

English, Chinese and ER diagrams

Peter Pin-Shan Chen*

Computer Science Department, Louisiana State University, Baton Rouge, LA 70803, USA

Abstract

This paper examines the similarities and correspondence between the Entity-Relationship Diagram (ERD) technique and the constructs of two natural languages: English and Chinese. The correspondence between English and ERD was based on the author's previous work, and it was found that the ERD constructs have direct correspondence to the English sentence structure. The correspondence between Chinese and ERD is new research work. It is found that the Chinese character construction has a direct correspondence to the ERD modeling principles. The main result, however, is philosophical in nature: it shows the universality of the primitives used in conceptual modeling no matter whether the primitives are used in data/information modeling or used in recognition/construction of Chinese characters. It also shows that these primitives are embedded in human thinking, and, therefore, their use is very natural in the conceptual modeling process. The results of this research can be useful for improving modeling methodologies, techniques and tools and for improving language translators or better ways to learn or understand a text written in English or Chinese.

Keywords: Entity-Relationship (ER) Model; Entity-Relationship Diagram (ERD); Requirements analysis; Data modeling; Information modeling; System modeling; Chinese character recognition; English sentence structure

1. Introduction

The Entity-Relationship (ER) diagram technique [2,7,10,11,12,13,14,28,29,32,33,38,39] is a graphical tool for information system designers and users in describing and communicating their understanding of the world. Since its formal introduction in 1976, it has been widely used around the world in many systems analysis and database design projects. Although the ER diagram technique is primarily used by information systems professionals and the users of information systems, its use has been spreading to other fields such as accounting and music composition.

On the other hand, natural languages are the daily tools for the general public in describing and communicating their understanding of the world. Because both the ER diagram (ERD) technique and the natural languages satisfy similar human needs, these two "human

* Email: chen@bit.csc.lsu.edu

communication" techniques should have something in common. Furthermore, if we can better understand the correspondence between "ERD" and "natural languages", it is very likely that we can use this knowledge to improve our modeling methodologies, techniques and tools.

In this paper, we will first start with a review of key concepts between "English grammar construct" and "ER diagram constructs". This section is a summary of the author's previous work on this topic. Then, the next section will describe the development history of the Chinese characters and the basic techniques for constructing a Chinese character. We will then describe a set of principles for constructing new characters from the existing characters. Some of these principles are well known by those familiar with the Chinese written language, but we will modify the concepts to fit better with computer and information systems professionals. The other principles are not well known, and they are synthesized by the author to fit the terminology of the modeling community. The purpose of this section is to show the reader that there is a system of guiding principles for constructing and interpreting Chinese characters and these principles could be useful for those in the design, development and use of conceptual/information modeling methodologies and tools. The final section states the conclusions and the future research/application directions. Throughout this paper, we assume that the reader has some basic understanding of the notations and concepts of the ERD technique. For further information on the ER model, ERD technique and their applications, refer to several papers and books listed in [16,17,19,21,25,30,31,35,36,37,40].

2. Review of the correspondence between English sentence structures and ERD constructs

The correspondence between the English sentence structures and the ERD construct was first presented at the 2nd ER Conference in 1981 [8] and later published in [9]. A summary of the basic translation rules are summarized in Table 1. For example, a "common noun" (such as "chair", "person") in English is a possible candidate for an entity type in an ERD. A "proper noun" (such as "John F. Kennedy") is a possible candidate for an entity (an instance of an entity type) in an ERD.

It turned out that this technique can be used in several ways. One application is to use it as an early stage requirement analysis tool: it can help users to identify entity types, relationship types, attributes and high-level ERD constructs based on the English sentence structure.

Table 1
Correspondence between English sentence structures and ERD constructs

English grammar structure	ERD structure
Common noun	Entity type (candidate)
Proper noun	Entity (candidate)
Transitive verb	Relationship type (candidate)
Intransitive verb	Attribute type (candidate)
Adjective	Attribute for entity
Adverb	Attribute for relationship
Gerund (a noun converted from a verb)	An entity type converted from a relationship type
Clause	A high-level entity type which hides a detailed ERD

Recently, researchers in OO (Object-Oriented) Analysis methods are starting to advocate the use of English “nouns” as a way to identify possible “objects,” and this is the same position we advocated in 1981 [8].

Another application is to use it as the basis for (manually or semi-automatically) translating a large amount of existing requirements specification of documents (in English) into ERD-like specifications. Several large consulting firms are practicing the technique in this manner. We can also use the technique in the reverse direction, that is, using ERD to assist the users to formulate a more English-like query. In other words, to use it as a basis for building a more English-like interface to database management systems. Some of the other applications or extensions of this technique (or similar techniques) can be found in [1,3,4,5,6,18,20,22,23,26,27,34,41].

In the following section, we will switch to the discussions on the correspondence between Chinese characters and ERD constructs.

3. The evolution of Chinese characters

3.1. The history of Chinese character development

Chinese written language is one of the earliest written languages in the world. A few decades ago, it was believed that the earliest development of Chinese characters started approximately in 4000 B.C. The recent discovery of some evidence put the date back to approximately 6000 B.C. when a few picture-like characters were carved onto turtle-back shells. Approximately 5500 B.C., some picture-like Chinese characters were carved on pottery. However, the number of characters discovered today for that period is very limited and can only be considered as a very early form of the current Chinese characters. In the Chou Dynasty (approximately 2000 B.C.) and Shang Dynasty (approximately 2000–1300 B.C.), more characters were developed and these characters were carved onto bronze instruments. Particularly, at the end of the Shang Dynasty (1300 B.C.), the character set had about 1000 characters. Also, a parallel development happened at the end of the Shang dynasty (approximately 1300 B.C.); that is, the characters were carved onto oracle bones. The character set discovered on oracle bone inscriptions today totals approximately 4500 characters, which, the anthropologists believe, should have been sufficient to communicate most daily life events during that time period. For more information on Chinese characters, language and culture, refer to [15,24,42].

3.2. The evolution of Chinese characters

Chinese characters are ideograms; each one represents an idea, a thing, etc. Initially, most characters were developed to imitate the image of the real world things. Let us use the author’s name as an example. The author’s middle name is “Pin-Shan”, which is the sound-based translation of two Chinese characters. The second character pronounced as “Shan,” by which the meaning is “mountain (or hill)”. The initial form of the character is shown in Fig. 1a. During the years, in order to simplify the effort of carving multiple strokes

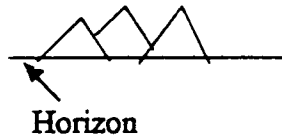

<u>Original Form</u>	<u>Current Form</u>	<u>Meaning</u>
 <p>(a)</p>	 <p>(b)</p>	Mountain (Hill)

Fig. An example of the original form of a Chinese character: (a) original form, and (b) current form.

on bronze or oracle bones, each two strokes to represent an individual hill in this character was reduced into one vertical stroke. The current form of the word is shown in Fig. 1b.

More examples are shown in Fig. 2. Fig. 2a shows the original and current forms of the character "eye". Fig. 2c shows the original and current forms of the character "bird". One of the main differences between the original forms and the current forms of these characters is: the round shapes have been evolved into squared or rectangular shapes while the circles are converted into straight lines. One possible explanation is that it was much easier to carve a straight stroke on bronze or oracle bones.

4. Principles on constructing a Chinese character

Most Western language characters (and even some modern Oriental language characters) are phonetic-based, while Chinese characters are mostly picture-based. How many Chinese characters exist today? The answer is: at least 50 000 characters are in circulation today. How







	<u>Original Form</u>	<u>Current Form</u>	<u>Meaning</u>
(a)			Mouth
(b)			Eye
(c)			Bird

Fig. 2. More examples of original forms of Chinese characters: (a) mouth, (b) eye, and (c) bird.

many characters does a person learn or know after college graduation? The answer is: at least 10 000 characters. How can an average college graduate memorize 10 000 ideograms together with their meanings, pronunciations, usage, synonyms, antonyms, phrases, etc.? How does the brain store and retrieve them? This is a very fascinating problem, but we will not have a complete answer to this question in the near future. However, one may only guess that there must be a way to relate these characters so that it will be easier for a person to recognize and understand them than if there were no relationships between the characters at all. In other words, it will be very difficult for anyone to learn a new character without the knowledge of a structured way of interpreting or guessing the meaning from its shape or its component (if the meaning of the component is known). Fortunately, there does exist a set of principles for constructing new characters from existing ones.

In the following, we will discuss a set of "principles" of special interest to the modeling community. Some of these principles are well known for those familiar with Chinese characters, while the others are either not well known or are synthesized by the author.

4.1. Principle I: Physical resemblance principle

"An ideogram may imitate the physical shape or major features of the 'thing' it tries to represent."











	<u>Original Form</u>	<u>Current Form</u>	<u>Meaning</u>
(a)			Sun
(b)			Moon
(c)			Person
(d)			Tortoise
(e)			Tree

Fig. 3. Physical resemblance principle: (a) sun, (b) moon, (c) person, (d) tortoise, and (e) tree.

For example, Fig. 3 has five characters constructed based on this principle. Fig. 3a shows the initial and current form of the character "sun". The original form shows the shape of the sun with a dot possibly representing the "sun spots". Fig. 3b shows the initial and current form of the character "moon". The initial form shows the shape of the moon with two dots possibly representing the various dark shadows detected by human eyes in the moon. Fig. 3c shows the initial and current form of the character "person". The initial form shows the shape of the person with a head and two legs. Fig. 3d shows the initial and current form of the character "tortoise". The initial form shows the shape of the tortoise with a back shell and legs. Fig. 3e shows the initial and current form of the character "tree". The initial form shows the shape of the tree with a horizontal line representing the earth surface, a three-pronged root and a vertical bar representing the body of the tree.

4.2. Principle II: The subset principle

"Apply or attach a restriction to an ideogram will create a new ideogram which represents a subset of the things represented by the original ideogram."

Fig. 4 shows that if we brace the ideogram, "person", by a rectangular box, we will create a new ideogram with the meaning "prisoner". (What type of person can it be if the person is confined to a place and is not allowed to move out?). In this case, the adding of the rectangular box implies that we imposing a restriction or selection criterion to get a subset of the things represented by the original ideogram.

4.3. Principle III: Grouping principle

"An ideogram, which is a duplex or a triplet of an existing ideogram, has a meaning of "many of" or "a group of" the original thing represented by the original ideogram."

Fig. 5 illustrates a set of new characters created by this principle. Fig. 5a shows the ideogram of a "tree". If two trees are together it implies a "forest". If three trees are together, it implies a "large forest". Fig. 5b shows the ideogram of a "fire". If two fires are together, it implies the concept of "very hot". Fig. 5c shows the ideogram of a "sun". If three suns are together, it implies the concept of "very bright" or "shining". Fig. 5d shows the ideogram of a "mouth". If three mouths are together, it implies the action, "to taste".

4.4. Principle IV: Composition (aggregation) principle

"The meaning of a new ideogram is the combination of the meaning of its component ideograms."

Fig. 6 depicts an example of this principle. The ideogram, "mouth" and the ideogram, "bird", combines into a new ideogram with the meaning of "bird's singing".

$$\text{人 (person)} + \square \text{ (movement is confined)} = \square \text{ (prisoner)}$$

Fig. 4. Subset principle.

	<u>One Instance</u>	<u>Two Instances</u>	<u>Three Instances</u>
(a)	木 (tree)	林 (forest)	森 (large forest)
(b)	火 (fire)	炎 (very hot)	
(c)	日 (sun/sunlight)		晶 (very bright/ shining)
(d)	口 (mouth)		品 (to taste)

Fig. 5. Grouping principle: (a) tree, (b) fire, (c) sun/sunlight, and (d) mouth.

口 (mouth) + 鳥 (bird) = 鳴 (Bird's singing)

Fig. 6. Composition (aggregation) principle.

4.5. Principle V: Commonality principle

“A new ideogram is formed by concatenation of two or more ideograms. Its meaning is the common property of these component ideograms.”

Fig. 7 illustrates an example of this principle. What does “sun” and “moon” have in common? The answer is: “bright” or “brightness by light”.

4.6. Principles VI: An instance-of principle

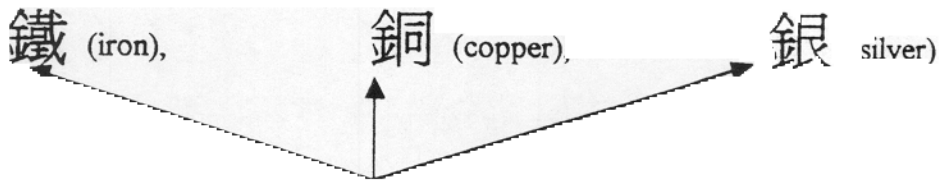
“An ideogram may be made of two component ideograms. Usually, the left one represents the (entity) type, while the other component ideogram indicates a special instance of this type.”

Fig. 8 shows all three ideograms, “iron”, “copper” and “silver”. They all have the same left-hand component, which is also an ideogram by itself with the meaning, “metal”. Therefore, all three characters indicate that they are special instances of the metal types. Currently, there are more than 100 characters, which are instances of the “metal” group. This principle is similar to the “root” concept in Western Languages.

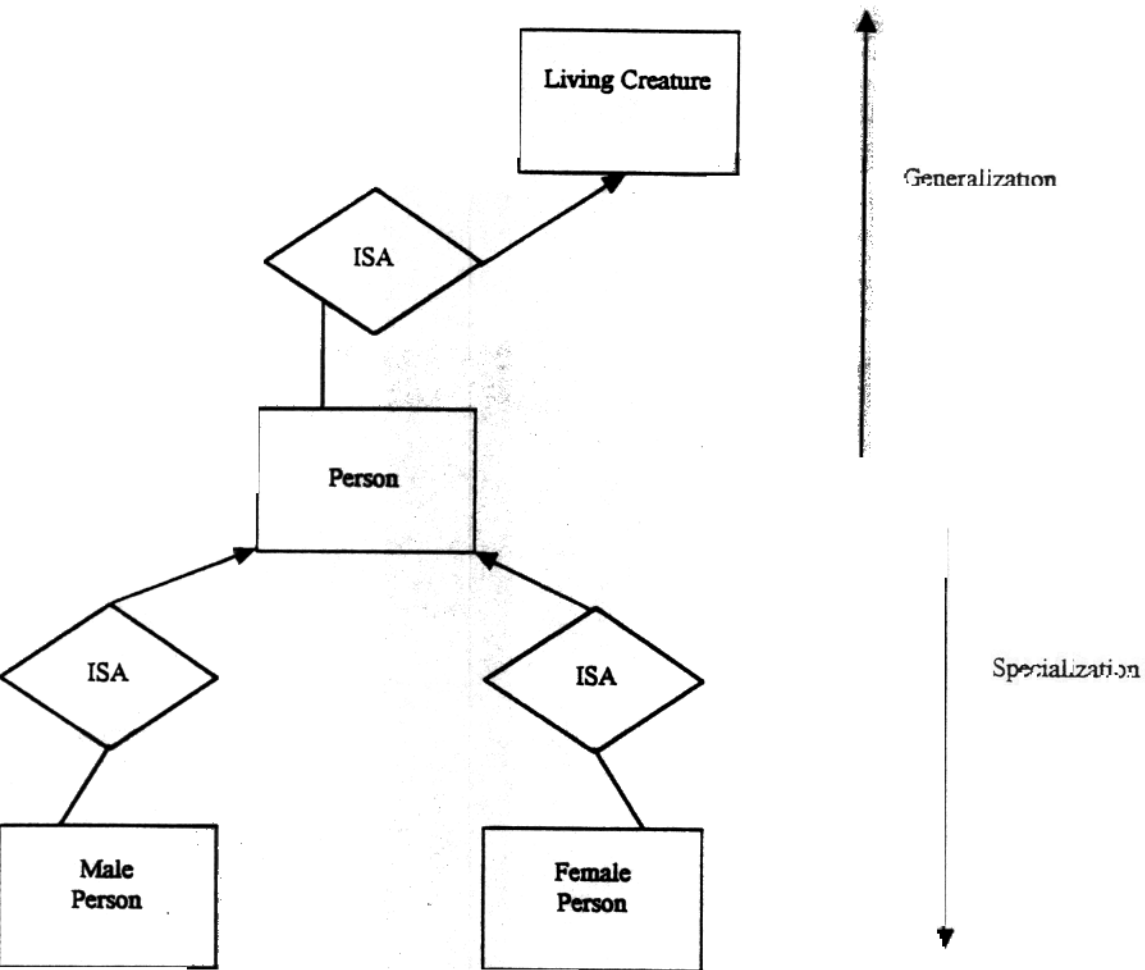
日 (sun) + 月 (moon) = 明 (Bright/ Brightness by light)

Fig. 7. Commodity principle.

Most of the metals have a special left-hand component.

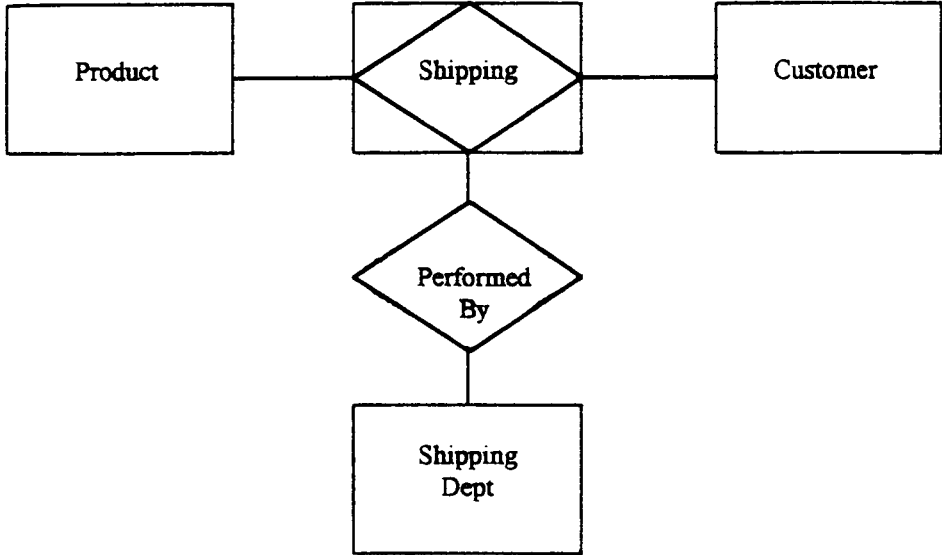


Indicates that metal type



Approach#1

Gerund



Approach#2

Assignment

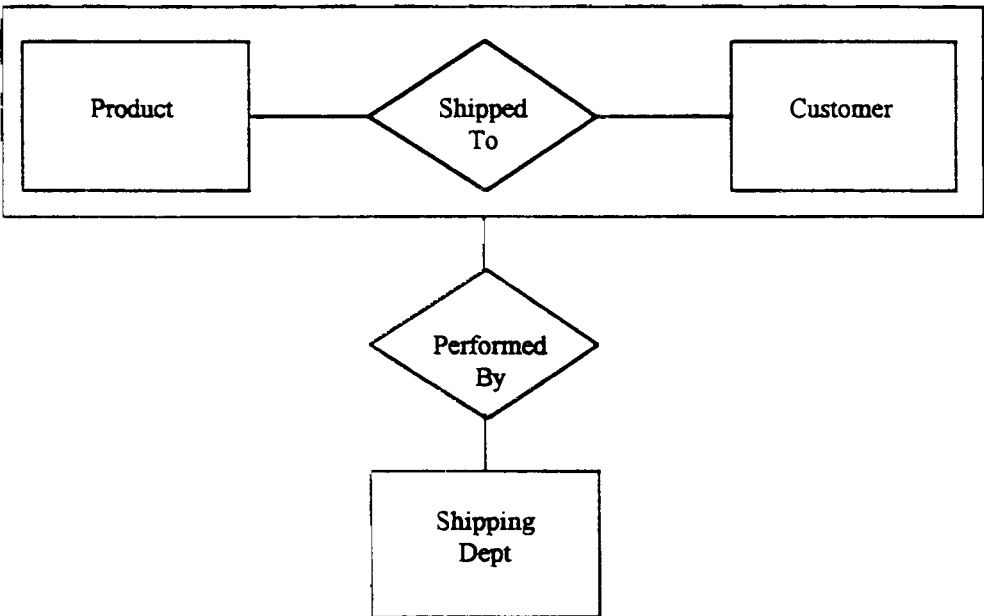


Fig. 10. High-level entity types ("relationship types defined on relationships").

5. Discussion, conclusion and future research directions

Fig. 9 shows the “subtype” relationship in ERD notation. Fig. 10 shows the “aggregation” (high-level entity type) concept in ERD notation. From these two diagrams, we can see some similarity between ERD constructs and Chinese character construction principles.

We can see that there is a lot of similarity between the principles used in constructing Chinese characters and the principles the information system professionals used in data or process modeling. We have demonstrated the universality of the primitives used in conceptual modeling no matter whether the primitives are used in data/information modeling or used in recognition/construction of Chinese characters. This is strong evidence that these primitives are embedded in human thinking and can naturally be used in the conceptual modeling process.

We believe that both camps can learn from each other. The information modeling people can learn from Chinese and English to improve the modeling methodologies and tools, while the people in the process of learning Chinese/English or building translators from one language to another can utilize some of the ERD constructs to help learn more quickly or to design a better translator.

There are several possible ways to extend this research work. One direction is to investigate other natural languages to see whether we can find something similar or different to English and Chinese. Another direction is to understand the rationale for the development and evolution of the character set and their construction principles. Another direction is to investigate how the human mind works in recognition of character sets with or without the help of this set (or another set) of construction principles. Therefore, it is possible to extend this research to the development of new methods of learning and the translation of natural language and characters and statements quickly. In the past, the Chinese characters are always the roadblock of people trying to learn more about the Chinese culture. Using this technique, it may be possible to eliminate or reduce the difficulty of this roadblock. In summary, this type of research can be very fruitful on many fronts.

References

- [1] C. Bagnasco, P. Bresciani, B. Magnini and C. Strapparava, Natural language interpretation for public administration database querying in the TAMIC demonstrator, in R.P. van de Riet, J.F.M. Burg and A.J. van der Vos (eds.), *Applications of Natural Language to Information Systems* (IOC Press, Amsterdam, 1996) 165-176.
- [2] C. Batini, S. Ceri and S. Navathe, *Conceptual Database Design: An Entity-Relationship Approach* (Benjamin/Cummings, Redwood, 1992).
- [3] A.T. Berztiss, Natural and formal languages in the development of information systems, in R.P. van de Riet, J.F.M. Burg and A.J. van der Vos (eds.), *Applications of Natural Language to Information Systems* (IOC Press, Amsterdam, 1996) 5-14.
- [4] E. Buchholz, A. Düsterhöft and B. Thalheim, Capturing information on behavior with the RADD-NLI: A linguistic and knowledge-based approach, in R.P. van de Riet, J.F.M. Burg and A.J. van der Vos (eds.), *Applications of Natural Language to Information Systems* (IOC Press, Amsterdam, 1996) 187-198.
- [5] J.F.M. Burg and R.P. van de Riet, The impact of linguistics on conceptual models: Consistency and understandability, *Data and Knowledge Eng.* 21 (1997) 131-146.

- [6] J.F.M. Burg and R.P. van de Riet, Analyzing informal requirements specifications: A first step towards conceptual modeling, in R.P. van de Riet, J.F.M. Burg and A.J. van der Vos (eds.), *Applications of Natural Language to Information Systems* (IOC Press, Amsterdam, 1996) 15–28.
- [7] P. Chen, The entity-relationship model: Toward a unified view of data, *ACM TODS* 1 (1976) 1–36.
- [8] P. Chen, English sentence structures and entity-relationship diagram, in P. Chen (ed.), *Entity-Relationship Approach to Information Modeling and Analysis* (North Holland, Amsterdam, 1983) in *Proceedings of 2nd International Conf. on Entity-Relationship Approach*, Washington, D.C. (1981).
- [9] P. Chen, English sentence structures and entity-relationship diagrams, *Information Sciences* (1983) 127–149.
- [10] P. Chen, The time-dimension in the entity-relationship model, in H.-J. Kugler (ed.), *Information Processing* (Amsterdam, 1986) 387–390.
- [11] P. Chen and A. Zvieli, Entity-relationship modeling of fuzzy data, *Proc. of 2nd International Conf. on Data Engr.*, Los Angeles (1987) 320–327.
- [12] P. Chen, N. Chandrasekaran and S.S. Iyengar, The denotational semantics of the entity-relationship model, *International Journal of Computer Mathematics* (1988) 1–15.
- [13] P. Chen, RER modeling for multimedia applications on the Internet, *Proceedings of 1995 Conference on Applications of Databases, ADB '95*, San Jose, CA (1995).
- [14] P. Chen and A. Yang, Efficient data retrieval and manipulation using Boolean entity lattice, *Data and Knowledge Eng.* 20 (1996) 211–226.
- [15] D. Crystal, *The Cambridge Encyclopædia of Language* (Cambridge University Press, Cambridge, 1987) 200–201.
- [16] C.J. Date, *An Introduction to Database Systems*, Vol. 1, 6th edition (Addison-Wesley, Reading, MA, 1995).
- [17] R. Elmasri and S.B. Navathe, *Fundamentals of Database Systems* (Benjamin/Cummings Publishing, 1996).
- [18] J.A. Gulla, A.J. van der Vos and U. Thiel, Retrieving conceptual models on the basis of word semantics, in R.P. van der Riet, J.F.M. Burg and A.J. van der Vos (eds.), *Applications of Natural Language to Information System* (IOC Press, Amsterdam, 1996) 117–128.
- [19] I.T. Hawryszkiewicz, *Database Analysis and Design* (Science Research Associates, Chicago, 1984).
- [20] J. Hoppenbrouwers, A.J. van der Vos and S. Hoppenbrouwers, NL structures and conceptual modeling: The KISS case, in R.P. van de Riet, J.F.M. Burg and A.J. van der Vos (eds.), *Applications of Natural Language to Information Systems* (IOC Press, Amsterdam, 1996) pp. 199–212.
- [21] J.G. Hughes, *Object-Oriented Databases* (Prentice-Hall, New York, 1991).
- [22] P. Johannesson, Supporting schema integration by linguistic instruments, *Data and Knowledge Eng.* 21 (1997) 165–182.
- [23] M.-S. Jun, S.-Y. Park and M.-S. Kim, The fast extraction using the keyfact, in R.P. van de Riet, J.F.M. Burg and A.J. van der Vos (eds.), *Applications of Natural Language to Information Systems* (IOC Press, Amsterdam, 1996) 141–152.
- [24] R. Kapp, *Communicating with China* (Intercultural Press, Chicago, 1983).
- [25] H.F. Korth and A. Silberschatz, *Database System Concepts*, 3rd Ed. (McGraw-Hill, New York, 1996).
- [26] G.J.H.M. Kristen, Understanding business languages, in R.P. van de Riet, J.F.M. Burg and A.J. van der Vos (eds.), *Applications of Natural Language to Information Systems* (IOC Press, Amsterdam, 1996) pp. 3–4.
- [27] G. Levreau and M. Bouzeghoub, HypER—an extended E/R model with hypertext facilities, in R.P. van de Riet, J.F.M. Burg and A.J. van der Vos (eds.), *Applications of Natural Language to Informations Systems* (IOC Press, Amsterdam, 1996) pp. 79–92.
- [28] T.W. Ling and M.-L. Lee, Overview of an entity-relationship based database management system, *Future Database '92* (World Scientific, Singapore, 1992).
- [29] P. Loucopoulos (ed.), *Entity-Relationship Approach—ER'94: Business Modelling and Re-Engineering*, Lecture Notes in Computer Science, No. 881 (Springer-Verlag, Berlin, 1994).
- [31] M. Modell, *A Professional's Guide to Systems Analysis* (McGraw-Hill, New York, 1988).
- [32] M.P. Papazoglou (ed.), *OOER'95: Object-Oriented and Entity-Relationship Modeling*, Lecture Notes in Computer Science, No. 1021 (Springer-Verlag, Berlin, 1995).
- [33] C. Parent, H. Rolin, K. Yetongnon and S. Spaccapietra, An ER calculus for the entity-relationship complex model, in F.H. Lochovsky (ed.), *Entity-Relationship Approach to Database Design and Querying* (North-Holland, Amsterdam, 1990).

- [34] A. Paula Ambrosio, E. Metais and J.-N. Meunier, The linguistic level: Contribution for conceptual design, view integration, reuse and documentation, *Data and Knowledge Eng.* 21 (1997) 111-129.
- [35] A.-W. Sheer, *Enterprise-Wide Data Modeling* (Springer-Verlag, Berlin, 1989).
- [36] T.J. Teorey, D. Yang and J.P. Fry, A logical database design methodology using the extended entity-relationship model, *ACM Computing Survey* 18(2) (June 1986).
- [37] T.J. Teorey, *Database Modeling & Design* (Morgan Kaufmann, San Francisco, 1990).
- [38] B. Thalheim, Foundations of entity-relationship modeling, *Annals of Mathematics and Artificial Intelligence* 7 (1993) 197-256.
- [39] B. Thalheim (ed.), *Conceptual Modeling—ER'96*, Lecture Notes in Computer Science, No. 1157 (Springer-Verlag, Berlin, 1996).
- [40] J.D. Ullman, *Principles of Database and Knowledge-base Systems*, Vol. 1 (Computer Science Press, Maryland, 1988).
- [41] B. van der Vos, J. Atle Gulla and R. van de Riet, *Data and Knowledge Eng.* (1997) 147-163.
- [42] K.C. Wu, *The Chinese Heritage* (Crown Publishers, New York, 1982).



Peter P. Chen is Foster Distinguished Chair Professor of Computer Science at Louisiana State University. Previously, he held regular faculty positions at M.I.T. and U.C.L.A. and visiting faculty positions at Harvard and M.I.T. He received a Ph.D. degree from Harvard University in 1973. His research interests include Entity-Relationship models, database design, analysis and design methodologies, CASE tools and the information superhighway.