Chung-Hwa Buddhist Journal (2012: 25:149-166)
Taipei: Chung-Hwa Institute of Buddhist Studies
中華佛學學報第二十五期　頁 149-166 (民國一百零一年)，臺北：中華佛學研究所
ISSN:1017-7132

A Relational Database for Text-Critical Studies

Wojciech Simson
Univ. of Zürich

Abstract

After a brief introduction to the scope of a project digitizing the Confucian *Analects* and a short explanation of the working principles of relational databases in general the architecture of the actual relational database used in the project is outlined. The database was designed to store, classify, compare and sort textual variants and to handle a considerable number of textual witnesses in such a way that strains of transmission could be compared with one another. Some attention is also paid to the handling of problems typical for manuscripts like illegible or doubtful characters, lacunae and non-standard characters that are not included in the *Unicode* standard. The database is further enhanced by a tagging system allowing to classify and to analyze different types of variants. Finally an evaluation of the whole system and suggestions for its further development are given.

Keywords:

Digitization, Relational Database, Textual Criticism, Stemmatology, *Lunyu*

經文鑑別研究之關係資料庫

Wojciech Simson
蘇黎世大學

摘要
在簡短的介紹完數位化孔子《論語》專案的概況及對於關係資料庫的一般工作原則之說明後，此計畫所使用的關係資料庫架構得由此可見其輪廓。此資料庫用以設計為貯存、分類、比較並排序文本的差異，同時處理相當數量文本見證，依此，不同的傳承系譜能夠有所比較。另外也關注一些處理寫本上的典型問題，例如難辨認的或可疑的字元，脫漏及未被包含在 Unicode 標準的非標準字元。此資料庫更進一步由標籤系統加強，可以歸類及分析差異的不同型態。另外，此文也提供一完整系統評估及進一步發展的建議。

關鍵詞：數位化、關係資料庫、經文鑑別、文獻系譜學、《論語》

Introduction

Whereas most papers in this volume deal with the digitization of East Asian texts by means of mark-up languages, the following project is quite different as to the digitization method and the kind of text that has been digitized. The text in question is not Buddhist but Confucian and is no other than the well known *Analects* or *Sayings of Confucius* (論語), and it was digitized not in a mark-up language but in a relational database. I think, however, that there is quite a bit of common ground: Among the copious textual witnesses of the *Analects* incorporated into the database there were, among others, more than 70 fragments of Tang time manuscripts stemming mostly from Dunhuang and partly from the ancient city of Gaochang near the modern village of Astana in Xinjiang province. They are very similar in age and in provenience to the Chan texts. The problems with digitization are, therefore, similar: We regularly have to deal with variant characters some of them not to be found even in the largest of dictionaries, we have many textual variants, hardly legible or even illegible passages, and large lacunae. These features are even more prevalent in the *Analects* manuscripts than in Buddhist texts, because the copies of the *Analects* were produced not by accomplished scribes but by children who underwent an elementary education in a more or less public school that must have been integrated into the monastery of Dunhuang. The *Analects* manuscripts were never intended to be treated as holy script that was to be preserved for future generations in a library. They look rather like the wastepaper left over from the school's daily practice. Due to the very frequent scribal errors committed by the young students the *Analects* manuscripts have to be treated with great caution as witnesses of the ancient text, nevertheless as very important witnesses, because they antedate by several centuries the earliest extant printed editions on which the *textus receptus* is based. Moreover, the Dunhuang manuscripts represent strains of transmission that are clearly distinct from the printed editions and, therefore, of great text-critical interest. The very frequent corruptions on the manuscripts were not regarded as a deficiency of this material but, on the contrary, came into a special focus of interest.

Scope of the Project

The primary goal of the project was not to produce a critical edition of the *Analects*, as might be expected from what has been said so far, but to provide the necessary material and methods for such a task. The scope of the project was therefore:

1) To gather the relevant material and to arrange it in a most flexible way for further investigation.
2) To study the mechanisms of textual corruption, i.e. to determine the conditions under which certain corruptions occur and, where possible, to establish rules that would enable a textual critic to discern original readings from errors. For this purpose it turned out to be a great advantage that the elementary students in

Dunhuang and Astana had produced a great amount of very obvious scribal errors that could not be regarded as valid textual variants but made it possible to determine with great certainty which reading is original and which a corruption. Without a clear and reliable identification of the original readings and the errors respectively it would have been simply impossible to develop an adequate understanding of the corruption process.

3) In the third place, one of the aims of the project was to test the applicability of stemmatology to Chinese manuscripts.[1] Stemmatology has been a major and in certain cases extremely efficient and reliable text-critical tool in bible studies and classical European philology.
4) Finally, the project resulted in a comprehensive textual history of the *Analects*[2] separating and describing the main strains of textual transmission and discerning within these strains the dependent from independent textual witnesses.

The Relational Database Approach

Most contributors to the present volume are concerned with the digitization of manuscripts, i.e. they produce digital representations of manuscripts that can be reproduced and read in standard Chinese characters and can be electronically searched or processed otherwise on a computer. Features of the manuscript that might be of interest but cannot be represented in standard Chinese characters like lacunae, uncertain readings, emendations and many others are usually represented by means of a mark-up language, a versatile and extendable code that has been designed to describe such features.

The present project was not so much concerned with the manuscripts themselves but with their differences. From the beginning it seemed to be a detour to digitize every single textual witness of the *Analects*. Though individual digital representations of the many text witnesses could have been easily produced by introducing their variants into an already digitized version of the text, these variants, however, would have been to be sieved out again from these digitized versions by collating them again by means of a specialized software. The whole procedure would have been very susceptible for input and processing errors, compatibility problems etc. It seemed, therefore, more straight-forward to store only the variants right from the start. To store them in a relational database allowed to maintain the data in a relatively flexible form that allowed further modifications and, most important, could be searched and sorted according to various criteria.

1 Cf. Simson (2002).

2 Simson (2006).

Relational Databases

For those not familiar with relational databases a short outline of their underlying working principles is given in this section and may be skipped by those who are well acquainted with them.

Relational databases were not developed to store whole texts of an unlimited length. They are mainly designed to store and handle different types of standardized information with a well defined length. Each type of information is assigned a so called field and types of information that belong inseparably together are stored together in the same table. In a bibliographical database, for example, we would group the title of the book together with its publishing year in the same table, and for each physical book we would always have one data set with the same structure.

Figure 1: Books table

There exists, of course, other very important information about the book, as the author or the publishing house, which is, however, better stored in separate tables. The reason for this splitting up of tables is that one and the same author may write several books as well as one and the same publisher will publish many different books. By separating the books and publishers from the book titles we need to store each piece of information only once. This saves a lot of input time and avoids typing errors because we don't have to retype the name of the author each time a book of his is entered into the database. Moreover, when the data of the author or publisher has to be updated or corrected, we have to change it only once. Ideally the database is built in such a way that it contains no redundant or contradictory data sets, this state is also called data consistency.

The different tables into which the information about the books in our bibliographic database has been subdivided must be related to one another, otherwise we will never find out which book belongs to which author or publishing house. This is achieved by keys.

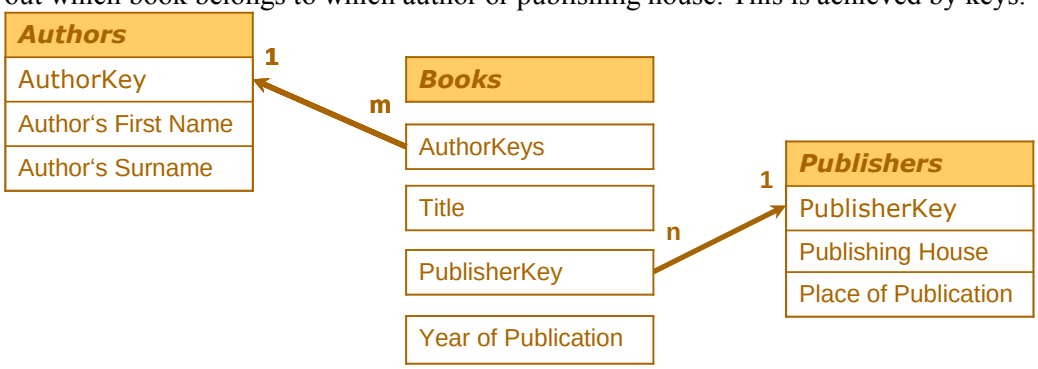

Figure 2: Authors, Books and Publishers tables

Keys are, in most cases, integer numbers generated by the database system automatically. Each author, for example, is assigned a number and this number is stored not only in the respective author's data set but also in the *Books* table to indicate the author of a book. While the *AuthorKey* is unique to the *Authors* table it can be stored in the *Books* table an unlimited number of times. This is called a one-to-many relation and is the mathematical representation of the fact that in real life one author may write several books. The *Publishers* are linked to the *Books* table accordingly. The relation between the *Authors* and the *Publishers* is called a many-to-many relation, and represents the real life situation where the same publishing house publishes books by various authors and the same author publishes his books in different publishing houses. Such many-to-many relations imply always the use of an intermediate or pivot table and can never be established between two tables immediately.

A relational database system provides not only the keys and safeguards their consistency but it is, furthermore, able to maintain a powerful indexing system which allows one to search millions of datasets within fractions of a second.

Chopping up the *Analects*

After this short introduction to relational databases it's time to ask how the text of the *Analects* covering around 15,000 Chinese characters can be stored in the limited fields of such a relational data structure. As already mentioned, it needs not to be stored there at all, because what is in the focus of interest is not the text as a whole but its variants. Collecting the variants alone, however, makes little sense without knowing to which place in the text they belong. It is, therefore, necessary to refer to text passages in an unambiguous way, and to do this it is necessary to lay a grid of coordinates over the text.

The Reference System

The grid follows the conventional way of referring to passages of the *Lunyu*, and moreover, has to be further refined as to be able to point unambiguously to short passages or even single characters within the chapters.

Traditionally the *Lunyu* is divided into 20 books (篇) and each book is further subdivided into chapters (章). Most of these chapters contain one saying of Confucius' and cover not more than a few dozens of characters. Both the books and the chapters have a conventional numbering generally accepted among western scholars. This reference system is taken over in the database and further refined by numbering the characters within each chapter. Because different versions of the text differ slightly in the total number of characters they contain, it is therefore necessary to stick to a certain text version to maintain the reference system unambiguous. This is an easily available electronic version of the *textus receptus*. This leads to the following data structure:

Figure 3: Chapters and Passages tables

The *Chapters* table contains all the chapter numbers of the *Lunyu*, 01.01 being the first chapter of the first book and so on. The reference text is included for practical reasons, but strictly speaking, it is but a help for the user and not an indispensable part of the database. The *Passages* table contains, of course, all the passages to which variants are found. Because a passage consists very often of only one character and the same character or even short phrases can possibly reappear several times within the same chapter, it is essential to store the beginning and end of the passage in the table. The wording of the passage is stored in the table too, but, strictly speaking, it could be also discarded as redundant information.

No overlapping passages are allowed. Otherwise consistency in the overlapping sections of the passages would be very difficult to maintain. Some other information, like transmitted commentaries referring to the passage, is also stored in the *Passages* table. They are skipped here, however, to keep the focus on the essentials.

The two tables are connected with one another by a one-to-many relation with the *Passages* table on the many side. This is the representation of the simple fact that there can be more than one passage within one and the same chapter.

The Variants

Having built a reference system that enables us to localize the variants within the text we can proceed to collect the variants. Of course we will have to attach a further table. It is linked to the *Chapters* table by a one-to-many relation, because there is always more than one variant for a certain passage. It is, of course, essential to know where this variant was found. Therefore we have to introduce another table storing all the textual witnesses of the text.

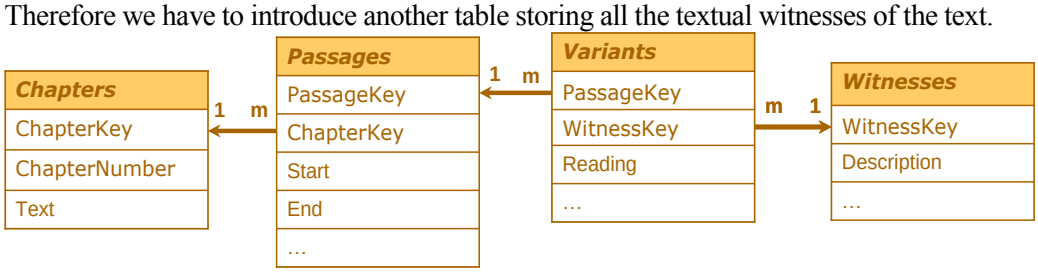

Figure 4: Chapters, Passages, Variants, Witnesses

One may ask here, why the *Witnesses* table is related to the *Variants* table by a one-to-many relation with the *Variant* table on the one side. One and the same witness bears usually not only a great number of variants referring to different passages, but the same variant is very often found on several witnesses. This seems to be a typical many-to-many relation. Basically this is true and it is possible to build the database in such a way. The system would thus just store each variant only once and ignore the readings that coincide with the *textus receptus*. This would make the data less redundant and more consistent. For text critical purposes it is, however, more convenient to have all the readings right at hand and not to have to determine first if a witness that is not listed among the variants has the same reading or no reading at all and has to be, therefore, counted as a lacuna. The *Variants* table stores the readings of all the textual witnesses whether they contain a deviation from the *textus receptus* or not. This entails a lot of redundant data but is more practical for text critical comparisons and for the presentation to the user who gets a good overview over all extant readings (see figure 5). It is, moreover, much easier to write queries with such an arrangement of data than with a mathematically more consistent one. To the user the hitherto established data structure is presented as follows.

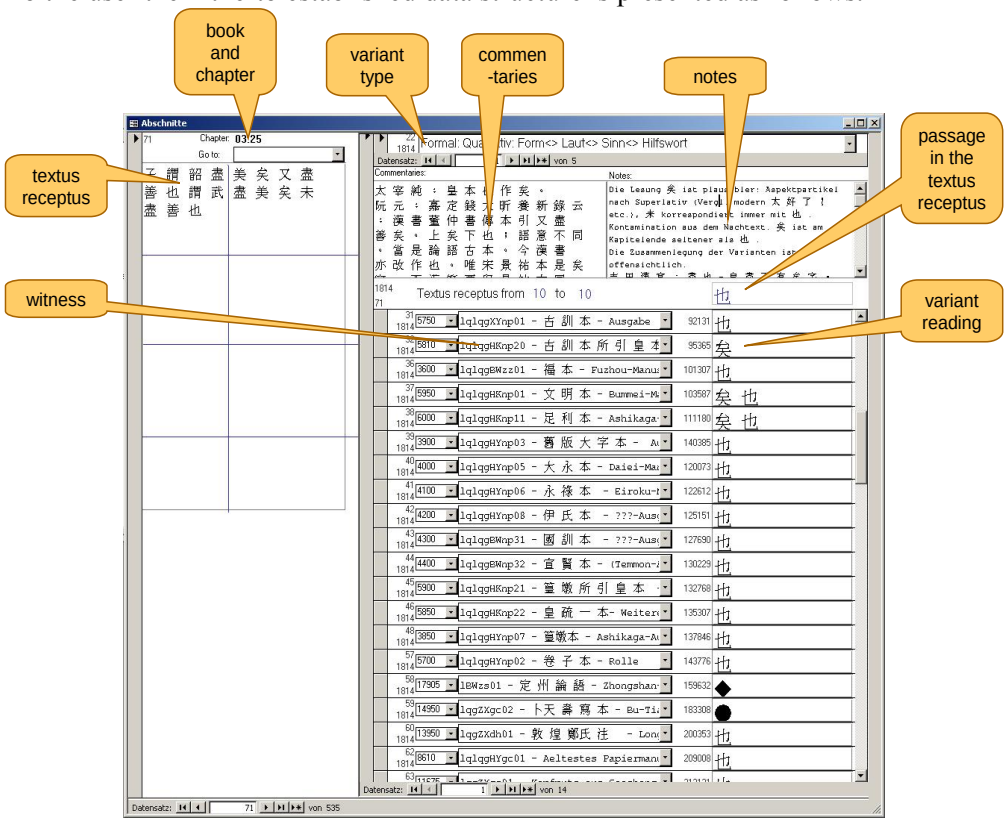

Figure 5: Main window of the user interface

The Representation of Lacunae, Doubtful Readings and Rare Characters

Two kinds of lacunae are differentiated: One type is represented by a black upright square (■). This is used when the number of lacking characters can be counted. This is the case with block prints, stone steles and Japanese manuscripts which have a regular number of characters per line. In cases where the number of missing characters cannot be determined a twisted black square (◆) is used. The reason for this differentiation is that in some cases the witnesses show variants that differ in the number of characters. In such cases we can still decide which version was followed by a certain witness if we count the missing characters.

Illegible characters are represented with an upright outlined square (□) when the illegible characters can be counted and by a twisted outlined square (◇) when they can be not. In doubtful readings every single character is put in brackets (〔 〕).

Another symbol, a black circle (●), is used to represent characters that are lacking from one version while are present in others. Strictly speaking, such a symbol is unnecessary, but it is much more conspicuous than just the lack of a character. This is of some practical importance when you have to scan tens and hundreds of variants (cf. figure 5).

Rare characters that are not included in the Unicode standard put a very serious difficulty to any attempt to digitize Chinese manuscripts. One rather elegant method to treat them is simply to extend the Unicode standard and to add these new characters to the Chinese font on the computer. Thus the non-standard characters can be processed by the computer without difficulties. The *Lunyu*-database takes, however, a different and more complicated approach.

Each character beyond the Unicode standard is put in braces (｛ ｝) which contain the assumed standard form of the character. In many cases this is not enough to identify the character in question unambiguously, because there often exists more than one scribal variant of one and the same character. Therefore, a special table with all the scribal variants was attached to the *Variants* table where all the characters beyond the Unicode standard are stored together with their pronunciation and a description of how they were written.

Example: The scribal variant 扵 for 於 is stored as ｛於｝ together with a description such as "扌on the left + 令 on the right hand side" in a separate table. At first glance the user sees only the form in braces ｛於｝ but can open by double clicking on the character the auxiliary form and read the description.

Though such a complicated treatment of rare characters ensues considerable programming effort to process the data correctly, it has also some advantages to offer. When comparing textual witnesses of the text in order to establish a genealogical tree of the textual witnesses one is usually not interested in orthographical variants, because they mostly give no reliable hints towards the lineage of manuscripts but can be used by

scribes at random. One is looking for so called significative errors instead, because they are, unlike most orthographical variants, irreversible and therefore can be interpreted as traces of the transmission process. By simply ignoring the braces in a query most of the ubiquitous scribal variants are ignored too and this helps a lot to focus on those variant readings that could be useful for stemmatological investigations.

Classifying the Witnesses

We have already several times touched upon the problem of tracing the lineage of textual witnesses. This is necessary in order to apply the stemmatological method which is able to decide between variant readings on the basis of their position on the pedigree of textual transmission and not on interpretative criterions. In order to establish the pedigree or the so called stemma it is very useful to have a reference system for the textual witnesses that makes it possible to group together witnesses that possibly belong to the same strand of transmission in order to check them for uniformity or to compare them against other branches of the pedigree. Though the result looks very logical and simple it took some time and several revisions of the data structure to establish a classifying system for the textual witnesses that is simple and efficient enough to satisfy these needs. The witnesses are classified according to three main criteria:

The Main Lines of Descent

The earliest historical report of the transmission of the *Analects* makes a distinction between three main traditions in which the *Lunyu* appeared in Early Han times. Namely the

L – (Lu 魯論) Tradition from the ancient state of Lu, the native state of Confucius.

Q – (Qi 齊論) Tradition from the ancient state of Qi.

G – (Guwen 古文論語) Tradition in ancient script found in the wall of Confucius' house during the Former Han.

None of these three has survived as an entire text to our days, but some readings could be identified as belonging definitely to the Guwen tradition and some fragments of two lost chapters of the Qi tradition have been handed down to us in commentaries and encyclopedias. All transmitted text versions of the *Lunyu* go back to one major collation:

LQ – A text based on the Lu version into which Qi readings were introduced by Zhang Yu, the Marquis of Anchang (張禹 † 5 BC), during the Former Han.

Another mixed version had to be introduced for the sake of a single but very important textual witness:

LG – The text of a bamboo manuscript found in modern Zhengzhou dating back to the first half of the first century BC.

The Commentarial Traditions

It can be expected that the text of the *Lunyu* was transmitted together with the commentaries which were written down together with the text on the same paper scroll from the second century AD on. By distinguishing the commentarial traditions we can also separate the lines of transmission. In analogy with the main lines of descent these commentarial traditions are abbreviated with the initials of the commentator's name. The earliest extant commentaries stem from the end of the second and the beginning of the third centuries AD:

ZX – Zheng Xuan (鄭玄 127-200)

HY – He Yan (何晏 190-249)

The He Yan commentary was subcommented three times:

HK – Huang Kan (皇侃 488-545)

XB – Xing Bing (邢昺 932-1010)

LDM – Lu Deming (陸德明 550?-630)

Two other later commentaries are included too, because they also contain variant readings that have to be taken into consideration:

HL – Han Yu (韓愈 768-824) and Li Ao (李翺 772-841)

ZZ – Zhu Xi (朱熹 1130-1200)

There is also a handful of manuscript fragments, one print and all the stone steles that don't carry a commentary, though they are all clearly derived from He Yan's version. These are abbreviated as BW for *baiwen* (白文) or plain text.

Types of Witnesses

Apart from the commentaries a further, rather vague, distinction was introduced to account for differences between straits of transmission due to geographical diversification:

dh – manuscripts from Dunhuang (敦煌)

gc – manuscripts from the ancient city of Gaochang (高昌)

np – for Nippon (日本), prints and manuscripts from Japan

Some other types of witnesses were introduced to distinguish groups of witnesses that show typical problems originating from different methods or circumstances of reproduction:

zj – for *zhujian* (竹簡) or bamboo slips

kb – for *keben* (刻本) or Chinese block prints

sj – for *shijing* (石經) or the stone classics erected by various dynasties in front of the imperial academy

Moreover, quotations of the *Lunyu* found with early authors are marked by the initials of the respective author.

Texts having all three criteria – i.e. descent, commentator and type – in common are further differentiated by adding a consecutive number to the abbreviation. Thus a short but meaningful label for every witness is provided.

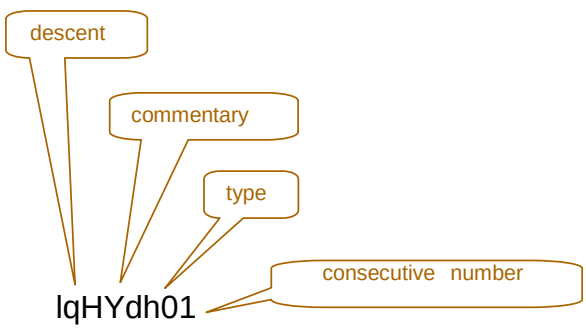

Figure 6: Labeling system

"lqHYdh01" for example denotes the first manuscript from Dunhuang with the He Yan commentary which goes back to the collation of the Lu and the Qi versions of the *Lunyu*. In queries single elements of such a caption can be skipped. We can refer to the whole group of Dunhuang manuscripts as "dh" or to the subgroup of Dunhuang manuscripts with the Zheng Xuan commentary as "ZXdh" and so on. This system allows us to build very precise queries that are able to pinpoint the data we are looking for.

Represented in database terms, this results in a four table structure:

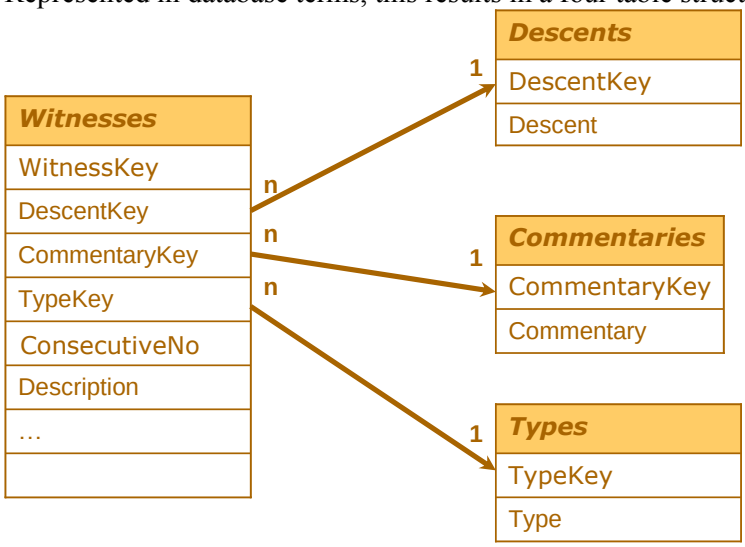

Figure 7: Witnesses with related Descents, Commentaries, Types

The *Witnesses* table is also used to store further information about the witness like a description of its physical appearance or which parts of the text are covered by the witness.

Organizing the data in such a way provides not only a unique label for each textual witness which can be used to refer to the witness for example in an *apparatus criticus*, but it also makes it possible to bundle or separate single strands of transmission and to compare them with one another, which is, of course, a necessary procedure when establishing a genealogical tree.

Classifying the Variants

Apart from the stemmatological investigation of the *Lunyu* tradition one major aim of the project was to provide a tool that could be used for the study of textual corruption. Understanding the mechanisms of corruption is a precondition for what every textual critic is aiming at, namely emendation. To understand these mechanisms it is necessary to sort out certain types of variants and to study them in their specific contexts in order to discover similarities that could be the reason for textual corruption or regularities that could help us to establish rules of emendation. What is needed here is a versatile and handy sieve for textual variants. Because the computer is but a stupid machine that processes mechanically binary data without even the slightest idea of what they are standing for it cannot be expected to sort the variants in a useful way, unless we implement some of our own intelligence into the machine. This is achieved by attaching

at least one label to each passage. This label indicates the type of variant that is found among the various witnesses of the text that cover the passage in question.

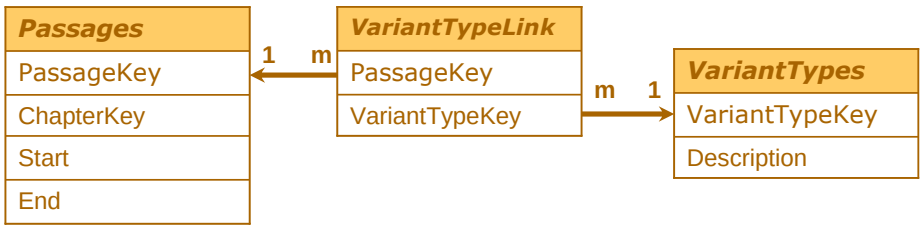

Figure 8: Passages, TypeLink, VariantTypes

As can be seen from figure 8, the relation between the passages and the variant types is many-to-many. This is because a potentially unlimited number of variants can occur for one and the same passage. Moreover, one and the same variant can be classified in more than one way for different purposes. On the other side of the many-to-many relation the same type of variant can occur in different passages. The intermediate *VariantTypeLink* table is necessary to link the *Passages* with the *VariantTypes* in a many-to-many relation, because only one-to-one and one-to-many relations are allowed between two tables.

What types of variants are distinguished then? – At first a very formal classification system for the variants was devised that does not contain any interpretative criteria, i.e. it does not suggest which variants are to be interpreted as errors and which as original readings.

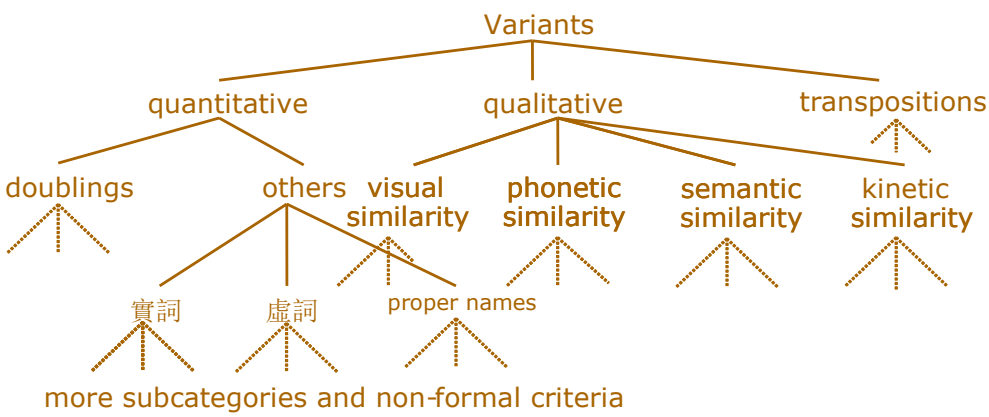

Figure 9: Hierarchy of variant types

The main distinction is between quantitative and qualitative variants and transpositions. Quantitative means that there are differences in the number of characters between the variant readings of a certain passage. These quantitative variants are subdivided into

doublings of single or more characters and others. These others can consist in non-repetitive differences of only one character or of whole phrases. The differences in single characters are further differentiated into those concerning particles and those concerning other words and so on.

The qualitative variants, on the other hand, describe variant readings differing not in the number of characters but in the characters themselves. This category comprises a long series of subcategories of which I want to mention only the major ones: Variants between characters that have a phonetic, visual, logical or kinetic similarity. As the investigation proceeds the need for more differentiation grows constantly and new categories and subcategories can be easily introduced into the system. Subcategories are usually defined in such a way that the label of a generic category is extended. The generic category called "quantitative" (variants) is extended into two subcategories "quantitative/doubling" and "quantitative/others". The former is extended into "quantitative/doubling/character" and "quantitative/doubling/phrase". The former can be further differentiated into "quantitative/doubling/character/particle", "quantitative/doubling/character/proper name" and so on. In queries wildcards can be used to refer to a generic category or to a whole group of subcategories. For example, we can search the database only for "*particle" variants and we will get all variants that concern particles, be they quantitative, qualitative or transpositions. Based on the results of this query we can make a statistic analysis of the types of variants that occur with particles and so on. The whole tagging system is very versatile and easily expandable in order to cope with new questions.

Evaluation of the System

Capabilities

The *Lunyu* is probably the best evidenced secular text in ancient Chinese literature. This means that we have not only a large amount of early textual witnesses at our disposal, but that these witnesses have also a great diversity in age, geographical distribution and commentarial tradition. Moreover, most of the many thousands of variant readings that could be collected in the database are obvious errors. To put it the other way around: In most cases we know with great certainty what the correct reading has to be. We have therefore a very rich source of material that allows us to study under which conditions errors occur and what shape they take. Such a study would provide an empirical data basis for a methodology of textual criticism of ancient Chinese texts. Such an empirically based methodology would be of great importance for every one dealing with ancient Chinese texts. Scholars rely on traditional Chinese emendation strategies and concepts of textual corruption instead. Some of their assumptions cannot be corroborated by empirical data at all or seem at least extremely farfetched when tested against real life material.

An empirically grounded methodology, though not yet fully materialized,[3] was a major goal of the database project. The afore mentioned labeling system of variant readings is the main device for such investigations. It allows us to extract variants of a certain type from the database in order to provide the necessary data for the study of textual corruption. We could focus for example on variants that show phonetic similarities and investigate which degree of phonetic similarity can occur. We can even easily confine our selection of phonetic variants to those coming from Dunhuang in order to investigate the local dialect spoken there in the middle ages. The whole classification system of variants can be adopted very easily to satisfy different needs.

The combination of the two classification systems for textual witnesses and variant categories respectively proves also to be a very powerful tool when it comes to trace the lines of transmission. The database can answer questions like "What variants have the Dunhuang manuscripts in common with the Japanese tradition that are not shared by other witnesses?" The result of such a query can be further confined to potentially significative variants by applying the variant categories to it. To sieve out and to evaluate such variants is the essential task of the stemmatological approach in textual criticism.

Strictly speaking, the database does not provide us with the desired variants themselves, but sieves out all the passages to which a certain type of variants can be found in a certain group of witnesses. For each passage we get therefore a long list of variant readings that we have to sieve again manually to get really at the variants we are looking for. This may look somewhat painstaking at first glance, but variant readings exist only in contrast with other readings and have to be viewed together with them in order to be understood as variants. We need, however, to accept that there may be several distinct variants to the same passage and that the result of a query may contain also some undesired material that we have to sort out manually. Though more than 130 witnesses were incorporated into the database this has never become really a problem.

To sum up we can say that the relational organization of the variants provides a capable tool for the study of textual corruption and stemmatological investigations. It stores also all necessary data for a critical edition. The text critical work can be even supported by adding notes and commentaries to the passages.

Problems

At this point it has first to be mentioned that the *Lunyu* database was never intended to be published or to be used by other people than the author himself. It has always remained a never ending construction site that was modified gradually in order to cope with uprising

3 Partial results will be published in Simson (2013).

questions and a shifting focus of interest. As with all software, a long list of known problems has to be added:

- A major inconveniency when processing Chinese characters with computers is that computers do not provide a useful sorting order for Chinese characters. At best they can be arranged according to their position in the Unicode code table which roughly follows the stroke number of the characters. An arrangement according to pronunciation or radical for example would be much more useful for most practical purposes. This is, however, not a problem of the database approach but of computing in general.
- As some readers may have noticed already, the database makes also use of mark-ups. The handling of rare, doubtful or illegible characters with their brackets and braces involves a special handling of such mark-ups that has to be implemented into the database. This means a lot of programming effort and a slowdown of the whole database because each time characters are processed the character string has to be checked for mark-ups. Moreover, the routines handling the mark-ups are compiled at run-time and this makes them much slower than precompiled programming code.
- Most variants of the text cover only one character and such one character variants can be handled easily by the system. Variants spanning over whole phrases are sometimes more cumbersome, especially when there is more than one variant to the passage and each has developed its own subvariants. The representation in the database becomes rather intricate and the results of queries contain a lot of redundant data that have to be sorted out manually. This has never become a real problem in the project, but the system would have serious difficulties to tackle texts that regularly differ in longer passages of several sentences in length.
- A special difficulty was the handling of lacunae. They were treated like a special sort of variants and required an even larger amount of programming than the mark-ups mentioned before.
- As already mentioned, the database is organized in such a way that the variant readings for a certain passage are stored for each witness separately. This involves a lot of redundant data, because the same variant reading is usually found on several witnesses. Moreover, it takes a lot of programming and computing time to maintain consistency among this redundant data, when manipulating them by introducing new witnesses or passages. A more consistent data structure should be, therefore, considered for the further development of the database.

References

Maas, Paul. 1950. *Textkritik* (2.verbesserte Auflage). Leipzig: B.G. Teubner Verlagsgesellschaft.

Pasquali, Giorgio. 1988. *Storia Della Tradizione e Critica del Testo*. Firenze: Casa Editrice Le Lettere.

Reenen, Pieter van; Mulken, Margot van, ed. 1996. *Studies in Stemmatology*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Reenen, Pieter van; Hollander, August den; Mulken, Margot van, ed. 2004. *Studies in Stemmatology II*; Amsterdam/Philadelphia: John Benjamins Publishing Company.

Simson, Wojciech Jan. Applying Stemmatology to Chinese Textual Traditions. *Textual Scholarship in Chinese Studies*. Ed. Vogelsang, Kai. Papers from the Munich Conference 2000; Asiatische Studien/Études Asiatiques 2002/3. 587–608.

Simson, Wojciech Jan. 2006. *Die Geschichte der Aussprüche des Konfuzius*. Bern: Peter Lang.

Simson, Wojciech Jan. 2013 (forthcoming). Contaminations in Chinese Manuscripts. *The Idea of Writing – Lapses, Glitches and Blunders in Writing Systems*. Ed. Behr, Wolfgang; Voogt, ALex de; Leiden: E. J. Brill.