Chapter 7

Discussion
7.1 General Summary

This dissertation is concerned with investigating SLD from a DST point of view. One purpose of this dissertation is to examine how a DST perspective can help us understand language development in written and oral productions in terms of lexicon and syntax. The other purpose is to explore new analytical methods to investigate the nonlinear dynamic development of language.

SLA studies aim at explaining the phenomena involved in second language learning (Jordan, 2004). These studies encompass “the simultaneous and sequential acquisition and loss of second, third, fourth, etc. languages and dialects by children and adults learning naturally or with the aid of instruction, as individuals or in groups, in second or foreign language settings” (Long, 1993, p.225). In this dissertation, the goal is not to come up with a theory of all of these aspects, but to focus on some specific aspects and to show how the principles used may be applied more widely. The perspective of choice is DST. With a DST perspective of SLD, it is assumed that language development is an iterative process in which the next state of the system is a modification of the previous state. During language development, the embedded linguistic subsystems (e.g. phonology, syntax, lexicon) change in different ways and interact with each other over time, which leads to the emergence of the new state of the whole language system. The language system is a subsystem embedded in a larger cognitive system; the cognitive system is a subsystem embedded in learner’s body and mind; the learner’s body and mind are subsystems embedded in the language environment. These open changing subsystems continuously interact with each other starting from the initial state of the subsystems. Due to the iterations of each subsystem and complex interaction of the subsystems, the language system of each language user is essentially different. Each learner’s language development should be seen as an individually owned process, which is shaped by nonlinear changing subsystems over time. Two dynamic growth models, the logistic model and the hidden Markov model, are used in the current dissertation, allowing us to deal with the iterative nature of developing complex dynamic systems and the non-linear nature of the embedded subsystems interacting over time. The longitudinal case studies of different developmental dimensions, especially the case of the identical twins, show clearly that language development is an individually owned process.

The first two studies in this dissertation (Chapter 3 and Chapter 4) investigated how the language environment influenced the embedded subsystem of vocabulary learning. The first study observed a group of beginning learners of English in Taiwan and found that natural exposure to English input had an effect on learning contextual vocabulary knowledge, collocations and associations over eight months. The second study took a closer look at the process of how the natural exposure to English input influenced vocabulary size. This study looks at the dense data of vocabulary size longitudinally with the dense corpus generated from four beginner learners of English in Taiwan, three of which were randomly selected from the
group study in the first study. Both studies provided valuable information on the
dynamics of second language acquisitions: the relation between language environ-
ment and vocabulary acquisition in size and the contextual knowledge. The group
study in the first study demonstrated the input factor affecting the acquisition of
contextual vocabulary knowledge at three moments in time; the case study in the
second study showed how input factor interacted with vocabulary size by observing
the developmental process at 56 moments in time. The second study, in addition,
evaluated the analytical DST method, which is the logistic model, in terms of the
goodness of logistic equation to quantify the lexical developmental patterns.

The last two studies in this dissertation investigated how the embedded subsys-
tems, writing and speaking, interacted with each other over time in terms of lexi-
con and syntax. These two studies found that the interaction between writing and
speaking showed a complex relationship but writing had higher lexical complex-
ity than speaking whilst speaking had higher syntactic complexity than writing.
These two chapters investigated two identical twins on their writing and speaking
with lexical and syntactic measures over eight months at 100 moments in time and
also explore a newly-applied DST analytical method, the hidden Markov model
(HMM).

7.2 Vocabulary learning and Input

One of the developmental subsystems that this study focuses on is the lexicon. In
1980, Meara argued:“Vocabulary acquisition is part of the psychology of second
language learning that has received short shrift from applied linguistics, and has
been very largely neglected by recent developments in research. This neglect is all
the more striking in that learners themselves readily admit that they experience
considerable difficulty with vocabulary, and once they have got over the initial
stages of acquiring their second language, most learners identify the acquisition of
vocabulary as their greatest single source of problems” (p.1).

The observed neglect of vocabulary in the 1980s has now been resolved to a large
degree. Over the last 30 years, the interest in vocabulary has grown considerably
with a small but very active group of researchers that include, in addition to
Paul Meara, researchers like Paul Nation, Norbert Schmitt and Bill Grabe and
Batia Laufer. Presentations on vocabulary at major conferences typically attract
a large audience, and there are articles on it in the major journals. Meara (2012)
presents a bibliometric analysis of the research literature on vocabulary in 2006
using author co-citation. With this technique, it can be shown who is co-cited with
whom and so provide a graph or mapping of the field with the main players. The
analysis shows that there were two clusters. One is a psycholinguistic cluster with
De Groot, Kroll and Green as the main scholars, and the other is a vocabulary
cluster lead by Nation and Laufer. More importantly, the analysis shows how large
the network of researchers on vocabulary is today.
With the growth of the field, the question or topic studies have grown as well. One of the main themes continues to be how learners manage to acquire a large set of words in relatively little time. The discussion has focused on two issues mainly: one is the role of active explicit word learning vs. implicit word learning (Schmitt, Schmitt & Mann 2011, Ortega 2014); the other one is what constitutes word knowledge, and in particular the depth of vocabulary (all the things we can know about a word) vs. width of vocabulary (how many words do we know). In this dissertation, these two issues, which are highly influenced by input, are dealt with from a DST perspective.

The strongest proponent for the role of input in vocabulary development is no doubt Stephen Krashen. His 1989 article “We acquire vocabulary and spelling through reading: additional evidence for the input hypothesis” presents his position very clearly. He argues: “In my view, the most promising hypothesis is that vocabulary and spelling are acquired in fundamentally the same way as the rest of the language is acquired.” (p. 440). The input hypothesis assumes that we acquire language by understanding messages. The input hypothesis is part of his larger theory on the natural approach which assumes incidental learning as the most important mechanism. While Krashen’s work has become extremely popular among teachers, it has met with serious critique on a number of assumptions. McLaughlin (1987) argues that most of the basic terms in Krashen’s model, like “acquisition”, “learning”, “incidental” and “input” are not well defined. Most of the critiques concern his ideas about grammatical aspects, not so much on vocabulary. Ellis (1994) points out that none of Krashen’s studies demonstrate the validity of the Input Hypothesis.

The research on various forms of input for vocabulary learning was summarized by Laufer (2009). In her overview of a large set of articles on this topic, intensity of contact and variation therein, one of the main points in the present dissertation by observing the vocabulary development through the source of learning (incidental learning by natural exposure to English input), have not been studied specifically so far according to her overview. There are many studies that focus on the amount and type of input, but there is no single solution or final conclusion on this issue. She concluded that “most recent approaches to vocabulary learning attach less importance to the source of learning and more to the quality of elaboration of word information, task involvement and frequent rehearsals.” (p. 341).

### 7.3 Components of word knowledge

Learners need adequate exposure to input in order to learn vocabulary, but learning vocabulary well requires more than extensive exposure as learners need a wide range of vocabulary knowledge to use the words well (depth of vocabulary). Three main types of vocabulary knowledge, word form, word meaning, and word use, as suggested by Nation (2001, p.27) explain how well learners “know” a word in
a thorough manner. Word form is composed of a written form, a spoken form, and morphological composition; word meaning is composed of linking word form and meaning, concept (e.g. associations), and referents; word use is composed of grammatical functions, collocations, and constraints on use. Generally speaking, among these three main aspects of vocabulary knowledge, word form is seen more as a foundation of starting to know/produce a word in isolation whilst word meaning and word use are relevant for knowing and producing a word in a context. For instance, knowing the L2 word “surprised” in L1 translation is the starting level of learning the word, but knowing that the word “surprised” is usually used together with “at” (collocations) is the contextual level (contextual vocabulary knowledge). Knowing a word “big” in its own L1 translation is the starting level of knowing the word, but knowing that the synonym of “big” is “large” and the antonym is “small” relates to another type of contextual vocabulary knowledge.

Apart from the challenge of learning a wide range of vocabulary knowledge, the other challenge of vocabulary learning is enlarging the vocabulary size (width of vocabulary)—vocabulary size refers to the number of vocabulary items a learner has acquired. This is relevant because one should know a certain stock of words in order to comprehend the content. For instance, 98% coverage of the written discourse (8000-9000 word families) has been shown to be sufficient for learners to comprehend authentic texts (e.g. novels or newspapers) (Nation, 2006). Many studies have been carried out to estimate the difficulty of the English input and to match the degree of coverage to comprehend such English input. However, there are few studies aiming at estimating learners’ vocabulary size for the purpose of providing appropriate input. For instance, a beginner learner should be provided with beginning level of learning material; a person who has an estimated vocabulary size of 1000 should be provided with texts with not many more than 1000 most frequent words for better learning results.

However, estimating a learner’s vocabulary size is more difficult than estimating the vocabulary size of the texts. The first challenge of estimating learners’ vocabulary size lies in the fact that it fluctuates constantly. The difficulty of the English input is generally static whilst learners’ vocabulary size fluctuates from time to time, day by day. Estimation of the vocabulary size of one learner cannot be determined by one point or two points of the time. If we want to do justice to the process-based nature of vocabulary development, it should be followed longitudinally with dense measurements over time. The second challenge of estimating one’s vocabulary size is that while the L2 English input (e.g. book) has a fixed context where the corpus is available for analysis, one learner’s corpus, gathered from his or her language productions, is generally not large enough for analysis. To measure one’s vocabulary size more precisely, it is required to have not only dense data of one’s language production, but also an appropriate analysis tool which is sensitive enough to detect or reveal the changes in vocabulary size within limited corpus. The last challenge of estimating one’s vocabulary size is that it is difficult to know when a “word” is known and to what extent a “word” is known. Is a word known when the reader recognizes it? Or is it known when the reader
knows all the contextual characteristics of a word?

In the case studies discussed in this dissertation, estimation of learners’ vocabulary size was dealt with in accordance with the challenges mentioned in the previous paragraph. Firstly, the learners’ written and oral productions were collected on almost daily bases (five times a week), and each production sample was at least 200 words in length. The dense data together with reasonable size of corpus allow us to make a better estimation of the vocabulary size than sparse data with small available corpus. The tool used in this study (V-size) clearly specifies the type of vocabulary size observed: productive vocabulary size. V-size, as a measure of productive vocabulary size, seems to be more sensitive to detect the changes of the productive vocabulary size in the nearly daily observations than the Lexical Frequency Profile (LFP).

7.4 Language environment and its embedded subsystems

A rich language environment provides the L2 learners' embedded vocabulary learning subsystem more chances to encounter the same word not only more frequently but also in different contexts. However, encountering a word may not be enough. In line with the involvement hypothesis (Hulstijn & Laufer, 2001), the more involvement the learner has with the words, the more likely it is that the words are learned. For instance, there is less involvement with a word only scanned by the learner in a text than when the word looked up in the dictionary. We found in the first study that pure incidental input had an effect on learning contextual vocabulary knowledge (vocabulary width) because the increasing frequency and various contexts made learners construct the conceptual representation of the word meaning in L2 and shaped the network of associations in L2. These two factors in turn facilitate the acquisition of contextual vocabulary knowledge through natural exposure to English input. Similarly, the second study showed that pure incidental input had an effect on vocabulary size. More specifically, it demonstrated how one learner moved from less difficult word use to more difficult word use where the trajectory of the change of the difficulty of the word use was closely observed. When learners have more chances to be frequently exposed to the English input, they are more likely to learn new words from the input and thus use more newly-learned words in their own writing.

As language is a dynamic system where the language environment and its subsystems interact with each over time, there are two dimensions of perspectives for describing the dynamic of language development: spatial averaging and time averaging. The first study in this dissertation (Chapter 3) observed the “spatial averaging” (Birkhoff, 1931) perspective to describe the average properties of many identical individuals comprising the system, which is what most of product-oriented studies do; the second study observed the “time averaging” perspective to
describe the average properties of a single individual over a period of time, which is what most of process-oriented studies do. These two perspectives of investigating the dynamic language system are not exclusive but, instead, complement each other. The choice between either a product-oriented or a process-oriented study depends on the research questions to be answered. If we wish to investigate the global patterns of second language learning and the linearly related factors affecting the language system at one moment in time, the spatial averaging perspective is more appropriate. If we wish to understand the developmental process with multiple subsystems interacting with each other over time, the time averaging perspective is more appropriate. As these two types of studies function differently in describing the dynamical system (van Geert, 2011), findings of group studies cannot be equated with any case study and vice versa. Therefore, the need to argue to what extent the case study can be generalized to the group does not seem to be very relevant.

One difficulty of investigating the impact of language environment on vocabulary learning is to control the language environment itself. In a large scale study carried out by De Bot et al. (2004) in Netherlands. Impact of input outside the classroom setting was found to have a strong impact on English proficiency among Dutch pupils, it is almost impossible to control the input outside the classroom setting as Dutch learners are exposed to English input via TV, Internet, or ads frequently. In comparison with the language environment in Netherlands, the participants in the first study are students from Taoyuang, a city in Taiwan, where English input is rarely seen or heard outside the classroom, and students either have no access to English input or do not ever think of using English input available to them according to the survey conducted before the start of the experiment. The limited English resources available to these participants allowed us to more confidently confirm the effect of the language environment, operationalized as amount of natural English input, on the acquisition of contextual vocabulary knowledge.

The language system is influenced not only by the language environment but also by several other interacting and embedded subsystems over time. With the continuous change of several interacting subsystems over time, it is very hard to predetermine the final result at the individual level. Even if we know and control the total amount of input, and all the types of input, we are still not able to precisely predict the final outcome of the language system. In the first study, we found that the sequence of presenting input plays an even more crucial role in learning contextual vocabulary knowledge than the pure incidental input itself. The relationship between the outcome (output) and the language exposure (input) is not proportional. Human beings do not develop their language as computers perform programming. There is no predictable linear relation between input and output. The relation between input and output is of a non-linear undetermined nature.

Since the relation between input and output is not proportional, it has been found in many of the longitudinal case studies that language development is an individually owned process. For instance, Larsen-Freeman (2006) showed that four
Discussions

Writings samples over a six-month time period of five Chinese learners of English demonstrate quite different developmental paths among five learners. In the second study of this dissertation, four beginning learners of English showed their individual language developmental patterns in terms of difficulty of word use, even though three learners from the same group in the first study performed more similarly than one learner not from the group of the first study. The inter-individual variability in the second study could be argued to result from initial condition of English proficiency, preferred input types (reading or movie), and preferred way of using newly-acquired words in writing (learning style) based on these learners’ writing productions. In the last two studies, we zoomed in on two of the participants, two female identical twins participating in the second study, and still found the inter-individual variability between them in their language development. This gives strong ground to the claim that every individual, even with the same level of English proficiency, the same family background, the same schooling education, and similar English exposure, no individual develops exactly the same language learning pattern as others.

As each individual has his or her own pattern of language development, the first study, even if it is a group study, has tried to eliminate such inter-individual variability by its experimental design. The comparison between the control group and the experimental group has been made within the same group to reduce inter-individual variability: there are two groups, each of which acts as its own control and experimental group. When a group acts as its own control group, there is no extra input outside the classroom. When a group acts as its own experimental group, there is extra input outside the classroom. Comparing the two conditions of the language environment within the same group enables us to eliminate the inherent inter-individual variability in the group study even though this design was not a complete success as it turned out that the groups were rather different.

Apart from the inter-individual variability, intra-individual variability is found to provide important information about language development: when there is more intra-individual variability, there is more likely to be a change in language development (Verspoor et al., 2008; Van Dijk et al., 2011). In the second study, there was a lot of intra-individual variability over the 56 measurements, especially for the participant from outside the group study, and the intra-individual variability of this participant mainly resulted from many of the trials of new words from the English exposure based on the observation of his writing productions. This reflected what was found by Spoelman and Verspoor (2010) that the increased intra-individual variability coincides with a developmental jump. When a learner tries out new words, he or she is making progress in his or her language system. Such progress can appear together with the high intra-individual variability. Therefore, intra-individual variability is relatively high in a period of language development when the language system is reorganizing whilst it is relatively low when the language system is in a stable state. This intra-individual variability is an inherent property of a self-organizing system, and the degree and of variability can inform us of the developmental process. In the longitudinal studies
of this dissertation, intra-individual variability was found in every single case for productive vocabulary size, in the lexical density, and in the syntactic complexity measures. However, the aim of these longitudinal studies was not to find how such intra-individual variability may inform us of the developmental process, but to observe either the general trend of the language development or the general learning stages. The intra-individual variability was carefully taken care of by the smoothing technique (Spline) without losing its important information as such intra-individual variability may hinder the observations of the general trend and the learning stages.

In summary, in terms of how the language environment influences its embedded subsystem, the first study has provided valuable information about the influence of the conditions of language environment on its embedded subsystem by involving three timings of measurement; it has also attempted to eliminate the inter-individual variability by the experimental design. However, this group study, which only looked at the overall developmental trends of a group of “homogeneous” learners, may not be representative for any individual learner as language development is an individually owned process and thus the inter-individual variability would inevitably exist within the group. Therefore, the second study was carried out with more dense data so as to more fully represent the individually owned process of the language development, which can be better approximated by extended time series than inferred by three points in time. The second study randomly selected three learners from the group of the participants in the first study and one learner from outside of the group and collected dense writing data from these four learners to extend our understanding of such individually owned language development.

7.5 Interaction among embedded subsystems

In the first two studies, the foci mainly lie in the interaction between the language environment and its embedded subsystem. A step further to investigate the dynamic language development is to explore the interactions among the embedded subsystems. The language system is integrated in a number of subsystems in complete interconnectedness, and language development is a complex dynamic process which is influenced by constantly changing interactions between subsystems at different timings and at different levels. This is in line with what has been emphasized by Thelen and Smith (1994, 2003) that ontogenetic processes emerge from the constantly changing complex interaction of subsystems rather than a predetermined route towards maturity.

Many of the previous studies investigating the interaction between the embedded linguistic subsystems have mainly focused on lexicon and syntax and found different relations, competitive or supportive, in different learning stages. For instance, Spoelman and Verspoor (2010) observed 54 writing samples of one Dutch
Discussion

learner of Finnish over a three-year time period and found that different linguistic complexity subsystems, lexical and syntactic complexity, interact differently over time in different stages of learning. Another cross-sectional study carried out by Verspoor et al. (2012) investigated 437 writing samples of five distinct levels of English proficiency and showed that the changes of the lexical and the syntactic subsystems may take place in different transitional learning stages.

However, the interaction between the two embedded linguistic subsystems, writing and speaking, has not yet been investigated even though these subsystems play crucial roles in language development. Firstly, the acquisition of the two production modes differs in L1 and L2. L1 learners firstly learn to speak long before they start writing whilst L2 learners develop their proficiency in the two modalities more or less at the same time. L1 learners, by the time they have become mature language users, show approximately the same levels of development for their writing and speaking productions while some L2 learners are able to write well but are not able to speak at all, or the other way round. For L2 learners, either their writing capacity precedes their speaking or the other way round. A better understanding of how L2 learners develop their writing and speaking skills would enable us to understand such discrepancies in the L2.

Secondly, based on the theoretical models of writing and speaking (Hayes & Flowers, 1981; Levelt, 1989), writing and speaking seem to be two closely related subsystems. Writing and speaking both function as a communicative tool in a less formal/organized form (natural meaning expression), and development of both subsystems goes through three distinct processes, planning (conceptualizer), translating (formulator), and revising (feedback loop). The relation between writing and speaking could be mutually supportive, since the development of written production can promote oral production, and vice versa. The major factor contributing to the difference between writing and speaking in L2 is the time dimension: speaking may not allow learners to have enough time or cognitive resources to search for appropriate words to express a certain meaning. The relation between writing and speaking could be competitive as these two production modes have to compete for the limited cognitive resources to develop.

Since the purpose of the final two studies is to understand how the two embedded subsystems of language production are shaped over time and since each changing subsystem is embedded in the whole language environment, writing and speaking tasks in the final two studies feature on spontaneous meaning expression: L2 learners produced writings on Facebook with given topics and recorded their speaking on their own. These two studies attempted to capture learners’ writing and speaking performances as naturally as possible. Therefore, in contrast to experiments taking place in the lab, our data collection took place in settings where learners feel free and comfortable to carry out their production tasks without time constraints. The advantages of language production in a natural setting compared to a lab setting are that learners are more likely to reveal their true feelings for the questions and that they are more likely to come up with new ideas during production as there is little time pressure. For instance, they do not have to worry
about not having enough time to express their ideas or not being able to fill up the
time with limited ideas of the topic, which is crucial in the comparison between
writing and speaking.

The data collection in a natural setting intends to induce the learners to use lan-
guage as a communication tool for themselves whilst the productions in the lab
setting would make learners tend to see language as an assessment tool. How-
ever, there are also disadvantages of language production in natural setting, where
learners mostly do not show full focus on their complexity, accuracy, and fluency
of their language production. More dense data collection is therefore needed to
detect the linguistic changes in their language production. Secondly, coding the
data consistently and accurately is more challenging in the natural setting due to
lower accuracy compared to the lab setting. When learners’ errors must be cor-
rected to process the estimation of the complexity measures, this means that the
more errors a language production contains, the more difficult it is for the coder
to correct the data consistently.

From a DST perspective, all the subsystems interact with each other and should
not be considered closed modules (e.g. lexicon and syntax) or modalities (e.g.
writing and speaking). Despite this fact, the last two studies investigated the
modules and modalities separately because not much was known about the pro-
cess of modalities (writing and speaking), and the observation of several modules
(lexical and syntactic complexity measures) would aggravate the difficulty of ob-
serving the differences for the modalities. The last two studies report on dense
data of 100 writing and speaking productions of two identical twins over eight
months and show which modality precedes which in several linguistic measures
in L2 and the supportive/competitive relation between writing and speaking over
time. Firstly, writing precedes speaking in terms of lexical diversity over time
whilst speaking precedes writing in terms of syntactic complexity over time. Sec-
ondly, competitive and supportive correlations between writing and speaking were
found over time, but such changing correlation did not have one clear direction,
where the identical twins show completely different directions of the correlations
over time. If the identical twins show such different directions of correlations be-
tween the subsystems, the difference between other individuals will undoubtedly
be more evident. This is why the inter-individual variability exists when observing
the developmental process, and this is why we need to study language from a DST
perspective: to understand the process instead of only the product.

From a DST perspective, several seemingly unrelated aspects of subsystems can
be expected to be related (Caspi, 2010). The two subsystem under investigation
in this dissertation, writing and speaking, were also found to have supportive
and competitive relations over time. However, the relations between writing and
speaking were compared separately in lexicon and syntax, which should not be
studied separately when using the observed data to infer the most likely learning
state because lexicon and syntax has been proved to be two interrelated subsystems
(Caspi, 2010; Verspoor, 2008, 2010). But so far, we have little knowledge about
the relation between writing and speaking over time. It is easier to study each
perspective separately in the preliminary studies and then further investigate it in a more thorough manner. Future studies should consider to combine the data from lexical and syntactic complexity measures to investigate the reality of interlanguage subsystems.

7.6 Modeling language development

The application of DST to SLA was initially only limited to metaphorical use of the terms and concepts, but more recent studies have started to use the DST analytical tools. In Verspoor et al. (2008) and Verspoor et al. (2010), the concept of moving correlations has been used to explore the changes of correlations over time, from supportive to competitive or vice versa, between linguistic subsystems over the time. Caspi (2010) and Caspi and Lowie (2013) further developed the analytical tools by applying the logistic growth function, based on van Geert’s work (1994), to explore the interaction between different types of L2 vocabulary knowledge. Sometimes, these vocabulary knowledge types show competitive relations due to the limited resources available; sometimes, they show supportive relations due to the interrelated connection of the vocabulary knowledge. The analytical tool has developed from merely visual observation of the correlations over time (moving correlations) to being able to quantify the supportive or competitive correlations to different degrees through modeling.

However, as suggested in the second study (Chapter 4), there are some limitations of the use of the logistic model for the explanation of the trajectories of the lexical development. The application of the logistic equation was used in Caspi’s study (2010) where she successfully quantifies the developmental patterns of four different types of vocabulary knowledge of four advanced learners of English. But the logistic equation in the second study is not able to model the trajectories of all four learners correctly. One participant does not develop his vocabulary size according to a logistic pattern. This implies that the parameters used in the logistic model cannot interpret this particular learner’s developmental trajectory. We could therefore argue that the pattern of vocabulary development is more complex than the logistic pattern can deal with. It seems that the logistic equation needs further modifications to enable the observations of a wider range of trajectories. From the present logistic model, we have three parameters to determine the developmental pattern: initial value, learning rate, and carrying capacity. Initial value refers to the beginning level of learners’ vocabulary; learning rate refers to the slope of the trajectory; carrying capacity refers to the final level of learners’ vocabulary. Future studies could either modify the logistic model by adding other potential parameters influencing the developmental pattern (e.g. the difficulty of the English input) or apply other mathematical models which could dynamically interpret the trajectories with wider range of participants.

Apart from the fact that the inter-individual variability makes modeling as the
general developmental trajectory more difficult, the intra-individual variability should also be dealt with. The large intra-individual variability of the data would hinder the observation of the general trend of the trajectory, but in order to select the most ideal mathematical model to interpret the trajectory, the general trend of the trajectory of the data should be visually recognized. Therefore, smoothing the data by removing some of the intra-individual variability without losing the major information is necessary (e.g. Caspi and Lowie, 2013). In this dissertation, a cubic smoothing spline is applied to all of our data. This smoothing technique seems ideal for obtaining the general trends of three participants from the same group but not the one from outside of the group. It is suggested that a more appropriate smoothing tool may have to be developed for the application in future studies.

Therefore, in order to construct the mathematical model to interpret the trajectory of the empirical data in language development, there are two important implications worth noticing. Firstly, there are other variables influencing vocabulary development, resulting in substantial inter-individual variability. It is difficult to include all the variables in one equation and to meaningfully interpret the value of each variable in linguistic terms. Secondly, due to high intra-individual variability, data smoothing is one important process before fitting the trajectory. There are many available smoothing tools, and the selection of the smoothing tool should depend on the characteristics of the raw data. Thirdly, to effectively fit the model with a number of variables, it is important to apply an effective and automatic tool. It is suggested in the second study that Amoeba could be one ideal program as it is commonly used to fit the complex multi-dimensional characteristics in physics. Technically speaking, future studies, based on these implications, could evaluate the smoothing technique and the automatic model fitting tool; theoretically speaking, a modified mathematical model, including more variables in the equation to describe the trajectory of language development can be developed.

It seems that the major challenge encountered in modeling the language developmental pattern lies in the inter-individual variability. One solution, as previously mentioned, is to modify the logistic mathematical model by including more possible parameters. The other solution is to search for an alternative mathematical model to investigate the longitudinal data. This alternative approach must meet two criteria: (1) Regardless of the trajectory, it can reveal the changes of the language development over time. (2) It must be compatible with the characteristics of DST.

The last two studies have applied a different analytical model for second language development, hidden Markov model (HMM), as an alternative mathematical model to interpret the individually owned learning process. The HMM is one of the most popular models used for analyzing multistage and time-varying sequential data. The word “hidden” implies that the state is not directly observable (e.g. level of English proficiency), but the data (e.g. lexical measures of writings), depending on the state, is observable. HMM uses the observable data to infer the hidden state through dynamic programming. Therefore, the advantage of HMM
is that we can use the hidden state inferred by the data, instead of looking at the trajectory of the data, to know the change of the development over time. Another advantage of applying HMM is that the state can be inferred from more than two variables at a time. As the purpose in our studies is to see the changes of language development over different states, multi-dimensional linguistic measures can give us more thorough information on the hidden states. For instance, the state of language proficiency defined by four variables and their interactive relations, both the lexical complexity and diversity in both writing and speaking is more complete than that defined by two variables and their interactive relation, like only the lexical complexity in writing. HMM seems to work more appropriately than the logistic model in the last two studies in the sense that it allows us to observe the changing state over time regardless of the developmental pattern and that it allows us to observe the interaction between more than two variables at a time.

7.7 Vocabulary learning

Schmitt and Meara (1997) found that learners were only able to correctly associate fifty percent of the known words with other words. This implies that knowing the word meaning does not necessarily guarantee the acquisition of the association and that knowledge of word meaning forms a foundation for learners to build up the contextual vocabulary knowledge. Therefore, contextual vocabulary knowledge is usually acquired later than non-contextual vocabulary knowledge. In the first study, the sequence of presenting non-contextual and contextual vocabulary knowledge is even more influential than the language environment itself. If a beginning learner intends to learn contextual vocabulary knowledge effectively, it is better to start from non-contextual vocabulary knowledge (e.g., word meaning) then to contextual vocabulary knowledge (e.g., collocations or associations).

Apart from the distinction between non-contextual and contextual vocabulary knowledge, it is also necessary to distinguish the two types of contextual vocabulary knowledge, collocations and associations as proposed by Nation (2001) that they represent different dimensions of vocabulary knowledge. Collocations reflect how learners can recognize or use the collocated words (or Conventionalized Ways Of Saying Things proposed by Smiskova, Verspoor, and Lowie in 2012); associations reflect how learners can associate the one particular word with other words (see Fitzpatrick, 2011 for summarized categories of associations.) In the first study, the acquisition of collocations and associations is found to be equally difficult, but the frequency (number of occurrences) and distribution (how the occurrence distributes in the input) of collocations and associations differ. The occurrences of associations in the extra English input are much higher than those of collocations, but the collocated words appear at a more visually noticeable distance than associations. For instance, one may find “big” and “small” (associated words) appearing within one movie/reading input many times, but the distribution of
“big” and “small” in the context is not within a visually noticeable distance. Consequently, it is difficult for learners to associate these two words. On the other hand, one may find “surprised” and “at” (collocated words) appearing within one movie/reading much fewer times than associations, but these two words are easy to connect visually. These two factors may explain why the acquisition of collocations and associations is equally difficult: in terms of frequency, associations seem to take advantage; in terms of distribution, collocations seem to take advantage.

Therefore, the acquisition of these two different types of contextual vocabulary knowledge may also depend on the type of input accommodating them as vocabulary knowledge is an embedded subsystem of language learning, which is an embedded subsystem of the language environment. As learners can sort out different types of vocabulary knowledge (e.g. collocations and associations) in different types of input (e.g. visual or audio or both), future studies could empirically investigate how the acquisition of collocations and associations could be enhanced in different types of input (e.g. audio input such as movies or written input such as readings).

To evaluate language production, there are at least two dimensions of lexical complexity, lexical diversity and lexical difficulty, to be included. Lexical diversity accounts for the number of different words used; lexical difficulty accounts for how these different words are distributed across different frequency bands. In the third study, writing precedes speaking in lexical diversity but not in lexical difficulty. This again proves the need to distinguish lexical diversity and lexical difficulty when it comes to measuring the lexical complexity in language productions, and it illustrates the complexity of the relationships discussed in the previous section.

7.8 Future studies

The four studies in this dissertation have demonstrated the influence of language environment on its embedded subsystem, the relation between the embedded subsystems, and have shown how modeling techniques can be used to interpret these dynamic changes. These studies also emphasize the importance of understanding the process of language development and the meaning behind the inter-individual variability and the intra-individual variability.

Future studies on comparing writing and speaking can also explore the relations between writing and speaking in more detail with the same set of dense data. The questions can be: How is oral language used to support written language? How is written language used to support oral language? How do oral language and written language compete with each other? One can also explore the real-time process of writing and speaking. When recording the process of writing productions, through the analysis of keystrokes, how learners produce their writing, revisions, corrections, selection of words, and pauses (hesitation) can be shown. These can
never be shown in the final product of writing, especially. With this recording approach, the processes of writing and speaking could be observed with many other interesting linguistic dimensions beyond lexicon and syntax (e.g. pause, revisions, corrections). As time is now seen as the major difference between writing and speaking tasks in these two studies, the timing of many other dimensions regarding the processes can be very relevant for future research. This means that the effect of timing in two tasks can be empirically tested. Finally, one can also investigate the effect of different types of input (written or oral input) on writing and speaking tasks. What is the comparative value of written (e.g. reading) and oral (e.g. movie) input on the development of writing and speaking in different linguistic measures? Do different types of learning also affect the effect of written and oral input on writing and speaking?

In a longitudinal study with such dense data, one will need mathematical tools to deal with such large amounts of information in the data sets. Only a limited number of mathematical tools is applied in interpreting the longitudinal data in the field of applied linguistics, but these tools are commonly applied in many other scientific fields. For instance, the model fitting tool used in the Chapter 4 is commonly used in physics; the hidden Markov model is commonly used in computer science. These mathematical tools can provide linguists solid ground of understanding how language is formed. More possibilities can be explored to advance our understanding of language development.

In case studies, the inter-individual variability cannot be neglected. Each individual develops his own language learning pattern; so do the identical twins who show contrasting developmental patterns of the correlations between the lexical measures and syntactic measures. Although this feature makes mathematical modeling more difficult, these case studies result in rich accounts of the language process of one single participant over a period of eight months. The in-depth examination of the process could potentially improve the process of language acquisition. For instance, knowing a learner’s preferred way of learning can help us to assist the learner to improve his/her language through providing him/her preferred way of learning.

This thesis looks at vocabulary acquisition from a Dynamic Systems Theory approach. In a sense the analysis of words as units as done in this dissertation is at odds with that kind of thinking: As argued by De Bot & Lowie (2013), there are serious doubts about the status of words in the mind. The old idea of the lexicon as a library in which words are ordered in a specific way, thematic, or alphabetic and can be used as stable, or fixed representations is seen as outdated: words change through use and they typically do that as part of larger units. It could be argued that words never occur in isolation apart from the strange kind of experiments with word recognition that psycholinguists do. Even uttering or reading a single word is influenced by all sorts of additional information: the topic of conversation, the conversational partners, assumed shared knowledge, noise in the room, previous use and so on. In the literature on vocabulary learning the focus has shifted from learning words in isolation to words in context. See for
instance the edited volume on fixed expressions by Schmitt (2004) or more recently Smiskova-Gustafsson (2013) on larger units, and what she calls CWOSTs (Conventional Ways Of Saying Things).

In future research on vocabulary acquisition, the context of use, the larger units of speech and the conversational setting and how these develop in interaction over time should be looked at, rather than the words in isolation. The natural data gathered in the present study already are an improvement over words tested in isolation, but for future research on embeddedness of lexical items should be up front rather than back stage.