (impossible) bigrams", that is, for configurations ***[First,Second]*** of tags which cannot occur as tags of two words following immediately each other in a correct text of a particular language (in English, e.g., the bigram *[ARTICLE, FINITE VERB]*). Such invalid bigrams as a rule occur in a realistic large-scale PoS-tagged corpus, for the following reasons:

- in a hand-tagged corpus, an invalid bigram results from (and unmistakeably signals) either an ill-formed text in the corpus body (including wrong conversion) or a human error in tagging
- in a corpus tagged by a statistical tagger, an "invalid bigram" may result also from an ill-formed source text, as above, and further either from incorrect tagging of the training data (i.e. the error was seen as a "correct configuration (bigram)" in the training data, and was hence learned by the tagger) or from the process of so-called "smoothing", i.e. of assignment of non-zero probabilities also to configurations (bigrams, in the case discussed) which were not seen in the training phase[9].

From a linguistic viewpoint, a *(linguistically) valid bigram* is a pair of tags ***[First,Second]*** in a certain natural language if and only if there exists a sentence (at least one) in this language which

[6] The definitions of positive and negative representativity are obviously easily transferable to cases with other definitions of a phenomenon. Following this, the definition of qualitative representativity holds of course generally, not only in the particular case of a corpus representative wrt. bigrams.

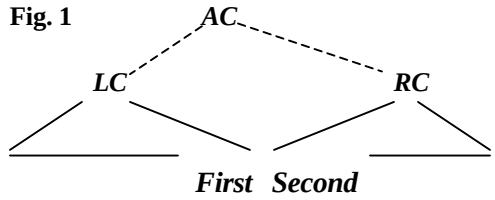
[7] This assertion holds only on condition that each sentence of the language is of length two (measured in words) or longer. Similarly, a corpus qualitatively representative wrt. trigrams is qualitatively representative wrt. bigrams and wrt. unigrams only on condition that each sentence is of length three at least, etc.

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

[9] This "smoothing" is necessary in any purely statistical tagger since – put very simply – otherwise configurations (bigrams) which were not seen during the training phase cannot be processed if they occur in the text to be tagged.

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").

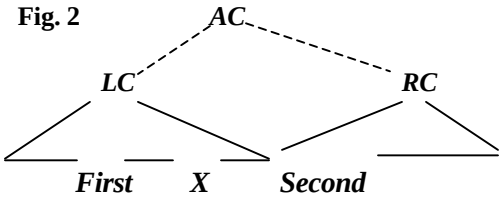
Fig. 1



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 simplified[10] syntactic 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 ***X*** (cf. Fig. 2)

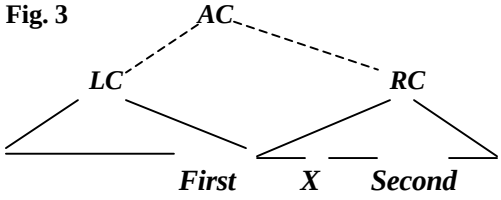
Fig. 2



(example: the bigram *[ARTICLE, FINITE VERB]* is impossible in German since in any ***LC*** – *NP*′s, *PP*′s, *S*′s etc. – an 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 ***X*** (cf. Fig. 3)

Fig. 3

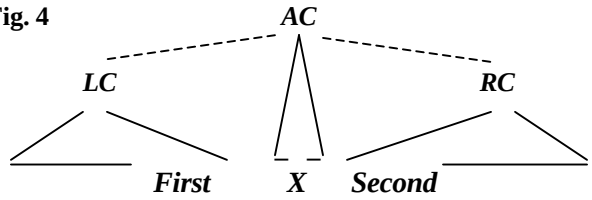


(example: the bigram *[SEPARABLE VERB PREFIX, POSTPOSITION]* is impossible in German since in any ***RC*** – *NP*′s, *PP*′s, *S*′s etc. – a postposition must combine with some preceding lexical material displaying (morphological) case before such a constituent can be combined with any other material into a higher unit

[10] This simplification is due to the implicit assumption that the syntactic structure of the language in question has a context-free backbone (the language does not allow for non-projective dependencies, on a dependency-based approach to syntax). Since this is not generally true, the results obtained on the basis of such considerations have to be revised – however, they constitute a very solid ground for a survey of the invalid bigrams in practice.

- the configuration from Fig. 1 is impossible because ***LC*** and ***RC*** can never occur as adjacent sisters standing in this order – cf. Fig. 4

Fig. 4



(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).

For a particular language with a particular tagset, the set of invalid bigrams[11] 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 (Schiller et al., 1999). 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 hundreds. 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 tens of 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 the common hand-tagging of a reasonably sized training corpus, and it is also 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, because:

- 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 (which, e.g., for a trigram-based tagger means that any trigram ***[First,Second,Third]*** containing an invalid bigram – i.e. if ***[First,Second]*** or ***[Second,Third]*** are invalid bigrams – should be assigned probability 0 (zero), and this also after smoothing or any similar actions are preformed).

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

- first of all, as it is obvious already from the wording, 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, since each such corpus came into being as

[11] The categorization of a particular invalid bigram into one of the classes depends obviously on the shape of constituent structure adopted. However, different categorization cannot change the fact of the invalidity of the particular bigram.

a collection of real texts gained and converted from newspaper publishers or publishing houses, and as such it contains typographical, grammatical or conversion errors.

Taking 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 an 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 case to occur 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 invalid bigram has to be hand-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 source of the invalid bigram is a tagging error, a legacy data error (i.e. error in the original text) or a metagrammatical usage, can be performed solely on the basis of (human) 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 training 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 training 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.

4. Extending the Invalid Bigrams

The invalid bigrams are a powerful tool for checking the correctness of a corpus, however, a tool which works on a very local scale only, since it is able to diagnose solely errors which are detectable as deviations from the set of possible pairs of tags standing adjacently. Thus, obviously, quite a number of "non-local" errors remain undetected by such a strategy. As an example of such an as yet "undetectable" error in German we might take the configuration where two words tagged as finite verbs are separated from each other by a string consisting of nouns, adjectives, articles and/or prepositions only. In particular, such a configuration is erroneous since the rules of German orthography require that some kind of clause separator (comma, dash, coordinating conjunction) occur inbetween two finite verbs[12].

In order to be able to detect also such kind of errors, the above invalid bigrams have to be extended substantially. The search for the generalization needed can be guided by the linguistic view on the invalid bigrams which has been introduced in the Figs. 2-4 above, in other words, by the deeper insights into the impossibility for a certain pair of PoS-tags to occur immediately following each other in any linguistically correct and correctly tagged sentence.

[12] At stake are true regular finite forms, exempted are words occurring in fixed collocations which do not function as heads of clauses. As an example of such usage of a finite verb form, one might take the collocation *wie folgt*, e.g., in the sentence *Diese Übersicht sieht wie folgt aus: …* Mind that in this sentence, the verb *folgt* has no subject, which is impossible with any active finite verb form of a German verb subcategorizing for a subject (and possible only marginally with passive forms, e.g., in *Gestern wurde getanzt*, or – obviously – with verbs which do not subcategorize for a subject, such as *frieren, grauen* in *Mich friert, Mir graut vor Statistik*).

The point is that an invalid bigram indeed does not come into being by chance but rather as a violation of a certain – predominantly syntactic[13] – rule(s) of the language. In particular, such a violation is usually a *violation of constituency.*

Thus, if the source of the invalidity of the bigram is missing the material ***X*** in situation as depicted in Fig. 2, it means that the constituent ***LC*** is incomplete (its constituency is violated). If the invalid bigram results from missing material ***X*** which should occur under ***RC***, as sketeched in Fig. 3, then the constituency of ***RC*** is obviously violated. Finally, if the source of the invalidity of the bigram is the absence of the material ***X*** depicted in Fig. 4, then it is the violation of the constituency of ***AC*** which is at stake.

As an example of a configuration breaking the constituency of ***LC*** (from Fig. 2), we might consider the bigram *[PREPOSITION,FINITE VERB]* (possible German example string: *...für-PREPOSITION reiche-FINITE VERB...*)[14]. From this it follows that either there is indeed an error in the source text (in our example, probably a missing word, e.g., *Der Sprecher der UNO-Hilfsorganisation teilte mit, für ~~Arme~~ reiche diese Hilfe nicht.*) or there was a tagging error detected (in the example, e.g., an error as in the sentence ... *für reiche Leute ist solche Hilfe nicht nötig...*). The source of the error in both cases would be a violation of the linguistic rule postulating that, in German, a preposition must always be followed by a corresponding noun (*NP*) or at least by an adjectival remnant of this *NP*[15].

The central observation lies then in the fact that the property of being an impossible configuration can often be retained also after the components of the "loosened invalid bigram" get separated in the string by other words occurring inbetween them. In particular, for an invalid bigram ***[First,Second]*** it holds that such a configuration remains incorrect also after the addition of some material inbetween the elements ***First*** and ***Second*** unless the material added is exactly ***X***, in other words, the configuration ***First - STRING - Second*** is invalid for ***STRING*** of any length on condition that ***STRING*** does not contain ***X*** (understood as the material depicted in Figs. 2-4).

Thus, e.g., in the example of the invalid bigram *[PREPOSITION, FINITE VERB]* immediately above, the property of being an impossible configuration is conserved if a conjunction is placed inbetween, creating thus an "invalid trigram". In particular, the configuration *PREPOSITION - CONJUNCTION - FINITE VERB* cannot be a valid trigram, exactly for the same reasons as *[PREPOSITION, FINITE VERB]* was not a valid bigram: *CONJUNCTION* is not a valid *NP* remnant. An additionaly important observation is then that not even two, three and in fact any number of conjunctions would make the configuration grammatical and hence would disturb the error detection potential of the "extended invalid bigram" *[PREPOSITION, FINITE VERB]*.

These linguistic considerations have a straightforward practical application. Provided a qualitatively representative (in the above ideal sense) corpus is available for training, it is possible to construct the set of invalid bigrams. Then, for each bigram ***[First,Second]*** from this set, it is possible to collect all trigrams of the form ***[First,Between,Second]*** occurring in the corpus, and collect all the possible tags ***Between*** in the set ***Possible_Inner_Tags***. Furthermore, given the invalid bigram ***[First,Second]*** and the respective set ***Possible_Inner_Tags***, the training corpus is to be searched for all tetragrams ***[First,Middle_1,Middle_2, Second]***. In case one of the tags ***Middle_1, Middle_2*** occurs already in the set ***Possible_Inner_Tags***, no action is to be taken, but in case the set ***Possible_Inner_Tags*** contains neither of ***Middle_1, Middle_2***, both the tags ***Middle_1*** and ***Middle_2*** are to be added into the

[13] Examples of other such violations are rare and are related mainly to phonological rules. In English, relevant cases would be the word pairs *an table, a apple*, provided the tagset were so fine-grained to express such a distinction, better examples are to be found in other languages, e.g. the case of the Czech ambiguous word *se*, cf. (Oliva, to appear).

[14] Unlike English, (standard) German has no preposition stranding and similar phenomena – we disregard the colloquial examples like *Da weiss ich nix von* – and hence, examples parallel to the English *The man Mary was waiting for-PREP came-VFIN late* are impossible in German.

[15] Again, this statement is not fully exact, since prepositions can create a *PP* also with certain (but by far not all) adverbs, e.g. *seit gestern, bis morgen, von dort*. This is to be taken care of lexically, since the class of such adverbs is strictly limited. Also, German prepositions can create *PP*'s with other prepositional phrases, cf. the example *eine Tonnage von bis zu über 200.000 BRT*. This, however, has no bearing on our example.

set ***Possible_Inner_Tags***. The same action is then to be repeated for pentagrams, hexagrams, etc., until the maximal length of sentence in the training corpus prevents any further prolongation of the *n*-grams and the process terminates.

If now the set ***Impossible_Inner_Tags*** is constructed as the complement of ***Possible_Inner_Tags*** relatively to the whole tagset, then any *n*-gram consisting of the tag ***First***, of any number of tags from the set ***Impossible_Inner_Tags*** and finally from the tag ***Second*** is very likely to be an *n*-gram impossible in the language and hence if it occurs in the corpus whose correctness is to be checked, it is to be signalled as a "suspect spot". Obviously, this idea is again based on the assumption of qualitative representativity of the training corpus, so that for training on a realistic corpus the correctness of the resulting "invalid *n*-grams" has to be hand-checked. This, however, is well-worth the effort, since the resulting "invalid *n*-grams" are an extremely efficient tool for error detection. The algorithmic implementation of the idea is a straightforward extension of the above approach to "invalid bigrams" – the respective bootstrapping algorithm in a semi-formal coating looks like as in Fig 5.

```
integer n, maximal_sentence_length_in_corpus;
set_of_tags possible_i_t, impossible_i_t, tagset;
forall invalid_bigram [First,Second]
  {  n := 3;
     possible_i_t := ∅;
     while  n =< maximal_sentence_length_in_corpus
      do {find all inner-sentential n-grams [First,V1,V2, ..,Vn-2,Second];
          for each n-gram found
           do if {V1, V2, .., Vn-2} ∩ possible_i_t = ∅
              then possible_i_t := possible_i_t ∪ {V1,V2,..,Vn-2};
          n := n + 1;
         };
     impossible_i_t([First,Second]) :=  tagset – possible_i_t;
  }
```

Fig. 5: Algorithm for bootstrapping negative *n*-grams

The above approach does not guarantee, however, that all "invalid *n*-grams" of a language are generated. In particular, any "invalid trigram" ***[First,Second,Third]*** cannot be detected as such (i.e. as invalid) if the ***[First,Second]***, ***[Second,Third]*** and ***[First,Third]*** are all possible bigrams. Such an "invalid trigram" in German is, e.g., *[NOMINATIVE NOUN, FINITE VERB, NOMINATIVE NOUN]* - this trigram is invalid[16] since no German verb apart from *sein/werden* (which are not tagged as main verbs in NEGRA) can occur in a context where a nominative noun stands both to its right and to its left, however, all the respective bigrams occur quite commonly (e.g., *Johann schläft, Jetzt schläft Johann, König Johann schläft*).

5. The Error-detection Potential of the Invalid Bigrams in Practice

Employing the invalid bigrams (including the extensions described) as an error-detection technique, we were able to correct 3.773 errors in the NEGRA corpus, and we can guarantee that the corrected version of the corpus is negatively representative wrt. bigrams based on the STTS tagset. Since we

[16] This is again a slight simplification. A genuine impossible configuration is only the tetragram *[BEGINNING OF SENTENCE, NOMINATIVE NOUN, FINITE VERB, NOMINATIVE NOUN]*. Even from such a configuration, quotations and other metalinguistic contexts, such as *Der Fluss heisst Donau, Peter übersetzte Faust - eine Tragödie ins Englische als Fist - one tragedy,* are to be exempted. These are, however, as a rule lexically specific and hence can be coped with as such.

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). The whole resulted in changes on 4.243 lines of the corpus; the rectification of errors in segmentation resulted in reducing the number of corpus positions by over 700, from 355.096 to 354.354.

Based on this, we were able to confirm experimentally the expected fact that the quality (i.e. representativity) of the training corpus has a paramount importance for the quality of a statistical tagger trained on this corpus. In particular, after finishing the corrections we experimented with training and testing the TnT tagger (Brants, 2000) on the "old" and on the "corrected" version of NEGRA. We used the same testing as described by Brants, i.e. dividing each of the corpora into ten contiguous parts of equal size, each part having parallel starting and end position in each of the versions, and then running the system ten times, each time training on nine parts and testing on the tenth part, and finally computing the mean of the quality results. In doing so, we arrived at the following results:

- if both the training and the testing was performed on the "old" NEGRA, the tags assigned by the TnT tagger differed from the hand-assigned tags within the test sections on (together) 11.138 positions (out of the total of 355.096), which yields the error rate of 3,14%
- if both the training and the testing was performed on the "correct" NEGRA, the tags assigned by the TnT tagger differed from the hand-assigned tags of the test sections on (together) 10.889 positions (out of the total of 354.354), which yields the error rate of 3,07%
- in the most interesting final experiment, the training was performed on the "old" and the testing on the "correct" NEGRA; in the result, the tags assigned by TnT differed from the hand-assigned tags in the test sections on (together) 12.075 positions (out of the total of 354.354), yielding the error rate of 3,41%.

These results show that there was only a negligible (and, according to the χ^2 test, statistically insignificant) difference between the results in the cases when the tagger was both trained and tested on "old" corpus and both trained and tested on the "corrected" corpus. However, the difference in the error rate obtained when the tagger was once trained on the "old" and once on the "corrected" version, and then in both cases tested on the "corrected" version[17], brought up a significant relative error improvement of 9,97%. This improvement documents the old and hardly surprising truth that – apart from the size – also the correctness of the training data is absolutely essential for the results of a statistical tagger.

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. As said above, there exist invalid trigrams ***[First,Second,Third]*** which cannot be detected as such (i.e. as invalid) by the method (even with the "generalized" invalid bigrams). 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 invalid trigram candidates ***[First,Second,Third]*** which contain an invalid bigram ***[First,Second]*** or ***[Second,Third]*** can be discarded automatically from the search space (these are invalid as bigrams, hence they are certainly also invalid as trigrams)
- all invalid trigram candidates ***[First,Second,Third]*** which have been discovered as "valid extended bigrams" (e.g., by the algorithm given in Fig. 5) are to be eliminated automatically from the search space, too, since they are already known to be possible trigrams.

Also, 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

[17] For obvious reasons, we did not even consider training on the "corrected" corpus and testing on the "old" one.

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 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 in German:

ohne wenn und aber	*Augen zu und durch*	*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*
dann und wann	*von einst*	*hin und wieder*
zu eigen machen	*dicht an dicht*	*von neuem*
Vorhang auf	*oben ohne*	

Given such idioms are dealt with properly, it is then possible to define the set of all invalid bigrams of German. In the following, we put forward the list of such (simple, non-generalized) 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 $X \neq$ *$, , $(, 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 $X \neq$ *$, , $(, KON, APPR, NN, NE*
- *[PRELAT,X]*: *PRELAT* must necessarily be followed by an *NP* (or at least by a remnant of an *NP*), so that *X* must be a tag marking a word which possibly can start an *NP*, hence tags *APPO, APZR, KOUS, PTKVZ, VVFIN, VVIMP, VVINF, VAFIN, VAIMP, VAINF, VMFIN, VMINF* are ruled out, and further impossible are also the following ones: (i) *$.* (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 at the beginning of a sentence (*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[18], 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

[18] 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. ☹

impossible; however, examples like *Er handelte den Vertrag mit aus* cast serious doubts on such statements

- *[X,VVIMP/VAIMP]:* Imperative[19] 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 $X \neq$ *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:* any of the pairs given by the Carthesian product
  {VAFIN,VMFIN,VVFIN,VAIMP,VMINP,VVIMP} x *{VAFIN,VMFIN,VVFIN,VAIMP,VMINP,VVIMP}*
  is impossible (it is an invalid bigram)
- *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. We did not pursue the search for such configurations systematically, but rather on an intuitive basis only. As examples of invalid trigrams we used 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}*

[19] STTS contains no tag for an imperative of a modal verb – hence only *VVIMP/VAIMP* is mentioned.

- *[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* !).

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 !).

6. Conclusions and Perspectives

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. 5 – 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[20] 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.

[20] even when trivial sounding

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

Hirakawa H., Ono K. and Yoshimura Y. (2000) *Automatic refinement of a PoS tagger using a reliable parser and plain text corpora*, in: Proceedings of the 18th Coling conference, Saarbrücken

Müller F.H. and Ule T. (2001) *Satzklammer annotieren und tags korrigieren: Ein mehrstufiges top-down-bottom-up System zur flachen, robusten Annotierung von Sätzen im Deutschen*, in: Proceedings der GLDV-Frühjahrstagung 2001, Gießen

NEGRA Corpus. For more information cf. `www.coli.uni-sb.de/sfb378/negra-corpus`

Oliva K. (2001) *The possibilities of automatic detection/correction of errors in tagged corpora: a pilot study on a German corpus*, in: 4th International conference "Text, Speech and Dialogue" TSD 2001, Lecture Notes in Artificial Intelligence 2166, Springer, Berlin 2001

Oliva K. and Květoň P. (2002) *Corpus Representativity, Bigrams, and PoS-Tagging Quality,* in: Proceedings of the Conference KONVENS 2002, Saarbrücken

Oliva K. (to appear) *Linguistics-based tagging of Czech: disambiguation of 'se' as a test case*, in: Proceedings of 4th European Conference on Formal Description of Slavic Languages held in Potsdam from 28th till 30th November 2001

Petkevič V. (2001) *Grammatical agreement and automatic morphological disambiguation of inflectional languages*, in: 4th International conference "Text, Speech and Dialogue" TSD 2001, Lecture Notes in Artificial Intelligence 2166, Springer, Berlin 2001

Schiller A., Teufel S., Stöckert C. and Thielen C. (1999) *Guidelines für das Tagging deutscher Textcorpora*, University of Stuttgart / University of Tübingen, also available at `www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html`

Skut W., Krenn B., Brants T. and Uszkoreit H. (1997) *An annotation scheme for free word order languages*, in: Proceedings of the 3rd Applied Natural Language Processing Conference, Washington D.C.