Controlling Gender Equality with Shallow NLP Techniques

M. Carl, S. Garnier, J. Haller
Institut für Angewandte Informationsforschung
66111 Saarbrücken
Germany
{carl,sandrine,hans}@iai.uni-sb.de

A. Altmayer and B. Miemietz
Universität des Saarlandes
66123 Saarbrücken, Germany
anne@altmayer.info
Miemietz.Baerbel@MH-Hannover.DE

Abstract

This paper introduces the "Gendercheck Editor", a tool to check German texts for gender discriminatory formulations. It relays on shallow rule-based techniques as used in the Controlled Language Authoring Technology (CLAT). The paper outlines major sources of gender imbalances in German texts. It gives a background on the underlying CLAT technology and describes the marking and annotation strategy to automatically detect and visualize the questionable pieces of text. The paper provides a detailed evaluation of the editor.

1 Introduction

Times of feminist (language) revolution are gone, but marks are left behind in the form of a changed society with a changed consciousness and changed gender roles. Nevertheless language use seems to oppose changes much stronger than society does. As the use of non-discriminatory language is nowadays compulsory in administration and official documents, a number of guidelines and recommendations exist, which help to avoid gender imbalance and stereotypes in language use. Although in some cases men may be concerned (as for example most terms referring to criminals are masculine nouns) the main concern is about adequate representation of women in language, especially in a professional context. Psychological tests demonstrate that persons reading or hearing masculine job titles (so-called generic terms allegedly meaning both women and men) do not visualize women working in this field.

In order to avoid this kind of discrimination two main principles are often suggested (e.g. as in (Ges, 1999; Uni, 2000)):

1. use of gender-neutral language, which rather "disguises" the acting person by using impersonal phrases.
2. explicit naming of women and men as equally represented acting persons.

Using and applying these guidelines in a faithful manner is time-consuming and requires a great amount of practice, which can not always be provided, particularly by unexperienced writer. Moreover these guidelines are often completely unknown in non-feminist circles. A tool which checks texts for discriminatory use of language is thus mandatory to promote written gender equality, educate and remind writer of unacceptable forms to avoid.

In this paper we describe the project "Gendercheck" which uses a controlled-language authoring tool (CLAT) as a platform and editor to check German texts for discriminatory language.

In section 2 we introduce three categories of gender discrimination in German texts and provide possibilities for their reformulation.

Section 3 introduces the technology on which the Gendercheck editor is based. The linguistic engine proceeds in two steps, a marking and filtering phase where gender discriminatory formulations are automatically detected. A graphical interface plots the detected formulations and prompts the according messages for correction .

Section 4 then goes into the detail of the marking and filtering technique. We use the shallow pattern formalism KURD (Carl and Schmidt-Wigger, 1998; Ins, 2004) first to mark possible erroneous formulations and the to filter out those which occur in "gendered" context. Section 5 evaluates the Gendercheck editor on two texts.

2 Gender Inequality in German Texts

Most prominent to achieve gender equality on a linguistic level in German texts is to find solutions and alternatives for the so-called generic

masculine: the masculine form is taken as the generic form to designate all persons of any sex. The major problem is to figure out whether or not a given person denotation refers to a particular person. For instance, in example (1a) "Beamter" (officer) is most likely used in its generic reading and refers to female officers (Beamtinnen) and masculine officers (Beamten). To achieve gender equality an appropriate reformulation is required as shown in example (1b).

(1a) Der Beamte muss den Anforderungen Genüge leisten.

(1b) Alle Beamten und Beamtinnen müssen den Anforderungen Genüge leisten.

Since we tackle texts from administrative and legal domains we principally assume unspecified references. That is, a masculine (or feminine!) noun will not denote a concrete person but rather refers to all persons, irrespectively of their sex.

A second class of errors are masculine relative, possessive and personal pronouns which refer to a generic masculine or an indefinite masculine pronoun.

(2a) Der Beamte muss seine Wohnung in der Nähe des Arbeitsplatzes suchen.

(2b) Jeder muss seinen Beitrag dazu leisten.

(2c) Wer Rechte hat, der hat auch Pflichten.

The possessive pronoun "seine" (his) in example (2a) refers to the preceeding "Beamte" (officer). The generic masculine use of "Beamte" and the referring pronoun will be marked. The same holds for sentence (2b) where the possessive pronoun refers to the indefinite pronoun "jeder" (every_{masc}). The indefinite pronouns "jemand" (someone) and "wer" (who) count as acceptable. However, masculine pronouns referring to it will be marked. In example (2c), the masculin relative pronoun "der" can be omitted.

A third class of gender inequality is lack of agreement between the subject and the predicative noun. Example (3a) gives an example where the masculine subject "Ansprechpartner" (partner_{masc}) occurs with the a female object "Frau Müller" (Mrs. Müller).

(3a) Ihr Ansprechpartner ist Frau Müller.

(3b) Ihre Ansprechpartnerin ist Frau Müller.

A solution for this class of errors is shown in example (3b) where the subject (Ansprechpartnerin) is adapted to the female gender of the predicate.

Suggestions to reformulate gender imbalances as shown in examples (1) and (2) can be classified in two main categories:

1. Whenever possible, use gender neutral formulations. These include collectiva (e.g. Lehrkörper (teaching staff) or Arbeitnehmerschaft (collective of employees)) as well as nominalized participles (Studierende (scholar)) or nominalized adjectives (Berechtigte).
2. Use both forms if gender neutral formulations cannot be found. That is, the feminine and the masculine form are to be coordinated with "und", "oder" or "bzw.". A coordination with slash "/" will also be suggested but should only be used in forms, ordinance and regulations.

Amendments should accord to general German writing rules. The so called "Binnen-I", an upper case "I" as in "StudentInnen" will not be suggested and also naming of the female suffix in parenthesis should be avoided. The same holds for the indefinite pronoun "frau" (woman) which was occasionally suggested to complement the pronoun "man".

3 The Gendercheck Editor

Controlled-Language Authoring Technology (CLAT) CLAT has been developed to suit the need of some companies to automatically check their technical texts for general language and company specific language conventions. Within CLAT, texts are checked with respect to:

- orthographic correctness
- company specific terminology and abbreviations
- general and company specific grammatical correctness
- stylistic correctness according to general and company specific requirements

The orthographic control examines texts for orthographic errors and proposes alternative writings. The terminology component matches the text against a terminology and abbreviation database where also term variants are detected

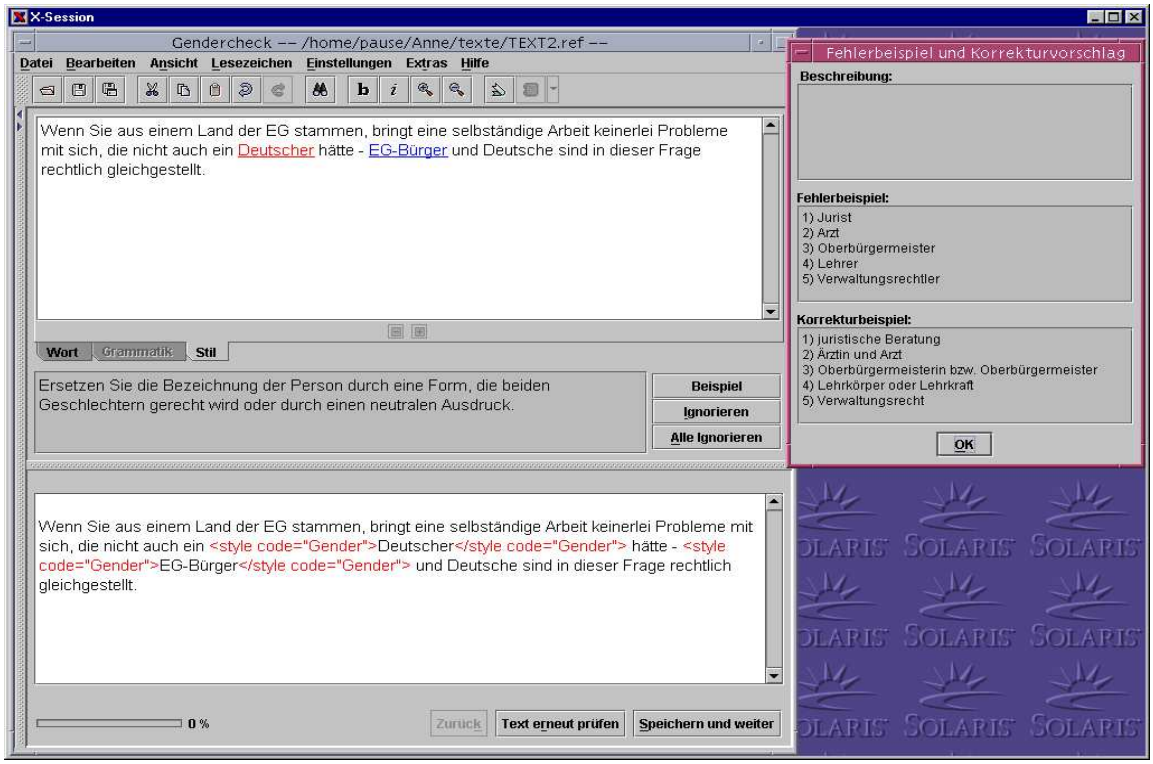

Figure 1: The Gendercheck Editor

(Carl et al., 2004). Grammar control checks the text for grammatical correctness and disambiguates multiple readings. Stylistic control detects stylistic inconsistencies.

The components build up on each other's output. Besides the described control mechanisms, CLAT also has a graphical front-end which makes possible to mark segments in the texts with different colors. Single error codes can be switched off or on and segments of text can be edited or ignored according to the authors need. CLAT also allows batch processing where XML-annotated text output is generated.

Figure 1 shows the graphical interface of the editor. The lower part of the editor plots an input sentence. The highlighted SGML codes are manually annotated gender "mistakes". The upper part plots the automatically annotated sentence with underlined gender mistakes.

As we shall discuss in section 5, gender imbalances are manually annotated to make easier automatic evaluation. In this example, the highlighted words "Deutscher" (German) and "EG-Bürger" (EU-citizen) are identical in the manually annotated text and in the automatically annotated text. The user can click on one of the highlighted words in the upper window to display the explanatory message in the middle part of the screen. Further information and correction or reformulation hints can also be obtained by an additional window as shown on the right side of the figure. The messages are designed according to main classes of gender discriminatory formulations as previously discussed.

4 Gender Checking Strategy

Gendercheck uses a marking and filtering strategy: first all possible occurrences of words in an error class are marked. In a second step "gendered" formulations are filtered out. The remaining marked words are assigned an error

code which is plotted in the Gendercheck editor.

According to the classification in section 2, this section examines the marking and filtering strategy for generic masculine agents in section 4.1, pronouns which refer to generic masculine agents (section **??**) and errors in agreement of predicative nouns (section **??**).

Marking and filtering is realized with KURD a pattern matching formalism as described in (Carl and Schmidt-Wigger, 1998; Ins, 2004). Input for KURD are morphologically analyzed and semantically tagged texts.

4.1 Class 1: Agents

4.1.1 Marking Agents

Two mechanisms are used to mark denotations of persons:

a) The morphological analysis of MPRO (Maas, 1996) generates not only derivational and inflectional information for German words, but also assigns a small set of semantic values. Male and female human agents such as "Soldat" (soldier), "Bürgermeister" (mayor$_{masc}$), "Beamte" (officer$_{masc}$), "Krankenschwester" (nurse$_{fem}$) etc. are assigned a semantic feature `s=agent`. Words that carry this feature will be marked `style=agent`.

b) Problems occur for nouns if the base word is a nominalized verb. For instance "Gewichtheber" (weightlifter) und "Busfahrer" (bus driver) will not be assigned the feature `s=agent` by MPRO since a "lifter" and a "driver" can be a thing or a human. Gender inequalities, however, only apply to humans. Given that the tool is used in a restricted domain, a special list of lexemes can be used to assign these words the style feature `style=agent`. The KURD rule `Include` shows some of the lexemes from this list. The list contains lexemes to cover a maximum number of words. For instance the lexeme `absolvieren` (graduate) will match "Absolvent" (alumnus$_{masc}$), "Absolventin" (alumnus$_{fem}$), "Absolventen" (alumni$_{plu,masc}$) and "Absolventinnen" (alumni$_{plu,fem}$).

```
1 Include =
2       Ae{c=noun,
3          ls:absolvieren$;
4             dezernieren$;
5             richten$;
6             fahren$;
7             administrieren$;
8             vorstand$}
9      :  Ag{style=agent}.
```

Lines 3 to 8 enumerate a list of lexemes separated by a semicolon. The column in line 3 following the attribute name `ls` tells KURD to interpret the values as regular expressions. Since the dollar sign `$` matches the end of the value in the input object, each lexeme in the list can also be the head of a compound word. Thus, the test `ls:fahren$` matches all lexemes that have `fahren` as their head words, such as "Fahrer" (driver), "Busfahrer" (bus driver), etc. The action `Ag{style=agent}` marks the matched words as an agent.

4.1.2 Filtering "gendered" Agents

The text then undergoes several filters to delete marks in words if the appear within gendered formulations.

a) Excluded are marked agents which preceed a family name. The marking of "Beamte" in example (4) will be erased since it is followed by the family name "Meier". "Beamte Meier" is likely to have a specific reference.

(4) Der Beamte Meier hat gegen die Vorschrift verstoßen.

In terms of KURD this can be achieved with the rule `AgentMitFname`: if a family name (`s=fname`) follows a sequence of marked agents (`style=agent`) the marks in the agent nodes are removed (`r{style=nil}`).

```
1 AgentMitFname =
2          +Ae{style=agent},
3           Ae{c=noun,s=fname}
4        :  Ar{style=nil}.
```

b) Also excluded are nominalized plural adjectives and participles since they are well suited for gender neutral formulations. In example (5), the nominalized plural adjective "Sachverständige" (experts) is ambiguous with respect to gender. The mark will thus be removed.

(5) Sind bereits Sachverständige bestellt?

c) Marked words in already gendered formulations are also erased. Pairing female and male forms by conjunction is a recommended way to produce gender equality. In example (6) the subject "Die Beamtin oder der Beamte" (the officer$_{fem}$ or the officer$_{masc}$) as well as the pronouns which refer to it "sie oder er" (she or he) and "ihrer oder seiner" (her or his) are gender equal formulations.

(6) Die Beamtin oder der Beamte auf Lebenszeit oder auf Zeit ist in den Ruhestand zu versetzen, wenn sie oder er infolge eines körperlichen Gebrechens oder wegen Schwäche ihrer oder seiner körperlichen oder geistigen Kräfte zur Erfüllung ihrer oder seiner Dienstpflichten dauernd unfähig (dienstunfähig) ist.

The KURD rule `gegendert` removes these marks. The description in lines 2 to 5 matches a conjunction of two marked agents (`style=agent`) which share the same lexeme `ls=_L` but which are different in gender. This latter constraint is expressed in two variables `ehead={g=_G}` and `ehead={g~=_G}` which only unify if the gender features "g" have non-identical sets of values.

```
1 gegendert =
2    Ae{style=agent,ls=_L,ehead={g=_G}},
3     e{lu=oder;und;bzw.;/},
4    *a{style~=agent}e{c=w},
5    Ae{style=agent,ls=_L,ehead={g~=_G}}
6 :  Ar{style=nil}.
```

The rule allows the conjunctions "und", "oder", "bzw." and "/".

d) Some nouns are erroneously marked even if no gender equal formulation is possible. For instance words such as "Mensch" (human being),a "Gast" (guest), "Flüchtling" (refugee) are masculine in gender, yet there is no corresponding female form in German. These words are included in an `exclude` list which works similar to the `include` list previously discussed.

```
1 exclude =
2       Aa{style=agent,
3             lu:mensch$;
4                flüchtling$;
5                säugling$;
6                gast$;
7                rat$}
8         : Ar{style=nil}.
```

4.1.3 Non marked Expressions

a) Currently, we do not mark compound nouns which have an agent as their modifier and a non-agent as their head. However, also words such as "Rednerpult" (talker desk = lectern) and "Teilnehmerliste" (participants list = list of participants) are suitable for gender mainstreaming and should be spelled as "Redepult" (talk desk) and "Teilnehmendeliste" (participating list).

b) We do not mark articles and adjectives which preceed the marked noun. This would be troublesome in constructions like example (7) where the article "der" (the) and the corresponding noun "Dezernent" (head of department) are separated by an intervening adjectival phrase.

(7) Den Vorsitz führt der jeweils für die Aufgaben zuständige Dezernent.

c) It is currently impossible to look beyond the sentence boundary. As a consequence, the reference of a agent cannot be detected if it occurs in the preceeding sentence. For instance "Herr Müller" is the reference of "Beamte" in the second sentence in example (8).

(8) Herr Müller hat die Dienstvorschrift verletzt. Der Beamte ist somit zu entlassen.

The word "Beamte" will be erroneously marked because information of the preceeding sentence is not available to resolve the reference.

4.2 Class 2: Pronouns

Also personal pronouns, possessive pronouns, relative pronouns and indefinite pronouns are marked. The strategy is similar to the one for agents above: first all pronouns are marked and in a second step markings in correct formulations are erased.

With the exception of indefinite pronouns ("Mancher", "Jemand", "Niemand" etc.), a marked referent agent must be available in the same sentence. Three different rules are used to mark relative pronouns, personal pronouns and possessive pronouns.

```
1 MarkRelativPronomen =
2         e{style=agent,ehead={g=_G}},
3        *a{lu~=&cm},
4         e{lu=&cm},
5        Ae{lu=d_rel,ehead={g=_G}}
6     :  Ag{style=agent}.
```

a) The rule `MarkRelativPronomen` detects a marked agent in line 2. Lines 3 and 4 search the next comma[1] that follows the marked agent and line 5 matches the relative pronoun[2] that immediately follows the comma. The relative

[1] commas are coded as "`&cm`" in the formalism.

[2] relative pronouns are assigned the lexeme "`d_rel`".

	Size of Test			Classes of errors			
Text	#sent.	#words	Errors/sent.	Class 1	Class 2	Class 3	$\sum$
ET_1	95	1062	1,83	97	62	15	174
TT_2	251	6473	0,46	95	21	—	116

pronoun must agree in gender with the agent (`ehead={g=_G}`). As we shall see in section 5, this is an error prone approximation to reference solution.

b) Personal and possessive pronouns are only marked if they refer to a male agent. The two rules `MarkPersonalPronomen` and `MarkPossesivPronomen` work in a similar fashion: in line 2 the marked masculine reference is matched. Lines 3 and 4 match the following personal pronoun (`c=w,sc=pers`) and possessive pronoun (`c=w,sc=poss`). In lines 5, the pronouns are marked.

```
1 MarkPersonalPronomen =
2         e{style=agent,ehead={g=m}},
3       1Ae{lu=er;er_es,c=w,sc=pers}
4        |e{s~=agent,sc~=punct}
5     :  Ar{style=agent}.
```

```
1 MarkPossesivPronomen =
2         e{style=agent,ehead={g=m}},
3       1Ae{lu=sein,c=w,sc=poss}
4        |e{s~=agent,sc~=punct}
5     :  Ar{style=agent}.
```

After the marking step, pronoun marks are filtered. Filtering of pronouns is similar to the previously discussed rule `gegendert`.

4.3 Class 3: Predicative Noun

Missing agreement between subject and predicative noun is detected with the following KURD rule:

```
1  Praedikatsnomen =
2   +Ae{mark=np,style=agent,ehead={g=_G}},
3   *Ae{mark=np},
4     e{ls=sein,c=verb},
5   *Ae{style~=agent},
6    Ae{mark=np,style=agent,ehead={g~=_G}},
7   *Ae{mark=np}
8 :  Ar{bstyle=Gen3,estyle=Gen3}.
```

Lines 2 and 3 detect the marked subject. Notice that noun groups are marked with the feature `mark=np` by a previous chunking module. Lines 5 to 7 match the predicative noun. Both parts of the sentence are connected by the copula "sein" (`be`). Similar to the rule `gegendert`, the rule only applies if both parts are different in gender.

5 Evaluation of Gendercheck

We evaluated the Gendercheck editor based on two texts:

ET_1 A collection of unconnected negative examples taken from the (Ges, 1999) and (Sch, 1996).

TT_2 The deputy law of the German Bundestag

Gender imbalances were manually annotated with a SGML code, where each different code refers to a different rewrite proposal to be plotted in the editor as in the lower part in figure 1. Table 4.1.3 shows the distribution of error classes in the two texts. Each error class had several subtypes which are omitted here for sake of simplicity.

In ET_1 every sentence has at least one error; on average one word out of six is marked as "ungendered". Since ET_1 is a set of negative examples, errors are uniformly distributed. Distribution of errors in text TT_2 is different from ET_1. TT_2 does not contain a single occurrence of a class 3 error. On average, only one word out of 60 is manually marked and — due to the long size of sentences — there are 0.46 errors per sentence on average.

Text ET_1 was used to develop and adjust the KURD rule system for marking, filtering and error code assignment. We iteratively compared the automatically annotated text with the manually annotated text and computed precision and recall. Based on the misses and the noise, we adapted the style module as well as the error annotation schema. Thus, in a first annotation schema we assigned more than 30 different error codes literally taken from (Ges, 1999) and (Uni, 2000). However, it turned out that this was too fine a granularity to be automatically reproduced and values for precision and recall were very low. We than assigned only one error class and achieved very good values for precision of over 95% and recall over more than 89%. Based on these results we carefully refined a number of subtypes of the three error classes.

Final results are shown in table 5. Results for the test text TT_2 are slightly inferior to those of

	Text ET_1				
Error	hit	misses	noise	precision	recall
Class 1	85	12	1	0.988	0.876
Class 2	55	7	5	0.917	0.887
Class 3	15	0	1	0.937	1.000
$\sum$	155	19	7	0.957	0.891
	Text TT_2				
Class 1	86	9	5	0.945	0.905
Class 2	14	7	4	0.778	0.667
$\sum$	100	16	9	0.917	0.862

the development text ET_1. We briefly discuss typical instances of misses and noise.

a) Noise in class 1 (generic use of masculine) are mainly due to "-ling" - derivations such as "Abkömmling" (descendant) which are masculine in German and for which no female equivalent forms exist. These words could be included in the `exclude` lexicon (see section 4).

b) In some cases nominalized participles such as "Angestellte" (employee) and "Hinterbliebene" (surviving dependant), which are usually very well suited for gendered formulations due to their ambiguity in gender, were erroneously disambiguated. These instances produced noise because filters did not apply.

c) Misses in class 1 can be traced back to some words which have not been detected as human agents such as "Schriftführer" (recording clerk) and "Ehegatte" (spouse). These words could be entered into the `include` lexicon. Both lexicon should be made user-adaptable and user extendible in future versions of the system.

d) Many of the misses in class 2 are due to a reference in the preceeding sentence. Since the system is currently sentence based, there is no easy solution in enhancing this type of errors. The possessive pronoun "**seiner**" in the second sentence of example (9) refers to "Bewerber" (applicant) in the first sentence. This connection cannot be reproduced if the system works on a sentence basis.

(9) Einem Bewerber um einen Sitz im Bundestag ist zur Vorbereitung seiner Wahl innerhalb der letzten zwei Monate vor dem Wahltag auf Antrag Urlaub von bis zu zwei Monaten zu gewähren. Ein Anspruch auf Fortzahlung **seiner** Bezüge besteht für die Dauer der Beurlaubung nicht.

e) An example for noise in class 2 is shown in example (10). The relative pronoun "der" (who,which) was detected by Gendercheck but erroneously been linked to "Beamte" instead of "Antrag" (application) which are both masculin in German.

(10) Der Beamte ist auf seinen Antrag, der binnen drei Monaten seit der Beendigung der Mitgliedschaft zu stellen ist, ...

Much more powerful mechanisms are required to achieve a breakthrough for this kind of errors.

6 Conclusion

This paper describes and evaluates the "Gendercheck Editor" a tool to check German administrative and legal texts for gender equal formulations. The tool is based on the Controlled Language Authoring Technology (CLAT), a software package to control and check technical documents for orthographical, grammatical and styptic correctness. A part of the Style component has been modified and adapted to the requirements of linguistic gender main-streaming.

The paper outlines a shallow technique to discover gender-imbalance and evaluates the technique with two texts. Values for precision and recall of more than 90% and 85% respectively are reported.

References

Michael Carl and Antje Schmidt-Wigger. 1998. Shallow Postmorphological Processing with KURD. In *Proceedings of NeMLaP3/CoNLL98*, pages 257–265, Sydney.

Michael Carl, Maryline Hernandez, Susanne Preuß, and Chantal Enguehard. 2004. English Terminology in CLAT. In *LREC-Workshop on Computational & Computer-assisted Terminology*, Lisbonne.

Gesellschaft für Informatik (Hg.), Bonn, 1999. *Gleichbehandlung im Sprachgebrauch: Reden und Schreiben für Frauen und Männer.*

Institut für Angewandte Informationsforschung, Saarbrücken, 2004. *Working paper 38.* to appear.

Heinz-Dieter Maas. 1996. MPRO - Ein System zur Analyse und Synthese deutscher Wörter. In Roland Hausser, editor, *Linguistische Verifikation, Sprache und Information.* Max Niemeyer Verlag, Tübingen.

Schweizerische Bundeskanzlei (Hg.), Bern, 1996. *Leitfaden zur sprachlichen Gleichbehandlung im Deutschen.*

Universität Zürich (Hg.), Zürich, 2000. *Leitfaden zur sprachlichen Gleichbehandlung von Frau und Mann.*