Brno Studies in English
Volume 38, No. 2, 2012
ISSN 0524-6881
DOI: 10.5817/BSE2012-2-9

Josef Schmied

Academic Knowledge Presentation in MA Theses: from Corpus Compilation to Case Studies of Disciplinary Conventions

Abstract
This article discusses discipline-specific conventions in presenting academic knowledge in academic writing, an important first research text at many universities. Its empirical basis is a new corpus of South African MA theses (the ZAMA corpus) from Stellenbosch University. A pilot corpus of 100 MA theses and over 4 million words was compiled, paying special attention to disciplinary breadth from Humanities and Social Sciences to Natural Sciences and Engineering, including some interdisciplinary subjects (like Geography) and Law. Altogether between four and eight texts were chosen from 18 disciplines, which were later categorised deductively into six discipline types. Other factors taken into consideration were the socio-biographical diversity of authors, esp. ethnic/language by author's name (distinguishing English, Afrikaans, Black South African names, and a few others). The decisions of corpus compilation are carefully documented in order to obtain a stratified and balanced corpus for research in the South African university context. Descriptive statistical analyses were carried out to test whether discipline-specific frequency patterns could be found in three case studies using variables indicating metadiscourse conventions, such as modal auxiliaries, personal pronouns, and cohesive linkers.[1]

Key words
Academic writing; disciplinary conventions; metadiscourse; modal auxiliaries; personal pronouns; cohesive linkers

1. Introduction: Knowledge presentation in academic discourse

In a modern constructivist perspective, academic knowledge is constructed in a social negotiation process, researchers (particularly young researchers, like

MA students) attempt to persuade other researchers (particularly more senior researchers, like their supervisors) as readers of the correctness of their claims, so that they gain community acceptance for their work as a contribution to disciplinary scholarship and knowledge generation. In this discourse approach, metadiscourse elements play a decisive role in the negotiation between author and reader. This means that knowledge is not simply generated by "presenting the facts" but by establishing a research discourse "connection" between the writers and their community of practice through disciplinary conventions in terminology, genre, and other research practices.

I examined three variables that are considered important in the relevant handbooks: modal auxiliaries, personal pronouns and cohesive linkers.

Personal pronouns are crucial for establishing a relationship between the writer (*I*), and the reader (*you*) implicitly or explicitly (inclusive *we*). The use of second person pronouns (*you/your*), for instance, encourages the reader to participate as an intelligent equal in the reasoning process. By including readers in this way, the writer credits them with possessing both in-group understanding and the intelligence to make the same reasonable inferences. The argument is thereby strengthened by claiming solidarity with the community and the mutual experiences needed to draw the same conclusions as the writer. The use of *I* is considered fair and honest (and not "subjective") in such a personal author – reader relationship, because it takes responsibility more explicitly than a majestic or humble *we* or a seemingly "objective", but formally complex, passive construction. Of course, personal pronoun use is a particularly well-known example of departmental conventions and modern trends, which leave the individual less freedom of choice than other features like linkers and modality.

Since Halliday and Hasan (1976) propagated the notions of textuality, cohesion and coherence, research and teaching at universities concentrated on these simple formal cohesive devices to support the reader to conceive coherence in texts. For argumentative texts like MA theses, the success of negotiating with the supervisor and examiner depends not only on the "facts" but also on the support the writers offer their readers to follow the argumentation, logic and persuasion of the presentation. Here the choice of "linkers" is wide (cf. Table 5 below). Cohesive linkers are considered reader-friendly devices that may also be employed to different degrees and in different types in discipline-specific argumentation conventions.

Finally, central concepts in the constructive debate on academic writing have been (author) stance and hedging, which included the traditional forms of modal auxiliaries, particularly in English. The distinction between "This must be the case." and "This may be the case." is important in intellectual debate, because the former makes it clear that the writer assumes that s/he has enough evidence for a good stance and cannot be attacked, whereas the latter may even invite the reader to contradict because the debate is still very open. Modal auxiliaries are extremely important for author perspective, commitment or hedging (cf. Schmied 2008a and 2008b), especially in epistemic usage as in the example above.

This emphasis on metadiscourse in academic writing is in line with recent discussions about the more explicit teaching of academic writing, in particular writing conventions based on genre analysis and metadiscourse (cf. Swales 1990 and 2004 and Hyland 2007 and 2012). The interrelationship between individuality and community in academic cultures has been analysed in detail:

> To project an identity as an academic means buying into the practices of a discipline and handling its discourses with sufficient competence to participate as a group member. How individuals exchange information, build alliances, dispute ideas and work together varies according to the group they belong to, so each discipline might be seen as a distinct academic *culture* (Hyland, 2004a) or *tribe* (Becher and Trowler, 2001), each with its particular norms and practices. (Hyland 2012: 15)

In this article, I try to prove that the variable "academic discipline type" is prominent to account for variation patterns in the genre of MA thesis from a wide range of departments within the same university context.

2. The database: ZAMA Corpus from Stellenbosch

The database this analysis exploits comes from the SUNScholar Research Repository from Stellenbosch University Library, South Africa. Stellenbosch seems a suitable university for our South African corpus, since it has a wide range of disciplines and attracts a wide range of students; despite its Afrikaans-medium tradition, it has a multilingual policy, so that enough English texts can be found in the Repository, which offers easy (and multilingual) access to the academic work there in a wide range of genres, even speeches ("written to be spoken") or conference announcements. "SUNScholar is an open access electronic archive for the collection, preservation and distribution of digital materials created by members of Stellenbosch University", as it says on the website http://scholar.sun.ac.za/ (28/02/13). "SUNScholar is built using open standards with open source software for the purposes of extremely long term digital preservation and sustainability" and also includes a useful Wiki help guide.

The pilot corpus consists of 100 MA theses, sampled from 18 disciplines (i.e. usually a department at Stellenbosch, occasionally an entire faculty) with at least 4 MA theses each. In a conscious attempt, this disciplinary breadth was later reduced (to gain more texts per category) by categorizing disciplines into discipline types from "soft" to "hard": English (Literature), Education/Curriculum Studies, Politics, and History were grouped in the Humanities (Hu); Psychology, Sports, and Sociology/Anthropology in Social Sciences (SS); Physics and Chemistry in Natural Sciences together with Medicine/Pharmacy and Genetics (from Agroscience, not from Medicine; NS); Mechanical and Civil Engineering in Engineering (EG). The texts from the Law Department were kept separate (Lw), although

there are too few texts to distinguish Public and Mercantile from Private Law as (sub-)disciplines. A special category was put together in Interdisciplinary (ID) cases like Economy, Journalism, Geography and (General) Linguistics. A few key words from the thesis titles may illustrate this: Economy ranges from advertising to business cycles, Journalism from blog ethics to the framing of climate change, Geography from physical geography in climate change to cultural approaches in wine marketing, and Linguistics from empirical error analyses to theoretical lingua franca discussions. A more empirical cluster analysis may be used to confirm or complement this top-down approach later-on. Then we will see, for instance, whether texts from Mercantile Law are closer to Economics and texts from Public Law are closer to Politics, or whether they are primarily in a class of their own.

The stratified sampling was achieved through close monitoring of the growing database until a pilot corpus of over four million words was collected. The lead variable was the discipline (usually 5 theses were selected, but some showed technical problems with down-load or transformation into simple text files, so that they could not be included). Within the discipline, the stratification included different subsections indicated explicitly or implicitly through the topic (such as natural versus social geography). Then, author names were analysed according to probable first language and ethnic background (this meant that clearly English, including possibly Scottish, and Black South African names were preferred to Afrikaans names, which were the majority in many departments). Finally, the year and the gender were taken into consideration, so that a balanced distribution was achieved over the last 10 years.

After the sampling, texts had to be anonymised and carefully edited. The main purpose of the editing process was to mark larger text elements not written by the author (i.e. quotations were placed between <quote> and </quote>). Since the usual "linear" corpus-linguistic retrieval tools (like AntConc and Wordcruncher) were to be used with the data, hypertext elements had to be inserted in the relevant place (i.e. footnotes were placed at the right position within the text and tagged with <fn> and </fn>). Since our standard corpus-linguistic tools take plain text as input, photos and large figures were removed automatically, only the respective title was left in the text. However, in all these careful interventions, borderline cases were noticed: names and personal references in prefaces and acknowledgements were usually not removed (since they often display interesting cultural differences), small citations were left untouched (since deletion would have changed the syntactic structure), and sometimes questionable text elements were not removed but simply "hidden" in angle brackets, so that the usual search programs would ignore them. These considerations have to be kept in mind when specific analyses are made (e.g. the word *submitted* is standard on every MA title page, which should be omitted from searches including such conventional items). It is clear that the number of words deleted from such standardized sections has a direct effect on the frequency figures per 1 million words displayed in the tables and figures below; their values are therefore relative.

As an initial analysis, we can compare the length of each text (in word tokens).

Figure 1 illustrates the length of the 100 texts in the (edited) pilot corpus; texts are arranged according to the discipline abbreviation in two parts: Figure 1a starts with the so-called "soft" side (Curriculum Development in Education, English [Literature], History, Political Science, Economics, Geography, Journalism, Linguistics and the three Law departments), Figure 1b with the "hard" side (Civic Engineering, Mechanical Engineering, Botany/Zoology, Chemistry, Genetics, Medicine/Pharmacy, Psychology, Sociology/Anthropology and Sports). Thus we have a simple visualisation of the differences within and between disciplines. These Figures 1a and 1b shows a few tendencies already: Generally, the texts in 1a are longer than in 1b, so that even the automatic scale is different. Specifically, the five chemistry texts, for instance, have a relatively similar length between 22,000 and 40,000 words; there is one exception which is so short (below 5,000 words) because the thesis consisted mainly of a long list of graphs, which were taken out during the transformation, so that few words remained (for the final composition of the corpus, a decision may be taken that such extreme cases may be replaced). By contrast, the texts from English (Literature) usually consist of 50,000 words, but there are also two extremes with more than 100,000 words. In fact, the longest text in the pilot corpus was almost double that size, because it contained an extensive appendix with four interviews, which were of course not part of the author's language and not in the genre-specific writing style either, but basically in interview style, the inclusion of which would distort the data considerably.

Figure 1a. "Soft science" texts in the ZAMA corpus according to text length (names starting with discipline acronyms)

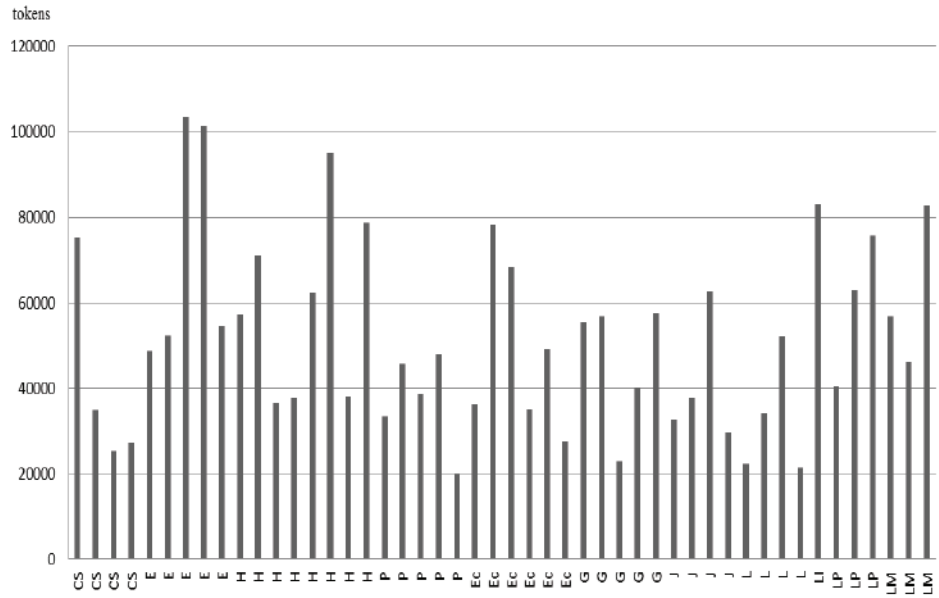


Figure 1b. "Hard Science" texts in the ZAMA corpus according to text length (names starting with discipline acronyms)

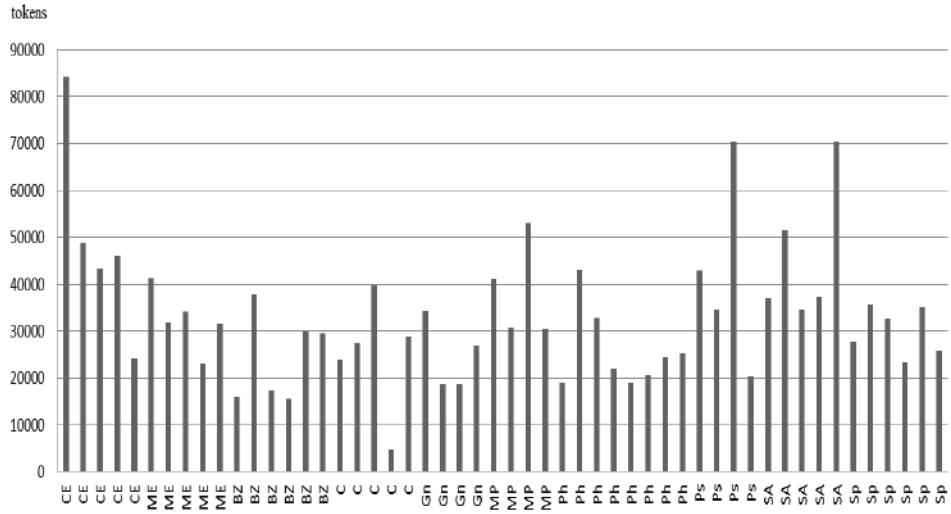
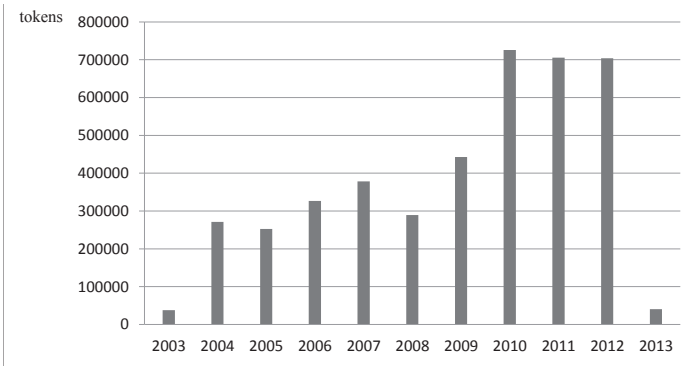


Figure 2 shows the number of words (not texts) by year of publication or submission (from 2003 to 2013). It illustrates that the year 2008 is under- and the year 2010 overrepresented in the current version; the oldest and most recent publication years are less well represented, but this can be remedied in the near future, as more MA theses (incl. older works) are made available. Of course, this diachronic variable was not included in the analysis, since other variables (such as discipline) were considered much more variable. Only an expansion over many more years would be a solid basis for an acceptable diachronic analysis.

Figure 2. Text length variation: tokens by year

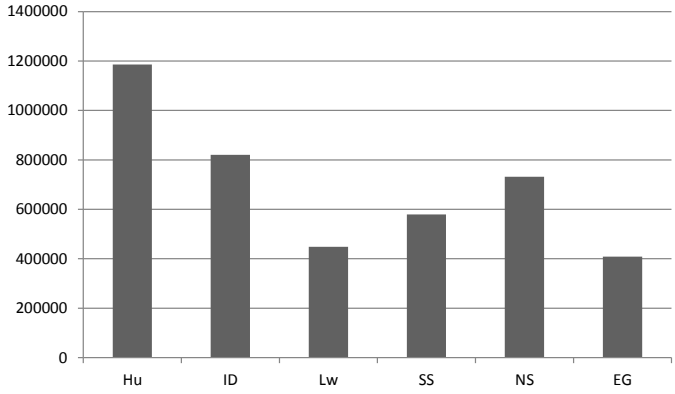


When coding the file names, I tried to construct telling names, so that the basic dimensions were visible at a glance. The discipline was indicated by one or two (initial) letters, and the date can be seen from the next two digits. The language was abbreviated as E for English, A for Africans and X for Black South African languages

(with an interesting rest category with Indian, Chinese and Luo), and finally one letter was added for the gender (male/female), so that G09AM would be a brief but sufficient mnemonic code (for an MA in Geography submitted in 2009 by a male author with an Afrikaans name) in tables and diagrams and after sample sentences.

Figure 3 shows that the spread of words across discipline types is not even. Humanities make up over a quarter of the entire ZAMA corpus, and engineering texts only one tenth (although they are also much shorter, as illustrated in Figure 1 above). Thus far-ranging conclusions in Engineering or Law may not (yet) be appropriate on this basis, although it is an excellent starting point for analysis and discussion. The "interdisciplinary" texts are long enough to let us hope that we will be able to identify conventions that allow us to categorize them in other discipline types (like Humanities or Social Sciences) at the end of our analyses.

Figure 3. Words in the ZAMA Corpus by discipline types



The main purpose of our stratified corpus compilation was to achieve a broad database for analysing metadiscourse (cf. Hyland 2007) variables that might be sensitive to the disciplinary variation and conventions in the texts. There are enough formal features that are accessible to corpus-linguistic extraction and analysis tools and quantitative comparison. But it must be emphasized that semantic distinctions in context (e.g. inclusive vs. exclusive *we*) and pragmatic usage were not possible this way. Different meanings indicating different cultural values would have to be distinguished "by hand"; i.e. researcher intuition for each single occurrence would have to decide whether to include or exclude an occurrence in our analysis. This could result in semantically tagged versions of the ZAMA corpus, e.g. according to modality type (root/deontic, epistemic, or dynamic; cf. below). Such a tagged version of the ZAMA could, of course, also be used with a counting tool.

The standard grammar model for the description of linguistic features was *A Comprehensive Grammar of the English Language* (Quirk et al. 1985), although later grammars were necessary for the more quantitative comparisons (like the corresponding Biber et al. 1999).

3. Presenting author commitment in modal auxiliaries

Modal auxiliaries are a very distinctive and very well researched section of English grammar even on a quantitative corpus-linguistic basis (cf. Warner 1993 or Collins 2009). Although many authors include semi- or quasi-modals in their analysis, only the nine generally accepted core modals *can, could, may, might, must, shall, should, will,* and *would* were included here. This was not only done for syntactic reasons (i.e. *got to*, *be to*, *ought to* or *need to* include an infinitive), but also because the frequency of other modal phrases is rather low. Even the frequency of *shall* was so low in many of the theses analysed that a detailed analysis was not possible on the basis of the 4 million word pilot corpus compiled so far. In academic writing, modal auxiliaries play an important metalinguistic role, because they modify authors' actions or commitment (from *it may be true* to *it must be true*).

Table 1 combines the normalised usage of modals according to gender with that according to the three main language groups. The figures per one million words show that female writers generally use considerably more modal auxiliaries than male writers, except for the most frequent auxiliary *can* and the infrequent auxiliaries *must* and *shall*. The highest figure according to mother tongue (as indicated by name) is that for the Afrikaans speakers, in particular the use of *can* again. Another interesting striking difference is the use of *would*, which has been proven to behave differently in many other varieties of English (e.g. Bautista 2004); in our data, it is the only auxiliary that is used clearly more often by native speakers of English than others. *Shall*, as usual, is so rare that the differences should not be overestimated. A relatively rare modal auxiliary among writers with Black South African language background is *could.* Their general tendency to use fewer auxiliaries may of course be influenced by their mother tongues, where modality is not as prominent in the verb system as in English.

Table 1. Modals by gender and language (per 1 million words)

	gender			language			
	female	**male**	**average**	**Afrikaans**	**English**	**SAfrLang**	**average**
can	1656	1916	1786	2160	1494	1455	1781
could	803	696	750	809	841	496	746
may	914	601	757	807	690	689	742
might	292	214	253	287	212	254	255
must	354	383	368	422	397	328	392
shall	26	67	46	14	45	75	38
should	855	560	707	737	647	614	679
will	1367	1164	1265	1458	1209	1123	1300
would	1150	783	967	864	1216	822	968
Sum	7417	6384	6900	422	376	313	381

As we know from other studies on modals (e.g. Collins 2009), there is a strong tendency for *can*/*could* to express about 80% dynamic modality, *may*/*might* are

almost the opposite, expressing almost 80% epistemic usage, whereas *shall* and *will* are more mixed. Epistemic usage is often related to politeness (cf. Gao 2012), which may explain the higher figures for Black South African languages, but only a detailed semantic usage analysis can provide evidence for this hypothesis. Generally, there is no consistent trend in the distribution according to gender and language in the ZAMA corpus.

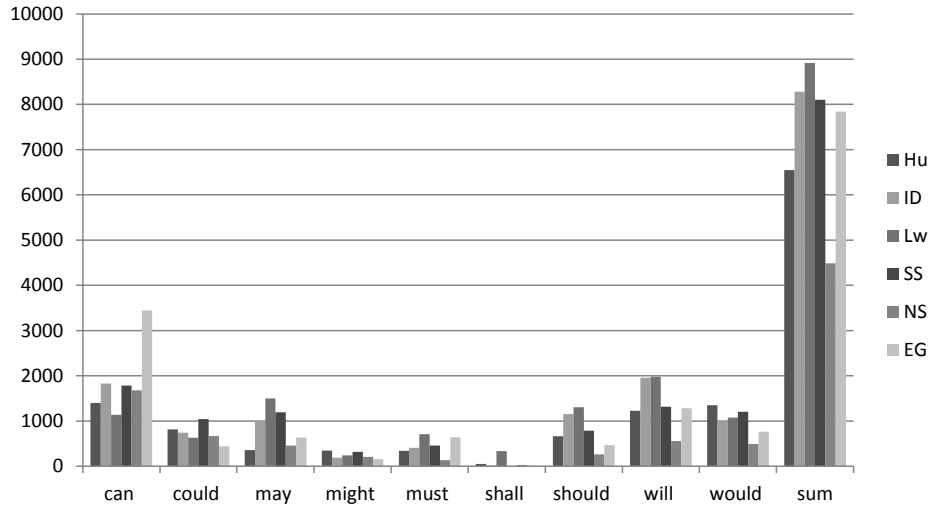
If we analyse the same data according to academic discipline type, we also receive a mixed picture (Table 2). The highest frequencies can be found in the Law texts, where only *can* is used relatively rarely. Similarly consistently high are the figures in the Social Sciences, especially for *could*. By contrast, Natural Sciences use few modals generally.

Table 2. Modals by discipline type (per 1 million words)

	Hu	**ID**	**Lw**	**SS**	**NS**	**EG**	**average**
can	1398	1827	1138	1784	1677	3442	1799
could	816	739	631	1043	666	439	744
may	360	1004	1502	1194	459	637	742
might	346	189	240	320	209	158	249
must	338	409	709	455	132	644	370
shall	52	18	332	8	29	11	48
should	664	1155	1307	787	263	466	693
will	1224	1957	1979	1313	557	1284	1255
would	1350	982	1076	1202	492	763	948
sum	6548	8282	8915	8105	4483	7843	7363

But modal auxiliaries are a very complex group of lexemes. If we look at individual auxiliaries (Figure 4), a few tendencies emerge: This analysis demonstrates again (like Collins 2009) that legal texts are particularly interesting (Sala 2013) and that today *shall* is only a Law modal, it occurs in this very discipline more often than in all the other 17 disciplines together, where it occurs so infrequently that a detailed distribution analysis could not be embarked on with such a small pilot corpus. *Should, may* and *will* are also most prominent in Law. *Would* seems to be the marker for Humanities and Social Sciences – and Law again, but not for Natural Sciences and Engineering. Similar tends can be observed for *might*, although the averages are surprisingly low. Whereas Engineering uses *can*, Social Sciences use *could* relatively often. The sum in Figure 4 visualises the really low figures for modals generally in Natural Sciences.

Figure 4. Modals by discipline type (per 1 million words)



4. Presenting author involvement and reader orientation in personal pronouns

Pronoun usage has been discussed in academic writing for a long time, because authorial presence has been an issue of diverging conventions (Hyland 2002). Personal (including possessive) pronouns include *I, me, my* (for first person singular), *we, us, our* (for first person plural) or *you* and *your* (for second person); others, like *mine*, are too rare in the pilot corpus. Of course, the "polyphony" of pronouns has been pointed out before (Fløttum et al. 2006: 108–110), the semantic differences between inclusive and exclusive *we*, the stylistic differences between *pluralis majestatis* and *modestiae*, and the different reference of *you* as audience and *you* as *(any)one* have been commented on in many style guides.[2]

Table 3 shows the variation of pronoun usage by gender and language again. It is obvious that first and second person pronouns are used only half as often by male as by female speakers. This is not in line with some other research results (such as Hyland 2012: 181 on book reviews), but it is in line with other sociolinguistic research, where women have always been shown to pay more attention to in-group communication and reader-writer relationship. The figures for the three languages analysed are not dramatically different, although English speakers seem to have a preference for first person singular and second pronouns, and speakers of Black South African languages a tendency towards plural pronouns (which could be related to the thesis topics as well as culturally-based collective thinking). Whereas the variation according to personal pronoun usage presents striking patterns for the factor gender, the variation according to mother tongue is not so consistently different.

Table 3. Personal pronoun usage by gender and language (per 1 million words)

	gender			language			
	female	**male**	**average**	**Afrikaans**	**English**	**SAfrLang**	**average**
1st sing	2716	1558	2079	1704	2461	1984	2013
1st pl	1154	795	957	909	782	1048	900
2nd	1082	227	612	599	838	500	653
sum	4953	2580	3767	3211	4080	3531	3607

This is why we concentrate again on the distribution of our pronouns according to discipline type and here we find much more consistent patterns than for modals auxiliaries: Whereas Humanities and particularly, but not surprisingly, Social Sciences are clearly first person singular disciplines, Engineering and Natural Science are much less interested in this personal perspective. This pattern occurs very consistently in all the linguistic features analysed (cf. Hyland 2002: 1098 has high frequencies of *I* in Economic student reports and low in Mechanical Engineering). In particular, first person plural and second person generally are extremely rare in Engineering and Law texts. The average figures also demonstrate that Natural Sciences use *we* and *our*, which can be explained by the discipline-specific group work.

In order to understand the relative proportion of the three pronoun types better, the relative percentage of the three groups was calculated in the lower part of Table 4. This visualises that generally first person singular forms (except in Natural Sciences) make up more than half of all the pronouns (especially in the Engineering subjects, as the others occur hardly at all). The Humanities, Social Sciences and Interdisciplinary category are characterised by reader-inclusive *you* and *your*. The Natural Sciences are clearly plural (*we/our*) and the Engineering subjects clearly singular (*I*) disciplines (if any such personal pronouns are used at all).

If we were to produce illustrative figures for these trends again, the Humanities would be towards the higher end again and the Engineering and Natural Sciences towards the lower as far as overall pronoun usage is concerned. The latter disciplines have complementary structures, alternative language options to express actions without including explicitly writers and readers, especially passive constructions and agentless subjects like *the data/results/analyses suggest*, etc.

This distribution of self-mention correlates with Hyland's results, who claims (2012: 17):

> One example of how academics and students use the resources of their disciplines to negotiate a self-representation is the preference for the use or avoidance of self-mention. Examples like these, from applied linguistics and electrical engineering articles, are commonplace and reflect the fact that explicit reference to the author is over four times more common in humanities and social science articles than those in the hard sciences.

Table 4. Variation in personal pronoun usage by discipline type (per 1 million words)

	Hu	ID	Lw	SS	NS	EG	average
1st sing	3167	1866	788	4128	1032	749	2079
I	2018	1269	663	2660	760	555	1391
me	346	225	41	692	92	79	258
my	803	372	84	777	180	115	430
1st pl	1142	713	193	1064	1464	18	957
we	608	395	75	650	914	9	559
us	195	85	19	196	132	5	126
our	339	233	99	218	418	3	272
2nd	824	948	100	1430	84	65	612
you	610	664	65	1095	56	57	450
your	213	284	35	335	27	8	162
sum	5133	3527	1081	6623	2579	831	3296
1st sing %	62	53	73	62	40	90	63
1st pl %	22	20	18	16	57	2	29
2nd %	16	27	9	22	3	8	19

5. Presenting cohesive linking in texts explicitly

Other language features used for the analysis of academic writing (Hyland 2007) are explicit cohesive devices (*and, but, because, then, therefore*, and *thus* as the most frequent), which I simply call linkers here. Haliday/Hasan (1976) and Quirk et al. (1985) made a comprehensive list of such devices, and about fifty items from these lists have been selected for analysis here. The differences in their use was enormous: some high frequency linkers occurred so often in the ZAMA Corpus that their individual usage could not be analysed by hand, others included in the list did not occur at all. It may be surprising, for instance, that in 100 MA theses the terms *at last* or *to sum up* were not found at all (in contrast to claims by Halliday/Hasan 1976: 238). Semantically, we can distinguish four types of conjunctions: *additive*, *adversative*, *causal* and *temporal* (here called *sequential*). Of course, there is an implicit hierarchy in this list, and I would claim that *causal* is the strongest semantic case, which presupposes some *temporal* element. Table 5 lists linkers (and their frequencies) according to type, and it demonstrates that there are many more *additive* than, for instance, *adversative* types and even (about ten times) more tokens.

The selection of linkers is not so easy since many types of conjuncts are polysemous. Thus, *rather* can be either a conjunct (*he would rather*) or an adverb (*rather big*) in English; these two functions cannot be distinguished in our automatic retrieval program and categorising individual occurrences by hand would be too laborious for our purposes. This is why only lexemes and phrases with a reasonably high proportion of linking functions were included in the list.

The quantitative comparison of types and individual linkers in Table 5 shows that the difference between individual linkers is enormous. By far the most fre-

quent tokens are *and* and (less than 10% of *and*) *also*; all other additive linkers are much less prominent. The most frequent adversative linker is obviously and expectedly *but*. The distribution of causal markers is much more even between *because*, *then* and *therefore*, and (in contrast to Halliday) we are inclined to include *thus* here, which would make the causal function almost as prominent as the adversative. The least important category in terms of frequency is obviously the temporal or sequential one, where only *previously* and *next* occur regularly in our corpus. Finally, it has to be mentioned that linkers are only (explicit) surface related devices that help the reader to establish coherence between elements of a text - there are other much more subtle and maybe powerful lexical and idiomatic devices that can fulfil the same function, i.e. to create coherence in the mind of the reader, but they are not so easy to analyse using corpus-linguistic tools.

Table 5. Frequency of English linkers by function type

additive	**tokens**	**adversative**	**tokens**	**causal**	**tokens**	**sequential**	**tokens**
also	1841	nevertheless	15	because	589	firstly	22
moreover	3	although	185	therefore	489	secondly	20
furthermore	25	yet	105	consequently	25	thirdly	6
and	19857	though	149	hence	57	previously	154
besides	8	but	1191	then	574	afterwards	9
actually	96	however	342	in this respect	2	eventually	48
alternatively	6	in fact	39	for this reason	2	finally	36
regarding	187	instead	79	on account of this	0	lastly	6
similarly	24	rather	252	as a result	111	anyhow	0
likewise	3			on this basis	2	anyway	3
namely	170			whence	0	next	207
in addition	27					at this point	6
incidentally	3					to sum up	0
thus	323					in short	4
for instance	40					in the end	4
in other words	14					ultimately	49
on the other hand	25					at last	0
for example	131						
sum	22786		2357		1850		575

Table 6 shows the linkers according to type by gender and language. Whereas there are few gender differences in our data, Hyland (2012: 181) found that female writers use clearly more “transition markers” (in book reviews). It is interesting to note that African mother tongue speakers obviously prefer the additive type at the expense of the sequential and others. Generally, the differences are not very prominent, neither according to gender nor according to language.

Table 6. Linkers by gender and language (per 1 million words)

	gender			language			
	female	**male**	**average**	**Afrikaans**	**English**	**SAfrLang**	**average**
additive	23517	21948	22654	22560	22554	23558	22791
adversative	2394	2328	2357	2621	2520	1987	2440
causal	2012	1718	1850	2034	1753	1860	1903
sequential	533	609	575	592	675	439	583
sum	28456	26603	27529	27806	27501	27843	27717

The differences in discipline types are more pronounced: this is immediately visible in table 7, where the distribution of linker types is displayed by discipline type. The table demonstrates that the vast majority of linkers are additive (esp. *and*), but the pattern is generally the same: there are no clear adversative, causal, or sequential discipline types (like Social Sciences, Engineering and Natural Sciences, respectively). Some disciplines like Social Sciences and Humanities use many (and all types of) linkers whereas Natural Sciences use only few. The use of explicit linkers may be, more than other language features, determined by the teaching: maybe MA students in Humanities are taught to use explicit devices in lower classes and find a more natural (implicit) balance later (cf. Schmied 2011). A higher academic level, i.e. in research articles, Hůlková (2011: 135) found considerably fewer linkers in Psychology than in Politics, for instance. Here, a cultural and a developmental dimension overlap. Bolton/Nelson/Hung (2002: 176f) calculate that Hong Kong and British students overuse the most frequent linkers from *however* to *thus*, compared to an academic usage norm (per sentence) based on 40 samples "taken from academic papers and books across a range of disciplines, published between 1990 and 1993 inclusively" (ibid: 173). There is obviously room for more empirical comparative research in this field.

Table 7. Linkers by discipline type (per 1 million words)

	Hu	**ID**	**Lw**	**SS**	**NS**	**EG**	**average**
additive	27012	25092	19183	27135	16600	20491	22654
adversative	3472	2670	2312	2420	1549	1432	2357
causal	2141	1949	2014	2143	1411	1656	1850
sequential	705	526	405	934	433	344	575
sum	33329	30237	23914	32633	19992	23924	27338

6. Conclusions

I hope to have shown that the ZAMA corpus of South African MA theses is a good database and that the analysis of disciplinary conventions in academic knowledge presentation shows some interesting variation. The simple descriptive statistics applied in this introduction to the corpus can be used as a simple student-friendly discovery procedure. So far, modal auxiliaries seem to display more

complex patterns than person pronouns and cohesive linkers. The first analysis of the three complex meta-linguistic features is a promising start to a more sophisticated analysis that seeks to measure the relative influence of the independent variables discipline, gender and language background on some of the key features illustrated here. The student-level discussion could start by presenting some concrete examples,[3] and specific figures and tables that visualise the different occurrences of the features discussed here; students could discuss extreme cases and this should be enough to initiate a discussion on departmental writing conventions and preferences in post-graduate teaching. Even without standard deviation and significance tests, students can mark easily over- and under-using disciplines in the tables above and draw practical conclusions for their own usage. For the specialists, the simple research perspective can be expanded into a more sophisticated multiple-regression analysis that could be carried out in Rbrul (Johnson 2012), since it has the convenient interface with R and its graphical capabilities. For such complex statistical procedures, the particular breadth of disciplines included in our pilot corpus may pose a specific problem.[4] We have to reduce the number of disciplines by grouping them into discipline types, but only an empirical cluster analysis would really show whether intuitive categorisations can be justified in this specific university context. This would also make research on the South African data comparable with other sociolinguistic data world-wide.

But even if we can identify all the disciplinary conventions in a wide selection of data, the work and academic discourse has just begun. For, all the departments have to decide whether they want to encourage or discourage discipline-specific features in MA theses - and all individual researchers have to negotiate their own usage conventions, if they are given the choice.

Notes

1 I wish to thank Prof. Arnold van Zyl for pointing out the SUNScolar Research Repository from Stellenbosch University Library as a suitable database for my research and to Sven Albrecht, Dunlop Ochieng, Cornelia Neubert, Matthias Hofmann and Susanne Wagner for the technical help and the discussion on statistics and analyses. Unfortunately, no sample sentences could be included in this analysis, because I focussed on illustrating the usage conventions of a few complex linguistic features in figures (esp. in the first sections) and tables as a discovery procedure and a basis for discussion, but this quantitative approach should also lead back to the texts and a detailed qualitative analysis.

2 Even this article makes a conscious distinction between *I* (taking responsibility for author's action) and inclusive *we* (offering to include the reader into the argumentation), but it uses no exclusive or majestic *we*; these semantic distinctions cannot be discussed in this survey – and are difficult to analyse using corpus-linguistic tools.

3 I am painfully aware of the space-limitations of this contribution: for pedagogical applications, the tables and figures in the article may be insufficient and hundreds of real-language examples of the metadiscourse features in context would make the practical issues much more evident.

4 The high number of disciplines makes a regression analysis impossible because the number of texts in this pilot corpus is still too small and the stratification always includes empty cells. So far the SUNScholar Research Repository does not provide enough MA theses written by male and female authors with Afrikaans, English and Black South African languages in all disciplines and all years.

References

Bautista, Maria Lourdes S. (2004) 'The verb in Philippine English: A preliminary analysis of modal *would*'. *World Englishes* 23, 113–127

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Ed Finegan (1999) *Longman Grammar of Spoken and Written English*. New York: Longman.

Bolton, Kinglsey, Gerald Nelson and Joseph Hung (2002) 'A corpus-based study of connectors in student writing'. *International Journal of Corpus Linguistics* 7 (2), 165–182.

Collins, Peter (2009) *Modals and Quasi-Modals in English*. Amsterdam: Rodopi.

Fløttum, Kjersti, Trine Gedde-Dahl and Torodd Kinn (2006) *Academic Voices*. Amsterdam: John Benjamins.

Gao, Quoping (2012) 'Interpersonal functions of epistemic modality in Academic English Writing'. *Chinese Journal of Applied Linguistics* 35, 352–364.

Halliday, M.A.K. and Ruqaia Hasan (1976) *Cohesion in English*. London: Longman.

Hůlková, Irena (2011) 'Conjunctive adverbials in academic written discourse'. In: Schmied, Josef (ed.) *Academic Writing in Europe: Empirical Perspectives*. Göttingen: Cuvillier Verlag, 129–142.

Hyland, Ken (2002) 'Authority and invisibility: authorial identity in academic writing'. *Journal of Pragmatics* 34, 1091–1112.

Hyland, Ken (2007) *Metadiscourse. Exploring Interaction in Writing*. London: Continuum.

Hyland, Ken (2012) *Disciplinary Identities: Individuality and Community in Academic Writing*. Cambridge: CUP.

Johnson, Daniel E. (2012) *Rbrul* (Version 2.05), http://www.danielezrajohnson.com/rbrul.html (accessed on 9 September 2012).

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik (1985) *A Comprehensive Grammar of the English Language*. London: Longman.

Sala, Michele (fc.) 'Language change in legal research article titles'. In: Poppi, Franca and Winnie Cheng (eds.) *The Three Waves of Globalization: Winds of Change in Professional, Institutional and Academic Genres*. Cambridge Scholars Publishing.

Schmied, Josef (2008a) 'Comparing complexity and interclausal sentence relations in african academic writing'. In Wolf, Hans Georg, Lothar Peter and Frank Polzenhagen (eds.) *Focus on English. Linguistic Structure, Language Variation and Discursive Use*. Leipzig: Leipziger Universitätsverlag, 173–188.

Schmied, Josef (2008b) 'Hedges in specialised vs. popular academic interaction: A case study of medical texts'. *Discourse and Interaction* 1 (2), 85–98.

Schmied, Josef (2011) 'Academic writing in Europe: a survey of approaches and problems'. In: Schmied, Josef (ed.) *Academic Writing in Europe: Empirical Perspectives*. Göttingen: Cuvillier Verlag, 1–22.

Swales, John (1990) *Genre Analysis: English in Academic and Research Settings*. Cambridge: CUP.

Swales, John (2004) *Research Genres: Explorations and Applications*. Cambridge: CUP.

Warner, Antony R. (1993) *English Auxiliaries*. Cambridge: CUP.

Josef Schmied has been Professor (Chair) of English Language and Linguistics at Chemnitz University of Technology since 1993. He wrote a PhD thesis on English in Tanzania and a post-doctoral thesis on Relative Constructions in the LOB and Kolhapur Corpora. He also wrote a textbook on English in Africa and (with colleagues in literature and cultural studies) an introduction to the study of English (in German). He has edited several volumes on academic writing and other applied issues in his series REAL (Research in English Language and Applied Linguistics) Studies recently. His main interests are in corpus- and sociolinguistic methodologies as well as in applications in Africa and (South-)East Asia. Over the last 20 years, he has compiled the International Corpus of English – East Africa (Kenya and Tanzania), the Lampeter Corpus (of 17th/18th century English tracts), the German – English ranslation corpus for the Internet Grammar, and several corpora on academic writing (e.g. the SPACE Corpus of Specialised and Popular Academic English, with Christoph Haase).

Address: Prof. Dr. Josef Schmied, English Language and Linguistics, Chemnitz University of Technology, Reichenhainer Straße 39/222, D-09107 Chemnitz, Germany. [email: josef.schmied@phil.tu-chemnitz.de]