Measuring lexical and syntactic complexity of the language production of the World-of-Warcraft gaming community

Simon Šuster

LCT master's student, U. of Groningen
Seminar in statistics and methodology

June 29, 2011

Contents

1 Introduction

This research paper presents topics from the automated measurement of the linguistic complexity on large texts samples. The motivation for this work comes from the field of computer-assisted language learning, where there is a relatively substantial body of work relating to the ways and possible usefulness of including video games in the language learning process. Thus, types and characteristics of in-game communication, the social interaction, and dynamics and communication patterns between gamers are well described. However, very little research has been done on the properties of the text production itself in games and of gaming community. In my opinion, the potential of games for language learning has been largely overlooked or underestimated by educators, and by researching this language more intensively – both its social aspects and on large collections of linguistic data – we can get to know more precisely in what ways the gaming language can be useful in language learning.

This paper focuses on a massively multi-player online game (MMOG), namely World of Warcraft (WoW).[1] I will put less emphasis on the use of games in language learning, and will instead focus more on describing automated ways of analyzing the WoW language production and potential problems associated with these. Naturally, various measures or categories of measures can be used to determine language complexity. Here, I will present measures from the following categories: lexical richness, lexical density, readability and syntactic complexity. The calculations are performed on text collections gathered from the Internet sites about the WoW, mostly discussion fora and wikis. Also, a general web corpus is included in the research and used as a comparison.

The results obtained from such an analysis will be useful, hopefully, in the context of language learning, mainly for the educators getting a clearer picture of the WoW-community text production, and, possibly, encouraging them to include the gaming activities in their class sessions.

[1] It is one of the most popular online games ever with presently over 11 million subscribers (see `http://www.curse.com/articles/world-of-warcraft-news/956087.aspx`), and it was chosen both because it features prominently in the research literature on the theme and because its meta- and in-game texts are diverse and numerous (esp. online texts about the game).

2 Background: games as a language-learning tool

MMOGs are game worlds characterized by the fact that they are played online, which means that gamers can interact not only with the game software but also with other players. Collaboration takes place and can be seen not only outside the game in various discussion fora and fan sites, but also, crucially, in the game itself. Gaming in the MMOGs is cognitively demanding; it involves exploration of complex problem spaces, constructing and analyzing models, and negotiating meaning within the gaming community; it involves coordinating people, virtual tools and artifacts, and it is also characterized by various forms of texts (Steinkuehler, 2008).

MMOGs such as WoW are rarely played by a gamer without any contact with other players throughout the game. Rather, connecting with co-gamers facilitates certain tasks and improves the outcome. WoW, specifically, allows text and voice communication in numerous chat channels with different privacy policies.

(Thorne, 2008; Thorne, Black, & Sykes, 2009) looked at the intercultural communication in massively multi-player online games, especially WoW. They notice that WoW is played in language-specific domains that concentrate together speakers of the same language. The advantage of this is that after installing a language pack, the game setting and most probably gamers as well will be language specific. However, this may have the disadvantage of hindering multilingual communication in the game. (Thorne, 2008) provides a case study in which two gamers with different levels of gaming proficiency exchange – in Ukrainian and English – expert knowledge, language-specific explicit corrections and requests for help. Involving in games often results in increased motivation for L2 learning, as also noted by Thorne.

Similarly, (Rankin, Gold, & Gooch, 2006) remark that MMOGs, especially their role-playing subtype, are particularly important for language learning because they represent immersive environments allowing fully-fledged experience of a virtual world, they abound with possibilities for social interaction, they are motivating and they create a virtual world as the context where language students concentrate on accurate and coherent language use, develop and test their practices.

Two examples of successful inclusion of MMOGs and their content into school context are well-known from the literature:

- A study of (Rankin, McKenzie, Shute, & Gooch, 2008) empirically evaluated Sony's Ever Quest II MMOG and found out that the game is especially appropriate for increasing vocabulary size: those students

who engaged in gaming activity four hours per week outperformed the students who only attended three-hour class instruction.

- (Bryant, 2006) describes with enthusiasm the language sessions with his students of German in WoW, and provides instructions and advice in his article to others willing to complement their regular language classes with in-game practicals.

(Gee, 2007) claims that children with early engagement in non-everyday technical language perform better later on in school and in situations demanding the use of complex specialist language. When using specialist language, children develop their "islands of expertise". Gaming, then, can be viewed as an activity of building, and adding to, a particular island of expertise. In addition, the kind of learning taking place in the games can facilitate situated verbal understanding – an understanding and meanings built bottom-up, i.e. starting with a relatively concrete case and rising to higher levels of abstraction. Gee discussed language use and learning in various game genres, each type presenting slightly different learning challenges to the gamers. However, in Gee's list of game types, none have been designed with language learning being game's primary goal. This is different to the new type of games, emerging now, that I would like to mention here. In these games, the communication between the gamer and the in-game characters is on the level of communication between-gamers. Such example is Bot Colony (`www.botcolony.com`), developed by the Canadian company North Side and advertised as a "real conversation game". It is an open-world adventure game encouraging the gamer to engage in conversation with game characters who are capable of human-like speech – and written production – by internally using advanced natural-language processing techniques. The progress through the levels correlates with the amount and quality of the gamer's language input, and the game promotes the language learning by fulfilling gamer's orders and wishes, but also by providing corrective feedback to the user ("Did you mean X?"). These games may prove particularly efficient for language learning, but their availability as of today is limited (Wilcox, 2011).

In this section, I have reviewed the literature that deals mostly with the social aspects of language learning in games. In the next section, I move on to discuss the array of tools and procedures used in this paper to provide a description of the text production in the gaming community.

3 Measuring language complexity

In the next sections, I first treat measures that put more emphasis on the lexical aspects of language production, and then I present a couple of measures that take into account also the more syntactic side.

3.1 Lexical measures

Lexical diversity Lexical diversity measures the size of vocabulary in a text or of an author. It is often used for describing and predicting vocabulary growth in a text (Baayen, 2008). Intuitively, lexical diversity should be quantifiable in terms of different word types. Since this measure is known to depend heavily on text size, number of tokens are also taken into account. One such measure that uses those two types of information from the text is type-token ratio (TTR). However, TTR is also dependent on the text size, it is bigger when texts are small and decreases as the texts get larger. Due to this fact, the comparison needs to be made either on same-sized texts or using statistical approximations (corrected ratios) that take this into account and try to remain constant by not depending on text length, or using probabilistic models. There exist several corrected measures, like Herdan's C, Guiraud's R, Yule's K and others (see (Baayen, 2008)). But since the within-text variability can be large – put differently, words are not used randomly in reality –, these measures may not be precise at differing sizes of text samples, as shown by (Tweedie & Baayen, 1998). More or less the same applies to probabilistic models, such as Sichel's generalized inverse Gauss-Poisson model and Zipf-Mandelbrot models (Evert & Baroni, 2008; Tweedie & Baayen, 1998). Especially the latter performed reasonably well and proved length-invariant in the Tweedie and Baayen's research. These are also called models for Large Number of Rare Events (LNRE) distributions, word frequency distribution being one such example.

To return to the TTR, mean segmental type-token ratio (MSTTR) is its simple transformation. It is calculated by first computing TTR on same-sized text samples, e.g. of 50 tokens, and then taking the mean (Lu, Thorne, & Gamson, 2011). As the mean here drowns out the individual differences in samples, all sample TTRs can be plotted to illustrate the ratio values changing as a function of position in the text.

Lexical density Lexical density has been shown to correlate strongly with the lexical diversity (Johansson, 2008). It measures the proportion of content, or lexical, words (verbs, nouns, etc.) to the total number of tokens,

and it is an indicator of information packaging in the text since the high proportions of content words mean bigger information load. It is not trivial to operationalize the definition of lexical density. First, one should decide what counts as a lexical word. Usually, lexical words are said to occur in closed word classes, like nouns, but there is some controversy whether it is reasonable to count (certain) adverbs, e.g., as lexical or non-lexical words. Secondly, an example like "turn up" may consist of a lexical word and a non-lexical, i.e. function, word. But a more advanced decision system could count the two elements as a single lexical item, see (Johansson, 2008) for further discussion.

Readability Readability formulas try to estimate the degree of text difficulty. There are over 200 of such formulas (Dubay, 2004). Many of those include a combination of sentence complexity – usually mean length of the sentence – and word complexity measures – usually average number of syllables or characters per word, or the proportion of complex words. (Lu et al., 2011) use different measures to predict appropriate grade levels of school reading materials. In their case, Flesch Reading Ease, Coleman-Liau and New Dale-Chall Readability formulas perform very well. Flesch Reading Ease, the formula used in this research, uses as one information the number of syllables and as the other the number of sentences. The reading ease is predicted on a scale from 0 to 100, with 30 being very difficult and 70-and-up being easy.

Style	Flesch Reading Ease Score	Average Sentence Length in Words	Average No. of Syll. Per 100 Words	Type of Magazine	Estimated School Grade Completed	Estimated Percent of U.S. Adults
Very Easy	90 to 100	8 or less	123 or less	Comics	4th grade	93
Easy	80 to 90	11	131	Pulp fiction	5th grade	91
Fairly Easy	70 to 80	14	139	Slick fiction	6th grade	88
Standard	60 to 70	17	147	Digests	7th or 8th grades	83
Fairly Difficult	50 to 60	21	155	Quality	Some high school	54
Difficult	30 to 50	25	167	Academic	High school or some college	33
Very Difficult	0 to 30	29 or more	192 or more	Scientific	College	4.5

Figure 1: (Flesch, 1974)'s analysis of the readability of adult reading materials

The index is calculated with the following formula:

$$FleschReadingEase = 206.835 - 1.015 \cdot \frac{N_{Words}}{N_{Sentences}} - 84.6 \cdot \frac{N_{Syllables}}{N_{Words}}$$

A modified formula to produce a US grade-level score is called Flesch-Kincaid:

$$Flesch - Kincaid = 0.39 \cdot \frac{N_{Words}}{N_{Sentences}} + 11.8 \cdot \frac{N_{Syllables}}{N_{Words}} - 15.59$$

3.2 Syntactic-complexity measures

The simplest measure of syntactic complexity is the mean length of sentences. Although it is insensitive to structural differences within sentences, it turned out as a reliable indicator of grade level in (Lu et al., 2011).

There exist more complex measures that define the developmental level of the sentences (D-level, DSS scoring, IPSyn) (Lu et al., 2011), and more descriptive techniques, as comparing the frequencies of POS tags for the most common n-grams between two texts (Wiersma, Nerbonne, & Lauttamus, 2011), capable of providing qualitative answers to the differences of the frequency of syntactic constructions in texts. D-level, used in this paper, is a measure based on child language acquisition premise, namely that the most complex sentence types are acquired last (Covington, He, Brown, Naçi, & Brown, 2006). According to (Cheung & Kemper, 1992), D-level has been shown to be a more adequate index of sentence complexity than other metrics, because it can differentiate between different types of clausal embedding.

4 Methodology

4.1 Research goal

1. Provide a frequency-based description of a WoW-community language sample, specifically texts from fora, wikis and news, in terms of language complexity and readability.

2. Positioning this language production on the complexity and readability scale – also in comparison with a general web corpus – would help educators determine the level of appropriateness for inclusion of game-related activities in school.

Level	Description	Example
0	Simple sentences, including questions	*John cried.*
	Sentences with auxiliaries and semi-auxiliaries	*This has solved it.*
	Simple elliptical (incomplete) sentences	*John did.*
1	Infinitive or *-ing* complement with same subject as main clause	*Try to smile.*
2	Conjoined noun phrases in subject position	*John and Mary left.*
	Sentences conjoined with a coordinating conjunction	*I came early but Peter arrived late.*
	Conjoined verbal, adjectival or adverbial constructions	*He sang and jumped.*
3	Relative or appositional clause modifying object of main verb	*John scolded the boy who stole his bicycle.*
	Nominalization in object position	*I understand his rejection of the offer.*
	Finite clause as object of main verb	*John knew that Mary was angry.*
	Subject extraposition	*It was hard for John to tell Mary.*
	Raising	*John seems to Mary to be happy.*
4	Non-finite complement with its own understood subject	*I saw him walking away.*
	Comparative with object of comparison	*John is older than Mary.*
5	Sentences joined by a subordinating conjunction	*They won't play if it rains.*
	Nonfinite clauses in adjunct position	*Having tried both, I prefer the second one.*
6	Relative or appositional clause modifying subject of main verb	*The man who cleans the room left early.*
	Embedded clause serving as subject of main verb	*For John to have left Mary was surprising.*
	Nominalization serving as subject of main verb	*His rejection of the offer was unexpected.*
7	More than one structure from Levels 1-6	*John decided to leave Mary when he heard that she was seeing Mark.*

Figure 2: Developmental levels according to (Covington et al., 2006). Table taken from (Lu et al., 2011).

4.2 Gathering the data

The texts upon which the analyses were run came from two sources. The first source were texts from the WoW fora, wikis and to a lesser extent news. For the compilation of the WoW corpus, the BootCat toolkit for querying and extracting web-based texts was used (Baroni & Bernardini, 2004). The corpus contains 261,000 tokens from 196 URLs.

The second source, which is the corpus chosen for comparison, was the ukWaC (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009). This is a large corpus built from diverse web-based texts by querying the .uk Internet domain. Its design is conceptually close to other reference corpora, such as British National Corpus, but its advantage is that it is freely available and that the text material is relatively current[2]. In addition, it only includes web texts, just like WoW. The original size of ukWaC is 1.9 billion tokens,

[2]UkWaC was built in 2007, whereas BNC's texts are at least two-decades old.

but due to limited computational resources and also to the fact that approx. same-sized corpus was needed to facilitate comparison, a 261,000-token randomized selection was used.

Once the plain-text corpora were prepared, they were lemmatized and POS-tagged with TreeTagger (Schmid, 1994).

4.3 Processing the data

In this research, the following measures were calculated on both corpora (unless indicated otherwise):

- TTR
- MSTTR
- N, V and Adj density
- Flesch Reading Ease and Flesch-Kincaid indexes
- Mean sentence length
- D-level (only for the WoW corpus)

In addition to calculating these measures, I try to fit a LNRE model to the data, which could allow the comparison of different-sized texts (although not needed here), since one could use interpolated (for smaller samples) or extrapolated (for larger samples) values in comparison.

For fitting, and plotting of, the probability models, the zipfR package (Evert & Baroni, 2008) was used. The measures listed above, except D-level, were calculated in R using the koRpus package (Michalke, 2011). For the test of significance for the difference between the two corpora in vocabulary size and vocabulary growth, the Z-statistic from (Baayen, 2010)'s languageR package was used.[3] When comparing different distributions, as in the case of TTR on text samples and Flesch Reading Ease index on samples, the Kolmogorov-Smirnov test was used to determine if the separation between two distribution curves occurred by chance, cf. (Baayen, 2008).

D-level scores were calculated with the D-Level Analyzer (Lu, 2009). To use it, raw texts need to be POS-tagged with the Penn Treebank tagset first, and then parsed with the Collins' parser. In my case, between these two steps, some modifications of the tagging output needed to be made, such

[3]Where the variance is estimated internally by fitting an LNRE model to two texts with a χ^2 test.

as counting the words per line and mapping certain POS tags in order to obtain successful parses from the Collins' parser. Still, the performance of the parser was often erratic, and some issues remained unresolved, i.e. in certain cases long sentences of ∼200 words or more kept returning errors. Due to this, only 2209 sentences out of 14491 were analyzed here. Following the described preparatory phase, Lu's D-Level Analyzer was run on the sentences, producing as output a number for each sentence ranging from 0 to 7, corresponding to the developmental levels described in figure 2.

5 Results

I start the Results part with the presentation and discussion of results for the lexical measures, continue with the readability, and conclude with the results of syntactic analysis.

5.1 Lexical measures

Lexical diversity The table (1) shows main results for WoW and ukWaC corpora. The type-token ratio for the entire ∼230,000-token[4] WoW corpus is .09, while the TTR for the ukWaC corpus is 0.12. In other words, the probability of encountering a new type at the end of the WoW corpus is 9 per cent, and in ukWaC, this is 3 per cent more likely.

	WoW corpus	ukWaC
TTR	.09	.12
MSTTR	.73	.74
N ratio	.16	.16
V ratio	.14	.15
Adj ratio	.06	.08

Table 1: Lexical diversity and density results.

However, having just one ratio value for the whole corpus only gives us a rough picture, and certainly does not reveal anything about the structure and distribution of TTRs on smaller samples or on different texts. For this purpose, the mean-segmental TTR was also calculated on samples of 100 tokens. Thus, the MSTTR (mean of 2,300 samples) for the WoW corpus is

[4]During the calculation of these scores, the punctuation is removed, that is why the total token count is smaller.

.73, and .74 for ukWaC. In a similar fashion as above, this can be interpreted as follows: on average, after a 100-token sample, the probability of encountering a new type is 73 per cent, and the chance of doing so in ukWaC is 1 per cent higher. Since MSTTR is a mean summary of TTRs of samples, it is useful to show the dispersion of the values graphically by estimating the density curve, cf. (Baayen, 2008).

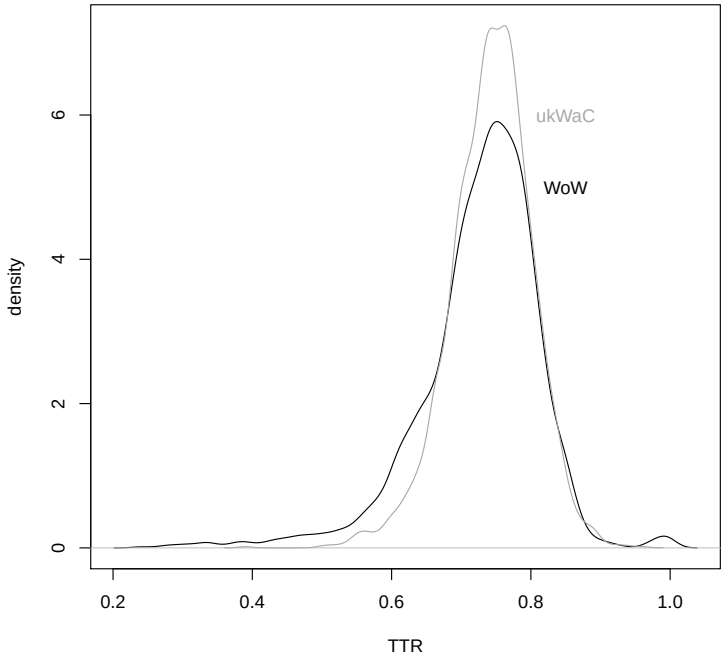


Figure 3: TTR for ukWaC and WoW 100-token samples

The graph (3) shows two curves – one for each corpus – which show the distribution of TTR values on small samples. It is easy to see that there is greater dispersion of TTRs in the case of WoW, whereas the ukWaC values are more uniform, i.e. the curve on the graph has a higher and narrower peak. The same difference is, of course, also witnessed in standard deviation values, $SD_{WoW} = .087$; $SD_{ukWaC} = .059$. This difference can be attributed to the fact that there is greater vocabulary variation in WoW corpus texts, i.e. more texts with lower vocabulary size, but it can also be attributed to

the peculiarities of sampling web texts for WoW corpus, e.g. there might be more repetition in some types of sites, such as Internet fora (referring to others' posts), which led to more samples having lower TTR.

In a parallel step, I checked if the displayed difference in distribution is statistically significant. This was done with the Kolmogorov-Smirnov test for two independent vectors, which was significant at $p < .001$ ($D = .1$).

Probabilistic model Sometimes, it may also be interesting or necessary to look at the vocabulary growth rate, because texts may have made use of the same number of types, but may differ with respect to the rate at which unseen types occur (Baayen, 2008). More broadly, introduction of new types, or counting the hapax legomena in a sample, is indicative of how productive the language or particular facets of language are.

In our case, following Baayen's work, the vocabulary growth rate is estimated by the ratio of the number of hapax legomena to the number of tokens. The results of the growth test for the two corpora are $growth.rate_{WoW} = .033, growth.rate_{ukWaC} = .059$. The difference between the results was determined with the Z-statistic, and was significant with p being effectively 0 ($Z = -50.2$). Since the size of the corpora compared is the same, the growth rate essentially tells us that there are much less (almost by half) hapaxes in the WoW corpus than in the ukWaC. This finding can be explained by the nature of corpora: ukWaC is a general, broad-coverage corpus, while WoW is a specialized, one-domain corpus, so a smaller number of new types is introduced. In graphs 5, I present the vocabulary growth, i.e. the number of new types encountered when moving linearly through the tokens of the corpus. The dotted line shows observed counts, and the smooth line the expected values, where its solid part indicates interpolated values and the dashed part extrapolated values. For both corpora, we can see that the expected curve is not very accurate. In these graphs, we fitted the inverse generalized Gauss-Poisson model to the data, which yielded a closer fit than the Zipf-Mandelbrot models, but the result is still not satisfactory: observed curves are mostly high above, and even move away from, the curve of expected values; also, interestingly, new types are sampled more slowly than expected if words were used randomly: the observed curve is under the expected one for some 40 thousand tokens in WoW and 100 thousand tokens in ukWaC, but following these tokens, the expected growth is clearly underestimated. These two observations may be due to the heavy use of topical words when the discourse is cohesive (words are not randomly chosen in cohesive discourse), but new texts are constantly being introduced when we

move from the beginning to the end of the corpus, which may explain the consistent introduction of more and more new types (see how the distance between the observed and expected curves increases).

Another noticeable fact is that the observed curve in the WoW corpus is less smooth than that of ukWaC. It is reasonable to assume that such sharp shifts in the curve cannot occur in "normally" structured cohesive texts, and that they could be attributed to inaccurate sampling procedure that might require some further modifications and filtering steps. From what we have seen, it would therefore be inappropriate in our case to try to predict the number of hapax (new types), dis-, etc. legomena for larger or smaller samples based on extrapolated or interpolated values obtained from fitting probabilistic models.

Lexical density Lexical density was computed separately for nouns, verbs and adjectives (see table 1), but cumulatively for all texts in each corpus. The proportion of nouns (excluding proper nouns) is .16 in both corpora. There are fewer verbs, .14 and .15 out of all tokens for WoW and ukWaC, respectively. The difference is bigger for the proportion of adjectives, .06 and .08. These data suggest that there is the same amount of words carrying "nominal" information, whereas the WoW corpus contains smaller proportions of verbs and adjectives. This can be interpreted as the communication in WoW texts (from the specific site types investigated in this paper) being closer to the spoken language, and ukWaC texts closer to the written language, although the difference is minimal. The observation that spoken texts tend to have lower lexical density than written texts was made already in (Halliday, 1989).

5.2 Readability

The Flesch Reading Ease index was calculated on 1000-token samples of complete corpora. From the graph (4), it can be seen that results look considerably different. The values for WoW texts are more uniform, while the ukWaC values show greater dispersion, $SD_{WoW} = 7.6$; $SD_{ukWaC} = 13$.

This is not surprising, since ukWaC as a general web corpus contains more diverse, and different-genre, texts with also varying degrees of readability. Compared to ukWaC, WoW corpus is then a homogeneous mass with more/most texts having similar readability. The difference between the means of Flesch scores for two corpora is 13 points, $M_{WoW} = 70$; $M_{ukWaC} = 57$. This difference is significant at $p < .001$ ($D = .52$). Translated into the Flesch-Kincaid grading system, this finding means that

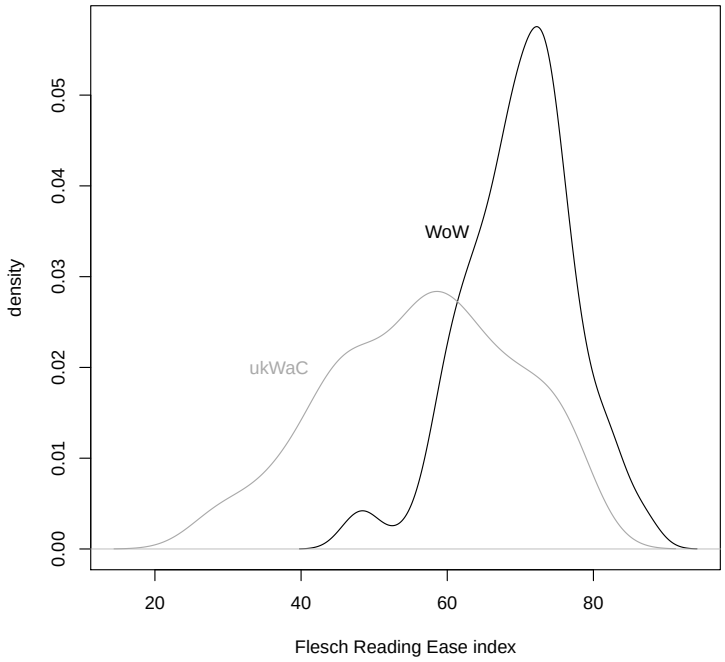
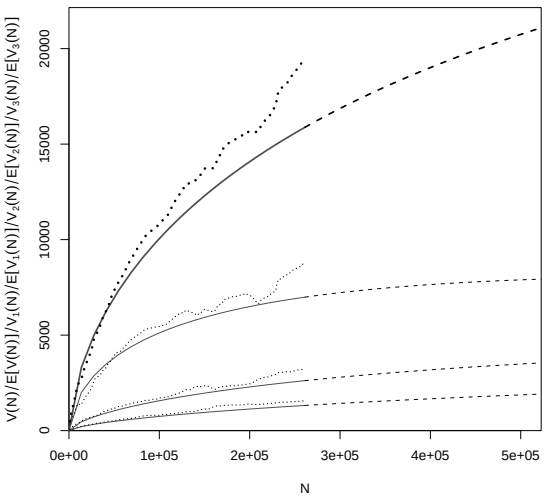


Figure 4: Flesch Reading Ease index for the WoW and ukWaC corpora

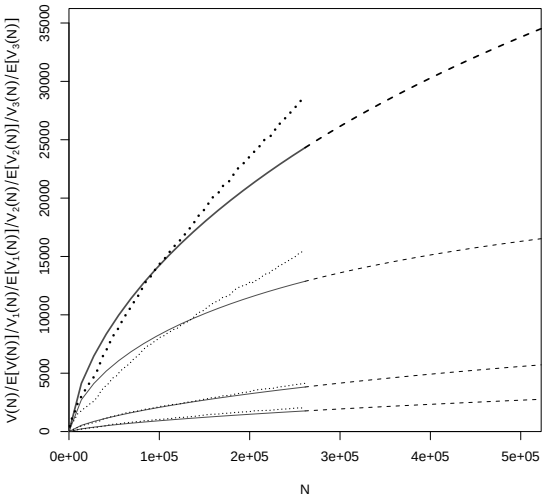
WoW texts are 2.2-grade levels below the ukWaC corpus, $M_{WoW} = 7.8$; $M_{ukWaC} = 10$. The Flesch's table (1) on readability of adult reading materials puts these results into a wider perspective. Thus, WoW text materials can be described as "standard-style", on average understood by ∼83% of US adults.[5] The ukWaC texts are characterized as "fairly difficult", presumably understood on average by 54% of US adults. The Flesch-Kincaid score for WoW, 7.8, thus sets WoW texts at the reading level of, and make them appropriate for, the age group 12–14 (or higher).[6]

[5]Of course, this is only a very rough estimate due to the datedness of the Flesch's research on readability.

[6]In other words, for students of US middle schools (or European secondary schools).



(a) WoW



(b) ukWaC

Figure 5: Vocabulary growth for the two corpora. The highest curve on the plot represents the total number of types (observed and expected), the lower curve represents the number of types occurring only once (hapax legomena), even lower curve the number of dis-legomena, and so on.

5.3 Syntactic complexity

Sentence length The Flesch index, as mentioned before, already incorporates one syntactic information, i.e. the mean sentence length (see formula 3.1). There are 19 words in a sentence on average in the WoW texts (see table 2), and two words more in an average ukWaC sentence. According to the table (1), this score for the WoW texts corresponds to style categories "standard" and "fairly difficult".

	WoW corpus	ukWaC
MSL (SD_{MSL})	19 (14)	21 (14)

Table 2: Mean sentence length for the WoW and ukWaC corpora.

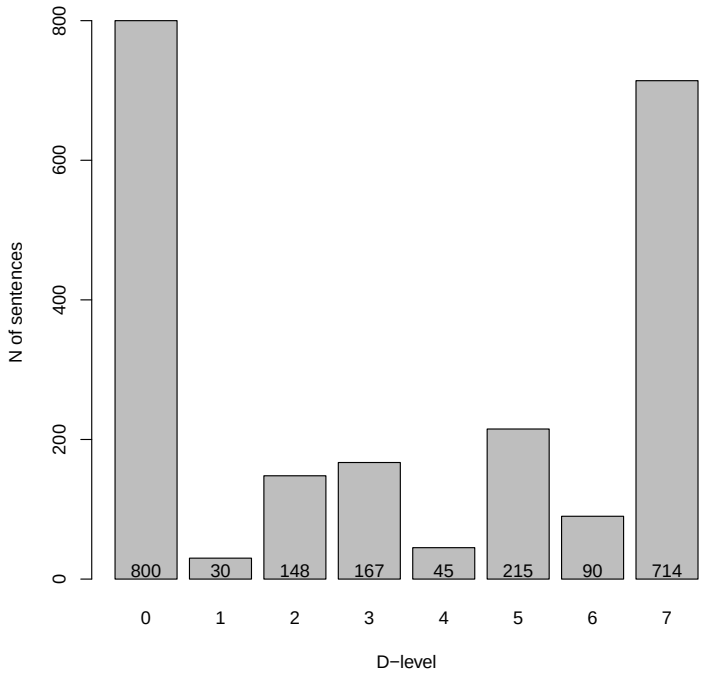


Figure 6: Distribution of D-level scores, expressed in number of sentences that fall into a particular level.

D-level The summary results for the D-level of 2209 WoW sentences are $M_{WoW} = 3.45; SD_{WoW} = 7.6$. Here, it is more interesting to look at the distribution of scores, which resembles a U-shape. From the graph 6, we see that most sentences can be described either as level 0 or 7. In other words, there are 800 simple sentences, but this figure needs to be taken with a certain amount of reserve due to the fact that upon manual inspection, several lists, incoherent productions, and errors from parsing and tagging were found. 714 sentences were analyzed as complex, i.e. containing more than one structure from the level 1-6. Other sentences at the bottom of this distribution were characterized as sentences with subordinating conjunctions and non-finite clauses in adjunct position (215); sentences with relative clauses modifying verb object, finite clauses as verb objects and a few other types (167), see figure 2; sentences with conjoined constructions and coordinations (148). Levels 1, 4 and 6 were rather infrequent which may be connected to the way the categories were formed – all, 1, 4 and 6, are rather narrow categories (subject nominalizations, appositions, non-finite complements with own subjects, ...).

6 Conclusion

In this paper, I have presented a brief language-complexity description of the WoW-community texts, and compared those results to the general web corpus ukWaC. Overall, results show that the language production in the WoW corpus is slightly less complex, but not for all measures. Both corpora display the same amount of information packaged in nouns, but the WoW language can be described as more spoken-like, because its lexical density is lower in general. The lexical diversity as measured by TTR on samples reveals that the vocabulary-size between texts in the WoW corpus differs to a larger extent than between texts in ukWaC, but on average, WoW texts reach a lower lexical diversity. The readability analysis has shown that WoW texts form a homogeneous group with less dispersion than ukWaC texts. On the basis of readability score and average sentence length, the WoW texts can be described as intermediate or "standard-style" texts, appropriate for use with the 12–14 year-old students or older. The analysis of developmental levels, which enabled syntactic analysis of types of sentence embeddings, showed that for the most part, sentence production is either very simple – doubts have been cast on the precision of results in this category, however – or very complex, consisting of sequences of structures that do not fall into the simplest category. To sum up, the analyses carried out in this

research showed very diverse results suggesting that the same is true for the WoW-community language. Even though the scores for the ukWaC – where computed – mostly indicate that a more complex language will be found in this corpus, this is understandable since a greater variety of texts is collected in ukWaC than in the WoW corpus, which is after all, a corpus of technical language – as claimed for the language in games in general by (Gee, 2007).

Directions for further work This study only looked at certain channels of communication of the WoW-community, i.e. fora, wikis and a few news sites. Strictly speaking, the data used in this research is, of course, not a representative sample of WoW-community language. Nothing has been said nor discovered about the several in-game ways of communicating and reading, which might differ with respect to the medium, place of occurrence of a text within the game software, or the level achieved by the gamer. There is also a vibrant fan-fiction community where – it is reasonable to suppose – a different picture might be discovered as we are dealing no longer with non-fiction but fiction material.

More time should be dedicated to preparing clean raw texts after collecting them from the web, although many of the problems related to data processing will likely remain. Communication on fora and discussion portals seems to be led by economical principles: misspellings, abbreviations, other spoken-like communication features and jargon words are all omnipresent, and pose a challenge for taggers and parsers. The in-game (chat) texts might be affected by these even to a larger extent. Much more could be found out about the language specifics by performing a more complete syntactic analysis of the language production, e.g. by examining most salient constructions (see 3.2). On the one hand, and especially if one starts an analysis of in-game communication, the tagging accuracy is likely to fall drastically, thus leading to many false salient constructions. On the other hand, this problem should not be very serious for the most frequently used word combinations, or the most commonly mistagged words could be added to the lexicon manually.

In this research, I chose a general web corpus as a point of comparison for WoW texts, however, other text types could be used, depending on the point one wants to make. By looking at a corpus of academic English, one could research how the complexity of WoW-gamers' language production compares to the learners' language production in formal settings.

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge University Press.

Baayen, R. H. (2010). languageR: Data sets and functions with "analyzing linguistic data: A practical introduction to statistics". [Computer software manual]. Available from `http://CRAN.R-project.org/package=languageR` (R package version 1.0)

Baroni, M., & Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of the 4th language resources and evaluation conference (lrec).* Lisbon, Portugal.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*, 209-226.

Bryant, T. (2006). Using world of warcraft and other mmorpgs to foster a targeted, social, and cooperative approach toward language learning. Available from `http://www.academiccommons.org/commons/essay/bryant-MMORPGs-for-SLA`

Cheung, H., & Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, *13*, 53–76.

Covington, M. A., He, C., Brown, C., Naçi, L., & Brown, J. (2006). How complex is that sentence? a proposed revision of the rosenberg and abbeduto d-level scale. *CASPR Research Report*.

Dubay, W. H. (2004). *The principles of readability.* Costa Mesa, CA: Impact Information. Available from `http://www.impact-information.com/impactinfo/readability02.pdf`

Evert, S., & Baroni, M. (2008). zipfR: Statistical models for word frequency distributions [Computer software manual]. Available from `http://zipfR.R-Forge.R-project.org/` (R package version 0.6-5)

Flesch, R. (1974). *The art of readable writing: with the flesch readability formula.* Harper & Row.

Gee, J. (2007). *Good video games + good learning: collected essays on video games, learning, and literacy.* P. Lang.

Halliday, M. (1989). *Spoken and written language.* Oxford University Press.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing. *Lund University, Department of Linguistics and Phonetics, Working Papers*.

Lu, X. (2009). Automatic measurement of syntactic complexity in child lan-

guage acquisition. *International Journal of Corpus Linguistics*, *14*(1), 3–28.

Lu, X., Thorne, S. L., & Gamson, D. (2011). *Toward a framework for computational assessment of linguistic complexity of grade-level reading materials* [Draft – work in progress].

Michalke, M. (2011). koRpus: Text analysis tool set [Computer software manual]. Available from `http://www.reaktanz.de` (R package version 0.01-7)

Rankin, Y., Gold, R., & Gooch, B. (2006). Playing for keeps: gaming as a language learning tool. In *Acm siggraph 2006 educators program.* New York, NY, USA: ACM.

Rankin, Y., McKenzie, M., Shute, M. W., & Gooch, B. (2008). User centered game design: evaluating massive multiplayer online role playing games for second language acquisition. In *Proceedings of the 2008 acm siggraph symposium on video games* (pp. 43–49). New York, NY: ACM.

Schmid, H. (1994). Probabilistic Part-of-Speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing* (pp. 44–49).

Steinkuehler, C. A. (2008). Cognition and literacy in massively multiplayer online games. In *Handbook of research on new literacies* (p. 1-38). Mahwah NJ: Erlbaum. Available from `http://labweb.education.wisc.edu/curric606/readings/Steinkuehler2005.pdf`

Thorne, S. L. (2008). Transcultural communication in open internet environments and massively multiplayer online games. In *Mediating discourse online* (pp. 305–327). Amsterdam: John Benjamins.

Thorne, S. L., Black, R. W., & Sykes, J. (2009). Second language use, socialization, and learning in internet interest communities and online games. *Modern Language Journal*, *93*.

Tweedie, F., & Baayen, R. (1998). How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, *32*.

Wiersma, W., Nerbonne, J., & Lauttamus, T. (2011). Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing*, *26*(1), 107–124. Available from `http://urd.let.rug.nl/nerbonne/papers/WierNerbLaut-LLC2010.pdf`

Wilcox, B. (2011). Beyond façade: Pattern matching for natural language applications. Available from `http://www.gamasutra.com/view/feature/6305/beyond_fa%C3%A7ade_pattern_matching_.php?page=1`