Originalveröffentlichung in: Gippert, Jost/ Gehrke, Ralf (Hrsg.):
Historical Corpora. Challenges and Perspectives.–
Tübingen: Narr, 2015. S. 65-76. (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 5)

ROLAND MITTMANN

Automated quality control for the morphological annotation of the Old High German text corpus

Checking the manually adapted data using standardized inflectional forms

Abstract

The project *Referenzkorpus Altdeutsch* ('Old German Reference Corpus') aims to establish a deeply-annotated text corpus of all extant Old German texts. As the automated part-of-speech and morphological pre-annotation is amended by hand, a quality control system for the results seems a desirable objective. To this end, standardized inflectional forms, generated using the morphological information, are compared with the attested word forms. Their creation is described by way of example for the Old High German part of the corpus. As is shown, in a few cases, some features of the attested word forms are also required in order to determine as exactly as possible the shape of the inflected lemma form to be created.

1. Introduction

When a morphologically annotated text corpus has been created without entirely automating the part-of-speech and inflectional annotation, an automated verification of these attributes appears to be a helpful approach for quality control. If the word tokens are already attributed to their lemmata, every lemma can be provided with the expected suffixes and endings (or also have its word stem altered according to the grammatical rules), and the result can be compared with the actually attested form. The choice of a consistent *sprachform* for the whole corpus would not only facilitate the creation of these standardized inflectional forms, but also facilitate search queries for specific inflectional forms of lemmata: an idealized Old High German word form *brungut* 'you (pl.) brought' – irrespective of its actual attested shape – is easier to key in than a lemma *bringan* plus the information that all forms of the second person indicative plural inflected according to the strong class IIIa are searched for. Moreover, for some kinds of corpus analyses, being able to neglect phonological and morphological variation will help to focus on the effective research question.

In the project *Referenzkorpus Altdeutsch* ('Old German Reference Corpus'), this course of action is applied. The project aims to produce a deeply-annotated corpus of all preserved texts from the oldest stages of German – Old High German (OHG) and Old Saxon (OS) –, which date from ca. 750 to 1050 CE. Comprising a total of about 650,000 word tokens, the corpus covers interlinear translations of Latin texts as well as free translations, adaptations and mixed German-Latin texts. These are complemented by a few texts composed entirely in an Old German language. The largest coherent sub-corpora are the OHG works of Notker and Otfrid of Weissenburg, an OHG translation of the Gospel harmony of Tatian, and the versified OS Gospel harmony known as the *Heliand*.

Since a consistent *sprachform* for OHG and OS could only be conceived in reconstructed Proto-West Germanic, and would thus differ too much from the attested word forms, the establishment of two different standards for the two languages was chosen. Below, the approach for OHG will be described.

2. The definition of a consistent *sprachform*

Whilst some OHG dictionaries use no consistent lemmatization, Splett (1993) does so for his *Althochdeutsches Wörterbuch*. As OHG is a "Sprachstufe, die keine allgemein gültige Leitvarietät besitzt",[1] Splett (1993: XXVIII) chooses "die Idealform des Ostfränkischen, das der Tatian überliefert"[2] (ibid.). It is therefore necessary to have a grammar that displays the inflectional endings in a language stage close to the one of Splett's lemmata. Fortunately, Braune (2004: 6) gives a description similar to Splett concerning his grammar: "In diesem Buch wird [...] die ostfrk. [ostfränkische] Sprache des ahd. [althochdeutschen] Tatian (2. Viertel 9. Jh.) zugrunde gelegt."[3] – Thus, these two works seem very appropriate to our task.

Notwithstanding the above, a dictionary cannot cover all grapho-phonematical particularities: in several declensional classes, the ending of the dative plural is *-um* (or *-om*) in the oldest OHG texts and appears as *-un* or *-on* in the Tatian (cf. Braune 2004: 185). Sievers (1892: LXVIII) shows that the preference for either form depends mainly on the individual scribe. In total, he counts *-un*

[1] I.e., a "language stage that possesses no universally valid ideal form".

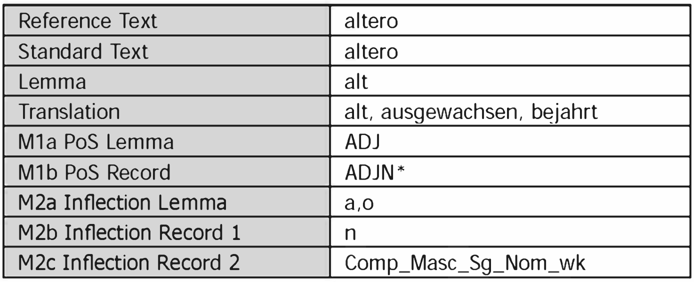
[2] I.e., "the ideal form of the East Franconian [language] that the Tatian passes down".

[3] "In this book, the East Franconian language of the Old High German Tatian (2nd quarter of the 9th century) is taken as a basis."

119 times and -*on* 113 times. Even the adverbialized dative plural forms in the Tatian have a similar distribution (ibid.). Due to the lack of more criteria, <u> as the older one of the two vowels (cf. Braune 2004: 62) was eventually chosen. In contrast to this, the merger of final *m* and *n* into *n* in inflexional endings – almost completed in the Tatian (cf. Sievers 1892: XXVIIf.) – was disregarded, as it could mark a semantic difference in some cases: e.g. *nāmum* 'we took' vs. *nāmun* 'they took; *later also:* we took' (cf. Braune 2004: 271).[4]

3. Description of the data at hand

The generation of the standardized inflectional forms starts from ELAN files, each containing a part of an OHG subcorpus. ELAN[5] is software for the manual adaptation of multilevel annotations developed by the Max Planck Institute for Psycholinguistics at Nijmegen (the Netherlands). One example of an annotated word token in ELAN is shown in Figure 1.



Reference Text	altero
Standard Text	altero
Lemma	alt
Translation	alt, ausgewachsen, bejahrt
M1a PoS Lemma	ADJ
M1b PoS Record	ADJN*
M2a Inflection Lemma	a,o
M2b Inflection Record 1	n
M2c Inflection Record 2	Comp_Masc_Sg_Nom_wk

*) ADJN = Adjektiv, nachgestellt (adjective, postpositive)

Figure 1: Sample word annotation from the Tatian in ELAN format – before adapting the standardized inflectional forms (simplified representation)

The reference text comes from the printed edition of the text. Before the creation of the standardized text, the second tier contains only padding, i.e. copies of the reference text. The lemma and its (Modern German) translation are given as listed by Splett (1993). If several possible translations are given, only the appropriate one is adopted. For the remaining five tiers, a specially developed tagset is used, and their information comes from a range of glossaries, each of which covers a section of the corpus.[6] They contain the part-of-speech

[4] Splett (1993) apparently made the same decision, documented by expressions like *io dèm wîlòm* 'on the spot' from *wîla* '(a) while' (p. 1124).

[5] www.lat-mpi.eu/tools/elan. All URLs have been checked and found valid as of late January 2015.

[6] Heffner (1961), Hench (1890, 1893), Kelle (1881), Sehrt (1955) and Sievers (1874, 1892).

types of lemma (M1a) and record (M1b), the lemma-dependent inflectional information pertaining to the lemma (M2a), the corresponding information of the individual record (M2b), and its record-specific inflectional information (M2c). The files have been automatically created using the OHG corpus on the TITUS website.[7]

The standardized inflectional forms are created from the standardized lemma, plus the inflectional information. The glossaries have been digitized and linked to the texts to save extensive manual work. However, their information is not always exhaustive and properly transferable. Because of this, the morphological information gained from the glossaries has to be disambiguated, corrected, and completed by hand before the standardized inflectional forms can be generated.[8]

4. Preparation of the data

To begin with the creation of the standardized inflectional forms, the files are again edited automatically. For every word token, its variant as occurring in the printed edition and its standardized lemma are extracted, as well as the morphological data with the exception of the lemma-dependent inflectional information pertaining to the lemma (M2a). The gender information concerning nouns, a part of M2b, is extracted to make M2b represent the inflectional class only.

In order to facilitate the replacements, two digraphs are substituted: <qu> is shortened to <q> to avoid <u> being treated as a vowel, and <ck> is replaced by <kk>, as doubled consonants are often simplified (cf. Section 5.1). When all replacement rules have been executed, these substitutions are undone. So, to the lemma *bok* 'buck', a genitive *bockes* will be created.

[7] http://titus.uni-frankfurt.de/texte/texte2.htm#ahd.

[8] A description of the digitization of the dictionaries can be found in Mittmann (2013). The extraction and processing of the glossary data, the adjustment of their lemmata to Splett (1993), the combination of the data with the TITUS text files, and the subsequent manual annotation of the ELAN files is treated of in detail in Linde/Mittmann (2013).

5. Generation of the standardized inflectional forms

The generation of the standardized inflectional forms is performed according to part-of-speech categories. There are two phenomena to be observed that cannot be attributed to a single category: consonantal simplification and the emergence of secondary vowels.

5.1 Consonantal simplification and secondary vowels

In the standard used, double consonants are simplified before consonants and at the end of a word. If, however, a word ends in a single consonant, it is not possible to determine whether this consonant is doubled in inflectional forms. Since Splett gives no information on that, a list of these words was needed.

The lemmata to be considered all feature a vowel or diphthong preceding the final consonant,[9] and the consonant is only simplified in final or pre-consonantal position. Within the automatically created list of these words, the test is then performed by hand, and the final list is set up: recorded inflectional forms and derived words that have a vowel following the original word's final consonant are checked as to whether the consonant is doubled or remains simple.

In some rare cases, several lemmata have an identical form but differ with regard to the doubling. If there is a categorial difference between them, this can be used: the <r> of *far* is only doubled if this is a masculine noun ('bull'), but not if it is neuter ('harbour', 'beacon') or an adjective ('going'). Otherwise, the decision is made according to the spelling in the printed edition. This is, however, considered as a makeshift solution, as it compromises the ambition to create the standardized inflectional forms only from the standard lemma and the morphological information of the record.

The emergence of so-called secondary vowels, as in *zeihhan* 'sign, token', developed from Proto-Germanic **taiknan* with regular apocope of the original ending *-an*,[10] leads to stem changes within the affected paradigms. However, it is often analogically transferred to other places in the paradigm: for the genitive plural, the Tatian has *zeihno*, *zeihhano* and *zeichano* (cf. Sievers 1892: 510). To

[9] Although double consonants were simplified after long vowels or diphthongs in the course of the OHG era, Splett (1993) still gives their doubled form – as with *lûttar* 'limpid, noble' (p. 577) –, so this case cannot be excluded.

[10] The secondary vowels arose before a resonant where the West Germanic resonant became syllabic due to loss of a vowel (cf. Braune 2004: 68).

avoid unnecessarily complicated paradigms for the inflected lemmata, the presence or absence of secondary vowels is assumed in accordance with the lemma.

5.2 Nouns (substantives)

With regard to the generation of the standardized inflectional forms, a first section converts the nouns, including the cardinal *dūsunt* 'thousand' and the pronouns *man* 'one' and *wiht* 'something', which inflect equally. The information on number and case is extracted from M2c. The addition of endings is performed depending on the inflectional class, and in some cases also on the gender. Some classes also trigger umlaut from <a> to <e>.[11] Here, the last <a> of the word, followed only by consonants (plus another vowel, if existing) before the end of the word, is converted, so that a nominative plural *gesti* is created from the masculine *i*-stem *gast* 'guest'. Where it applies, consonantal doubling is performed, changing *bal* 'ball' into a genitive *balles*, but converting *wal* 'whale' into *wales*.

Since the rules needed for the genitive singular are very similar to those for the dative, they are treated together – and subsequently, only two rules are needed to transform any genitive into a dative form: final *-s* is removed, and final *-a* is replaced by *-u*. The accusative and instrumental cases are again created from the nominative. In the plural, a couple of rules can be collectively applied to all case forms; e.g. the addition of the *-ir*-suffix to the *z*-stems. For pluralia tantum, a correct inflection is also taken care of. Proper nouns of Latin origin ending in *-us* may keep or lose this ending when inflected: *Petrus* 'Peter' has both *Petre* and *Petruse* as its dative (cf. Sievers 1892: 298). Here, the attested records have to serve as the decisive aspect.

5.3 Adjectives, participles, determiners, pronouns, and numerals

A second section converts the remaining adjectives, participles, determiners, pronouns, and numerals. Grade, gender, number and inflectional type are extracted from M2c. The creation of the base forms of the participles is treated together with the verbs, but their inflection together with that of the adjectives.

[11] Henceforth, the term "umlaut" will be used only to describe an umlaut from <a> to <e>. Non-umlauted /e/ is written as <ë>.

Before the inflectional endings are attached, consonantal doubling is performed where required, and the comparative and superlative suffixes are added to the stems. In the case of lemmata with irregular comparison, the stems are replaced. Since there is no definite rule to determine whether *-ōr-* or *-ir-* (the latter also triggering umlaut) are added as the comparative suffix – and *-ōst-* or *-ist-* for the superlative, respectively –, this has once more to be decided according to the word form in the text, depending on whether the token contains *-or-/-ōr-* or *-ost-/-ōst-*. Thus, in Figure 1, *eltiro* is generated from *alt* 'old' and inserted as the standardized inflectional form, as the record *altero* does not show <ōr> or <or>. This criterion has been chosen because /ō/ usually retains its quality in middle syllables, whilst short /i/ is not uncommonly rendered as <e> (cf. Braune 2004: 64-66). Some stems had to be adjusted to the attachment of inflectional endings; e.g., *gëlo* 'yellow' has a stem *gëlaw-*. Consonantal doubling also occurs: *smal* 'small' remains *smal-*, but *snël* 'quick' becomes *snëll-*.

De-adjectival adverbs are formed by adding the ending *-o* to the stem; *-i*-stem adjectives lose this vowel and their umlaut (*herti* 'hard' becomes *harto*). In the comparative and superlative, there is no ending at all. Apart from one adverbial positive that deviates from the corresponding adjective (*wola* 'well' from *guot* 'good'), comparative and superlative forms are created similarly to those of the adjectives. One exception is that all adverbial comparatives feature *-ōr-* (cf. Braune 2004: 232).

Various pronouns and some of the cardinal numbers inflect divergently and are dealt with separately. The other cardinals do not inflect at all. *ein* 'one, a', *beide* 'both' and the ordinals behave like adjectives. The possessive pronouns *unsēr* 'our' and *iuwēr* 'your' feature both long and short stems – *unser-/iuwer-* and *uns-/iuw-* – and have to be treated according to the attested word forms.

5.4 Verbs

For the verbs, from M2c, mood, tense, number and person are extracted. As the imperative only exists in the present and has forms differing from the indicative only in the singular, the two moods are treated together, and only in the singular are their forms determined separately. All modifications are carried out on the highest possible level, so the umlaut-like vowel changes of strong verbs from <io> to <iu> and from <ë> to <i>, which hold for the whole singular of the indicative and the imperative, need to be declared only once.

In the present tense, the first person indicative singular shows exemplarily that individual replacement rules for preterit-presents and other irregular verbs are required. The rules for weak and strong verbs are rather straightforward. The second and third person forms exhibit umlaut (except for weak verbs inflecting according to class II or III) as well as simplification of stem-final double consonants with some verbs featuring an infinitive ending in *-en*. For most consonants, the application of this rule depends on the cluster's history (cf. Braune 2004: 295f.), and as the evidence is again numerous for both cases, the word form from the printed edition has to be checked for a doubled consonant. So, *stellen* 'put' results in a form *stellis*, but *zellen* 'count, tell' gives *zelis*. Here, this approach is especially risky, as consonantal doubling or simplification is not rare in OHG inflectional morphology.

The imperative shows the same rule for the consonantal simplification of verbs in *-en* before the ending *-i*, but furthermore, verbs in *-an*, featuring a zero ending, exhibit a general simplification of final double consonants. In the plural and the whole subjunctive, fewer rules are needed, as only the personal endings deviate from the lemma.

In the past tense, weak verbs and preterit-presents have an *-(i)t*-suffix, and weak class Ia verbs feature consonantal simplification and absence of umlaut: *brennen* 'burn' becomes *branta*. The stem changes in the past tense require various replacement rules: not only do the stem vowels of all strong verbs change, but alternations also affect the consonants of some of these verbs. Apart from consonantal doubling and simplification, the most frequent reason for this is Verner's Law,[12] as shown by the example of the past tense form *wārun*, belonging to *wësan* 'be'.

For most inflectional classes, Verner's Law does not affect the first and third person indicative singular. Nevertheless, it is more straightforward to apply the same rules to all other cases than to apply them to the whole past and undo them in these two cases – although the opposite would also be possible,[13] given that all cases of two lemmata, one of which is affected by Verner's Law and has

[12] Verner's Law describes a peculiar consonantal mutation of Proto-Germanic: voiceless fricatives were voiced unless being word-initial or immediately preceded by the Indo-European word accent (cf. Verner 1877: 114). In OHG, the voiced fricatives have evolved to plosives, and /z/ has become /r/.

[13] In stem classes where the mentioned two forms are unaffectable by Verner's law, they also have a deviating stem vowel.

the same past tense stem as the other one, are ruled out in OHG.[14] These cases constitute some of the instances of the elimination of Verner's Law within verbal paradigms in OHG, raising the question of whether it is more useful for possible queries in the corpus to keep the standard word forms close to the degree of perpetuation of Verner's Law observable in the Tatian; or to consider its execution in its original extent as the standard, so that all deviations from this could be more easily examined.

Most past participles have a prefix *gi-*, but there is no definite rule to discern whether this applies; so again, the word form appearing in the printed edition must dictate. The prefix may appear with <k>, <c> or <ch> instead of <g> and with <e> or <a> for <i>; this has to be considered as well. As some verbs have another prefix before the *gi-*, the correct position for the possible insertion has to be determined. First, it is checked whether the prefix appears within the word and the lemma lacks the digraph <gi>. If not, no insertion is performed. But if so, as for *nidargiuualzten* (cf. Sievers 1892: 486) from *nidarwelzen* 'bow down' (cf. Splett 1993: 1091), it is checked whether the grapheme of the word form directly after the prefix *gi-* is the same as the grapheme of the lemma at the same position, but without the prefix *gi-*. If this is the case, (-)*gi-* is added; otherwise – that is, if the initial grapheme of the verbal stem is different or the other prefix differs in length – the same check is performed again one by one with a list of initial graphemes or digraphs that often differ between record and lemma. Here, it is checked whether the altered grapheme has the same position in the record (after the *gi*-prefix) as the original one in the lemma without the *gi*-prefix. So from the lemma *nidarwelzen*, the M2b information wk1a, the M2c information P_Pos_Masc_Pl_Dat_st[15] and the record *nidargiuualzten*, a standard word form *nidargiwalztēm* is created. Needless to say, if the record, but not the lemma, begins with the prefix *gi-*, it is added in any case. If this still does not help, another test ascertains whether the initial grapheme is shifted within the record to either direction (for instance, if the prefix *aba-* 'off' appears as *ab-*), and the variants of initial graphemes are tried out also. For the verbs that contain <gi> and are thus not covered by the test, it is performed again completely, taking this aspect into account. In the event that no possibility is found of placing the *gi-*, it is omitted and left to the manual annotation.

[14] *fliohan* 'flee' should have †*flug-*, but this is normalized to *fluh-*, since *fliogan* 'fly' also has *flug-* (cf. Braune 2004: 279).

[15] P: pronominal inflection.

Apart from the prefix, the past participles behave similarly to the past tense stems, and inflect like adjectives. Those of the weak class Ia (and adjectives based on them) have an -*it*-suffix, but lose the <i> and undergo the same stem changes as in the past tense if they acquire an inflectional ending: *gibrennit* 'burnt' becomes *gibrant*-. The base form of the present participle is generated from the infinitive plus -*ti* (with umlaut). Substantivized infinitives inflect like *a*-stem nouns, with their final <n> being doubled.

6. Final adaptation and insertion of the standardized inflectional forms

For the remaining part-of-speech categories, the lemmata remain unaltered by inflection. In a following step, identical standardized inflectional forms of double lemmata are reduced: e.g., a weak class III preterit of *sagen*,*sagēn* 'say' is reduced from *sagēta*,*sagēta* to a single *sagēta*.

With separable preverbs, a contextual approach is needed: in the phrase *ni ges thû thanan ûz* 'thou shalt not come out thence' (Sievers 1892: 51), the second person present subjunctive *ges* is attributed to the lemma *ūȝgān* 'to go out, to end', as the meaning of verbs with separable prefixes differs from those of their constituents. To make the program generate a standardized inflectional form *gās*, not *ūȝgās*, comparable to *ges*, the prefix has to be deleted. As separable preverbs have their own part-of-speech tag, this is achieved by taking them as a basis, and searching in both directions until a verb that begins with them is found.

Once the standardized inflectional forms have been generated, they can be inserted into the text files. To do this, they are assembled in a list, as are all language tags within the text file, after having been extracted. A third list is made of the internal index of every standard word form, unless it contains a punctuation mark and therefore does not have to be replaced. Then, for every element in the index list, the subsequent dummy for the standardized inflectional form – if the corresponding element in the language list is goh ("German, Old High", according to ISO 639-3[16]) – is replaced by the real one. The altered code is written into a new ELAN file.

[16] cf. www.sil.org/iso639-3/codes.asp.

7. Facilitating the data checks

In all cases where (a) the spelling of the printed edition is also considered, (b) a separable preverb is identified, or (c) double lemmata are reduced, information on this and the affected word forms is stored in a log file during the program execution to enable a manual check of potentially incorrectly generated word forms. Another program then compares the attested word forms and the standardized inflectional form: if they deviate considerably from each other, a mistake in the part-of-speech or the morphological annotation is likely. To this end, the *Relative Levenshtein Distance* is calculated: the average length of both word tokens is divided by the *Levenshtein Distance*, i.e. the minimal number of insertion, deletion and replacement operations required to convert one character string into the other (cf. Levenshtein 1966). The deviating word token pairs are output into another log file, sorted according to the largest deviance.

8. Conclusion

The automated generation of standardized inflectional forms for the Old German text corpus allows for a verification of the part-of-speech and morphological annotation, and offers an opportunity to do research on the corpus without having to consider cross-text variation within phonology or morphology. The creation of the standardized inflectional forms on the basis of standardized lemmata and inflectional information proves to be a manageable task in most cases, although some of them need complex examinations to generate the appropriate standard word forms. Sometimes, even the record from the printed edition has to be taken into account, as the inflectional information does not cover some specific cases. Nonetheless, enhancing the data quality of the corpus and facilitating its utilization using a high degree of automation justifies the effort of having to scrutinize the Old German grammars in every detail.

References

Braune, Wilhelm (2004): Althochdeutsche Grammatik. Band I: Laut- und Formenlehre. 15th edition, ed. Ingo Reifenstein. Tübingen: Niemeyer.

Heffner, Roe-Merill Secrist (1961): A word-index to the texts of Steinmeyer. Die kleineren althochdeutschen Sprachdenkmäler. Madison: The University of Wisconsin Press.

Hench, George Allison (1890): The Monsee fragments. Straßburg: Trübner.

Hench, George Allison (1893): Der althochdeutsche Isidor. Straßburg: Trübner.

Kelle, Johann (1881): Glossar der Sprache Otfrids. Regensburg: Manz.

Levenshtein, Vladimir I. (1966): Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet Physics Doklady 10(8): 707-710.

Linde, Sonja/Mittmann, Roland (2013): Old German Reference Corpus. Digitizing the knowledge of the 19th century. Automated pre-annotation using digitized historical glossaries. In: Bennett, Paul/Durrell, Martin/Scheible, Silke/Whitt, Richard J. (eds.): New Methods in Historical Corpora (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP) 3). Tübingen: Narr, 235-246.

Mittmann, Roland (2013): Digitalisierung historischer Glossare zur automatisierten Vorannotation von Textkorpora am Beispiel des Altdeutschen. In: Journal for Language Technology and Computational Linguistics 27: 39-52. www.jlcl.org/2012_Heft2/3Mittmann.pdf.

Sehrt, Edward (1955): Notker-Wortschatz. Halle: Niemeyer.

Sievers, Eduard (1874): Die Murbacher Hymnen. Halle: Buchhandlung des Waisenhauses.

Sievers, Eduard (1892): Tatian. Lateinisch und althochdeutsch mit ausführlichem Glossar. 2nd edition. Paderborn: Schöningh.

Splett, Jochen (1993): Althochdeutsches Wörterbuch. Berlin: de Gruyter.

Verner, Karl (1877): Eine Ausnahme der ersten Lautverschiebung. In: Zeitschrift für vergleichende Sprachforschung auf dem Gebiete der indogermanischen Sprachen XXIII: 97-130.