Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Conglomerate Industry Choice and Product Language

Gerard Hoberg, Gordon Phillips

To cite this article:
Gerard Hoberg, Gordon Phillips (2017) Conglomerate Industry Choice and Product Language. Management Science

Published online in Articles in Advance 31 Mar 2017

. http://dx.doi.org/10.1287/mnsc.2016.2693 http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–21
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Conglomerate Industry Choice and Product Language

Gerard Hoberg,[a] **Gordon Phillips**[b,c]

[a] Marshall School of Business, University of Southern California, Los Angeles, California 90089; [b] Tuck School of Business, Dartmouth College, Hanover, New Hampshire 03755; [c] National Bureau of Economic Research, Cambridge, Massachusetts 02138

Contact: hoberg@marshall.usc.edu (GH); gordon.m.phillips@tuck.dartmouth.edu (GP)

Received: August 13, 2015
Accepted: October 10, 2016
Published Online in Articles in Advance: March 31, 2017

https://doi.org/10.1287/mnsc.2016.2693

Abstract. We analyze the words that firms use to describe their products so we can examine the determinants of which industries conglomerate firms operate within. Our central finding is that multiple-industry firms operate across industries with higher product language overlap. Multiple-industry firms also avoid industries with more distinct language boundaries and those with more specialized within-industry language. We also find evidence linking these results to specific synergies such as potential entry into new markets and realized synergies in the form of higher 10-K product description growth. These findings are consistent with multiple-product firms operating primarily in industries that lack language specialization. Our findings show that most conglomerates are not true diversified conglomerates with little overlap in their lines of business, as most firms that operate across multiple industries choose industries with high language overlap and potential synergies. Our results support theories of firm organization and organizational language.

History: Accepted by Amit Seru, finance.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2016.2693.

Keywords: conglomerates • product market language • synergies • firm organization • organizational language • text analytics

1. Introduction

Why do firms choose to produce across particular industry combinations and not others? The literature has postulated both benefits (Stein 1997) and costs (Scharfstein and Stein 2000) of the multiple-industry organizational form. However, the literature has taken the existing multiple industry choices as given.[1] The literature has not shown why multiple-industry firms choose some industry combinations and not others or explained the role of potential firm synergies and product market language in the choice of organizational form.

This paper studies the determinants of which industries conglomerates choose to operate within. Firms that choose to operate in multiple industries face a trade-off between choosing complementary industries—which may enhance overall efficiency—and choosing industries that are different in order to diversify. Historically, producing in multiple industries has been viewed as a way to reduce the variance of cash flows by producing products with uncorrelated cash flows, as studies by Lewellen (1971) and recently by Hann et al. (2013) illustrate. We show that most conglomerate formation is related to choosing complementary industries whose products are related—not unrelated. Given that diversification and reducing the variance of firm cash flows is one of the main arguments in the literature for conglomerate formation, the fact that most conglomerates are producing in related industries is surprising.

The choice of which related industries to produce in involves trade-offs between specialization and coordination across industries. From a theoretical standpoint, Becker and Murphy (1992) model how firms trade off between the costs of coordination across different tasks and the gains to specialization in determining which tasks and products are grouped together. Hart and Moore (2005) focus on how agents within the firm are either coordinators or specialists in determining the optimal hierarchy within an organization. Crémer et al. (2007) focus on product language, the words that firms use within and across industries, and the extent to which these languages can be broad enough to allow coordination across industries. They predict that broader and less specialized product languages can lower the cost and increase the benefits of organizing across industries. Their theory shows that it is more likely that a firm chooses to be broader and to communicate across industries when the degree of language overlap and potential synergies across industries are high and the cost of imprecise communication is low.

We analyze firm industry choice and test the predictions of Crémer et al. (2007) by considering the degree of within-industry specialization and the extent to which firms in different industries share common product market language, which can allow them to develop across-industry synergies. To analyze firm product language, we use computational linguistics to analyze the words that firms use in the business descriptions of the 10-Ks they file with the U.S.

Securities and Exchange Commission (SEC). Analyzing the cross-industry structure of words from firm 10-K filings, we test hypotheses on how synergies and asset complementarities relate to industry configuration choice for multi-industry firms.[2]

Crémer et al. (2007) focus on the key trade-off between facilitating internal communication and encouraging communication with other organizations. They conclude that distinct sets of technical words place a limit on firm scope. A broader scope allows for more synergies to be captured, but this has to be weighed against the cost of less precise communication in each unit. We find direct support for this link along three dimensions. First, firms that operate in multiple product markets are more likely to operate in markets with more across-industry language overlap. Second, multiple-industry firms avoid industries with strong language boundaries and industries with a high degree of within-industry specialization and focus. Third, we find evidence of links to specific synergies in the form of potential entry into related product markets, as well as evidence of realized synergies in the form of greater ex post product description growth when firms are producing in industries with higher cross-industry product language overlap.

These contributions extend the research of Hoberg and Phillips (2010), which examines *within*-industry relatedness and shows that merging firms with high ex ante relatedness have high future product growth consistent with synergies for *within*-industry mergers. However, this existing research does not study industry choice and, moreover, does not use any industry information or information about groups of industries. We extend this previous work by examining the fundamental industry factors that drive conglomerate industry choice and where conglomerate firms produce at the industry level. We also extend that paper by considering that conglomerate firms are less likely to operate in highly specialized industries and more likely to operate in industry pairs with high-value, less competitive industries, residing between the given pair so that related synergistic new products may be produced. These two hypotheses are particularly distinctive to this paper.

In our analysis, we first convert firm product text into a spatial representation of the product market, following Hoberg and Phillips (2016). In this framework, each *firm* has a product location in this space, based on its product text, that generates an informative mapping of likely competitors. A central innovation of our article is to illustrate that *industries* also have locations in the product space, and relatedness analysis at the industry level can be used to examine theories of multiple-industry production. Our spatial framework thus allows an assessment of how similar industry languages are to each other and which industries in the product market space are "between" any given pair of industries, providing unique measures of potential asset complementarities.[3]

Apple Inc. is an example of a firm that illustrates our key ideas. Its multiple-function products enable it to compete in multiple markets and to offer differentiated products competing with cell phones, computers, and digital music—industries that are highly related today. Apple was successful in its decision to operate jointly in these industries and uses language that focused firms use in each of these markets. It likely utilizes synergies found across previous industry boundaries.

Although our main tests use a framework that relies on the validity of industry classifications, an additional innovation is that we also examine the links between product vocabulary and asset complementarities using a framework that is invariant to industry classifications. In particular, we consider the degree of transitivity in product language overlap among rival firms, as well as among rivals of rival firms.[4] This concept of transitivity is related to the concept of industry boundaries and the degree of vocabulary specialization, as the ability to develop communication that can cross industry boundaries is essential in the realization of product scope benefits.

Our paper makes four main contributions. First, we examine in which industry combinations multiple-industry firms choose to operate based on industry product language. We find that product language overlap and potential synergies, within-industry language specialization, and the existence of industries lying between two industries can explain conglomerate industry choice.

Second, we show how fundamental industry characteristics, including economies of scale and vertical relatedness, differ in their effect on organizational form. Multiple-industry firms are less likely to operate in industries with high economies of scale and more likely to operate across industries that are vertically related. These respective industries exhibit high ex ante measures of language overlap, consistent with the existence of potential synergies.

Third, we show that conglomerates are less likely to produce in industries with high language complexity. Moreover, when conglomerates do produce in industries with high language complexity, they operate in industries that are more tightly clustered in the product space. These results support the prediction in Crémer et al. (2007) that firms favor a narrower operating profile when the cost of imprecise communication (in our case, resulting from complexity) is high. Fourth, we show evidence consistent with increases in product offerings by multiple-industry firms when their respective industries exhibit high ex ante measures of language overlap, consistent with the existence of potential synergies.

In related work, and consistent with this view, Hoberg and Phillips (2015) document that multiple-industry firms that have distinct product offerings trade at stock market premia. In all, our findings support theoretical links to organizational language, language boundaries, and synergies. Our results also help explain why so many firms continue to use the conglomerate structure despite potential negative effects on valuation as noted by past studies.

Our evidence is also consistent with the conclusion that multiple-industry production, as identified by the Compustat segment tapes, does not fit the historical view that multiple-industry firms operate unrelated business lines under one corporate headquarters, with diversification being the primary aim.[5] Rather firms choose industry pairs in which to operate based on industry language overlaps and potential asset complementarities. For example, we find that roughly 69% of Compustat multiple-industry pairs are in industries that satisfy one of the two following conditions: (a) the language overlap of the pair is similarly as high as industry pairs in the same SIC-2, or (b) the industry pair is above the 90th percentile of vertical relatedness among all industry pairs. The magnitude of this finding suggests that studies aimed at explaining the behavior of diversified multiple-industry firms need to reduce the sample of Compustat multiple-industry firms to the much smaller subsample that plausibly has diversification of cash flows as a primary motive.

The rest of our paper proceeds as follows. In Section 2, we present new measures of industry relatedness based on product language and develop our key hypotheses. In Section 3, we discuss our data, variables, and the methods we use to examine industry choice. Section 4 presents the results of our analysis of industry choice. Section 5 presents our analysis of competitor firm product-market transitivity based on product language used by firms. Section 6 presents our analysis of subsequent product growth. Section 7 examines how our results change as language complexity increases, and Section 8 concludes.

2. Industry Fundamentals and Firm Organization

We ask whether there are certain fundamental industry characteristics—distinct from vertical relatedness—that make operating in two different industries valuable. The central hypotheses we examine are whether product market overlap across industries, within-industry language specialization, and the industries lying between a given pair of industries impact which industries firms operate within and what types of firms operate across these industries.

Our research foundation is related to the trade-off between specialization and coordination. Historically, producing in multiple industries has been viewed as a way to reduce the variance of cash flows by producing products with uncorrelated cash flows. Existing studies (e.g., Hann et al. 2013) thus take the industries chosen as given and examine the properties of the cash flows of conglomerates. We focus on whether and how much conglomerate formation is related to choosing complementary industries whose products are related—not unrelated—and whether there are other motives. Becker and Murphy (1992) model how firms trade off the costs of coordinating workers across different tasks versus the gains to specialization across industries. In their analysis, specialization among complementary tasks links the division of labor to coordination costs, knowledge, and the extent of the market. Workers invest in specialized knowledge until the costs of coordinating specialized workers outweigh the gains from specialization. Theories of the impact of communication on this trade-off have been studied by Bolton and Dewatripont (1994) and Alonso et al. (2008).

Our analysis captures the extent that different industries use different sets of specialized words and a common language across products when the firm wishes to capture synergies, as is theoretically modeled by Crémer et al. (2007). They write, “A broader firm, which must use a common code, is more likely when the degree of synergy among services is high, when the cost of imprecise communication is low, and when the types of problems faced by the services are similar” (p. 376). A broader scope of language thus allows for more synergies to be captured but at the cost of less precise communication within each unit.

Our focus on the potential for asset complementarities also relates to the proposition from Teece (1980), who writes, “[I]f economies of scope are based upon the common and recurrent use of proprietary knowhow or the common and recurrent use of a specialized and indivisible physical asset, then multiproduct enterprise (diversification) is an efficient way of organizing economic activity” (p. 223). Industry economies of scale, as Maksimovic and Phillips (2002) emphasize, exert the opposite force, as economies of scale increase the optimal size of a firm. Higher economies of scale reduce the incentive to produce across industry pairs, as the relative advantage of operating within a single industry increases with economies of scale.

We discuss our key hypotheses through the lens of a spatial representation of the product market (see Hoberg and Phillips 2016 for a discussion of the text-based product market space).[6] In this representation, all firms have a “location” on a high dimensional unit sphere that is determined by the overall vocabulary used in the given firm’s 10-K business description.

We extend the previous firm-specific work of Hoberg and Phillips (2016) by constructing new *industry*-based measures of how groups of firms are related to each other. Thus, the new measures in this paper capture how industries have a simple but highly informative representation in the product market language space that can be used to examine how industries relate to one another. Intuitively, an industry should be viewed as a cluster of firms in the product market space, and hence each industry has both a location and a degree to which it is spread out in the product market space. For example, industries that are highly spread out have a low degree of within-industry product similarity.

The new fundamental measures of industries that are constructed in this paper allow us to assess how every pair of industries relates to one another, capturing potential synergies and how products differ within industries. We first measure how close industries are in the product space using the extent of language overlap, *Across-Industry Language Similarity* (AILS). We measure the extent of transitivity of language across competitors, *TransComp*, to measure the strength of product market boundaries. We also measure the extent of within-industry language specialization and focus, *Within-Industry Language Similarity* (WILS), and the extent to which other industries lie between a given industry pair, *Between Industries* (BI).

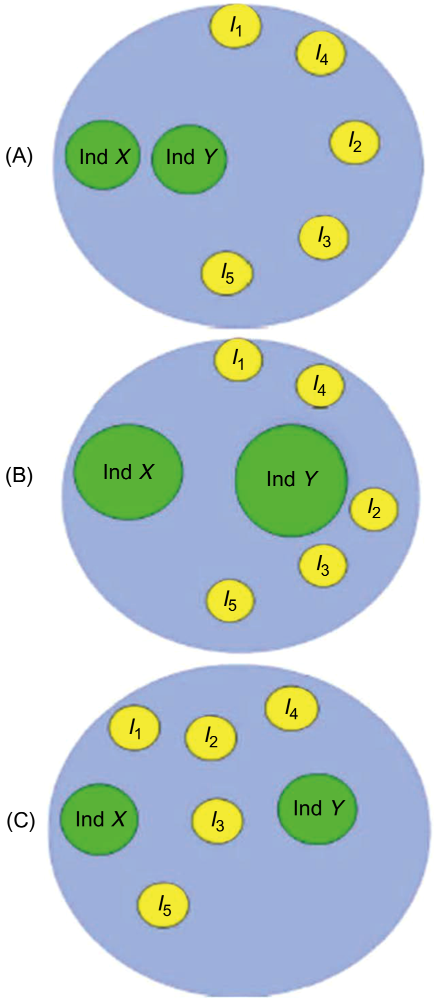
We use these new industry-relatedness measures from firm product text to test three hypotheses. These hypotheses are illustrated in Figures 1(A)–1(C), where each circle represents an industry in the product market space, and the size of the circle illustrates the degree to which the given industry is spread out (low within-industry similarity).

Hypothesis 1 (**H1**; Synergies, Asset Complementarities, and Cross-Industry Production). *Multiple-industry firms are more likely to produce in industry pairs that have higher product language overlap and thus higher potential for cross-industry synergies.*

The intuition behind this hypothesis relates to Crémer et al. (2007). Industries with more organizational product language overlap are more amenable to multiple-product production because more communication is possible across the divisions. Thus, these industries have a greater potential for feasible synergies. Hence, we should observe more multiple-industry firms operating across industries with greater product market overlap, and we should also observe higher levels of realized synergies for these same firms. Figure 1(A) depicts industries *X* and *Y* as having a high degree of cross-industry similarity compared with other industry pairs, and H1 predicts that more multiple-product firms will choose to jointly operate in *X* and *Y* relative to other pairwise configurations. Note that this hypothesis and the evidence provided complements Hoberg and Phillips (2010), which analyzes potential synergies in mergers inside industries.

Hypothesis 2 (**H2**; Within-Industry Similarity). *Multiple-industry firms are less likely to produce in industries with high within-industry language similarity.*

Figure 1. (Color online) Illustration of Hypotheses



Notes. Panel (A) depicts the concept of across-industry similarity (potential asset complementarities). Panel (B) depicts the concept of WILS. Industries with low levels of WILS occupy a larger volume of the product market space; for example, industries *X* and *Y* have lower WILS than I_1 or I_4. Panel (C) depicts the concept of BI. Industry I_3 lies between industries *X* and *Y*.

The main idea underlying this hypothesis is best summarized by this quote from (Crémer et al. 2007, p. 373): "The need to learn specialized codes constrains the scope of the organization: a specialized code facilitates communication within a service or function, but limits communication between services, and thus makes coordination between them more difficult." Figure 1(B) depicts industries *X* and *Y* as having a low degree of within-industry similarity compared to other industries, and hence firms residing in *X* and *Y* likely have greater potential to coordinate across-industries because of less specialization, and H2 predicts that more multiple-product firms will choose to jointly operate in *X* and *Y* relative to other pairwise configurations.

Hypothesis 3 (H3; Between Industries). *Multiple-industry firms are more likely to operate in an industry pair when the pair of industries has more high-value, less competitive industries residing between the given pair.*

This hypothesis is related to the first hypothesis regarding potential synergies but tests the more refined prediction that synergies in the form of potential entry into related markets are independently relevant. The intuition underlying this hypothesis is that the existence of high-value, less competitive industries between a given pair of industries may allow a multiple-industry firm to potentially enter these highly valued product markets. Figure 1(C) depicts industries X and Y as having a third industry, I_3, residing between them. If firms in I_3 are highly valued, then H3 predicts that multiple-product firms will more frequently choose to operate in industries X and Y.

3. Data and Methodology

3.1. The Compustat Industry Sample

We construct our Compustat sample using the industrial annual files to identify the universe of publicly traded firms, the Compustat segment files to identify which firms are multiple-industry producers, and the industry of each segment. We define a conglomerate as a firm having operations in more than one SIC-3 industry in a given year. To identify segments operating under a conglomerate structure, we start with the segment files, which we clean to ensure we are identifying product-based segments instead of geographic segments. We keep conglomerate segments that are identified as business segments or operating segments. We only keep segments that report positive sales. We aggregate segment information into three-digit SIC codes and identify firms as multiple-industry firms only when they report two or more three-digit SIC codes. We identify 34,218 unique multiple-industry firm years from 1996 to 2013 (we limit our sample to these years because of required coverage of text-based variables), which have 88,578 unique conglomerate-segment-years. We also identify 70,503 unique pure play firm-years (firms with a single segment structure).

When we examine how multiple-industry firms change from year to year, we further require that a multiple-industry structure exists in the previous year. This requirement reduces our sample to 29,777 unique conglomerate years having 78,533 segment-years. Because we use pure play firms to assess industry characteristics that might be relevant to the formation of multiple-industry firms, we also discard conglomerate observations if they have at least one segment operating in an industry for which there are no pure play benchmarks in our sample. We are left with 25,541 unique multiple-industry firm-years with 69,355 unique segment multiple-industry firm-years. This final sample covers 3,344 unique three-digit SIC industry-years. As there are 18 years in our sample; this is roughly 186 industries per year.

We also consider a separate database of pairwise permutations of the SIC-3 industries in each year. We use this database to assess which industry pairs are most likely to be populated by multiple-industry firms that operate in the given pair of industries. This industry-pair-year database has 382,494 total industry pair × year observations (roughly 21,250 industry pair permutations per year).

3.2. The Sample of 10-K Filings

The methodology we use to extract 10-K text follows Hoberg and Phillips (2016) and (2010). The first step is to use web-crawling and text-parsing algorithms to construct a database of business descriptions from 10-K annual filings on the SEC EDGAR website from 1996 to 2013. We search the EDGAR database for filings that appear as "10-K," "10-K405," "10-KSB," or "10-KSB40." The business descriptions appear as Item 1 or Item 1A in most 10-K filings. The document is then processed using APL for text information and a company identifier, the Central Index Key (CIK).[7] Business descriptions are legally required to be accurate, as Item 101 of Regulation S-K requires firms to describe the significant products they offer, and these descriptions must be updated and representative of the current fiscal year of the 10-K.

3.3. Word Vectors and Cosine Similarity

Using the database of business descriptions, we form word vectors for each firm based on the text in each firm's product description. To construct each firm's word vector, we first omit common words that are used by more than 25% of all firms. Following Hoberg and Phillips (2016), we further restrict our universe in each year to words that are either nouns or proper nouns (excluding geographical terms such as countries, states, and the top 50 cities).[8] Let M_t denote the number of such words. For a firm i in year t, we define its word vector $W_{i,t}$ as a binary M_t-vector, having the value 1 for a given element when firm i uses the given word in its year t 10-K business description.[9] We then normalize each firm's word vector to unit length, resulting in the normalized word vector $N_{i,t}$.

Importantly, each firm is represented by a unique vector of length 1 in an M_t-dimensional space. Therefore, all firms reside on a M_t-dimensional unit sphere, and each firm has a known location. This spatial representation of the product space allows us to construct variables that more richly measure industry topography, for example, to identify other industries that lie between a given pair of industries.

The cosine similarity for any two word vectors $N_{i,t}$ and $N_{j,t}$ is their dot product, $\langle N_{i,t} \cdot N_{j,t} \rangle$. Cosine similarities are bounded in the interval $[0, +1]$ when both vectors are normalized to have unit length and when they do not have negative elements, as will be the case for the quantities we consider here. If two firms have similar products, their dot product will tend toward 1.0, while dissimilarity moves the cosine similarity toward 0. We use "cosine similarity" because it is widely used in studies of information processing (see Sebastiani 2002 for a summary of methods). It measures the cosine of the angle between word vectors on a unit sphere.

3.4. Firm Restructuring Over Time

We examine whether our spatial industry variables can explain how multiple-industry firms restructure over time, and we classify restructuring in three different ways. Because we consider the role of industry topography, the unit of observation for these variables is a pair of segments operating within a conglomerate. We define *New Segment Pairs* as a new pair observed in a conglomerate in year t that did not exist in the conglomerate in the previous year $t-1$. We then define *New Segment Pairs Likely Obtained Through Growth* as pairs that did not exist in the conglomerate's structure in the previous year, and the conglomerate had fewer segments in year $t-1$ relative to year t. Finally, we define *New Segment Pairs Linked to SDC (Securities Data Company) Acquisitions* as segment pairs that did not exist in the conglomerate's structure in the previous year, and the conglomerate was the acquirer of an acquisition of at least 10% of its assets between year $t-1$ and year t.

3.5. Industry Variables

The primary dependent variable we seek to explain is the fraction of multiple-industry firms producing in an industry. Our primary four explanatory industry variables are *Across-Industry Language Similarity* (potential synergies), *Within-Industry Language Similarity*, the fraction of industries that are *Between Industries*, and the *Transitivity of Competitors.* In this section, we discuss these variables and the additional industry variables we consider both as control variables and as variables of individual interest.

Because we seek to examine the industry pairs in which multiple-industry firms produce, to avoid any mechanistic relationships, we focus only on single-segment firms to calculate the characteristic industry-relatedness variables we later use as explanatory variables. We then use the Compustat segment tapes to examine how observed conglomerate industry configurations relate to these text-based industry characteristics computed from single-segment firms.

Because conglomerate segments are reported using SIC codes, our initial analysis relates to industry configurations and their incidence based on three-digit SIC code industry definitions. In later analysis, we consider industry groupings using the fixed industry classifications (FIC) from Hoberg and Phillips (2016), where firms are identified as competitors using text-based methods.

3.5.1. Text-Based Industry Variables. ***Across-Industry Language Similarity of Product Language***: The AILS measure is based on industry product language overlap. It captures the extent to which product descriptions of firms in two different industries use overlapping language. The AILS measure is meant to capture the similarities between the products that two industries produce and thus the potential for synergies. Specifically, across-industry similarity is the average textual cosine similarity of all pairwise permutations of the N_i and N_j firms in the two industries i and j, where textual similarity is based on word vectors from firm business descriptions (see Section 3.3 for a discussion of the cosine similarity method). This measure captures the average overlap in product words that two randomly drawn firms from industries i and j will have in common.

Within-Industry Language Similarity: The WILS measure captures the language specialization of industry i. It is the average cosine similarity of the business descriptions for all pairwise word permutations of the N_i firms in industry i (i.e., the degree of language overlap within industry).

Between Industries: For the BI measure, we use the across-industry similarity measure (described above) to assess which other industries lie between any given industry pair. Specifically, a third industry is between two industries in a given industry pair if the third industry is closer in textual distance to each industry in the pair than the two industries in the pair are to each other.

The across-industry similarity measure based on across industry language overlaps discussed above is instrumental in computing the fraction of industries between a given pair. More formally, where $AILS_{i,j}$ denotes the across-industry product language similarity of industries i and j, we define a third industry k as being *between* industries i and j if the following relationship holds:

$$AILS_{k,i} \geq AILS_{i,j} \quad \text{and} \quad AILS_{k,j} \geq AILS_{i,j}. \qquad (1)$$

The fraction of industries between a given pair of industries i and j is therefore the number of industries k (excluding i and j) satisfying this condition divided by the total number of industries in the database in the given year (excluding i and j).

Transitivity of Competitors: TransComp is a measure of how weak a given product market's language boundaries are (the degree to which its language is not specialized). This measure is computed for each firm and is based on the Text-Based Network Industry Classification (TNIC) of Hoberg and Phillips (2016). TNIC industries are derived from the 10-K text in firms' business descriptions. The industry classification identifies, for each given firm, the set of rival firms having the most similar business descriptions to the given firm using the cosine similarity method. TNIC industries are calibrated to be as granular as the widely used three-digit SIC industry classification. To compute TransComp, for each focal firm, we first identify the set of TNIC rivals. We then also use TNIC to identify the set of rivals of the rivals. TransComp is the fraction of firms in the set of rivals of rivals that are also in the set of rivals of the focal firm, as explained above. Because TNIC links are direct estimates of language overlap, TransComp measures the degree to which language overlap is transitive in a given product market. This variable by design lies in the interval $[0,1]$. TransComp is a particularly stark measure of language specialization because it does not rely on the quality of the Compustat segment tapes and their potentially questionable SIC code designations. Markets with strong language boundaries (high transitivity) likely use highly specialized languages that do not overlap with neighboring industries. Hypothesis 2 predicts that such markets are less likely to be chosen by multiple-industry firms.

3.5.2. Non-Text-Based Industry Control Variables. As is the case for AILS and the fraction of industries between a given pair, our first set of three additional control variables is a property of a pair of industries. These include a key control for industry-pair relevance, a measure of vertical relatedness, and a dummy identifying which industries are in the same two-digit SIC code. Because we aim to examine conglomerate incidence rates across industry pairs, controlling for industry pair relevance is important. For example, if multiple-industry firms were formed by randomly choosing among available pure-play firms in the economy, then the incidence of conglomerate operating pairs would be related to the product of the fraction of firms residing in industries i and j. Therefore, we define the *Pair Likelihood If Random* variable as the product $(F_i x F_j)$, where F_i is the number of pure-play firms in industry i divided by the number of pure-play firms in the economy in the given year.

We consider the input-output tables to assess the degree to which a pair of industries is vertically related. The inclusion of this variable is motivated by studies examining vertically related industries and corporate policy and structure, including Fan and Goyal (2006), Kedia et al. (2011), and Ahern and Harford (2014). We consider the methodology described in Fan and Goyal (2006) to identify vertically related industries; we use the closest proceeding fifth year, given these tables are only available every fifth year, of the "Use Table" of Benchmark Input-Output Accounts of the U.S. Economy to compute, for each firm pairing, the fraction of inputs that flows between each pair.

We also consider economies of scale and measure the gains to scale within each industry. This measure is captured by estimating a traditional Cobb–Douglas production function.[10] As with our measure of across-industry similarity, we estimate this measure for both traditional SIC industry groupings and the new text-based FIC of Hoberg and Phillips (2016). We estimate the production function using firm-level data from Compustat. We use 10 years of lagged data for each firm in a given industry and use sales as the dependent variable. We include the following right-hand-side variables: net property plant and equipment for capital, the number of employees, the cost of goods sold, and firm age. All variables are in natural logs, and variables except for age and the number of employees are deflated to 1987 real dollars using the wholesale price index. An industry's *Economies of Scale* is measured as the sum of the coefficients on net property, plant, and equipment and the cost of goods sold.

We also consider two additional control variables that are a property of a single industry: patent applications and industry instability. We compute patent applications at the industry level as the fraction of total patents applied for by firms in the given industry (as a fraction of all patents applied for in the given year) scaled by the total assets of firms in the given industry in the given year. We multiply this quantity by 10,000 for convenience. We compute industry instability as the absolute value of the natural logarithm of the number of firms in the industry in year t divided by the number of firms in the same industry in year $t-1$. Industries with higher instability experience changes in the industry's membership over time.

3.6. Summary Statistics

Table 1 displays summary statistics for our conglomerate and pure play firms, as well as industry pair databases. Panel A shows that multiple-industry firms are generally larger than the pure-play firms in terms of total value of the firm.

Panel B of the table compares randomly drawn pairs of SIC-3 industries to the SIC-3 industries comprising a conglomerate configuration. The panel shows that a randomly drawn pair of three-digit SIC industries has 0.169 multiple-industry firms having segments operating in both industries of the given pair. Hence, most randomly chosen industry pairs do not have multiple-industry firms operating in the pair. The average across-industry similarity or "language overlap" of *random* pairs is 0.014, which closely matches the average firm similarity reported in Hoberg and Phillips (2016). This quantity more than doubles for actual multiple-industry firms to 0.037, indicating that multiple-industry firms are perhaps less diversified than previously thought.

Table 1. Summary Statistics

Variable	Mean	Std. dev.	Minimum	Median	Maximum
Panel A: Multiple-industry (25,541 obs.) and pure-play firms (70,503 obs.)					
Firm Value (multi-industry)	13,173.2	47,661.7	0.009	1,216.76	1,036,340
Firm Value (pure-play)	2,672	25,931	0.001	239	3,307,572
Panel B: Industry pairs (382,494 obs.) and multiple-industry firms (25,541 obs.)					
Number of Multiple-Industry Firms in Pair (ind. pairs)	0.169	0.977	0.000	0.000	62.000
Across-Industry Language Similarity (ind. pairs)	0.014	0.014	0.000	0.010	0.205
Across-Industry Language Similarity (multiple-industry firms)	0.037	0.027	0.001	0.029	0.160
Economies of Scale (ind. pairs)	0.812	0.358	0.000	0.958	1.381
Economies of Scale (multiple-industry firms)	0.005	0.064	0.000	0.000	1.111
Between Industries (ind. pairs)	0.321	0.263	0.000	0.252	0.992
Between Industries (multiple-industry firms)	0.088	0.129	0.000	0.032	0.954
Within-Industry Language Similarity (ind. pairs)	0.097	0.043	0.000	0.091	0.625
Within-Industry Language Similarity (multiple-industry firms)	0.088	0.034	0.020	0.081	0.253
Vertical Relatedness (ind. pairs)	0.003	0.014	0.000	0.000	0.536
Vertical Relatedness (multiple-industry firms)	0.026	0.067	0.000	0.005	0.536
Patent Applications (ind. pairs)	0.127	0.301	0.000	0.000	4.546
Patent Applications (multiple-industry firms)	0.356	0.486	0.000	0.140	3.108
Industry Instability (ind. pairs)	0.258	0.209	0.000	0.210	2.000
Industry Instability (multiple-industry firms)	0.463	0.203	0.000	0.453	1.500
Same 2*-digit SIC Dummy* (ind. pairs)	0.018	0.135	0.000	0.000	1.000
Same 2*-digit SIC Dummy* (multiple-industry firms)	0.219	0.376	0.000	0.000	1.000

Notes. Summary statistics for firm value are reported for our sample of conglomerate and pure-play firms (panel A) for our sample from 1996 to 2013. Summary statistics for key variables of interest are reported for both multiple-industry and single-segment firms in panel B. These variables are discussed in detail in Section 3.5. *Across-Industry Language Similarity* is the average pairwise similarity of firms in one of the industries in the pair with firms in the other industry. *Between Industries* is the fraction of all other industries that lie between the given pair of industries. *Vertical Relatedness* is the degree of vertical relations based on the input-output tables. *Economies of Scale* is based on the estimation of a Cobb–Douglas production function over 10 years, with sales being the dependent variable. *Within-Industry Language Similarity* is the average pairwise similarity of firms in the given industry. *Patent Applications* is at the industry level and is the fraction of total patents applied for by firms in the given industry. *Industry Instability* is the absolute value of the logarithmic change in the number of firms in the given industry over the past year.

To further illustrate this point, we note the historical perspective taken by many studies is that most conglomerates are highly diversified. If we divide all industry pairs into quartiles based on across-industry similarity, we find that 71% of all conglomerates reside in the highest similarity quartile and just 5% in the least similar quartile, indicating that the converse of the historical perspective is perhaps more accurate. That is, most conglomerates operate in less diversified, and more highly related, industry pairs.

This conclusion is further reinforced by comparing the fraction of all other industries lying between the given pair, which is 32.1% for random pairs and just 8.8% for actual multiple-industry firms. Consistent with our central language overlap hypothesis (H1), conglomerate industry pairs are in regions of the product space that are substantially closer together than randomly chosen industries. The average within-industry similarity, intuitively, is much higher, at 0.097. Consistent with our language specialization hypothesis (H2), this quantity is somewhat lower, at 0.088 for actual multiple-industry firms.

We calculate a Pearson correlation table of our primary industry variables. In the interest of space, we present the correlation table in the online appendix as Table EC.1. Foreshadowing a main result, Table EC.1 shows that there is a high correlation (0.243) between the ratio of multiple- versus single-industry firms (a key variable we explain) and AILS. Other correlations are generally modest. Our independent variables also generally correlate little except for the larger correlation for our BI variable and the AILS variable. We thus examine whether multicollinearity might pose a problem for these variables in our regressions. We find that variance inflation factors are less than 2.0, so the correlation of −66.8 between these variables does not pose any multicollinearity concerns in our regressions. These results for variance inflation, along with our very large database of 382,494 observations, indicate that multicollinearity is not a concern in our analysis. The reason AILS and BI are negatively correlated is that if two industries are far apart, then most other industries are "between" them spatially. For this reason, we include both BI and AILS in our regressions to ensure that each is held fixed when examining the other.

Table 2 displays the mean values of our three key text variables for various conglomerate industry pairings. One observation is an industry pair permutation of an actual conglomerate. In panel A, we find that multiple-industry firms populate industries with high across-industry similarity of 0.0344, which is 142% higher than the 0.0142 of randomly chosen industry

Table 2. Conglomerate Multiple-Industry Firm Summary

Subsample	AILS	WILS	BI	No. of obs.
Panel A: Overall				
All multiple-industry firms	0.0344	0.0914	0.1125	58,976
Randomly drawn SIC-3 industries	0.0142	0.0970	0.3209	382,494
Panel B: By conglomerate size				
Two segments	0.0402	0.0908	0.0721	13,947
Three segments	0.0358	0.0896	0.1001	17,944
Four or five segments	0.0328	0.0925	0.1246	19,560
Six or more segments	0.0244	0.0939	0.1853	7,525
Panel C: Shrinking, stable, and growing multiple-industry firms				
Shrink by two or more segments	0.0296	0.0949	0.1493	684
Shrink by one segment	0.0334	0.0917	0.1160	3,868
Stable conglomerate	0.0352	0.0916	0.1083	45,939
Add one segment	0.0316	0.0901	0.1261	6,485
Add two or more segments	0.0285	0.0883	0.1461	2,000
Panel D: Vertical and same SIC-2 multiple-industry firms				
Vertically related segments	0.0378	0.0862	0.0652	20,967
Same SIC-2 segments	0.0583	0.0980	0.0195	11,454

Notes. Summary statistics for various industry pairs from 1996 to 2013. Panel A compares observed multiple-industry pairs to randomly drawn industry pairs. Panel B displays observed multiple-industry pairs for firms with varying segment counts. Panel C displays industry pairs for multiple-industry firms that are growing, stable, or shrinking, as noted. Panel D displays conglomerate industry pairs for vertically integrated segments and for segments that are in the same two-digit SIC code.

pairs. Hence, multiple-industry firms are more likely to operate in industry pairs with higher levels of language overlap, likely capturing higher potential synergies. Multiple-industry firms also tend to populate industries with lower-than-average within-industry similarity and industries having a lower-than-average number of other industries between them.

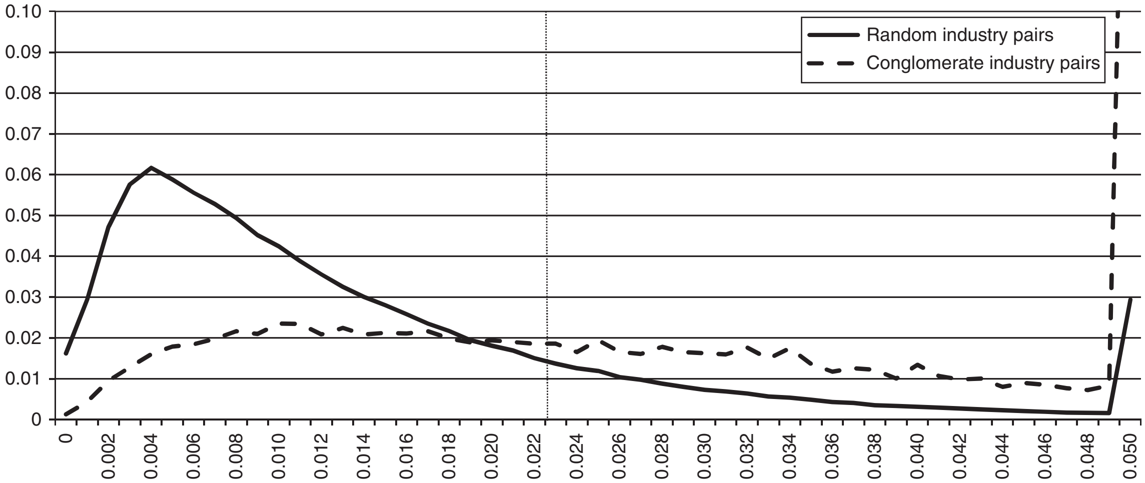
In panel B of Table 2, we report results for smaller multiple-industry firms (two or three segments) compared with those of larger multiple-industry firms. The table suggests that larger multiple-industry firms tend to produce across a wider area of the product market space, as they have lower across-industry similarity. They also tend to produce in industries with more industries between them and in industries that have higher within-industry similarity. In panel C, we observe that most multiple-industry firms (45,939) are stable from one year to the next, although 3,868 of them reduce in size by one segment, and 684 multiple-industry firms reduce in size by two or more segments. Analogously, 6,485 firms increase in size by one segment, and 2,000 firms increase in size by two segments.

In panel D, we observe that vertically related multiple-industry firms have average across-industry similarities that are close to the average for all conglomerate pairs. However, the panel also shows that across-industry similarities are higher for industries having the same two-digit SIC code. Both vertical industries and those in the same two-digit SIC code also have fewer industries between them than do randomly drawn industries or the industry pairs in which most conglomerates operate.

In Tables EC.2–EC.4 of the online appendix, we also present tables that show the top 10 and bottom 10 industries for our primary industry variables in 2013, the last year of our sample. Tables EC.5–EC.7 present the same statistics for 1997, the first year of our sample. This provides intuitive examples our new text-based industry variables. We present these in the online appendix because of space constraints. Table EC.2 shows the top 10 and bottom 10 industries for AILS. The table shows that the industries are indeed quite related. Pairs with different SIC two-digit codes include pipelines (SIC 461) and natural gas transmission (SIC 492) and also pipelines and wholesale petroleum bulk stations and terminals (SIC 517). The BI pairs presented in Table EC.3 are also intuitive. Pairs include footwear (SIC 314) and retail—women's clothing stores (SIC 562). Table EC.4 shows the industries with the highest and lowest WILS. Transportation industries, collection agencies, and tires are sample industries with high WILS.

Figure 2 displays the large economic magnitude of the link between across-industry product language similarity and conglomerate firm industry choice. In particular, the solid line displays the distribution of across-industry product language similarity scores for randomly drawn industry pairs, and the dashed line displays this distribution for observed conglomerate firm industry pairs. The figure shows that the dashed line has a distribution that (a) is strongly shifted to the right relative to the solid line and (b) has a very large right tail, as evidenced by the higher level of density on the right side of the figure and the large amount

Figure 2. Distribution of AILS Scores for Randomly Drawn Industry Pairs vs. Conglomerate Industry Pairs



Notes. Across-industry similarity is the average pairwise 10-K textual similarity of firm pairs in each SIC-3 industry based on the text in each firm's business descriptions. The *x* axis depicts the level of across-industry similarity ranging from 0 to 0.05 (values above this level are in the last data point), and the *y* axis depicts the fraction of industry pairs with the given level of AILS. The dashed line depicts the median across-industry language similarity (0.023) for conglomerate industry pairs. This median is reached at the 85.5th percentile of across-industry similarity for randomly drawn pairs (solid line).

of mass to the right of 0.05. To put this in perspective, the median across-industry similarity of conglomerate industry pairs is at the 85.5th percentile among randomly drawn pairs.

4. Firm Industry Choice

We examine whether our hypothesis can explain in which industry pairs multiple-industry firms produce. We test whether potential synergies and asset complementarities measured through across-industry product language similarity, the fraction of industries between a particular industry pair, and within-industry similarity matter for the likelihood that multiple-industry firms will produce in a particular industry pair. We also consider economies of scale and vertical relatedness.

Table 3 presents ordinary least squares (OLS) regressions where each observation is a pair of three-digit SIC industries in a year derived from the set of all pairings of observed SIC-3 industries in the given year in the Compustat segment tapes. The dependent variable is the ratio of multiple-industry firms to single-segment firms operating in the given industry pair. This is computed as the number of multiple-industry firms operating in the given industry pair divided by the total number of single-segment firms operating in the two industries of the given pair. Panel A displays results based on the entire sample of industry pairs. Panels B and C display results for various subsamples that divide the overall sample based on the competitiveness or the valuations of industries lying between the industry pair. All regressions include industry and year fixed effects, and all standard errors are clustered by industry.

Consistent with our central language overlap hypothesis (H1), panel A shows that higher across-industry language overlap is associated with a higher fraction of multiple-industry firms producing in a particular industry. Consistent with our specialization hypothesis (H2), we find that average within-industry similarity is negatively associated with multiple-industry firms producing in a particular industry. Consistent with H3, panel A also shows that the fraction of industries between a given industry pair also matters positively.

To put the economic magnitude of these results into perspective, we first note that the average ratio of conglomerate segments to pure-play firms for a given industry pair is 0.66. Computing economic impact as the regression coefficient multiplied by each variable's standard deviation, we find that a one-standard-deviation increase in across-industry language similarity would increase this ratio from 0.66 to 1.66. A one-sigma increase in the fraction of industries between a pair would increase this ratio to 0.92, and a one-sigma increase in within-industry similarity would decrease this ratio to 0.55. In all, these economic magnitudes (particularly that for AILS) are very large.

Panels B and C show that especially when high value industries and industries that have high levels of product differentiation (measured using TNIC product similarities as in Hoberg and Phillips 2016) are between the given pair, a higher fraction of multiple-industry firms operates in the given pair. This does not hold for competitive low value industries, as the fraction of industries between the pair becomes insignificant in row (9).

Table 3. Where Multiple-Industry Firms Exist

	Sample	AILS	BI	WILS	Economies of scale	Vertical relatedness	Patent applications	Industry instability	Pair likelihood if random	Same two-digit SIC	No. of obs. [R-SQ]
					Panel A: Full sample						
(1)	All industry pairs	69.907	0.977	−2.588	−1.170	22.758	−0.069	0.344	−0.008	6.006	382,494
		(6.01)	(3.17)	(−5.60)	(−4.94)	(3.09)	(−1.46)	(4.52)	(−0.78)	(5.91)	[0.135]
					Panel B: Univariate subsamples						
(2)	High differentiation	91.249	1.677	−2.866	−0.529	28.402	−0.046	0.287	−0.006	5.250	186,673
	ind. pairs	(9.30)	(6.03)	(−4.23)	(−2.05)	(2.48)	(−0.85)	(3.69)	(−0.37)	(5.89)	[0.166]
(3)	Low differentiation	55.605	0.616	−2.111	−1.541	13.009	−0.052	0.242	0.006	5.646	191,254
	ind. pairs	(3.83)	(2.01)	(−4.33)	(−5.67)	(2.15)	(−0.87)	(2.66)	(2.71)	(2.67)	[0.085]
(4)	High firm value	68.654	1.056	−1.882	−1.180	13.472	−0.063	0.137	0.007	7.002	188,970
	ind. pairs	(4.08)	(2.62)	(−6.21)	(−5.20)	(1.73)	(−1.29)	(2.87)	(3.28)	(3.68)	[0.112]
(5)	Low firm value	61.252	0.531	−2.741	−0.994	27.229	−0.074	0.393	−0.002	4.845	188,957
	ind. pairs	(5.38)	(1.57)	(−4.36)	(−3.37)	(3.29)	(−1.43)	(4.20)	(−0.19)	(6.03)	[0.122]
					Panel C: Bivariate subsamples						
(6)	High diff + High value	82.515	1.767	−3.218	−0.767	24.332	−0.084	0.215	0.015	5.951	55,311
		(6.07)	(4.13)	(−4.58)	(−2.29)	(2.08)	(−1.61)	(2.00)	(1.08)	(5.53)	[0.176]
(7)	Low diff + High value	65.943	0.872	−1.613	−1.269	7.461	−0.052	0.104	0.006	7.781	133,659
		(3.20)	(1.92)	(−4.78)	(−4.77)	(1.10)	(−0.68)	(2.05)	(2.77)	(2.17)	[0.099]
(8)	High diff + Low value	93.867	1.586	−2.439	−0.423	29.789	−0.060	0.331	−0.016	5.035	131,362
		(9.65)	(6.19)	(−2.98)	(−1.30)	(2.52)	(−1.04)	(3.52)	(−0.68)	(5.58)	[0.166]
(9)	Low diff + Low value	52.797	0.593	−3.057	−2.174	18.440	−0.114	0.571	0.001	3.552	57,595
		(3.36)	(1.40)	(−3.43)	(−3.71)	(3.42)	(−1.51)	(2.28)	(0.15)	(3.02)	[0.106]

Notes. OLS regressions with year and industry fixed effects and standard errors (in parentheses) are clustered by industry for our sample of 382,494 industry pairs from 1996 to 2013. One observation is one pair of three-digit SIC industries in a year derived from the set of all permutations of feasible pairings. The dependent variable is the ratio of multiple-industry firms operating in the given industry pair (relative to the total number of single segment firms operating in the two industries in the given pair), expressed as a percentage for convenience. Panel A displays results based on the entire sample. Panels B and C display results for subsamples based on product differentiation and valuations of industries lying between the given industry pair.

This result shows how industry boundaries might be crossed or redrawn using product market synergies to lower the cost of entry into previously differentiated or difficult-to-enter product markets. In particular, multiple industry firms operating in industries spatially located on both sides of a differentiated industry likely have a technological advantage to enter the BI. These results support H3 and document a role of entry synergies in multiple-industry production.

Table 4 examines how industry characteristics influence which new industry pair operations are added to multiple-industry firms in a given year. We also separately consider segment additions by firms having large transactions in the SDC mergers and acquisitions (M&A) database. One observation is one pair of segments in an existing conglomerate in year t.

The dependent variable varies by panel in Table 4. The dependent variable in panel A is the relative fraction of new multiple-industry operating pairs. It is computed as the number of new multiple-industry segments (where the conglomerate did not have this segment in the previous year) operating in each three-digit SIC code pair in the given year divided by the total number of single segment firms operating in the two industries in the given pair. In panel B, we restrict attention to new segments in firms that were the acquirer in an acquisition in the SDC database for a transaction amounting to at least 10% of the firm's assets. In panel C, we restrict attention to segments that were likely created through reclassification. Likely reclassified segments are those that newly appear in years where the total number of segments reported by the given firm is less than or equal to the past-year number of segments (indicating that these new segments were likely classified as being in a different industry in the previous year). The independent variables include various product market variables characterizing the industry pair.

The results in panel A of Table 4 show that segment pairs are likely to be added if product market language overlaps and potential synergies are high. The panel also shows that the coefficient on the across-industry product similarity variable is higher when the industries between two industry pairs have highly differentiated products and are highly valued (and the lowest coefficient when the converse is true). This result is consistent with multiple-industry firms using complementary industry assets to extract

Table 4. New Firm-Industry Segments

	Sample	AILS	BI	WILS	Economies of scale	Vertical relatedness	Patent applications	Industry instability	Pair likelihood if random	Same two-digit SIC code	No. of obs. [R-SQ]
	Panel A: Dep. var = New segment pairs										
(1)	All industry pairs	20.965	0.325	−1.027	−0.485	5.308	−0.019	0.102	−0.001	1.582	382,494
		(6.08)	(3.51)	(−6.69)	(−3.83)	(2.66)	(−1.34)	(4.60)	(−0.45)	(6.20)	[0.066]
(2)	High diff + High value	34.008	0.873	−1.117	−0.411	7.680	0.006	0.088	0.008	1.721	55,311
		(3.96)	(3.02)	(−3.82)	(−2.02)	(2.27)	(0.24)	(2.00)	(1.47)	(3.81)	[0.083]
(3)	High diff + Low value	18.290	0.253	−0.709	−0.333	1.124	0.025	0.032	0.003	1.718	133,659
		(3.44)	(2.19)	(−6.21)	(−2.85)	(0.59)	(1.41)	(1.51)	(3.26)	(1.97)	[0.039]
(4)	Low diff + High value	27.864	0.505	−0.748	−0.426	8.234	−0.017	0.121	−0.008	1.201	131,362
		(7.78)	(5.04)	(−2.60)	(−3.23)	(2.52)	(−1.12)	(3.24)	(−1.02)	(4.75)	[0.078]
(5)	Low diff + Low value	13.194	0.042	−1.131	−0.703	4.531	0.004	0.087	0.001	1.139	57,595
		(3.43)	(0.38)	(−3.96)	(−3.37)	(2.48)	(0.18)	(1.69)	(1.19)	(3.41)	[0.060]
	Panel B: Dep. var = New segment pairs linked to SDC acquisitions										
(6)	All industry pairs	0.978	0.019	−0.040	−0.015	0.197	0.000	−0.001	0.000	0.042	382,494
		(4.68)	(3.57)	(−3.78)	(−1.94)	(1.97)	(0.10)	(−0.30)	(0.15)	(2.91)	[0.007]
(7)	High diff + High value	2.821	0.084	−0.057	−0.060	0.458	0.002	−0.007	−0.000	0.052	55,311
		(2.33)	(2.06)	(−1.80)	(−1.81)	(1.41)	(0.48)	(−1.22)	(−0.27)	(1.48)	[0.015]
(8)	High diff + Low value	0.601	0.010	−0.015	−0.002	0.071	0.001	−0.001	0.000	0.060	133,659
		(2.87)	(2.20)	(−2.63)	(−0.31)	(0.93)	(0.59)	(−0.41)	(3.30)	(1.48)	[0.007]
(9)	Low diff + High value	1.462	0.034	−0.030	−0.001	0.014	−0.000	−0.003	0.001	0.040	131,362
		(2.80)	(2.11)	(−1.26)	(−0.05)	(0.17)	(−0.10)	(−0.54)	(1.82)	(2.35)	[0.008]
(10)	Low diff + Low value	1.002	0.029	−0.068	−0.030	0.090	−0.001	0.004	0.000	0.004	57,595
		(2.44)	(2.48)	(−2.19)	(−2.24)	(0.54)	(−0.26)	(0.39)	(1.38)	(0.17)	[0.018]
	Panel C: Dep. var = New segment pairs created by likely reclassification										
(11)	All industry pairs	1.335	0.015	−0.090	−0.045	0.174	−0.005	0.010	−0.000	0.106	382,494
		(5.08)	(2.29)	(−3.52)	(−4.04)	(1.25)	(−1.69)	(2.14)	(−0.60)	(3.46)	[0.009]
(12)	High diff + High value	1.133	0.026	−0.093	−0.061	0.401	−0.003	0.005	−0.000	0.035	55,311
		(1.48)	(1.08)	(−2.42)	(−2.66)	(1.22)	(−0.57)	(0.62)	(−0.18)	(1.19)	[0.008]
(13)	High diff + Low value	1.538	0.016	−0.080	−0.053	−0.128	−0.000	0.013	0.000	0.255	133,659
		(2.27)	(1.04)	(−2.52)	(−3.06)	(−0.52)	(−0.05)	(2.21)	(1.55)	(1.65)	[0.013]
(14)	Low diff + High value	2.442	0.047	−0.099	−0.008	0.412	−0.005	0.003	0.001	0.077	131,362
		(4.85)	(2.88)	(−2.02)	(−0.47)	(2.09)	(−1.50)	(0.44)	(1.17)	(2.88)	[0.008]
(15)	Low diff + Low value	0.947	0.017	−0.096	−0.074	0.069	−0.009	0.007	0.000	0.073	57,595
		(3.46)	(1.87)	(−2.22)	(−3.30)	(0.50)	(−1.44)	(0.64)	(0.75)	(2.12)	[0.018]

Notes. OLS regressions with year and industry fixed effects and standard errors (in parentheses) are clustered by industry. The dependent variable is the relative fraction of new multiple-industry segments, which is the number of new multiple-industry segments in each three-digit SIC code pair in the given year divided by the total number of single-segment firms operating in the two industries in the given pair, multiplied by 100 for convenience. Panel A counts the number of new multiple-industry firms operating in both industries of an industry pair. Panel B restricts attention to new segments of multiple-industry firms that were the acquirer in a transaction amounting to at least 10% of the firm's assets. Panel C restricts attention to likely reclassified segments, which is the number of new segments that appear in years where the total number of segments reported by the given firm is less than or equal to the past-year number of segments.

product market synergies that allow them to lower the cost of entry into profitable highly differentiated industries. We also observe that multiple-industry firms are more likely to add new segments when the fraction of industries between the conglomerate pair is high and the average within-industry similarity is low.

The results in panel B further show that conglomerate segments are more likely to be added through growth or acquisition when highly differentiated and highly valued industries lie between the segment pairs. In particular, multiple-industry firms add such segments when the resulting industry pairs have high across-industry similarity, low within-industry similarity, and a high fraction of industries that lie between the industry pair.

The results in panel B are consistent with the following two-stage mechanism for how firms might enter neighboring high-value industries that might be protected by barriers to entry. First, panel B shows that the firm might strategically acquire a segment in an industry that is spatially located on the other side of the targeted industry. The second stage would be to potentially combine the technologies of the spatially bracketing industry segments and enter the BI. This two-stage strategy can explain why the acquirer in the first stage might see adequately high product market synergies to justify the acquisition. We believe future research further examining this potential mechanism could be valuable.

The results in panel C are also consistent with firms reclassifying segments into industries that have higher across-industry product language similarity, more BI, and lower within-industry similarity. The only difference in panel C is that the subsample tests reveal that panel C results are stronger when BIs have higher valuations and *lower* differentiation rather than higher valuations and *higher* differentiation. This difference might occur because without acquisitions, as discussed in panel B, penetrating BIs that are highly differentiated might be difficult. In particular, without acquiring another firm, this form of entry through organic reclassification might not be possible because of frictions such as patents, which are more likely to exist when an industry is differentiated.

We conclude overall that our results are broadly consistent with multiple-industry firms choosing to expand into industries that give them the most potential for related-industry synergy gains. These results are especially relevant for those industries that also have a lower degree of specialized language and are thus consistent with the theory of organizational language in Crémer et al. (2007).

4.1. Text-Based Industry Classifications

In this section, we replicate the multiple-industry firm choice analysis in Table 3 using text-based industry classifications from Hoberg and Phillips (2016). In particular, we focus on the fixed industry classification with 300 industries (FIC-300), which is a set of 10-K-based industries chosen to be roughly as granular as SIC-3. To implement this calculation, we first need to reassign each firm to a set of FIC-300 segments as a substitute for the SIC-3 segments indicated by Compustat. This is achieved using the textual decomposition of each conglomerate firm into its respective segments from Hoberg and Phillips (2015). This decomposition generates a full set of single-segment peers for each segment of each conglomerate, with associated weights that sum to 1, and that best replicates the product offerings of the given conglomerate. For a conglomerate with N segments, we assign it to the N FIC-300 industries having the highest total weight in the Hoberg and Phillips (2015) decomposition. This methodology is parsimonious and fully accounts for the documented improvements in conglomerate benchmarking illustrated in the paper. We refer readers to Hoberg and Phillips (2015) for details regarding the weighted conglomerate decomposition.

The main impetus for this analysis is to establish robustness using an alternative classification, and to also establish robustness using an industry classification based on text-based industry-relatedness variables. We do not include this analysis as our primary analysis because many variables are not as readily available using text-based classifications in this system, as text-based classifications only became available starting in 1996. As a result of these limitations, our sample is restricted to 264,781 industry-pair-years rather than the 382,494 available in Table 3. Furthermore, we do not have measures of vertical relatedness in this setting, and variables requiring multiple years to compute such as economies of scale especially limit the sample size available using FIC-300 industries.

Table 5 displays the results of this test using FIC-300 industries. The table shows that most of our key findings are robust to using FIC-300 instead of SIC-3 despite the smaller sample size. For example, multiple-industry firms are far more likely to operate in industry pairs with a high potential for synergies (across-industry product language similarity), with a larger fraction of BIs, and with less specialized languages (lower within industry similarity).

However, three results in Table 5 differ from those in Table 3. First, multiple-industry firms are less likely to operate in high-patenting industries using FIC-300 industries but are not significantly linked to high-patenting industries using SIC-3 industries. We find this result interesting, especially given that FIC-300 industries are fully updated each year, whereas SIC-3 industries change little. Second, the economies of scale variable is negative using SIC-3 and either positive or insignificant using FIC-300. We believe the reason for this difference is likely technical. The economies of scale variable requires a longer time series to properly estimate, and inadequate long-term FIC-300 data exist to make this possible. Third, the industry instability variable is positive for SIC-based industry pairs and negative for FIC-based industry pairs. Regardless of these changes for the control variables, we note the text-based variables are stable across the two specifications.

5. Product Market Boundaries

In this section, we examine the robustness of our findings relating to across-industry product language similarity and potential synergies using a framework that does not rely on the Compustat segment database. This test is important for two reasons. First, Hoberg and Phillips (2016) show that SIC-based classifications cannot adequately capture information about industry memberships.[11] Villalonga (2004) has also questioned the reliability of the Compustat segment database, showing that it does not capture multiple-industry production. Second, we view the results regarding potential synergies across industries to be the primary contribution of the current article. Hence, reexamining the same predictions through a more refined framework can offer a highly discriminating test of robustness regarding our primary contribution.

Table 5. Redefined Segments Using Text-Based Classifications

	Sample	AILS	BI	WILS	Economies of scale	Patent applications	Industry instability	Pair likelihood if random	Vertical relatedness	Same two-digit SIC code	No. of obs. [R-SQ]
					Panel A: Where multiple-industry firms exist (as in Table 3)						
(1)	All industry pairs	39.251	0.166	−2.417	0.253	−0.000	−0.140	0.206	N/A	N/A	264,781
		(6.63)	(1.71)	(−7.35)	(1.94)	(−2.06)	(−5.33)	(5.04)			[0.097]
					Panel B: New conglomerate segments (as in Table 4) overall						
(2)	All industry pairs	25.201	0.081	−1.925	0.132	−0.000	−0.134	0.156	N/A	N/A	264,781
		(6.97)	(1.36)	(−7.90)	(1.36)	(−2.08)	(−6.00)	(5.59)			[0.083]
					Segments likely obtained through acquisition						
(3)	All industry pairs	1.348	0.005	−0.120	0.012	−0.000	0.002	0.007	N/A	N/A	264,781
		(7.00)	(1.45)	(−6.18)	(1.68)	(−1.02)	(0.58)	(5.68)			[0.012]
					Segments likely created through reclassification						
(4)	All industry pairs	10.638	0.009	−0.963	0.145	−0.000	−0.008	0.054	N/A	N/A	264,781
		(7.42)	(0.43)	(−8.03)	(3.40)	(−2.15)	(−0.80)	(6.23)			[0.058]

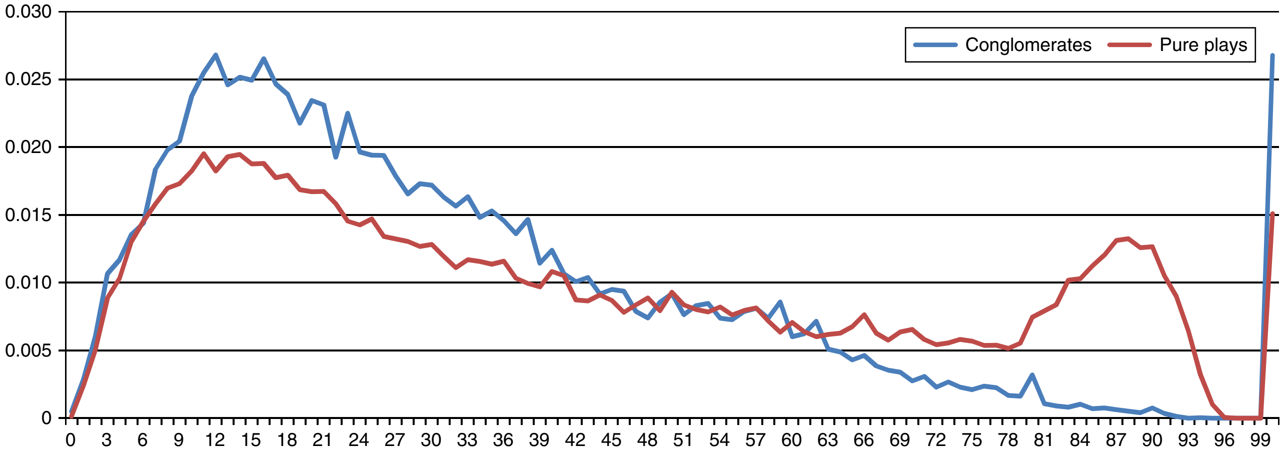
Notes. OLS regressions with year and industry fixed effects and standard errors (in parentheses) are clustered by industry for our sample of 264,781 industry pairs from 1996 to 2013. One observation is one pair of FIC industries in a year derived from the set of all permutations of feasible pairings. The dependent variable is the number of multiple-industry firms operating in the given industry pair, multiplied by 100 for convenience. Panel A displays results for existing segments. Panel B displays results for newly added segments in three categories: (1) overall, (2) those linked to major acquisitions, and (3) likely reclassified segments (new segments that appear in years where the total number of segments reported by the given firm is less than or equal to the past-year number of segments).

Our alternative measure of the potential for synergies is the degree of product market language overlap transitivity. This is a measure of how strong a given product market's language boundaries are. Markets with weak boundaries, for example, are likely susceptible to entry by firms in neighboring markets at relatively low cost because of asset complementarities. We define product market transitivity at the firm level, and for a given focal firm, we start by identifying its set of rival firms as indicated by the 10-K-based TNIC industry classification from Hoberg and Phillips (2016). (TNIC industries identify a set of rival firms for each firm as those having the most similar 10-K business descriptions to the given focal firm.) For each rival, we also use TNIC to identify the set of rivals of the rivals. TransComp (our measure of transitivity) is then the fraction of firms in the set of rivals of rivals that are also in the set of rivals of the focal firm as explained above. Figure 3 displays the distribution of this variable for firms with more than one segment in the Compustat database and separately for firms that have just one segment. It is also important to note that although we compute language transitivity for both multiple-industry firms and single-segment firms, we only use single-segment firms as reference peers for the purposes of the calculation itself, to maintain consistency with the rest of our study, and to ensure no mechanistic differences affect transitivity scores for multiple-industry firms.

Figure 3 shows a high degree of variability in the transitivity of product markets; it also shows that multiple-industry firms are fundamentally different from single-segment firms regarding the degree of transitivity faced in their respective markets. In particular, single-segment firms lie within a sharply bimodal distribution, and multiple-industry firms lie within a sharply unimodal distribution and generally have much lower transitivity than single-segment firms. We interpret this in terms of product market boundaries. We conclude that multiple-industry firms almost universally operate in product markets with weak boundaries and greater potential for cross-industry communication, whereas single-segment firms operate both in markets with weak boundaries and in markets with stronger boundaries (which have more specialized languages).

We also note that the degree of transitivity varies widely across industries. On the basis of the Fama–French 48 industries, for example, the beer industry exhibits high industry transitivity, as firms share a strong common language. Construction and insurance have lower transitivity, indicating that firms in these markets speak a broader language that can be applied in other markets. These results would suggest that cross-industry synergies are more relevant in construction and insurance than in the beer industry. Potential complementarities are also more likely in business services and retail than in textiles. This is consistent with the emergence of broad retail empires such as Amazon.com, which likely benefit from asset complementarities. Indeed, we confirm that Amazon.com has weak product market boundaries, with a transitivity score averaging less than 20%. Apple also has a transitivity score close to 20%, supporting our earlier

Figure 3. (Color online) Density of Product Market Transitivity for Reported Multiple-Industry Firms and Single-Segment Firms



Notes. Product market transitivity is the observed probability that firms A and C are rivals, given that A and B are rivals and that B and C are rivals. A pair of firms is defined as being rivals if they are classified as such using the TNIC-3 industry classification. The graph reports the probability density on the y axis and the percentage level of transitivity (which is bound between 0 and 100) on the x axis.

conjecture that Apple likely benefits from strong synergies across previously disparate industries.

Table 6 formally examines the association between industry transitivity and organizational form using three panels. Panel A examines highly transitive versus low transitivity industries and shows that more competitors are multiple-industry firms in industries with low transitivity (52.8% versus 39.2%) and that this difference is large. Panel B examines this relationship across subsamples based on firm size and firm age, two variables that are also strongly linked to whether a firm is a conglomerate. Panel B shows that smaller, younger firms in highly transitive industries are especially less likely to be multiple-industry firms (just 15%). By contrast, segments are likely to be in multiple-industry firms if they are larger, older, and in weakly transitive industries (75%). Panel B also shows that all three variables (size, age, and transitivity) are distinct and that each is separately economically important in explaining whether a firm is likely to be a conglomerate.

Panel C displays the results of logistic regressions, where one observation is one firm in one year, and the dependent variable is a dummy equal to 1 if the firm is a multiple-industry firm (defined as a firm having more than one segment in the Compustat tapes) and 0 for a single-segment firm. The independent variables include the degree to which the given firm is in a transitive product market and control for firm age, size, and profitability. The results show that multiple-industry firms are more likely to be in industries with weak boundaries (lower transitivity). Conglomerate multiple-industry firms are also more likely to be older, larger firms. Our finding that multiple-industry firms are producing in product markets with weaker product market boundaries is consistent with these firms

Table 6. Product Market Transitivity

Sample	Transitivity subsample	Fraction multiple-industry	No. of obs.
	Panel A: All firms		
All firms	Weakly transitive	0.528	65,431
All firms	Highly transitive	0.392	65,044
	Panel B: Subsamples based on size and age		
Small young firms only	Weakly transitive	0.303	18,050
Small young firms only	Highly transitive	0.150	23,762
Small old firms only	Weakly transitive	0.497	14,191
Small old firms only	Highly transitive	0.336	9,229
Large young firms only	Weakly transitive	0.497	11,361
Large young firms only	Highly transitive	0.430	11,678
Large old firms only	Weakly transitive	0.750	21,829
Large old firms only	Highly transitive	0.679	20,375

	Fraction transitive	Log sales	Log firm age	OI/Sales	Industry fixed effects	No. of obs. [R-SQ]
			Panel C: Logistic regressions			
(1)	−1.650				No	130,475
	(−20.74)					[0.050]
(2)	−1.401	0.286	0.741	0.002	No	130,475
	(−15.86)	(22.29)	(22.48)	(0.79)		[0.226]
(3)	−0.719				Yes	130,475
	(−7.35)					[0.240]
(4)	−0.638	0.216	0.715	0.008	Yes	130,475
	(−6.12)	(16.53)	(21.01)	(1.91)		[0.317]

Notes. Summary statistics and logistic regressions with year and industry fixed effects and standard errors (in parentheses) are clustered by industry for our sample of 130,475 Compustat firms from 1997 to 2013. Panels A and B report summary statistics regarding the average fraction of multiple-industry firms for various subsamples as noted. Panel C displays the results of logistic regressions where the dependent variable is a dummy equal to 1 for a multiple-industry firm and 0 for a single-segment firm. Product market transitivity is the fraction of peers of a given firm that also consider the given firm itself to be a peer, as computed using the TNIC-3 industry classification. OI/Sales is operating income plus depreciation divided by sales.

Table 7. Product Market Transitivity and Divesting Conglomerate Segments

	Fraction transitive	R&D/ Sales	CAPX/ Sales	OI/ Sales	Log assets	Document length	Industry fixed effects	No. of obs. [R-SQ]
(1)	−0.010						No	42,374
	(−2.74)							[0.002]
(2)		−0.001					No	42,374
		(−1.38)						[0.002]
(3)			0.003				No	42,374
			(1.33)					[0.002]
(4)				−0.000			No	42,374
				(−0.39)				[0.002]
(5)					0.004		No	42,374
					(2.66)			[0.002]
(6)						−0.002	No	42,374
						(−1.69)		[0.002]
(7)	−0.010	−0.004	0.002	−0.001	0.004	−0.003	No	42,374
	(−2.64)	(−1.84)	(0.93)	(−1.26)	(2.92)	(−1.89)		[0.003]
(8)	−0.008						Yes	42,374
	(−2.09)							[0.017]
(9)		−0.000					Yes	42,374
		(−0.40)						[0.017]
(10)			0.002				Yes	42,374
			(0.93)					[0.017]
(11)				−0.000			Yes	42,374
				(−0.82)				[0.017]
(12)					0.003		Yes	42,374
					(2.18)			[0.018]
(13)						−0.005	Yes	42,374
						(−3.04)		[0.018]
(14)	−0.007	−0.003	0.000	−0.001	0.004	−0.005	Yes	42,374
	(−1.87)	(−1.37)	(0.22)	(−1.28)	(2.52)	(−3.08)		[0.018]

Notes. OLS regressions with year and industry fixed effects and standard errors (in parentheses) are clustered by industry for our sample of multiple-industry firms from 2000 to 2013. The dependent variable is negative component of the logarithmic growth in the number of segments of the given conglomerate from year t to year $t+1$. Hence the dependent variable is the relative decline in the number of segments. We note that the results here are asymmetric, and we do not find analogous results for conglomerate segment additions (and hence they are not reported). All independent variables are measures of change in the given quantity from year $t-3$ to year t. Product market transitivity is the fraction of peers of a given firm that also consider the given firm itself to be a peer, as computed using the TNIC-3 industry classification. All regressions include year fixed effects and three-digit SIC industry fixed effects (when specified). OI/Sales is operating income plus depreciation divided by sales.

choosing to operate in markets where potential synergies are likely; that is also consistent with the theory of organizational language in Crémer et al. (2007).

We now examine whether these results hold in differences and whether ex ante changes in industry transitivity are linked to ex post changes in conglomerate organization.

Table 7 examines whether multiple-industry firms drop segments following changes in transitivity. The dependent variable is a negative component of the logarithmic growth in the number of segments of the given conglomerate from year t to year $t+1$. Hence, the dependent variable is the relative decline in the number of segments. We note that the results here are asymmetric, and we do not find analogous results for conglomerate segment additions (hence, they are not reported). All independent variables are measures of change in the given quantity over the three prior years from year $t-3$ to year t. In addition to three-year changes in product market transitivity, we consider three-year changes in research and development (R&D), capital expenditure (CAPX), profitability, and firm size. Specifications also include time fixed effects and industry fixed effects when noted, and standard errors are clustered by industry.

Table 7 shows that multiple-industry firms decrease the number of reported segments when transitivity increases. The results are consistent with multiple-industry firms responding to any strengthening of product market boundaries by dropping segments. Because stronger product market boundaries indicate that firms cannot easily expand their scope, this finding is also consistent with firms reacting to changes in the potential for product market synergies by changing their overall operating configuration.

This section offers a robustness check that is independent of the potentially unreliable SIC code links provided in the Compustat segment tapes. These results are instead based on a more direct measure of the potential for synergies, and moreover, they are based on measures that are updated each year (they are constructed from yearly firm 10-K filings). A stark test of this nature is prohibitively difficult using existing SIC or NAICS-based data because not much cross-industry relatedness data are available, and moreover, these classifications are generally updated little over time.

6. Growth of Product Offerings

Given our findings in earlier sections, we examine a finer prediction of H1 (potential synergies through asset complementarities) in this section. In particular, if multiple-industry firms indeed operate in some markets to act on potential synergies, we should observe a positive link between potential asset complementarities and increases in firm product offerings over time. We thus examine whether multiple-industry firms operating in industry combinations with greater across-industry product language similarity increase their product offerings over time. We consider the size of a firm’s 10-K business description as a measure of the depth of a firm’s product offerings in a given year. Because Form 10-K is filed annually, we can assess the degree to which a firm increases its product offerings in a given year by examining the extent to which its business description grows from one year to the next. We can then examine whether this growth is related to ex ante measures of potential synergies.

Table 8 presents the results of this test. The dependent variable is the firm’s product description growth, defined as the natural logarithm of the number of words in the firm’s business description in year $t+1$ divided by the number of words in the firm’s business description in year t. We consider the same explanatory variables as in Table 3, although there, we focus our attention on the across industry similarity variable. Panel A displays results based on raw firm-level product description growth. Panel B displays results based on TNIC industry adjusted product description growth.

The results show that conglomerate product description growth is highly related to ex ante measures of potential synergies, as measured by across-industry product language similarity. The findings are consistent with H1, which predicts that potential cross-industry synergies provide opportunities for multiple-industry firms to increase their product market offerings. The results are also consistent

Table 8. Product Description Growth

	AILS	BI	WILS	Econom. of scale	Vertical relatedness	Patent applications	Industry instability	Same two-digit SIC	Pair likelihood if random	Document length	R&D/ Sales	CAPX/ Sales	OI/ Sales	Log assets	No. of obs. [R-SQ]
						Panel A: Product description growth									
(1)	0.321	0.028	−0.038	−0.041	−0.027	0.000	−0.012	−0.011	−0.000	−0.000	0.067	0.016	0.037	0.000	15,515
	(2.37)	(1.42)	(−0.38)	(−2.37)	(−0.58)	(0.50)	(−0.81)	(−1.68)	(−0.00)	(−20.76)	(1.72)	(1.47)	(2.96)	(3.89)	[0.112]
(2)	0.188									−0.000	0.071	0.015	0.036	0.000	15,515
	(1.83)									(−20.54)	(1.81)	(1.38)	(2.89)	(3.87)	[0.111]
						Panel B: Industry-adjusted product description growth									
(3)	0.308	0.003	−0.071	−0.036	0.020	−0.000	−0.016	−0.010	0.000	−0.000	0.071	0.015	0.027	0.000	15,157
	(2.21)	(0.11)	(−0.69)	(−2.10)	(0.43)	(−0.56)	(−1.04)	(−1.43)	(0.47)	(−15.05)	(1.68)	(1.40)	(1.85)	(2.13)	[0.042]
(4)	0.261									−0.000	0.077	0.015	0.027	0.000	15,157
	(2.50)									(−14.84)	(1.78)	(1.39)	(1.85)	(2.07)	[0.041]

Notes. OLS regressions with year and industry fixed effects and standard errors (in parentheses) are clustered by industry for our sample of multiple-industry firms from 1997 to 2013. One observation is one conglomerate in one year. The dependent variable is the firm’s product description growth, defined as the natural logarithm of the number of words in the firm’s business description in year $t+1$ divided by the number of words in the firm’s business description in year t. Panel A displays results based on raw firm-level product description growth. Panel B displays results based on TNIC industry adjusted product description growth. OI/Sales is operating income plus depreciation divided by sales.

with the fundamental characteristics of asset complementarities as outlined by Teece (1980) and Panzar and Willig (1981).

7. Language Complexity

Because our measures of industry relatedness are based on verbal content in business descriptions of the firms in our sample, our empirical laboratory is a natural fit for testing the Crémer et al. (2007) theory of firm organization and organizational language. However, one concern is that our measures of similarity might relate to potential operational asset complementarities across industries (e.g., see Hoberg and Phillips 2010). To further solidify our conclusion that language specifically drives our results, at least in part, we examine whether our results are stronger in product markets where language barriers are likely to be more binding.

We categorize product markets using two established measures of readability of 10-K business description text: the Gunning Fog Index (Gunning 1952) and the Flesch–Kincaid readability index (Kincaid et al. 1975). Both indices are established measures of language complexity and are computed using formulas that take as input quantities such as the number of syllables per word and the number of words per sentence. In our framework, for each index, we define a "language complexity dummy" as 1 if a given firm's 10-K business description has above-median language complexity and 0 otherwise. We then reconsider our main regressions in Tables 3 and 4 with just one addition: we add the language complexity dummy, and cross terms with our key language distance variables, to the regression. We include all control variables and fixed effects that are currently in the existing models in Tables 3 and 4, although we do not report the full set of coefficients, to conserve space.

Table 9 displays the results of these tests. In panel A, we consider the Gunning Fog Index as our measure of language complexity, and in panel B, we consider the Flesch–Kincaid readability index. Results in both panels are similar. The first row in each panel examines the industries in which conglomerates operate, as in the first row of Table 3. The remaining rows in each panel display results for new conglomerate segments based on the first row in each panel in Table 4.

In the first specification (row (1)) in Table 9, we find that conglomerates are less likely to produce in industries with high language complexity as captured by the variable above-median language complexity. The interaction effect of language complexity with AILS shows a positive coefficient, indicating that when conglomerates do produce in industries with high language complexity, the industries are clustered closer together in the product space. These results support the prediction in Crémer et al. (2007) that firms favor a more narrow operating profile when the cost of imprecise communication (in our case, caused by complexity) is high. The finding is also economically large, as the baseline AILS coefficient is 61 in row (1), and the cross term shows that this increases by 67% to 102 (61 baseline plus 41 from the cross term) when language is more complex.

We conclude that conglomerates are less likely to choose industries with complex language but are likely to choose industries that cluster closer in product space when language is more complex. These findings provide rather unique support for Crémer et al. (2007). Because this result is significant at the 5% level in both rows (1) and (5) of Table 9, we conclude that it is robust to either measure of language complexity. The remaining rows in each panel show that this result also obtains for likely segment reclassifications, but we do not find analogous results for newly added segments based on acquisitions.

Regarding BIS, row (1) shows an analogous positive interaction coefficient, suggesting that firms again choose industries that are closer in the product space when language is complex, as was the case for AILS. However, the between coefficient in row (1) just misses the cutoff for significance at the 10% level. The remaining rows based on newly added segments show that the between cross term is positive and significant for segments likely added through reclassification, especially in row (8). Although some results regarding BIs are not significant, we view these results as suggestive that our results for BIs are likely driven at least in part by issues specifically related to language.

Finally, the table also shows that our results for WILS are uniformly less negative in markets where language is more complex—suggesting that within-industry language similarity can mitigate the aforementioned problems associated with language complexity. In general, conglomerate firms avoid industries with high within-industry similarity, but language complexity mitigates this strong negative effect. This is consistent with conglomerate firms providing expertise to help mitigate the problems of specialization when language is complex.

Overall, our results based on language complexity illustrate that industry choice is indeed affected by the level of language complexity. This in turn supports our central thesis that language itself can influence important corporate decisions.

8. Conclusions

We examine product language overlaps across industries using text-based analysis of business descriptions from 10-K filings with the SEC. We examine industry configuration choices for multiple-industry firms and the extent that fundamental industry

Table 9. The Role of Language Complexity

Sample	AILS × Language complexity	BI × Language complexity	WILS × Language complexity	Same SIC-2 × Language complexity	Above-median language complexity	AILS	BI	WILS	Same two-digit SIC code	No. of obs. [R-SQ]
				Panel A: Gunning Fog Index						
(1) Existing conglomerates	40.837	0.733	3.396	−3.939	−0.739	61.163	0.725	−4.608	7.325	382,494
	(2.22)	(1.65)	(3.58)	(−4.46)	(−2.16)	(5.47)	(1.88)	(−5.38)	(6.25)	[0.140]
(2) New segment pairs	5.803	0.046	1.127	−1.036	−0.112	20.136	0.321	−1.714	1.926	382,494
	(1.19)	(0.35)	(3.56)	(−3.63)	(−1.16)	(5.35)	(2.51)	(−5.83)	(6.19)	[0.067]
(3) New M&A pairs	0.227	−0.002	0.055	−0.031	−0.002	0.982	0.022	−0.075	0.051	382,494
	(0.52)	(−0.21)	(2.99)	(−1.40)	(−0.27)	(4.26)	(2.83)	(−4.39)	(2.81)	[0.007]
(4) Likely segment reclass.	1.326	0.026	0.101	−0.009	−0.026	1.080	0.007	−0.149	0.108	382,494
	(2.33)	(1.91)	(2.48)	(−0.23)	(−2.49)	(4.59)	(0.94)	(−4.18)	(4.30)	[0.010]
				Panel B: Flesch–Kincaid readability index						
(5) Existing conglomerates	24.064	0.465	2.229	−2.695	−0.508	58.336	0.677	−3.683	7.491	382,494
	(2.25)	(1.36)	(3.19)	(−2.13)	(−1.83)	(4.24)	(1.72)	(−5.35)	(4.95)	[0.137]
(6) New segment pairs	5.938	0.078	0.577	−0.754	−0.112	18.179	0.273	−1.307	2.000	382,494
	(1.56)	(0.68)	(2.48)	(−1.91)	(−1.23)	(4.37)	(2.23)	(−5.79)	(4.74)	[0.067]
(7) New M&A pairs	−0.291	−0.012	0.050	−0.048	0.006	1.165	0.026	−0.067	0.071	382,494
	(−0.99)	(−1.24)	(2.80)	(−1.68)	(0.81)	(3.93)	(2.81)	(−3.89)	(2.52)	[0.007]
(8) Likely segment reclass.	0.639	0.026	0.105	−0.047	−0.025	1.004	−0.001	−0.147	0.131	382,494
	(2.32)	(3.29)	(2.96)	(−1.19)	(−3.69)	(3.14)	(−0.12)	(−4.16)	(2.90)	[0.009]

Notes. OLS regressions with year and industry fixed effects and standard errors (in parentheses) are clustered by industry for our sample of 382,494 industry pairs from 1996 to 2013. One observation is one pair of three-digit SIC industries in a year derived from the set of all permutations of feasible pairings. In rows (1) and (5), the dependent variable is the fraction of multiple-industry firms operating in the given industry pair (relative to the total number of single-segment firms operating in the two industries in the given pair), multiplied by 100 for convenience. In the remaining rows, the dependent variable is the relative fraction of new multiple-industry segments, which is the number of new multiple-industry segments in each three-digit SIC code pair in the given year divided by the total number of single-segment firms operating in the two industries in the given pair, multiplied by 100 for convenience. In rows (2) and (6), the dependent variable is based on all new segments. In rows (3) and (7), the dependent variable restricts attention to new segments of multiple-industry firms that were the acquirer in a transaction amounting to at least 10% of the firm's assets. In rows (4) and (8), the dependent variable restricts attention to new segments that were likely created through reclassification (new segments that appear in years where the total number of segments reported by the given firm is less than or equal to the past-year number of segments). See Table 1 for a complete description of the independent variables. In this table, we focus on two language complexity measures and their cross terms with our key conglomerate industry variables. In panel A (panel B), language complexity is defined based on the Gunning Fog Index (Flesch–Kinkaid readability index), and we include a dummy indicating whether the industry pair on average has readability that is above median in terms of difficulty of readability of its 10-K business description text.

characteristics—not diversification—drive conglomerate industry choice. We find that multiple-industry firms are more likely to operate in industry pairs with higher language overlap, in industry pairs that have highly valued product markets "between" them, and firms are *less* likely to operate in industries with high within-industry product similarity. These findings are consistent with firms using the multiple-industry structure and language overlaps to take advantage of cross-market product synergies and asset complementarities. These results are robust both when examining existing multiple-industry firm industry configurations and when considering changes in these configurations.

We construct a more general test measuring the extent to which product markets have strong language boundaries. This test is based on the degree of transitivity of firm language within industry groups. This relaxes the need to rely on the quality of the Compustat segment tapes or any particular industry classification. Low levels of language transitivity indicate strong language boundaries and are consistent with a lower potential for synergies and a reduced potential for scope. Multiple-industry firms are more likely to operate in product markets with weak language boundaries, and these results are economically large. We show directly that conglomerates are less likely to produce in industries with high language complexity. When conglomerate firms do produce in these industries with complex language, we show that the industries are clustered closer in product space.

Last, we find that industries with high ex ante measures of across-industry product language overlap experience increased product description growth. These results are consistent with fundamental product overlaps facilitating product market synergies that result in new products and features. In all, our findings support theories of organizational language and product market synergies and help explain why many firms use multiple-industry structures despite potential negative valuation effects suggested by prior studies. These findings show that choosing complementary industries with related products is a primary motivation for conglomerate firm industry choice. Our findings call into question the previous major reason for conglomerate firm industry choice—that conglomerate firms choose unrelated industries to diversify their cash flows.

Acknowledgments

The authors thank department editor Amit Seru, an anonymous associate editor, and two anonymous referees for excellent comments. They also thank Ricardo Alonso, Peter MacKay, Andrea Prat, Merih Sevilir, Albert Sheen, and Denis Sosyura, as well as seminar participants at the Western Finance Association meetings, the University of Chicago, City University of Hong Kong, Columbia University, Duisenberg School of Finance and Tinbergen Institute, Erasmus University, the Rotterdam School of Management, Rutgers University, Stanford University, Tilburg University, University of Illinois at Champaign–Urbana, and the University of Mannheim for helpful comments. All errors are the authors' alone.

Endnotes

[1] A large literature has examined ex post outcomes subsequent to firms choosing to produce in multiple industries, comparing multiple-industry firms to single-industry firms. Lang and Stulz (1994) and Berger and Ofek (1995) examine whether diversified multiple-industry firms trade at discounts. Subsequent literature, including Shin and Stulz (1998), Maksimovic and Phillips (2002), and Schoar (2002), examines ex post investment and productivity to understand the potential reasons for this discount. Santalo and Becerra (2008) show that the discount only exists when there are a large number of single-segment firms operating alongside conglomerate segments (see Maksimovic and Phillips 2007 for a detailed survey).

[2] Panzar and Willig (1977), Teece (1980), and Panzar and Willig (1981) provide an early analysis of economies of scope and multiple-industry production. For more recent work on multiple-product firms, see Bernard et al. (2010) and Goldberg et al. (2010) for an analysis of changes to multiple-product firms in a developing country context.

[3] *Between industries* are industries that are closer to each industry of a given industry pair than the industry pair is to each other, based on product language similarity. We formally define this measure in the next section.

[4] This spatial representation does not impose transitivity on competitor networks. Similar to a Facebook circle of friends, each firm has its own set of competitors, and competitors need not be overlapping with other firms' competitors even within industry groups. This flexibility allows us to measure the degree to which a product market has strong boundaries (more transitivity indicates strong boundaries). Standard Industrial Classification (SIC) and North American Industry Classification System (NAICS) industry groupings do not permit such an analysis because they mechanistically impose transitivity: if firm A and firm B are competitors, and if firm B and firm C are competitors, then firms A and C are also competitors.

[5] Many earlier studies are rooted strongly in the assumption that conglomerates are highly diversified (such as Lang and Stulz 1994 and Berger and Ofek 1995) and that there are key costs and benefits stemming from this fact, such as the dark side (Scharfstein and Stein 2000) versus the bright side (Stein 1997) of conglomerates. The view of diversification as being a major consideration in conglomerate formation dates back at least to Gort (1962) and Lewellen (1971).

[6] Note that the product market space is a full representation of the products that firms offer and the extent to which they are similar, and the space should not be interpreted as a geographic space.

[7] We thank the Wharton Research Data Service for providing us with an expanded historical mapping of SEC CIK to Compustat gvkey, as the base CIK variable in Compustat contains only the most recent link.

[8] We identify nouns using Merriam-Webster.com as words that can be used in speech as a noun. We identify proper nouns as words that appear with the first letter capitalized at least 90% of the time in the corpus of all 10-K product descriptions. Previous results available from the authors did not impose this restriction to nouns. These results were qualitatively similar.

[9] Our use of binary vectors follows Hoberg and Phillips (2016), who show that using frequencies reduces the power of analogous industry measures.

[10] We also estimate the industry economies of scale using a translog production function for robustness, and results are similar.

[11] It is very telling that Apple was classified as a single-segment firm in the Compustat segment database until 2007, five years after it introduced the iPod.

References

Ahern K, Harford J (2014) The importance of industry links in merger waves. *J. Finance* 62(2):527–576.

Alonso R, Dessein W, Matouschek N (2008) When does coordination require centralization? *Amer. Econom. Rev.* 98(1): 145–179.

Becker GS, Murphy KM (1992) The division of labor, coordination costs, and knowledge. *Quart. J. Econom.* 107(4):1137–1160.

Berger P, Ofek E (1995) Diversification's effect on firm value. *J. Financial Econom.* 37(1):39–65.

Bernard A, Redding S, Schott P (2010) Multiple-product firms and product switching. *Amer. Econom. Rev.* 100(1):70–97.

Bolton P, Dewatripont M (1994) The firm as a communication network. *Quart. J. Econom.* 109(4):809–839.

Crémer J, Garicano L, Prat A (2007) Language and the theory of the firm. *Quart. J. Econom.* 122(1):373–407.

Fan J, Goyal V (2006) On the patterns and wealth effects of vertical mergers. *J. Bus.* 79(2):877–902.

Goldberg P, Khandelwal N, Pavcnik N, Topalova P (2010) Multi-product firms and product turnover in the developing world: Evidence from India. *Rev. Econom. Statist.* 92(4): 1042–1049.

Gort M (1962) *Diversification and Integration in American Industry* (Greenwood Press, Westport, CT).

Gunning R (1952) *The Technique of Clear Writing* (McGraw Hill, New York).

Hann R, Ogneva M, Ozbas O (2013) Corporate diversification and the cost of capital. *J. Finance* 68(5):1961–1999.

Hart OD, Moore J (2005) On the design of hierarchies: Coordination versus specialization. *J. Political Economy* 113(4):675–702.

Hoberg G, Phillips G (2010) Product market synergies in mergers and acquisitions: A text based analysis. *Rev. Financial Stud.* 23(19):3773–3811.

Hoberg G, Phillips G (2015) Product market uniqueness, organizational form and stock market valuations. Working paper, University of Southern California, Los Angeles.

Hoberg G, Phillips G (2016) Text-based network industry classifications and endogenous product differentiation. *J. Political Econom.* 124(5):1423–1465.

Kedia S, Ravid A, Pons V (2011) When do vertical mergers create value? *Financial Management* 40(4):845–877.

Kincaid J, Fishburne R, Rogers R, Chissom B (1975) Derivation of new readability formulas. Research Branch Report 8-75, Naval Air Station Memphis, Millington TN.

Lang L, Stulz R (1994) Tobin's q, corporate diversification, and firm performance. *J. Political Econom.* 102(6):1248–1280.

Lewellen W (1971) A pure financial rationale for the conglomerate merger. *J. Finance* 26(2):521–537.

Maksimovic V, Phillips G (2002) Do conglomerate firms allocate resources inefficiently across industries? Theory and evidence. *J. Finance* 57(2):721–767.

Maksimovic V, Phillips G (2007) Conglomerate firms and internal capital markets. Eckbo BE, ed. *Handbook of Corporate Finance: Empirical Corporate Finance* (North-Holland, Amsterdam), 423–480.

Panzar J, Willig R (1977) Economies of scale in multi-output production. *Quart. J. Econom.* 91(3):481–493.

Panzar J, Willig R (1981) Economies of scope. *Amer. Econom. Rev.* 71(2):268–272.

Santalo J, Becerra M (2008) Competition from specialized firms and the diversification-performance linkage. *J. Finance* 63(2): 851–883.

Scharfstein D, Stein J (2000) The dark side of internal capital markets: Segment rent seeking and inefficient investments. *J. Finance* 55(6):2537–2564.

Schoar A (2002) The effect of diversification on firm productivity. *J. Finance* 57(6):2379–2403.

Sebastiani F (2002) Machine learning in automated text categorization. *ACM Comput. Surveys* 34(1):1–47.

Shin HH, Stulz RM (1998) Are internal capital markets efficient? *Quart. J. Econom.* 113(2):531–552.

Stein J (1997) Internal capital markets and the competition for corporate resources. *J. Finance* 52(1):111–133.

Teece DJ (1980) Economies of scope and the scope of the enterprise. *J. Econom. Behav. Organ.* 1(3):223–247.

Villalonga B (2004) Does diversification cause the diversification discount? *Financial Management* 33(2):5–27.